LOCAL PARSIMONIOUS DATA-DRIVEN MODELS IN STREAMFLOW FORECASTING

DIMITRI P. SOLOMATINE

Associate Professor, UNESCO-IHE Institute for Water Education
P.O. Box 3015, 2601 DA Delft, The Netherlands
(Tel: +31-15-2151-715, Fax: +31-15-2122-921, e-mail: d.solomatine@unesco-ihe.org)

Streamflow forecasting in typically based on the extensive use of models. Usually a mathematical model is based on the description of behaviour (often physics, or first-order principles) of a phenomenon or system under study. They are often referred as process, simulation, or *physically-based* models. Another approach is based on the analysis of all the data characterising the system under study. A model can then be defined on the basis of connections between the system state variables (input, internal and output variables) with only a limited knowledge of the details about the "physical" behaviour of the system. Statistical models, like a linear regression model, follow this approaches well as more complex models like artificial neural networks. Such models can be called *data-driven models* (DDM).

One can see the increasing popularity of DDM due to the larger amount of data available to river managers; the advances in machine learning, and in difficulties in calibrating the physically-based models. A typical application of a DDM is in collecting all available data characterizing the process important for flow modeling, and using it to train a model. Many natural phenomena, however, are multi-stationary, are composed of a number of processes and their accurate modeling is not possible by building of one single ("global") model. When a DDM is built the training set can be split into a number of subsets, so that the separate, local models can be trained. The overall model can be referred to as the *modular model (MM)* (Fig. 1).

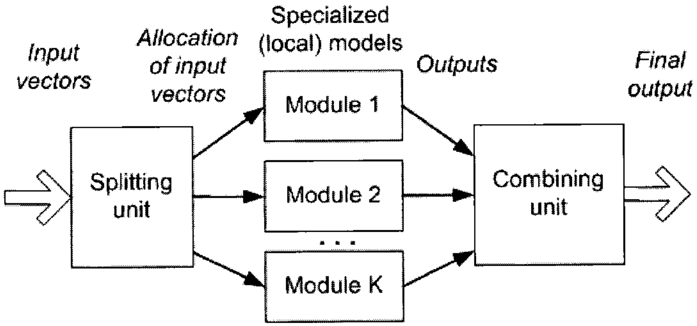

Fig. 1 Modular model

Modular models include components each of which is responsible for a particular hydrological condition, and they are typically more accurate than the overall global models. The component local models can be made simpler than the overall model; such parsimonious (often linear) models are better accepted by domain experts. The use of the

M5 model tree algorithm for flow forecasting was demonstrated by Solomatine and Dulal (2003) and Solomatine and Xue (2004).

In real-life situations managers and domain experts are interested in incorporating more domain knowledge when building models. In the context of MMs such role is in bringing in hydrological knowledge into the rules for filtering the data subsets to each model, in the choice of these models and in the way they are combined.

Incorporation of domain knowledge into the training process of the integrating unit is an important problem. A typical machine learning algorithm minimizes the training (cross-validation) error seeing it as the ultimate indicator of the algorithms performance, so is purely data-driven. Domain experts, however, may have other considerations in judging the model quality, and want to have certain control over the modeling process. These models could be not only DDMs based on machine learning, but also physically-based models based on the description of the underlying processes.

Flow forecasting and the related physically-based (process) modelling activities typically involve domain experts (hydrologists, hydraulic engineers and managers);.they pose the modeling problem, identify the data sources, to set up a model and calibrate it, perform the model runs, interpret the results and, possibly, to make decisions. The involvement of experts in DDM, however, is much lower, and this does not help the acceptance of such models. Challenge here is to integrate the background domain knowledge into a DDM by allowing the experts user to determine some important structural properties of the model based on the physical insight, and leaving more tedious tasks to machine learning.

In building modular models, a possibility to include a domain expert is to make it possible for him/her to construct the rules performing the data split for local models. Solomatine and Siek (2004) developed and implemented a version of M5 regression algorithm called *M5flex*. It makes it possible for a domain expert to make decisions about the splits at each node (as opposed to the standard M5 algorithm where this is done automatically). This method enables the user to determine split attributes and values in the top-most nodes, and then the M5 (or M5opt) machine learning algorithm takes care of the remainder of the model tree building. In the context of flow prediction, for example, the expert user can instruct the M5flex to separate different hydrological conditions to be modeled separately. Hence, the M5flex model trees can be more suitable for hydrological applications, than ANNs.

Three case studies (catchments in China, Nepal and Italy) were considered. A global method (ANN) and modular local modelling methods (M5', M5opt, M5flex) were compared. The following can be concluded:

- the modular models allow for building accurate local specialized models that can capture the details of the processes characterized by the certain hydrological conditions;
- local models can be made parsimonious and more transparent for decision makers;
- the acceptance of data-driven models depends on how well the domain knowledge is incorporated into a model, and the M5flex algorithm helps in this;
- the presented modular models are data-driven, but various types of models can be combined: physically-based models with DDM. The choice of a particular model type should be determined by the data availability and the desired forecast horizon.

REFERENCES

Solomatine, D.P. and Dulal, K.N. (2003). Model tree as an alternative to neural network in rainfall-runoff modelling. *Hydrological Sciences J.* 48 (3), 399-411.

Solomatine, D.P. and M.B. Siek (2004). Flexible and optimal M5 model trees with applications to flow predictions. *Proc. 6th Intern. Conf. on Hydroinformatics*, Singapore, June 2004. World Scientific, 2004.

Solomatine, D.P. and Xue, Y. (2004). M5 model trees compared to neural networks: application to flood forecasting in the upper reach of the Huai River in China. *ASCE J. Hydrol. Engineering*, 9(6), 491-501.