8-1-2014

Development Of A Repository For Hydrologic Model Uncertainty Data

Elius Etienne

Michael Piasecki

DEVELOPMENT OF A REPOSITORY FOR HYDROLOGIC MODEL UNCERTAINTY DATA

ELIUS ETIENNE (1), MICHAEL PIASECKI (2)

(1) NOAA CREST Center, PhD Student Civil Engineering Department, 160 Convent Ave., NYC, NY 10031, corresponding author, eetienn02@gmail.com
(2) Environmental Crossroads Initiative, Dept. of Civil Engineering, 160 Convent Ave, NYC, NY 10031, corresponding author, mpiasecki@ccny.cuny.edu, 212-650-8321, http://daisy.ccny.cuy.edu/~michael

Abstract

Hydrologic processes are complex and when modeling them using a deterministic or stochastic approach one invariably introduces errors because of simplifications and assumptions made. However, not all assumptions and simplifications in the approach chosen produce the amount of errors; in fact the impact of deviations from the truth on a final output set of variables varies greatly. In addition, not every catchment behaves alike adding another layer of complexity to the modeling effort. Hence, every approach exhibits a degree of uncertainty in their results. While this uncertainty can be examined systematically in this technical note we focus on the development of a repository for modeling uncertainty data. We store information about the model used (lumped, semi distributed, fully distributed), the objective function (Nash Sutcliff, Root Mean Square Error, …) used to calculate the fitness of an approach, a Pareto best parameter combination, and also some statistical values that arrive from a specific approach and its ensemble such as median, max and min values. We describe the development of a database to store this data and also an online based submission system (based on the DRUPAL environment) that can be used to submit, explore, and retrieve uncertainty data. Finally, we use a sample data set from the 392 Model Parameter Estimation Experiment (MOPEX) catchments as an initial submission to our system which we use to show some of the features of our Uncertainty DB that will be accessible through http://uncert.net.
Keywords: hydrologic modeling uncertainty, database, web-services

I- MOTIVATION

The uncertainty in model outputs reflects in general the inaccuracies of the inputs such as measurement errors, the variability in the parameter values, and the imprecisions in the model's structure. Differences between the observed values compared to the simulated values can be caused by known and unknown errors in the inputs, the parameters values, and the model's structure as well as natural variability of the observed quantities (UNESCO, 2005).

Performance of mathematical models in general and hydrological models in particular, is evaluated statistically by calculating one or more performance criteria, also known as "goodness-of-fit" measures (Elshamy, 2008). Some model calibration processes involve the use of different parameter sets, one at each run of the model, for the derivation of the model performance measure. P. Krause (2005) demonstrated the importance of using multiple criteria in model performance studies. Those practices generate many values of the performance indicator, which can take on many forms, for example the Nash-Sutcliffe model efficiency coefficient (Nash and Sutcliffe, 1970) or the Root Mean Square Error, RSME. The performance indicators values, likely to be different for each run, give an indication that the model in general produces uncertain results and also how far "off" the model is. If one uses many different input parameter combinations one can compile an ensemble of performance measures using the same measure, or even several ensembles using a multitude of performance measure approaches that in turn give an indication of how good or appropriate a specific performance measure is. This in turn can be interpreted as compiling a set of parameter estimation uncertainty values(Rasouli et al., 2012) since the difference is the result of inputting a different set of parameter values at each run.

Knowing the uncertainty of the performance of the model and having the parameter values corresponding to that performance are valuable information for model calibration studies. The final and best set of parameters should correspond to the best performance measure, i.e. the least amount of deviation from the true results. In fact, if one measures the change of output parameter to magnitude of input parameter variation one can compute a set of sensitivity information that helps in identifying those parameters that have a large impact on the output versus those that do not and as such can be estimated with much less uncertainty.

This work will use a statistical approach to describe the set of performance values generated from a model's application over a wide set of input parameter combinations, and to represent and communicate the model's performance indicator uncertainty. Because there are different models, different performance measures, and thousands of catchments in just the continental US, computing a broad set of performance measures becomes a multi-dimensional problem (Kollat et al, 2012) very quickly generating vast amount of output data that even when reduced by one dimension (into a single performance measure value) produces thousands if not millions of data points. This in turn strongly suggests the development of a repository in which one can insert the uncertainty values, annotate these datasets with metadata describing how they were derived, and then adding some statistical added value data products to each ensemble to better (and quickly) transmit some basic characteristics of the ensemble data set.

In this technical note we will describe and outline the development of a database schema to store hydrologic modeling uncertainty data, the metadata framework to properly annotate the data, and also the development of an online submission/search/retrieval system for this data. We will outline some of the functionality having submitted uncertainty data from the MOPEX effort that used two different hydrologic models, 4 different performance measures, and was derived or computed from 392 watersheds distributed throughout the US.

III- DATA STORAGE

3.1- Data model

Our initial intention was to stay as close to those appliances developed by the Consortium of Universities for the Advancement of Hydrologic Sciences Inc., (CUASHI), namely the Observations Data Model database schema, to better match already existing schemata, and also to reduce the amount of new development. Consequently, our data model conceptually reflects the CUAHSI ODM design (CUAHSI, 2014). The ODM data model contains a table named '*Data Values*' which contains all hydrologic observations. All other tables contain the metadata associated with that value. A similar concept was adopted in the design of our data model. Two tables contain the data values (the '*objective function data*' table and the '*uncertainty data*' table) and the other tables contain the metadata about those values. Two tables must be used because the system stored two types of "data values": the objective function data and the uncertainty data about them. The '*uncertainty data*' table should have a particular structure to make possible the plot of the boxplots on the web site using the jpGraph package which is an Object-Oriented Graph creating library for PHP (jpGraph, 2014). This library requires that the data retrieved be a multiple of 5 for the boxplot plot. Therefore the five statistics must be stored in the same record, and each record retrieved will represent a box plot.

Because of the need for this special structure of the '*uncertainty data*' table, two tables are used to store the "data values" – instead of one in the original ODM model.

A database is designed to store the data and metadata. A relational database format is used to allow the ease of querying and retrieval of data.

Because the data will be made available through the SOS web service, it is important that the data model complies with the SOS service implementation standards (Bröring et al, 2012). Since the schema was designed to store sensors' data, there is the need to add the necessary tables for the storage of the performance measures data, the parameters data, and the uncertainty data. Changing the name of the existing tables of the SOS schema would impair the functionality of the service. Therefore the names of the existing relations remain the same. However, a particular table is chosen so its name makes sense to the metadata that will be stored in it. For instance the '*procedure*' table, which contains information about the sensors in a SOS instance where sensors data are concerned, will be filled with the models' information. Models and sensors are considered to be similar. The '*offerings*' table will be used for the objective functions; and the '*features of interest*' table will be used for the catchments. The main purpose of the SOS service is to publish sensors' data. Those similarities will allow the data to make sense in a sensors' data context. The modified schema is presented in Figure 1.

3.2- Data Stored

Five descriptive statistics are supported in the system and should be provided when a user is submitting data. They are the minimum, the maximum, the first and third quartiles, and the median of the set of values of the performance indicator also often called "objective function".

Our database currently contains 392 Model Parameter Estimation Experiment (MOPEX) catchments, 4 performance indicators, and 2 hydrological models. Those data were used in a study by Kollat et al (2012) about the meaningfulness of multiobjective calibration trade-offs in hydrological models. The number of parameters involved in the calibration was 8 for one model and 14 for the other one. Many values for each of the 4 objective functions were obtained from the calibration process. Those data constitute the raw data for our system and the models, catchments, objective functions are the metadata. The parameter sets corresponding to the minimum and maximum values of an objective function are also stored in the database.

IV- SYSTEM ARCHITECTURE AND PROCESS FLOW

The system provides data access via a web portal and via an implementation of the SOS service. Data submission can seamlessly be done through the web portal. It could as well be done via the SOS service using the SOS feeder framework or the SOS transactional function (SOS Data Modeling, 2014). Since the development of the SOS feeder has been stopped and the transactional operations of the SOS are at an early state (SOS Data Modeling, 2014), data submission is only handled using the web interface. The SOS service is only used as an open connection to request data.

The web portal is built using Drupal, an open source framework distributed under the GNU GPL license, written in PHP, which allows web development in an easy and powerful way (Drupal, 2014). The plotting library jpGraph allows the visualization of the uncertainty data and is used inside a custom module within Drupal. The data are stored in and PostgreSQL database and is accessed via the PHP Data Objects (PDO) interface. The Openlayers and

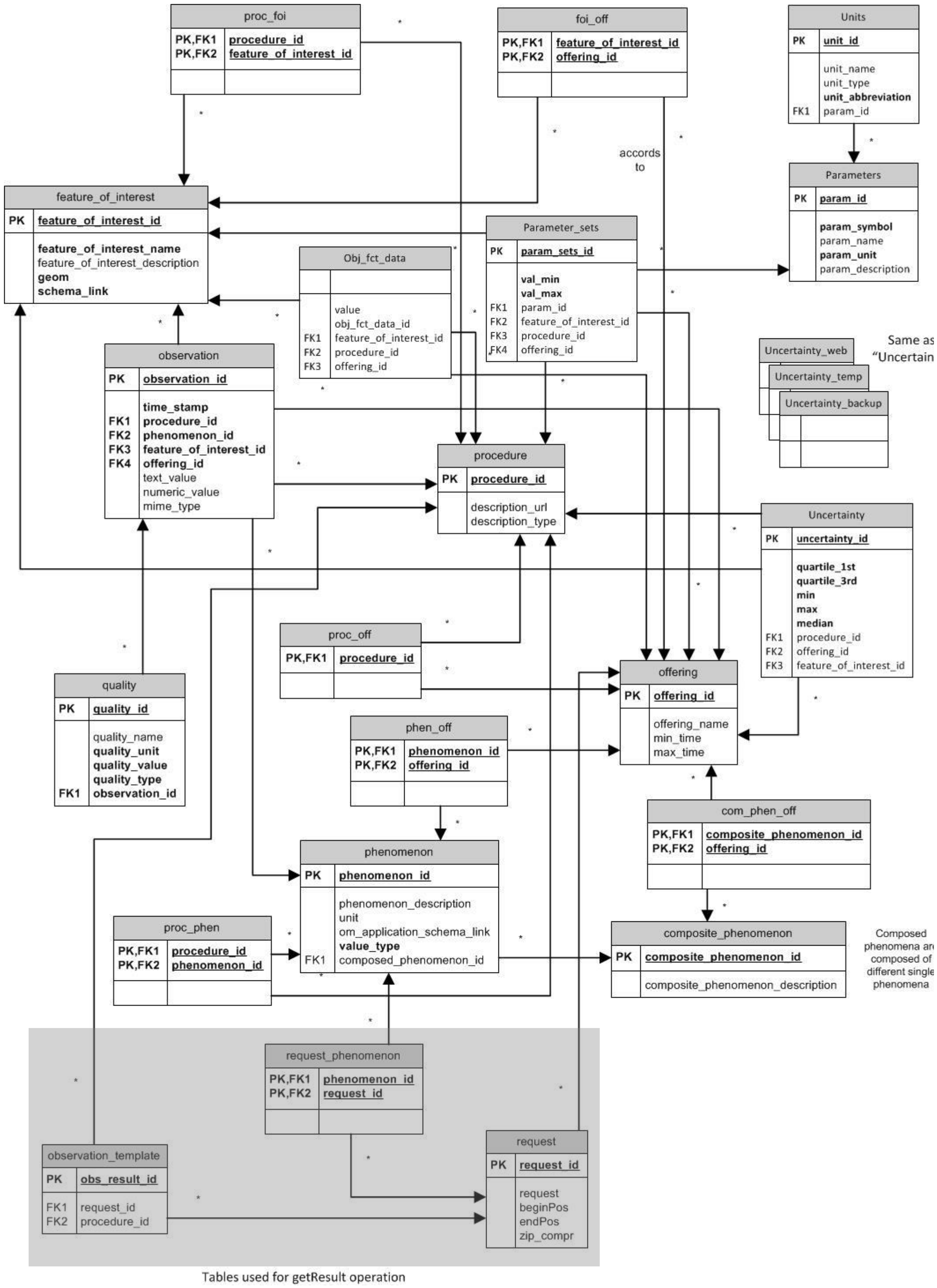
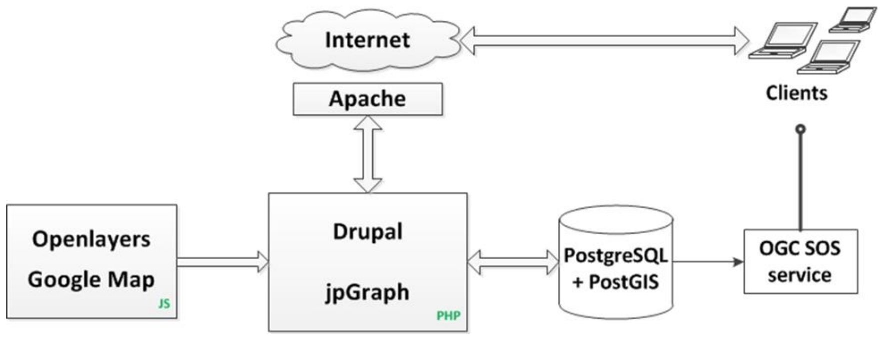


Figure 1: Data model

Google Map APIs are used to display the catchment gages layer overlaid on Google Map. This allows the user to locate the catchment for which data are to be requested or submitted.



V- SUMMARY

We have outlined the development of a prototype hydraulic modeling uncertainty data store that can host uncertainty data together with some basic statistical characteristics about it. This development consists of a data model that we modified from CUAHSI's ODM schema and have implemented as a postGreSQL instance. The associated metadata is derived in part from the Sensor Observation Service, SOS, requirements which we use to publish the uncertainty data, in part from what the ODM considers basic requirements, and in part what the UncertML conventions require, which we use to encode the uncertainty data.

We have also developed an online system that permits submission of new uncertainty data and searching and viewing already submitted uncertainty data. The system is flexible enough to permit the submission of a new performance measure dataset to an already existing watershed and existing models, or to add a new model to an already existing watershed for which a set of performance measures have already been computed, or both. Or one can submit a completely new watershed with new uncertainty data.

We are currently limited to using lumped models such as the one mentioned earlier (HBV and HYMOD), which compute performance only at the outlet point (discharge). A next step forward would target the inclusion of semi distributed models, and ultimately fully distributed models that use grids and meshes as a discretization framework on which to integrate the governing equations. We would not anticipate major work to extent our current data model to include distributed models where one would compute potentially many performance measure values (instead of just one) per parameter set run.

We are also targeting the possibility of including time variant uncertainty data and cluster these data by month or season to display sub annual characteristics that may change over the course of the year. Additional action items concern the expansion of the statistical representation to include confidence levels and probability distributions. These would then also mean that we need extend our visualization capabilities both on the statistics as well as the added dimensions on the data (time axis).

Finally, we have loaded uncertainty data for the 392 catchments of the MOPEX experiment that used 4 different performance measures and 2 different hydrologic models. The system has been set up on The City College of New York servers at the CrossRoads Initiative and will be accessible at http://uncert.net.

VII- References

1. Nash, JEa, and J. V. Sutcliffe. "River flow forecasting through conceptual models part I – A discussion of principles." *Journal of hydrology* 10.3 (1970): 282-290.
2. Kollat, J. B., Kollat, J. B., P. M. Reed, and T. Wagener. "When are multiobjective calibration trade-offs in hydrologic models meaningful?." *Water Resources Research* 48.3 (2012).
3. Mohamed E. Elshamy, “Assessing the Hydrological Performance of the Nile Forecast System in Long Term Simulations”, *Nile Basin Water Engineering Scientific Magazine*, 1 (2008)
4. Rasouli, Soora, T. A. Arentze, and H. J. P. Timmermans. "Analysis of Uncertainty in Performance Indicators of a Complex Activity-Based Model: The Case of the Albatross Model System." *Proceedings innovations in travel demand modelling conference*, Tampa, FL. 2012.
5. CUAHSI (2014), Consortium for the Advancement of Hydrologic Sciences, Inc., web-site accessed April 2014.
   <http://www.cuahsi.org>
6. Bröring, Arne, Christoph Stasch, and Johannes Echterhoff. "OGC Sensor Observation Service interface standard, version 2.0." *OGC Document* 12-006 (2012).
7. Arne Bröring, Christoph Stasch, Johannes Echterhoff. "OGC Sensor Observation Service Interface Standard." *OGC* 12-006 (2007).
8. Krause, P., D. P. Boyle, and F. Bäse. "Comparison of different efficiency criteria for hydrological model assessment." *Advances in Geosciences* 5.5 (2005): 89-97.
9. "JpGraph." - *Most Powerful PHP-driven Charts*. N.p., n.d. Web. 16 Apr. 2014.
   <http://jpgraph.net/>
10. "SOS Data Modeling" - *52°North SOS data models*. N.p., n.d. Web. 16 Apr. 2014.
    <https://wiki.52north.org/bin/view/SensorWeb/SosDataModeling>
11. "Drupal." - Open Source CMS. N.p., n.d. Web. 20 Apr. 2014.