3 Modelling in the EVOp cloud

The EVOp portal deploys interactive web applications to model local flooding and diffuse pollution. From a user perspective, web-based environmental models do not require any software installation as they are accessed using standard web protocols using an internet connection (e.g., http). As a result, they are platform independent and can be easily integrated in workflows as part of the process of environmental data retrieval, manipulation, visualisation, and communication. To maximize efficiency and interoperability with storage, interaction and visualisation tools, the EVOp team have adopted the Web Processing Services (WPS) standard of the Open Geospatial Consoritum (OGC), which consists of a wide range of industry and academic actors. Additionally, several stable software implementations of the OGC standards exist, which greatly facilitated the development of

3.1 Selected models and web service implementation

For the purpose of this project, three different environmental models were implemented as web services:

- TOPMODEL (Beven and Kirkby, 1979) was chosen to demonstrate the local exemplar. Topmodel is a widely used and well documented semi-distributed, conceptual hydrological model. It was originally developed for small, upland catchments in a humid environment, and is therefore very suitable for the catchments selected for the local case studies. The model produces a time series of river discharge and spatial patterns of soil moisture.
- FUSE (Clark et al., 2008) is a state-of-the-art modelling toolbox which includes well established algorithms including VIC and the Stanford Watershed Model. As a toolbox, it allows combination of many different algorithms, thus creating over 1000 possible model structures. As such, it provides a very flexible solution for the large number of catchments to be modelled in the national case study. The model produces a time series of river discharge.
- Export Coefficient Model (ECM) (Johnes et al., 2007) is an established approach to allow prediction of nutrient (nitrogen and phosphorus) flux from land to inland and coastal waters. The predictions are based on agricultural practices, such as stocking densities, fertiliser use and crop types, human population density and atmospheric deposition, hence making the model suitable for rapid mitigation scenario development and testing. The model produces predictions from field to national scale and generates mapped outputs and summary tables of nutrient export by source and practice.

These models were written in different programming languages (C, Fortran, C++, and MS Excel). In order to integrate them into a common web service infrastructure, TOPMODEL and FUSE were integrated in the R data analysis platform, either by wrapping R code around the original implementation (for TOPMODEL), or by re-implementing them directly in R (for FUSE). TOPMODEL is already available as an R package, and FUSE has also been made available. A major advantage of R is its convenient integration with PyWPS, a Python implementation of the OGC WPS. The communication between Python and R is facilitated by an existing connector (RPy2). A local PostgreSQL database was used for data storage because a web-based implementation such as the OGC Sensor Observation Service was deemed unfeasible within the context of the pilot.

The ECM was originally implemented in Excel and hence a new implementation of the model was written for this project in a combination of PHP (on the cloud side) and JavaScript (on the desktop side). This language and platform split in the implementation means that the cloud was able to perform the query of the spatial database for the area of interest and the bulk of the model calculations. The summary tables are then passed to the final calculation in the browser. Keeping the final stage of the calculation on the desktop means that the user can undertake scenario testing, such as altering the livestock numbers in a catchment, with the results being very rapidly calculated and displayed. The main spatial database is stored in a MySQL database and the outputs of the modelling can be rapidly mapped at any spatial scale for visual interpretation of model predictions.

3.2 Workflow integration and applications

The modelling web services were integrated in the workflows of two EVOp applications to showcase their potential. These applications are accessible via the main EVOp portal and focus on simulating the impact of land-use changes on flood risk, and diffuse pollution and water quality.

The communication between client (e.g. modelling web portal) and server is defined by the OGC WPS standard. The web client sends the server a HTTP GET request to execute the process. Once the execution terminates, the server sends back an XML response which is parsed at the client side extracting the simulated time series. Data are retrieved by querying the database in real time while the visualization of time series is based on the Google-charts API and similar libraries.

The flood modelling tool is available at the following four locations:

- River Eden catchment:
  - ❑ Dacre Beck at Dacre Bridge, Cumbria, England.
  - ❑ Blind Beck, Cumbria, England.
- Dyfi catchment: Dyfi Bridge, Machynlleth, Wales.
- Tarland catchment: Coull Bridge, Aberdeenshire, Scotland.

It consists of the following user-interactions:

i. Selection of the location of interest: Based on the user input, observed river flow and precipitation data, as well as pre-calibrated model parameters are queried from the database.

ii. Modification of the land-use scenario: The user input is then mapped onto modifications of the optimal model parameters. At this stage of the EVOp, this conversion is based on an empirical parameter translation table. However, theoretical research on how to parameterise user scenarios is on-going.

iii. Subsequently, the model is run with the modified parameter values and the results are visualised in the web interface. The model can be run many times with different scenarios. The simulated discharge time series are kept while the user session is active. Multiple simulations can be selected and visualised together to facilitate the comparison of different scenarios. The plotting functionality also allows for the visualisation of an estimate of the streamflow threshold above which the flow spills over the banks causing possible flooding.

The diffuse pollution modelling tool has been implemented at the full UK scale at a spatial resolution of 4km^2. To run the model, the user makes a series of choices:

- Select an area of interest: These are based on countries, OSPAR zones, coastal drainage zones, River Basin Districts and river or water quality catchments based on the locations of gauging sites.
- For the area of interest, specify various options connected with the model setup, although not all options are enabled in the demo:
  - ❑ Selection of the model.
  - ❑ Selection of the data set to run of the model, such as land cover data from different years.
  - ❑ Selection of the hydrological model.
- For the selected model configuration, the user can choose to run the model with the observed historic land management, or to test a set of mitigation options. The mitigation options are 'Good Agricultural Practice', 'Catchment Sensitive Farming', 'Mitigation through on-farm measures', 'Farming for WFD compliance' and 'Urban Waste water Treatment Directive Plus (UWwTD+)'. Each of these mitigation scenarios makes suitable adjustments to the configuration of the landscape to represent the change. These mitigation options can be spatially targeted at the level of the geoclimatic region (targeting different measures according to the key characteristics of each region).
- The results are then presented as tables, maps or charts for both the observed historic, and the developed scenario.

3.3 Uncertainties

Uncertainties remain a major issue to tackle. The project reviewed the relevant technologies that are available to document uncertainties in a web services context. One of the most promising initiatives is UncertWEB which aims to provide data models and mark-up tools for the propagation of uncertainties in modelling workflows and communication of uncertainties to the end-user. Experiments are on-going to integrate uncertWEB (Williams et al., 2010) in the FUSE modelling toolset by means of adding UncertWEB data models to the object-oriented class definitions in the R toolbox.

At the same time, efforts were focused on integrating the major uncertainty quantification methods in the modelling toolbox. The following methods are currently available:

- Generalized Likelihood Uncertainty Estimation (GLUE, Beven and Binley, 1992, Beven and Binley 2013).
- Differential Adaptive DREAM (Vrugt et al., 2009).
- Bayesian Total Error Analysis (Thyer et al., 2009).

These implementations are a combination of efforts from the R development community and the EVOp project.

Lastly, efforts have been aimed at developing an objective model structure selection algorithm, based on methods for data mining of model performances, and the extrapolation of model structures in space and time.

Many of the abovementioned activities need further attention. Even though functional implementations of each of the algorithms exist, major technological and conceptual challenges are still to be overcome. These, and others, should be explored further in eventual follow-up projects to EVOp.

3.4 Opportunities

The future opportunities for linking the hydrological and biogeochemical models can be seen in both research and decision support focused areas. For example, the tighter linkage of the model components or models could help to investigate the grand scientific environmental challenges on, for example, the influence of hydrology on processes controlling losses of nutrients (N, P and other macronutrients) from land to surface waters, the location of critical source areas for diffuse pollution in the landscape and how these export processes and locations may alter under projections of climate change. This need for tighter coupling has been recognised in the EU SEAMLESS project, which is aiming to overcome "fragmentation in research models and data in Europe for assessing agricultural systems". EVO has an opportunity to integrate more closely with this, and other, projects.

The modelling community is moving from an approach where one model approach / structure was deemed sufficient to understand and predict system properties to approaches that embrace ensembles of approaches and model structures. This change in approach creates an opportunity for EVO since cloud computing offers a solution to the greater computational resources required, and simplifies the testing of different models (approach and structure) against the same datasets to provide new insights. EVOp could have a significant role in supporting the community and advancing the science by providing a database (and metadata) on a wide range of catchment studies that could be used by the wider community. Through better sharing of data and approaches (as has been shown in molecular and marine sciences) EVO can act to accelerate progress in the environmental sciences.

Taking a high level view of the possible future opportunities, there are many possible ways in which EVO could add value to science and society:

- The range of predictions could be extended through linking other catchment hydrological and biogeochemical models into the framework and allowing the cross comparison between the different predictions. As each new model is included, the potential number of combinations significantly increases, requiring the computation resource available in the cloud. Examples include CAPRI DNDC, PSYCHIC, MITERRA, INITIATOR, INCA, RiverStrahler, CRUM3, SCIMAP and PIHM. Research is required to understand which model couplings yield the greatest information gain and how to interface different type of model output (resolutions and types of predictions)
- Time and space scales. Finer spatial scale models, such as SCIMAP, at the 5x5 metre resolution to compliment national models, such as the current ECM. Temporally dynamic models such as HBV light, INCA or CRUM3. The temporally dynamic models could enable real time flow and or WQ predictions for water companies which could be used in abstraction or discharge planning.
- Testing old models with new data. DTC N and P time series are the required level of resolution to test detailed process catchment models. The DTC datacentre could provide data series as web services (mid 2014) and used to drive simulation models. Linking the data stream with the modelling tools could create a powerful platform to test approaches and hence encourage developers to make their tools compatible with the EVO approach.
- There are opportunities to develop mobile applications that make use of the cloud hosted models to deliver support services for CSFOs or for teaching and learning at all levels of education (flood models, mitigation scenarios and augmented reality).
- There is a growing trend towards open source models and publishable workflows (myExperiment, R and iPython notebook), which share best practice and help others to use techniques. Linking the EVO hydrological and biogeochemical models into these tools increases their accessibility and value.
- Stakeholder engagement earlier in the modelling process. Through having a range of different hydrological and biogeochemical models available, it is possible to work closely with stakeholders to understand their problems and build a tailored solution rather than focusing the current limited set of tools that may be available to that user.

3.5 Barriers

The barriers that have been identified are:

- IP on models and datasets. This cannot be understated.
- The required data may not exist, may be protected by IP, not be publicly available (governmental or commercial issues), or be lost. There is therefore the need to make as much data available as possible and to ensure that the datasets that are collected in the future have maximum information content and well documented metadata.
- Modellers and modelling groups may see the conversion of existing models to a web service approach difficult or not worthwhile. If there could be a modelling portal that had licences for the key datasets (e.g. NextMap 5m product, LCM 2007, geology, soils, rainfall, discharge and water chemistry data), then this would be a significant pull for developers to move their tools to the platform.
- There are significant challenges with linking models together to create a solution. This relates

to the issue that models may have been written for a different reason, work at different space and time resolutions, and the feedback between processes represented in different models may not be captured.

- As models become more advanced, the computational cost increases significantly. The computational resource is there via cloud computing technologies but there needs to be a method to cover the costs of using the resource. Precedents for this exist in HPC (e.g. Hector) although a more dynamic 'economy' could be created with 'rewards' for sharing and creating tools and computational resources and 'costs' for using tools and computational resources.
- There is the issue of standards for model description, parameterisation and output. There are a few standards for model output (e.g. WaterML) but there is no currently defined standard for the communication of the uncertainties in predictions and measurements between models.
- The need for further development of methods to propagate uncertainties within modelling workflows.
- The need for methods to communicate uncertainties to end-users.

3.6 References

Beven, K., & Binley, A. (1992). The Future of Distributed Models: Model Calibration and Uncertainty Prediction. Hydrological Processes, 6, 279-298.

Beven, K. J., and Binley, A. M., 2013, GLUE, 20 years on. Hydrol. Process. DOI: 10.1002/hyp.10082.

Beven, K. J. and M. J. Kirkby, (1979), A Physically Based Variable Contributing Area Model of Basin Hydrology, Hydrological Sciences Bulletin, 24(1): 43-69.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735.

Johnes, P.J., Foy, R., Butterfield, D. and Haygarth, P.M. (2007) Land use for England and Wales: evaluation of management options to support 'good ecological status' in surface freshwaters. Soil Use and Management, 23 (Suppl 1). pp. 176-196. doi: 10.1111/j.1475-2743.2007.00120.x

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research, 45, W00B14.

Vitolo, C., Buytaert, W., McIntyre, N, Reusser D., and the EVOp team. Data Mining of Hydrological Model Performances with the FUSE framework. In preparation.

Vrugt, J., Ter Braak, C. J. F., Gupta, H. V., & Robinson, B. A. (2008). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling ? Stoch Environ Res Risk Assess. doi:10.1007/s00477-008-0274-y.

Williams, M., Cornford, D., Bastin, L., & Pebesma, E. (2010). Uncertainty Markup Language, OGC Discussion paper 08-122r2.

Related links

EU SEAMLESS project

http://www.seamlessassociation.org

FUSE

https://f-forge.r-project.org/projects/r-hydro

TOPMODEL R-package

http://cran.r-project.org/web/packages/topmodel/

Uncertainty Markup Language

http://www.uncertml.org/

Uncertweb discussion paper

http://www.uncertweb.org