Centre for Geospatial Science
OSGIS Nottingham 2010
Monday 21st / Tuesday 22nd June
OPEN SOURCE GEOSPATIAL

SECOND OPEN SOURCE GIS UK CONFERENCE – OSGIS 2010 WORKSHOP ON WORKFLOWS ON EARTH OBSERVATION 21ST JUNE 2010

EXTENDED ABSTRACT

Hydro-climatic monitoring in the Guyana Rainforest, South America – a real-world challenge for scientific workflow management

C. Isabella Bovolo[1, 2], Geoff Parkin[2], Thomas Wagner[2], Philip James[2]

[1] Iwokrama International Centre for Rainforest Conservation & Development, Guyana
[2] School of Civil Engineering & Geosciences, Newcastle University, UK

Introduction

A new hydrology and climate instrumentation program located in the remote Guyana rainforest and adjacent savannah, on the northern rim of the Amazon basin in South America, exposes several challenges and opportunities for scientific workflow and information management. Here, experiences and strategies in collecting, processing, storing and using earth observation data from *in situ* sensors in the tropics is discussed in terms of workflows, and challenges and opportunities for improvements in incorporating satellite-based and airborne remote sensing data are highlighted.

Workflows define a sequence of steps or operations necessary to carry out a complex combination of tasks such as raw data acquisition and retrieval from sets of sensors, data quality assurance, organised data storage and access, data processing and analysis and computational modelling. Following a set workflow minimises error, ensures repeatability and transparency, and allows a certain degree of efficiency and good scientific practices to be established. In scientific workflows, the assumption is that many steps or operations can be specified and automated over a distributed set of computing resources and infrastructures, thereby reducing the reliance on human inputs for repetitive tasks which can lead to errors. It is usual practice in the computing science community to develop workflows management systems for data processing, analysis and computational modelling, but the data acquisition and quality assurance stages are usually neglected. The challenge is to integrate all of these stages into a workflow and combine the use of multiple types of data.

Ideal workflow system

Ideally, *in situ* instrumentation measuring hydrological and meteorological parameters within the rainforest would operate autonomously by long-term battery or solar power,

would record data internally and send it via regular (or perhaps real-time) wireless telemetry systems, such as through a VSAT satellite link, to a computer hub located in a safe, controlled environment with a regular power supply and high speed internet link, where data would automatically be quality controlled, processed, analysed, stored and made securely available to a defined set of local or remote users. The computing hub would also allow users to integrate data collected *in situ* with all other available digital data such as historical observations, satellite-based and airborne remote sensing data, stored in useful, pre-defined, compatible formats with associated metadata in structured data archives. The combined datasets would then be able to be automatically visualised and used within hydrological modelling environments. Human inputs at the various stages would therefore be minimised except at the analysis and interpretation stage and for general instrument upkeep.

Challenges

Such a system, however, is costly and there are several practical issues to deal with making it a challenge to implement in the tropics. In particular, the high costs involved in setting up, maintaining and operating the infrastructure and instrumentation in a suitable long-term site are strong barriers.

Information & Communication:

In the Guyanan rainforest, information and communication options are limited. Satellite based internet access is available but signal strength is subject to interference by precipitation and moisture so is often poor in quality and cannot be relied upon. A good line-of-site between the instrument antenna and satellite improves performance and communication range but signal strength is reduced from obstructions such as walls, trees, foliage and concrete, and eliminated with metal objects or structures. There are no fixed-cable telecommunication infrastructure or cellular networks, rendering the usual GSM modems (requiring network coverage) and PSTN modems (requiring national telephone network) unusable; although telegraphy (Morse code), radio-based telephony (requiring antenna towers and repeater sites) and expensive mobile satellite phones are possible.

Power supply

Power is required for running the instrumentation (and any associated computing equipment) but different power requirements may be needed for each instrument (e.g. 240V or 120V). Generally batteries and renewable energy solutions such as solar energy provide the easiest power options, although solar panels can only be installed in clearings or above the forest canopy. Small scale hydropower is generally not an option in Guyana, despite the presence of several rivers because of the low head present in the relatively flat environment, whilst wind energy is generally more unreliable. A diesel generator can provide a limited, interruptible and variable power supply but requires a regular supply of fuel.

Other considerations

Other practical challenges include access to the area, dealing with lightning, heavy rainfall, high temperatures and high humidity. Instrumentation, infrastructure (including fencing), wiring, computing and power facilities are all at risk from insect, animal and human interference (including theft) and may be inundated by flood water during the wet-season. Furthermore a working system needs local expertise (requiring training and capacity building) and regular upkeep. Finally, care has to be taken to deal with the local socio, economic and political framework, to involve the local communities and agencies and to have a regard for intellectual property rights, especially regarding remote or international data transfers and storage.

Instrumentation Program

Current point-based observations in Guyana (such as precipitation timeseries streamflow) are discontinuous in time and spread unevenly geographically. Traditional methods of monitoring weather and river levels rely heavily on volunteers to manually collect and record data one or more times a day. In remote but inhabited areas, monitoring datasheets have to be sent by plane, boat or vehicle to the data archiving centre where information is digitised and stored.

The new climate and hydrology instrumentation program at the Iwokrama International Centre for Rainforest Conservation and Development, a 1 million acre rainforest site in central Guyana, goes some way towards automating the data collection process, allowing data to be recorded digitally (thereby removing digitisation errors) at a finer timescale than previously possible and in a more reliable way. Bearing in mind a limited budget and the limitations and challenges mentioned above, the following instruments have been newly acquired and installed within Iwokrama.

- *Two Automatic Weather Stations* (AWS) measuring rainfall, humidity, wind, solar energy, temperature and evaporation have been set up at the main Iwokrama rainforest research station and at the Bina Hill Research Institute further south in the savannah.
- *Five Tipping Bucket Raingauges* (TBRG) have been installed at various accessible sites within the rainforest. These are standalone units.
- *Five River Level Monitors* (Frogs) have been installed below bridges for ease of access. Two of these sites also include water quality probes.

The data derived from these instruments will form part of a long-term dataset used to establish current conditions and monitor environmental change. The data will also be used to model current hydrological processes and for climate change and landuse change impact studies.

Current Workflow System

The new monitoring system is still in its infancy and consequently the workflow processes are still in development. Currently we can define the following workflow.

Data is recorded internally on the AWS and is also downloaded automatically in real-time to a computer housed in a nearby building using a permanent cable link with a combined solar and diesel generator power supply. This land-line interface supplies both 24V DC power and communication signals to the remotely located AWS. The AWS can also run independently following disconnection or remote power failure using its own backup 12V battery supply powered by a solar panel array. Unfortunately, at this stage, if the external power supply is discontinued, the real-time download has to be manually re-started.

On standalone instruments, such as the TBRG (powered by 4 AAA batteries), Frogs (powered by Lithium battery packs lasting approximately 7 years) and water quality probes (4 AA batteries), data is recorded internally and has to be manually downloaded once a month or so once the data storage system is full, onto a rugged, waterproof handheld computer via a serial (RS232) cable. At this stage it may also be necessary to manually re-calibrate the instruments and replace batteries. Data then has to be manually downloaded from the handheld computer to a computer with storage capacity.

Once the raw data has been retrieved from the various instruments, the data has to be quality assured and particular care taken to ensure the clocks of different instruments are synchronised and that seemingly similar data from different instruments is equivalent (for example one instrument may record a day starting at 8am whilst another would record a day starting at midnight). Obvious erroneous values can be removed using basic software whilst other errors or problems are currently identified using graphing packages and user know-how.

Once the data has been quality controlled and has been associated with metadata (either automatically or with human input), it needs to be stored securely and accessed remotely for processing. Due to the poor communication options in the rainforest, this aspect of the workflow is particularly difficult. When available, the internet can be used to send data to a data hub, otherwise data has to be physically sent (on a storage device) to the data hub where it can be uploaded to the main data archive (still to be made operational). Climatic and hydrological datasets can rapidly reach several gigabytes in size. Options are currently being sought to host the archive in Georgetown (the capital city) where fast internet, a continuous power supply and good computing resources would be a challenge or remotely in the UK where issues of data access, permanent storage and ownership (IPR) are being raised. However, given the principle that those with the most to lose are the best custodians, a sensible approach would be to enable a data centre to be maintained within Guyana. Although hardware and bandwidth costs will always be applied, one area where costs can be managed is through the implementation of open source products that have no purchase cost and no ongoing maintenance costs. For Earth Systems Engineering an added benefit of open source data platforms is that these products have more in-built support for international data and interface standards than their proprietary rivals. An example of such is the Geoserver product (www.geoserver.org) that is the reference implementation for the Web Map Services and Web Coverage Services standards specification that provide geospatial coverage such as remote sensed imagery, climate data etc. These tools have in-built capabilities for data format conversion (e.g. shape files to KML) and re-projection on the fly, removing some of the potential bottlenecks in data re-use. Adopting an open source platform and formats based on international standards may require extra processing steps to provide inputs into modelling programs but should provide a solid platform for longer term collection, archive, deployment and delivery of environmental data in the region.

Once it has been made available to a user through a remote data archiving system, data can be assimilated with other types of datasets for analysis. Each instrument stores data in different formats, although comma-separated values are often available thereby aiding automatic data processing. Combining various data from the field site instrumentation is, therefore, not a major problem. However combining digital point-based observations with other historical site observations or geospatial data automatically is more of a challenge due to different spatio-temporal scales, different formats and various licensing issues and is still an area that needs development.

Future challenges and opportunities

Temporal data (e.g. climate, river levels, discharge) and geospatial data (e.g. DEMs, land cover types, soil types) are used as inputs to various hydrological models. The initial basis for such models is a DEM from which catchment boundaries and drainage networks can be automatically derived using GIS software. However, good quality digital geospatial data are not generally available for Guyana. Satellite remote sensing

systems for the tropics need to penetrate cloud and canopy cover in order to provide the necessary ground elevation data for hydrological applications. Remote sensing imagery also needs to be ground-truthed before use which is difficult in this type of environment. Although some data exists, good quality, fine resolution data is generally costly. Likewise, maps of vegetation cover are available, but need ground-truthing and soil maps need verifying and further soil geochemical analysis is needed. Hydrology models are usually forced by time and space varying climatic (e.g. precipitation, temperature, evapotranspiration) data. Precipitation in Guyana is at times highly localised so interpolating point observations over a wide geographic area may therefore result in error. Automatically combining satellite or radar images with *in situ* climate observations (as is the case in more developed countries where data is available) would be particularly useful. (In Guyana, a new Doppler radar system has been installed near Georgetown but its coverage does not unfortunately extend as far as the field-sites.) New technologies such as Google Earth allow geospatial data to be combined with some point-observations for visualisation and analysis purposes but workflows integrating data retrieval, quality control, storage and access, processing and analysis using multiple datasets for hydrological modelling and catchment management purposes still need to be developed.

Environmental and technical issues notwithstanding, a framework for acquiring and publishing sensor data such as the OGC Sensor Web Enablement (SWE) framework could provide a starting point for managing heterogeneous sensors. The benefits of this type of approach are that the data is standardised (in structure) and self describing such that other processes can be built more easily to handle the data. Technology candidates such as the OGC Web Processing Service (OGC WPS) can also provide relatively simple means of wrapping geospatial processes around sensor outputs, although the WPS specification needs to be modified to manage asynchronous processing. Standards based architectures like these, whilst potentially increasing initial development time, provide a potential route to a rich, service based process architecture that allows for the rapid incorporation of new data and processes. An additional benefit of this approach will be the utilisation of the wide range of tools and services that already support geospatial standards to deliver technology such as GeoServer, OpenLayers and PostGIS.

Conclusion

Scientific workflows in hydrology (as for many earth science applications) are still in their infancy, but there is a great potential for workflow technology to enhance scientific practices, minimise error and ensure repeatability and transparency. Scientific workflow tools are generally developed and used to perform complex analysis on scientific data, however workflows also need to incorporate raw data acquisition and processing. Data telemetry can simplify and speed the acquisition of critical information from remote locations but this is not always possible. Furthermore workflows need to be developed to allow multiple datasets to be automatically processed and combined. Data also needs to be stored remotely and accessed from different geographical locations. The main challenges therefore, include the automatic retrieval and quality control of datasets, managing and processing of large volumes of data (including dealing with access issues), the integration of point-based and geospatial heterogeneous datasets and automatically visualising, modelling and analysing these datasets. A new climate and hydrology monitoring program in Iwokrama, Guyana, demonstrates many of the challenges and opportunities involved in defining workflows in the tropics.

Document reference:
Bovolo, C. I., Parkin G., Wagner T., James P. Hydro-climate monitoring in the Guyana Rainforest, South America: A real world challenge for scientific workflow management 2010. *Proceedings of the Second Open Source GIS UK Conference (OSGIS2010), Workflows for Earth Observation Systems workshop* (available from http://www.staff.ncl.ac.uk/isabella.bovolo/)