Provenance in Streamflow Forecasting

Heiko Müller, Chris Peters, Yanfeng Shu, Andrew Terhorst
Intelligent Sensing and Systems Laboratory
Commonwealth Scientific and Industrial Research Organisation (CSIRO)
GPO Box 1538, Hobart TAS 7001, Australia
{heiko.mueller, chris.peters, yanfeng.shu, andrew.terhorst}@csiro.au

ABSTRACT

Within this extended abstract we describe a data set of provenance traces that we collected over the past two years for a continuous streamflow forecast in the South Esk river catchment in Tasmania, Australia.

1. MOTIVATION

Allocation of water resources in Australia is highly regulated. The Australian Bureau of Meteorology is currently implementing continuous short-term streamflow forecast models in selected river catchments across the country. Such models predict the amount of water flowing through a catchment over a period of several hours to several days. Model outputs can be used to inform water resource management decisions. Model runs are executed every few hours. Forecasts may be based on observations coming from sensors operated by different agencies or other predictive models. Provenance tracking is critical to improve the confidence and decision-making capacity of water managers.

Figure 1 shows the workflow for a streamflow forecast model set up for the South Esk River catchment in Tasmania, Australia. The South Esk Flow Forecast (SEFF) is driven by data from hydro-meteorological sensors integrated into the South Esk Hydrological Sensor Web [2], and 72 hour rainfall predictions from the CSIRO Conformal-Cubic Atmospheric Model (CCAM) [3]. The first step in the workflow transforms rainfall observations into a regular one kilometre grid structure using a Kriging algorithm [1]. CCAM data is provided directly in gridded format. Both data sets are then interpolated into rainfall data for 49 sub-catchments in the South Esk. In the final step the gridded rainfall data and streamflow observations from five gauging stations are input into a rainfall-runoff model (GR4J) [5]. The model predicts streamflow at each of the five gauging stations over a period of 72 hours. The workflow is executed every hour.

2. EXPERIENCE

Different parts of the South Esk Flow Forecast (SEFF) workflow are executed on different machines. The workflow is "orchestrated" by a collection of scripts that run on local machines and copy files over a network. Input data, output data, and configuration parameters for every step are stored in files on local disks. Information about dependencies and data flow in SEFF can only be gained from knowledge about the process structure and local log files.

We implemented a provenance management system for SEFF to provide systematic access to provenance information. The system collects data from different machines into provenance traces for individual workflow runs. We use a centralized web service for storage and retrieval of information. A major challenge is to maintain a large number of overlapping time series data. On average, there are almost 200,000 data points in the input data for each SEFF run. The average total number of data points in each run is 450,000. There is a high overlap in the data of consecutive runs since the time period for input and output data is usually much longer than the frequency with which the workflow is run.

We address the storage problem by using the archive management system XARCH [4]. XARCH allows for compact storage of evolving time series. That is, we maintain all versions of input data, intermediate results, and output data in a compact data structure. Data in XARCH is stored in XML format. Documents in XARCH follow a key structure that provides a canonical identification for each element. We make use of this feature by implementing a web service that exports archived data (either as XML, JSON, or RDF) with resolvable URIs for individual elements. Provenance information is stored as labeled graphs that connect documents and elements within different archives to reflect dependencies and data flow. We also provide links to external data sources providing information about sensor assets. Details of the implementation are outside the scope of this paper.

3. APPLICATIONS AND QUERIES

Water managers base their decisions on forecast results. Provenance for individual forecasts allows users to verify that the right data was used. Typical user queries are: *Which sensors were used*, *Which sensors reported new data*, and *Have all sensors been calibrated within the last 6 month*? Comparing provenance traces is another important application area. Users are interested in the differences between runs, especially if the quality of the results differs significantly. Keeping provenance information is helpful not only to highlight the differences but also to identify how differences in the input data relate to differences in the predicted values.

Sets of historic model runs are helpful to identify problems in a sensor network, e. g., *Which sensors fail frequently to report data on time*? For us, however, the main motivation for collecting these provenance traces is to assess and improve the quality of future forecasts. The performance of a forecast model is assessed by comparing the predicted values with the later on observed ones. Note that we provide references to these observations in our data set as

EDBT/ICDT '13, Mar 18-22 2013, Genoa, Italy

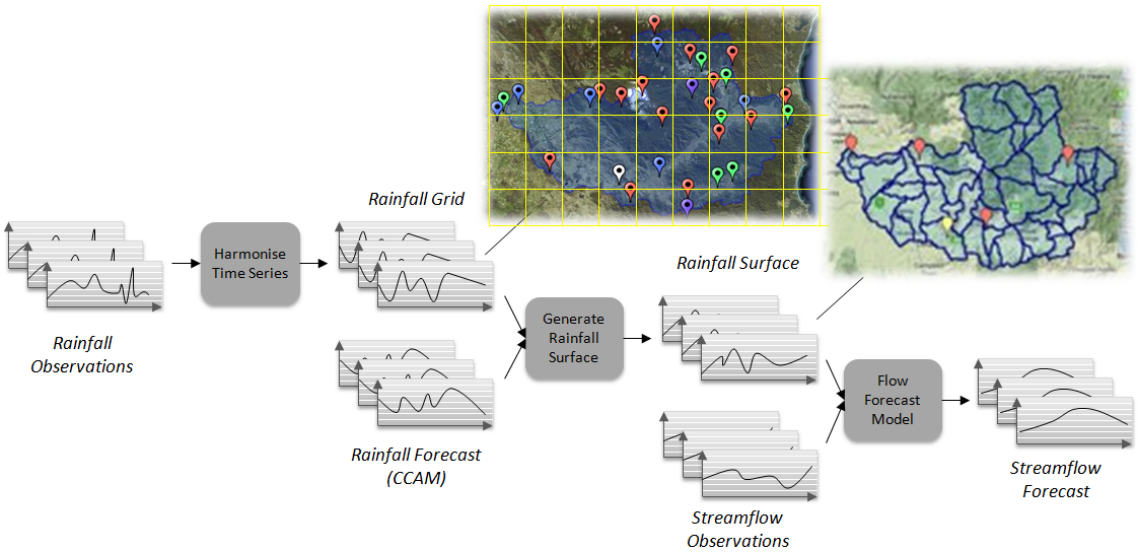

Figure 1: Workflow for streamflow forecasting in the South Esk river catchment in Tasmania, Australia.

Table 1: Summary of submission

Data format	RDF/XML
Data model	PROV-O, O&M, and domain concepts
Size	Average number of triples per trace is approx. 1,000,000
Tools	XARCH, Java, Apache Jena
Application domain	Hydrology
Submission group	Intelligent Sensing and Systems Laboratory, CSIRO
Contact	Heiko Müller heiko.mueller@csiro.au
License	Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)

well. Having access to input data and intermediate results of each model run allows the use of machine learning algorithms to identify patterns in the data that are characteristic for scenarios of good or poor model performance. This information is used to provide users with confidence values for future model runs based on the presence or absence of certain patterns in the data. The insights gained by the machine learning algorithms are also important to modify and improve the forecast model.

4. SUMMARY

SEFF provenance data will be made available in RDF, XML, and JSON format on the World Wide Web. The provenance database is updated as new traces become available. In this submission we provide a current snapshot of the data in RDF format. This snapshot contains over 4,000 provenance traces collected over a period of almost ten month. Table 1 gives a general summary of the data set.

Whereas formal ontologies are still being developed for the hydrology domain, we use different ontologies to represent provenance information, namely PROV-O and a streamflow forecasting provenance ontology. The latter one is an adaption of the ontologies published in [6]. We also use concepts from the OGC's Observations and Measurements specification (O&M) [1]. The use of PROV-O concepts and properties is summarized in Table 2.

[1]http://www.opengeospatial.org/standards/om

Table 2: Coverage of PROV-O concepts and properties

PROV-O terms	Covered (Y/N)
prov:Activity	Y
prov:Agent	Y
prov:Entity	Y
prov:actedOnBehalfOf	Y
prov:endedAtTime	Y
prov:startedAtTime	Y
prov:used	Y
prov:wasAssociatedWith	Y
prov:wasAttributedTo	Y
prov:wasDerivedFrom	Y
prov:wasGeneratedBy	Y
prov:wasInformedBy	Y

5. REFERENCES

[1] F. P. Agterberg. *Geomathematics. Mathematical background and geo-science applications.* Elsevier Scientific Pub. Co., Amsterdam, New York, 1974.

[2] S. M. Guru, P. Taylor, H. Neuhaus, Y. Shu, D. Smith, and A. Terhorst. Hydrological sensor web for the south esk catchment in the tasmanian state of australia. In *IEEE International Conference on eScience*, pages 432–433, 2008.

[3] J. L. McGregor and M. R. Dix. An updated description of the conformal-cubic atmospheric model. In *High Resolution Simulation of the Atmosphere and Ocean*, pages 51–67, 2008.

[4] H. Müller, P. Buneman, and I. Koltsidas. XArch: archiving scientific and reference data. In *ACM SIGMOD*, 2008.

[5] C. Perrin, C. Michel, and V. Andreassian. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279(1-4):275–289, Aug. 2003.

[6] Y. Shu, K. Taylor, P. Hapuarachchi, and C. Peters. Modelling Provenance in Hydrologic Science: a Case Study on Streamflow Forecasting. *Journal of Hydroinformatics*, 14(4):944–959, 2012.