# Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content

Amin Elshorbagy and Ibrahim El-Baroudy

## ABSTRACT

Soil moisture has a crucial role in both the global energy and hydrological cycles; it affects different ecosystem processes. Spatial and temporal variability of soil moisture add to its complex behaviour, which undermines the reliability of most current measurement methods. In this paper, two promising evolutionary data-driven techniques, namely (i) Evolutionary Polynomial Regression and (ii) Genetic Programming, are challenged with modelling the soil moisture response to the near surface atmospheric conditions. The utility of the proposed models is demonstrated through the prediction of the soil moisture response of three experimental soil covers, used for the restoration of watersheds that were disturbed by the mining industry. The results showed that the storage effect of the soil moisture response is the major challenging factor; it can be quantified using cumulative inputs better than time-lag inputs, which can be attributed to the effect of the soil layer moisture-holding capacity. This effect increases with the increase in the soil layer thickness. Three different modelling tools are tested to investigate the tool effect in data-driven modelling. Despite the promising results with regard to the prediction accuracy, the study demonstrates the need for adopting multiple data-driven modelling techniques and tools (modelling environments) to obtain reliable predictions.

**Key words** | evolutionary polynomial regression, genetic programming, prediction, soil moisture, tool uncertainty

**Amin Elshorbagy** (corresponding author)
**Ibrahim El-Baroudy**
Department of Civil & Geological Engineering,
Centre for Advanced Numerical Simulation
(CANSIM),
University of Saskatchewan,
Saskatoon SK,
Canada S7N 5A9
E-mail: *amin.elshorbagy@usask.ca*

## INTRODUCTION

### Evolutionary data-driven techniques

The advancement of the field of hydroinformatics in the past decade has been capitalizing and building on the emergence and maturity of a variety of soft-computing techniques adopted for prediction purposes e.g. neural networks (NNs) and genetic programming (GP). The widespread application of NNs in hydrology has not yet resulted in a negation of all their shortcomings, nor addressed all raised concerns. One of the major concerns related to the use of NNs is the lack of a systematic way to decide on the optimum set of inputs and model structure/configuration (Maier & Dandy 2000; Elshorbagy & Parasuraman 2008). This problem has been partly addressed by the emergence of the GP technique (Koza 1992; Babovic & Keijzer 2000) as an evolutionary technique for constructing populations of models using stochastic search methods. In GP, both the variables and constants of the candidate models are optimized. It is therefore not required to choose the model structure *a priori* (Parasuraman *et al.* 2007*a*).

In hydrology-related studies, GP has been applied to model the rainfall–runoff process (Whigham & Crapper 2001), runoff forecasting (Khu *et al.* 2001), temperature downscaling (Coulibaly 2004), hydraulic engineering (Minns 2000) and the rainfall–recharge process (Hong *et al.* 2005).

doi: 10.2166/hydro.2009.032

Giustolisi & Savic (2006) developed the evolutionary polynomial regression (EPR) technique and user-friendly tool, which can be considered a restricted version of GP. EPR is a data-driven technique that incorporates the main features of numerical regression together with symbolic regression. It produces flexible structure polynomial models where each monomial can include user-defined functions. EPR was designed so that it avoids producing functions that grow in length over time (Davidson *et al.* 1999). In this paper, the authors aim to challenge both the GP and the EPR techniques using one of the most complicated hydrological processes for prediction, i.e. soil moisture content.

## Soil moisture and its influence on all hydrological processes

Soil moisture has a crucial role in both the global energy and hydrological cycles. It controls the partitioning of the available surface energy into sensible and latent heat fluxes; it also affects the amounts of evapotranspiration, runoff, infiltration and deep percolation. Accumulated literature on soil system heterogeneity confirms the spatial variability of the soil system properties and its flow parameters (Warrick 2003). This spatial variability affects the local atmospheric dynamics to the extent that it can cause significant changes in the precipitation intensity and temperature fields (Entekhabi *et al.* 1996). Moreover, soil moisture affects different ecosystem processes such as carbon assimilation and nitrogen mineralization, which in turn have a vital effect on the biomass production (Williams & Albertson 2004).

Entekhabi *et al.* (1996) demonstrated that the dissipation of the land surface energy is achieved through turbulent flux and thermal radiation, which depend on both soil surface temperature and the near surface atmospheric conditions. Soil surface temperature and the near surface atmospheric conditions are directly controlled by the soil moisture (D'Odorico *et al.* 2000). This control is realized through the effect of the soil moisture on the evapotranspiration process, especially in arid and semiarid regions and consequently dissipation of land surface energy into sensible heat flux, causing the soil surface temperature to rise.

The spatial variability of the surface soil moisture (neighbouring dry and wet soil patches) induces local circulations, which improves the transport of heat from the soil surface to the near surface layer and affects the temperature fields (Brutsaert 1982; Karl 1986; Segal & Arritt 1992). Soil moisture is therefore becoming the focus of many climate studies. The effect of the soil moisture on the boundary layer and consequently global climate fluctuations has been recognized (Lawford 1992; Daly & Porporato 2006). The link between the soil moisture and its thermal properties extends to controlling the thermal inertia and shortwave 'albedo' of the surface (Entekhabi *et al.* 1996). The influence of the soil moisture on the global energy cycle is mutual, i.e. the surface soil moisture responds to the changes in the global energy cycle. The mutual influence between the soil moisture and the near surface atmospheric conditions adds to the soil moisture complex and nonlinear behaviour caused by the spatial and temporal variability (Munro *et al.* 1998). Different hydrological processes accompany the soil moisture response to precipitation. This results in a significant lag in this response, which is also dependent on the soil layer physical properties e.g. overall layer thickness and texture (Entekhabi *et al.* 1996). Therefore, soil moisture can be considered as the memory of the hydrological system (Small & Kurc 2003).

Many studies have highlighted the role of the soil moisture on the partitioning of the precipitated water into infiltration and surface runoff (Fernández-Gálvez *et al.* 2007; Rollenbeck & Anhuf 2007). This confirms the dominant role of the soil moisture on the hydrological system response to the physical environment (Yoo *et al.* 2001). Warkentin (1992) demonstrated the use of soil moisture in predicting river flow in Manitoba, Canada, emphasizing the strong correlation between runoff and soil moisture estimate. Yamaguchi & Shinoda (2002) noted that land surface processes used to model water and heat fluxes do not accurately simulate soil moisture due to the concentration on predicting evapotranspiration more than soil moisture itself. As a result, some hydrological models incorporate soil moisture to reduce model predictive uncertainty in addition to the enhancement of their prediction capabilities (Goldman *et al.* 1990; Todini 1996; Tokar & Markus 2000; Donker 2001; Aubert *et al.* 2003).

The significant effect of the soil moisture on the hydrological cycle is accompanied by an indirect but effective role in many ecosystems processes. Rodriguez-Iturbe *et al.* (1999) explained the various mechanisms through which soil moisture affects the diversity of the ecological systems. They emphasized the importance of the quantitative description of the soil moisture dynamics to quantify its strong effect on the nutrient supply to the plant roots and, consequently, the biomass production.

The mutual influence between the soil moisture and the near surface atmospheric conditions (feedback mechanism) adds to the spatial and temporal variability of the soil moisture nonlinear behaviour, which in turn undermines the reliability of most of the current measurement methods (Albertson & Montaldo 2003). As a result, soil moisture modelling techniques are necessary to supplement measurements. These modelling techniques could add to the understanding of the land-atmosphere interaction and enable a clear and a concise understanding of the linkage between soil moisture state and the surface and atmospheric processes to be developed.

The objectives of this paper are (i) to investigate the capabilities of evolutionary data-driven techniques in modelling the soil moisture contents at various depths; (ii) to gain some insight into the impact of the data-driven modelling technique and tool on the prediction results; and (iii) to identify dominant key variables that affect the soil moisture dynamics and its response to different indigenous and exogenous factors, such as soil layer thickness, texture and near surface atmospheric forcing.

This research employs the EPR and GP as evolutionary data-driven modelling techniques to capture the highly nonlinear and complex behaviour of soil moisture response, which may not be adequately characterized by simple water balance models. The utility of the proposed models is demonstrated through the understanding of the (daily) depth-averaged soil moisture response of peat and till layers of three experimental soil covers–D1 (50 cm), D2 (35 cm) and D3 (100 cm)–used for the reclamation of watersheds that are disturbed by the oil sands industry. These watersheds are located in the Mildred Lake mine, and referred to as South Bison Hill (SBH) to the north of Fort McMurray, Alberta, Canada (Elshorbagy & Barbour 2007).

## PREDICTING SOIL MOISTURE

Remote sensing techniques are used to estimate the soil surface moisture status using active and passive microwave imagery together with the scattered in-field measurements for the calibration of the estimated values (Mohanty *et al.* 2000; De Lannoy *et al.* 2007). For example, Detto *et al.* (2006) proposed the use of the micrometeorological measurements together with the ground-based infra-red (IR) and high-resolution observations to estimate land surface fluxes for different vegetation types to derive a relationship between evapotranspiration and soil moisture. The problem with these methods is that they only provide estimates of the near surface soil moisture (up to 5 cm below soil surface). This renders these estimates insufficient for the hydrological and ecological studies, which require soil moisture estimates at deeper soil profiles. Several attempts have been made to integrate remotely sensed estimations together with the plant-soil-atmosphere modelling to obtain root zone soil moisture estimates (Wigneron *et al.* 1999). Remote sensing techniques also suffer from some limitations due to the problems associated with geophysical calibration.

The timescale of the considered soil moisture response also plays a significant role in the complexity of the expected response (Mahmood 1996). In other words, higher resolution of the required soil moisture response (from seasonal to hourly time scales) requires a nonlinear increase in the complication of the considered processes that affect this response. This is due to the increased storage effect on the soil moisture response (in smaller time scales) from exogenous factors such as precipitation, solar radiation and boundary layer conditions and indigenous factors such as soil texture, physical properties and thickness (Daly & Porporato 2005).

## EVOLUTIONARY DATA-DRIVEN TECHNIQUES

In this study, the utility of GP and EPR for modelling the soil moisture response to meteorological variables, such as the net radiation, precipitation, air temperature and soil temperature, is investigated and demonstrated. Modelling this response enables the identification and quantification of the storage effect of the soil layers as well as the effects of the various atmospheric conditions. Using different techniques/tools that use different types of random search approaches will enable the proper investigation of the utility of each technique/tool. It is also important to take into consideration that there is no single data-driven technique/tool that is capable of capturing all aspects/realizations of complex nonlinear behaviour at all times.

## Genetic programming

Genetic Programming (GP), developed by Koza (1992), is a widely used machine learning (ML) technique. GP is an evolutionary algorithm that mimics the biological evolution process (of natural selection) in an effort to build computer models capable of simulating complex physical processes e.g. non-linear, spatially and temporally variable processes. GP uses a tree-like structure, as decision trees, to represent its concepts and its interpreter as a computer program (Banzhaf *et al.* 1998). It is therefore considered a superset of all other ML representations; this may enable GP to produce any solution that is produced by any other ML system. It uses different genetic operators such as crossover and mutation, together with beam search to reach candidate solutions from the overall population of solutions.

Although GP is computationally intensive (as most soft-computing techniques are), especially for generating programs that are capable of simulating complex processes, its major advantage is that it handles symbolic expressions. Babovic & Abbot (1997) demonstrated the utility of the GP and Evolution Strategies (ES) in extracting the 'semantic content' of the hydraulic data to develop a representative model of complicated physical processes. Four different hydrologic applications, including rainfall–runoff modelling, of GP were presented in their work. Similar to any data-driven technique, GP has its own limitations. The major problem is the deterioration of the prediction ability of the developed model with longer prediction horizon, which is a common problem in any modelling method. The adverse consequences of this problem can be mitigated by combining GP technique with knowledge-based techniques that depend on the accumulated knowledge of the process under consideration. This will enhance the quality of the developed models and add to the understanding of the complicated hydrological processes (Babovic & Keijzer 2002).

Several applications of the GP technique in hydrology exist in the literature. Parasuraman *et al.* (2007*a*,*b*), explored the utility of the GP technique to develop pedotransfer functions (PTFs) for estimating the saturated hydraulic conductivity ($K_s$) from soil texture (sand, silt and clay) and the bulk density. Babovic & Keijzer (2002) addressed the utility of GP in developing rainfall–runoff models on the basis of hydro-meteorological data, as well as in combination with other conventional models i.e. conceptual models. It was reported that the GP models provided more insight into the functional relationships between different input variables resulting in more robust models. Similarly, Jayawardena *et al.* (2005) compared the GP technique in modelling rainfall–runoff process to the traditional modelling approaches. They used the GP technique to predict the runoff from three catchments in Hong Kong and two catchments in southern China, and showed that the GP technique evolved simple models that enabled the quantification of the significance of different input variables for prediction. In this study, two different GP implementation tools were tested: GPLAB (Silva 2004), a GP toolbox for MATLAB that provides the evolved equation in the form of a parse tree, and Discipulus (Francone 2001).

## Evolutionary polynomial regression

Evolutionary Polynomial Regression (EPR) is another data-driven technique that models time series data containing information about physical processes (Giustolisi & Savic 2006). EPR combines the power of evolutionary algorithms with numerical regression to develop polynomial models combining the independent variables together with the user-defined function as follows (Laucelli *et al.* 2005):

$$\hat{Y} = \sum_{i=1}^{m} F(X, f(x), a_i) + a_o \tag{1}$$

where $\hat{Y}$ is the EPR-estimated dependent variable, $F(\cdot)$ is the polynomial function constructed by EPR, $X$ is the matrix of the independent variables, $f(\cdot)$ is a user-defined function, $a_i$ is the coefficient of the $i$th term in the polynomial, $a_o$ is the bias and $m$ is the total number of the polynomial terms.

Inclusion of the user-defined function is provided to enhance the characterization of the response (dependent) variable. As the developers of the EPR tool state, “EPR is a two-stage technique for constructing symbolic models: (i) structure identification; and (ii) parameter estimation”, where it uses a Genetic Algorithm (GA) simple search method to search in the model structure space. EPR uses the Least Squares (LS) method to estimate the parameters of the selected model structure based on the performed GA search.

Applications of EPR are found in Savic *et al.* (2006), Giustolisi *et al.* (2007) and Doglioni *et al.* (2008). The search proceeds by using the standard GA operators: crossover and mutation. It is noted that this type of search is not exhaustive as it is practically impossible to conduct such a search on an infinite search space (Laucelli *et al.* 2005).

This study makes use of the EPR toolbox (Laucelli *et al.* 2005), which is based on “homonymous modelling methodology based on a hybrid evolutionary paradigm”. It is a multi-objective implementation of EPR in the sense that it produces several models which are the best trade-off considering fitness to training data versus parsimony. The EPR tool performs three types of regression, i.e. dynamic, static and classification.

Dynamic modelling is used to model systems that have memory or, in other words, when the present state of the system depends on the previous states of other input variables. On the other hand, static systems are systems that are not influenced by the previous states of input variables. Classification modelling is a special type of static modelling in which the model output is an integer (Laucelli *et al.* 2005). The reader may refer to the user manual for the details of the EPR toolbox and the different components of its simple interface (Laucelli *et al.* 2005). It has to be noted that the choice of the EPR is based on the different successful hydrological applications, in addition to the simplicity of its tool and the possibility of evolving an explicit mathematical model for the application under consideration.

## CASE STUDIES

The soil moisture response of an experimental reclamation site, named the South Bison Hill (SBH), is used as a case study. This research site is one of the six experimental fields of the oil sands reclamation sites located at Syncrude Canada Ltd. Mildred Lake mine site north of Fort McMurray, AB, Canada.

The climate is a sub-humid continental, which is characterized by cold winters and warm summers. The mean annual precipitation is 456 mm, of which 342 mm occurs as rainfall mostly between June–August. The SBH is constructed with waste rock material from oil sands mining in the period from 1980 to 1996 (Parasuraman *et al.* 2007*a*).

The area of SBH is 2 km$^2$; topography rises 60 m above the surrounding landscape and has a large flat top several hundred metres in diameter. The underlying shale is covered by a 20 cm layer of peat mineral mix on top of an 80 cm layer of glacial till to reclaim the overburden soil surface. This area was then fertilized and seeded to agronomic barley and planted to white spruce and aspen in the summer of 1999. The top of the SBH is dominated by foxtail barley and other minor species of fireweed (Carey 2006).

This study focuses on three inclined experimental soil covers D1, D2 and D3, which were constructed in 1999 as part of the SBH site. The thicknesses of the three covers are 50 cm, 35 cm and 100 cm, respectively. D1 cover is comprised of a peat mineral mix layer of 20 cm thickness overlying 30 cm of glacial till layer. The thinnest cover D2 consists of a 15 cm peat layer overlying a 20 cm till layer. The thickest cover D3 consists of a 20 cm peat layer overlying an 80 cm till layer. Each soil cover has an area of 1 ha, covering a sloping sub-watershed with a 20% slope where each sub-watershed is practically independent from the adjacent sub-watersheds.

A wide variety of hydrological and meteorological measurements is conducted in these experimental sites to monitor the evolution of the reconstructed watersheds. TDR sensors are used to measure soil moisture content twice a day for different soil depths for each soil cover in both peat and till layers: $SM_p$ and $SM_t$. Soil temperature of peat and till layers, $ST_p$ and $ST_t$, are also measured on an hourly basis for the corresponding soil moisture measurement depths. A weather station and a Bowen station are used to provide hourly meteorological measurements of air temperature (AT), precipitation (P), net radiation (NR) and other energy fluxes.

The current research uses the daily averaged soil temperature, for each layer, together with the daily averaged AT, P and NR to predict the corresponding daily soil moisture content for all covers. The records cover the years 2000–2005, considering time periods from May–October to avoid winter periods when the soil is frozen. The total number of available records amounts to 792 instances, divided equally into two sets for training and testing, respectively. The training and testing data sets include alternating records (every other instance) to allow the data-driven tools to properly capture the evolving hydrological change of the reclaimed (reconstructed) watersheds.

## RESULTS AND ANALYSIS

EPR toolbox and Discipulus$^{\mathrm{TM}}$ were used to model the daily soil moisture response to daily NR, P, AT and ST. To investigate the soil storage effect, both techniques/tools were used to model the soil moisture content as a response to: (i) time-lag inputs and (ii) summation inputs. The first experiment uses all input variables and their time-lags up to 20 preceding days as independent variables to model the depth-averaged daily soil moisture content. In the second experiment, the summation of the previous 20, 15, 10 and 5 days of each input were used as independent variables.

Both EPR tool and Discipulus had the freedom to select the most suitable input variable(s) which contribute to the characterization of the daily soil moisture response. The analysis of the results is therefore based on two main aspects: (a) the identification of the most influential input variables that affect the soil moisture response of different soil covers and the quantification of the soil storage effect, and (b) the ability of the techniques/tools to identify the storage effect of the soil layer on its moisture response.

The performance of the modeling tools was evaluated based on three error measures: (i) the Root Mean Squared Error (RMSE), (ii) the Mean Absolute Relative Error (MARE) and (iii) Correlation Coefficient ($R$). Both RMSE and MARE are overall error measures, where the first is a real valued metric while the latter is a relative value metric. RMSE is biased towards high soil moisture peaks which tend to produce high error values, while MARE is less sensitive to high values as it does not square the error magnitude. Due to these limitations, $R$ is used as a complementary error measure that quantifies the overall agreement between the observed and calculated datasets.

### The effect of input configuration and manipulation

The EPR tool and Discipulus$^{\mathrm{TM}}$ produce several models and provide the user with the best performing models based on the least sum of squared errors (SSE). Table 1 provides the average values (training and testing) of the various error measures for both experiments, for the peat and till layers. As shown in the table, Discipulus$^{\mathrm{TM}}$ produced better models than EPR, which is evident by the average value of $R$ in the Discipulus$^{\mathrm{TM}}$ models. The $R$ values range from 0.67–0.60 and 0.58–0.68 in training and testing phases, respectively. This is higher than average $R$ values of the EPR models which range from 0.31–0.51 and 0.30–0.53 in training and testing phases, respectively. This also applies to the other two error measures, where the ranges of the average MARE values for Discipulus models were 0.06–0.08. The corresponding values of the EPR models range from 0.08–0.09, which are relatively higher. The same remark applies to the average RMSE values.

In both techniques (EPR and Discipulus$^{\mathrm{TM}}$), the summed (cumulative)-input models produced better performance than time-lag models. This indicates that the storage effect of the soil moisture response can be quantified using cumulative inputs better than time-lag inputs, which can be attributed to the effect of the soil layer moisture holding capacity. For example, the accumulated precipitation of the previous few days, together with other near surface atmospheric conditions, determines the behaviour of the soil moisture in addition to the physical characteristics of the soil cover.

The scatter plots, shown in Figure 1, are the results of the Discipulus$^{\mathrm{TM}}$ models of the upper peat and the underlying till layers for the three soil covers, using cumulative inputs. Inclusion of the effect of the preceding

**Table 1** | Training and testing phases average results of the EPR and Discipulus$^{\mathrm{TM}}$ models for the time-lag and summation experiments

| | ***R*** | | **MARE** | | **RMSE** | |
|---|---|---|---|---|---|---|
| **Tool (experiment)** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| EPR (time-lag) | 0.31 | 0.30 | 0.09 | 0.09 | 0.04 | 0.09 |
| EPR (summation) | 0.51 | 0.53 | 0.08 | 0.09 | 0.05 | 0.05 |
| Discipulus (time-lag) | 0.60 | 0.58 | 0.07 | 0.08 | 0.04 | 0.04 |
| Discipulus (summation) | 0.67 | 0.68 | 0.06 | 0.06 | 0.03 | 0.03 |

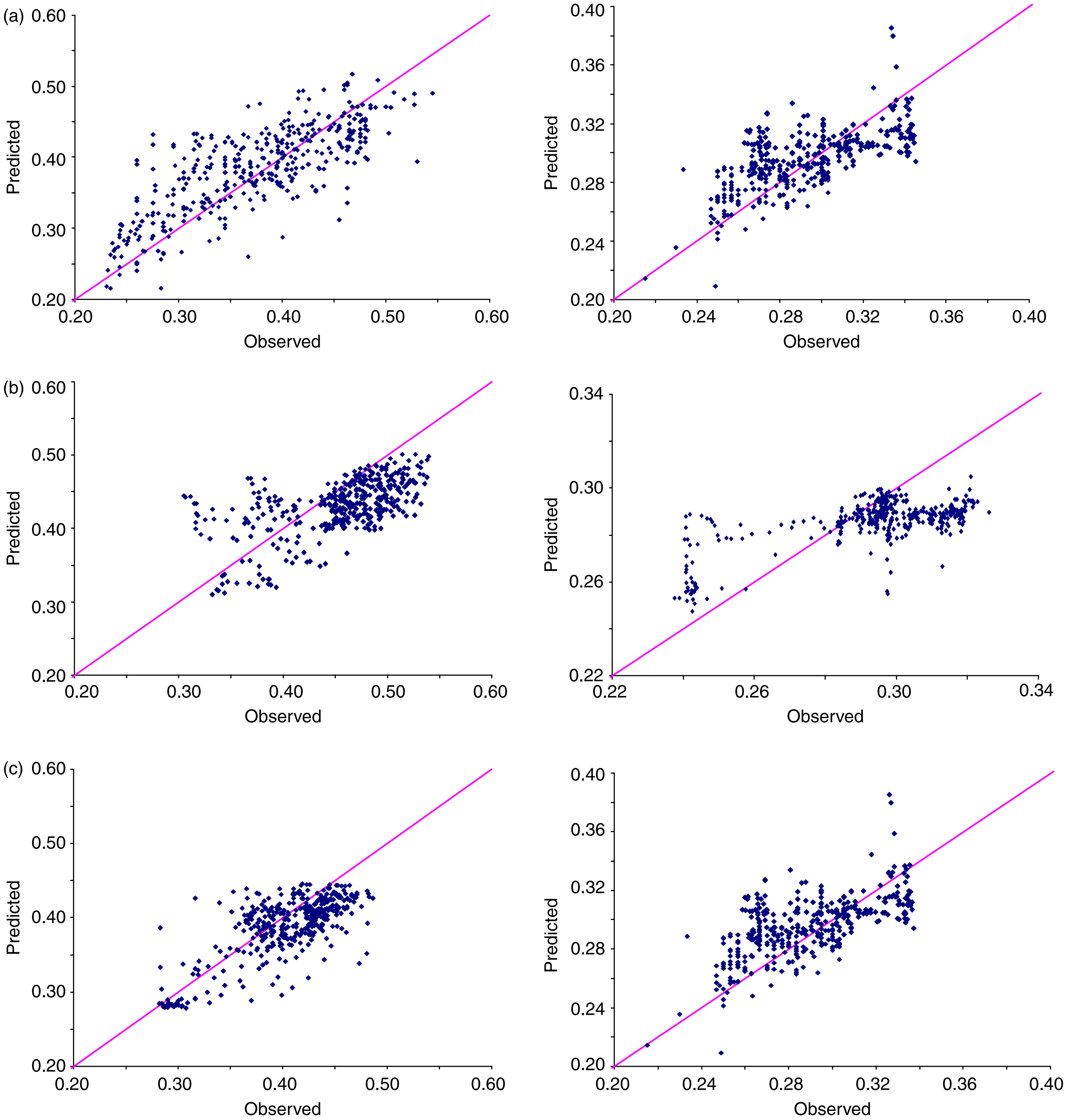

**Figure 1** | Discipulus$^{TM}$ models of upper peat layer soil moisture content (testing phase) using summation inputs for the three soil covers: (a) D2 cover (35 cm); (b) D1 cover (50 cm); and (c) D3 cover (100 cm).

days as cumulative values, in relatively thick covers, produced efficient models e.g. linear trend in the scatter plot between the predicted and observed values of the volumetric soil moisture content. As shown in Figure 1, the storage effect of the relatively thick cover (50 cm) controlled its moisture response by including previous days' inputs. As a result, the 50 cm cover model performed better (with the exception of the dry conditions of the till layer) than the

35 cm cover. This effect increases with increasing soil cover thickness as the Discipulus$^{TM}$ models managed to characterize, efficiently, the response of the thickest soil cover D3 (100 cm).

### The soil storage effect

Table 2 demonstrates the performance of the GP models for different soil layers in the training and testing phases. As shown in the table, the peat layer models have higher *R* values compared to the till layer values, while the till layer models have better MARE and RMSE values. These discrepancies highlight the importance of evaluating models based on the proper error measure(s). In case of soil moisture content with a narrow range (0.25–0.50), the RMSE and MARE could be misleadingly small making the comparison difficult, especially with varying layer thicknesses. In this case, the *R*-statistic is a better representative of the model performance. The peat layer models managed to characterize the dynamics of the soil moisture response, indicated by better *R* values, more than the till layer models. This indicates the higher sensitivity of the peat layer to the surrounding atmospheric conditions. The lower sensitivity of the till layer can be attributed to the buffering effect of the overlying peat layer, which trims the effect of the near surface atmospheric conditions on the till layer moisture response.

Generally, the soil moisture response of relatively thick cover exhibits a clear storage effect due to its relatively high moisture holding capacity and consequently better memory (effect of previous days' inputs). This is indicated in the increasing *R* values from the D2 cover (35 cm) to D3 cover (100 cm) in case of predicting the underlying till moisture content (Table 2), with a corresponding decrease in the MARE and RMSE values. For the peat layer, this trend is significantly less notable as the upper peat layer is exposed and more sensitive to the exogenous factors. Discussion of the storage effect is deferred to the next section.

### Results of the GP technique

Although GP models evolved by Discipulus software outperformed the EPR models, it has to be noted that Discipulus does not provide a mathematical formula of the developed model. Instead, it provides a code in different programming languages such as Assembly or Java. This code contains, in addition to the ordinary mathematical operators such as addition and subtraction, conditional and comparison instructions which makes it difficult to extract the mathematical formula that represents the developed code. Equation (2) is a simplified example of one of the evolved models by Discipulus$^{TM}$ for the till layer of soil cover D1. This formula was deduced after removing all logical comparison functions to provide a comprehendible mathematical formula:

$$\mathrm{SM_p} = \frac{0.32\sum \mathrm{NR_{20}}(B+1)\mathrm{ST_p} + 0.03\left(\sum \mathrm{NR_{20}}\right)^2}{\mathrm{ST_p^2}(B+1)^2} \tag{2}$$

where

$$B = -\mathrm{INT}\left| -18.81A^{3.70} + \frac{1.73\sum \mathrm{P_{10}} - 61.80A^{1.85} - 50.75 + 13.83\mathrm{ST_p}}{\left(\sum \mathrm{NR_{20}}\right)}\right|$$

and

$$A = \mathrm{INT}\left(\left(10^{-4}\frac{\mathrm{ST_p}}{\left(\sum \mathrm{NR_{20}}\right)}\right) \times \left(\begin{matrix} -4.00 - 12.00\left(\sum \mathrm{P_{20}}\right)^2 + 43.00\left(\sum \mathrm{P_{20}}\right) - \left(\sum \mathrm{AT_{20}}\right)^2 \\ -43.00\left(\sum \mathrm{AT_{20}}\right) + 23.00\,\mathrm{ST_p}\left(\sum \mathrm{P_{20}}\right)\left(\sum \mathrm{AT_{20}}\right) \end{matrix}\right)\right)$$

where $\mathrm{NR_{20}}$ is the cumulative net radiation over the preceding 20 days, $\mathrm{P_{10}}$ and $\mathrm{P_{20}}$ are the cumulative precipitation over the preceding 10 and 20 days and $\mathrm{AT_{20}}$ is the cumulative air temperature over the preceding 20 days.

Equation (2) is not easy to understand and it is a tedious task to use it to analyze the soil moisture response model. However, Discipulus$^{TM}$ provides two proxy measures of the importance of each input variable. These two measures can be used as an alternative to deducing the mathematical formula and therefore provide insight into the complex hydrological process. These proxy measures are the *impact* and the *frequency*. The average *impact* of a certain input is defined as the average effect of removing all instances of the considered input from the best thirty programs developed by Discipulus$^{TM}$. *Frequency* of an input is the percentage

**Table 2** | Performance measures, training (testing) of the GP models for the various soil layers

| | D2 (35 cm) | | D1 (50 cm) | | D3 (100 cm) | |
|---|---|---|---|---|---|---|
| **Measure** | **Peat** | **Till** | **Peat** | **Till** | **Peat** | **Till** |
| *R* | 0.79 (0.77) | 0.63 (0.58) | 0.66 (0.74) | 0.68 (0.66) | 0.86 (0.78) | 0.71 (0.70) |
| MARE | 0.09 (0.10) | 0.05 (0.06) | 0.04 (0.04) | 0.04 (0.04) | 0.05 (0.06) | 0.02 (0.02) |
| RMSE | 0.05 (0.05) | 0.02 (0.02) | 0.07 (0.06) | 0.01 (0.01) | 0.02 (0.03) | 0.01 (0.01) |

of the models (out of the best thirty models developed by Discipulus$^{TM}$) that contained the considered input (Francone 2001). Figure 2 shows the two proxy measures produced by the Discipulus$^{TM}$ for the D1 soil cover for both the peat and till layers.

As shown in Figure 2, the bars indicate the *frequency* of different input variables (on the *x* axis), while the triangles and circles indicate the maximum and average *impacts* of each input variable, respectively. In the peat layer model, two inputs (AT and $AT_{20}$) were included in all thirty best programs, i.e. frequency value of 100%. The same two inputs in the till layer also had high frequency, 83% and 67%, respectively. This highlights the buffering effect of the peat layer on the till layer in trimming the atmospheric forcing effect.

Although these two variables have occurred almost in all the best thirty programs, they were not equally important which is indicated through the impact values. $AT_{20}$ has high maximum and average impact values (38% and 46%), while AT has only 5% and 12% respectively, for the peat layer. This is also shown in the till layer, where $AT_{20}$ has high maximum and average impact values of 25% and 47% while AT has only 12% and 39%, respectively. These high values point out the importance of the previous 20 days of AT over the current values on the characterization of the soil moisture state which apparently carries information about the evapotranspiration. It has to be noted that the almost equal importance of all input variables, especially summation of previous days, convey the joint effect of storage and evapotranspiration processed in this relatively medium thick cover.

The storage effect is also shown in high impact values for the previous precipitation events. However, variables such as $P_5$, $P_{10}$, $P_{15}$ and $P_{20}$ do not have similar frequency values, as they have a range of frequency values of 63–80%. Some of these variables have very low impact values such as $P_{10}$ and $P_{20}$, which have maximum impact values of 5% and 20%, respectively, for the peat layer. The till layer exhibits less sensitive behaviour considering the effect of the overlying peat layer, which acts as a buffer reducing the effect of the atmospheric forcing on the till layer.

The case study area is located in a semi-arid region, in which evaporative demand plays a significant role in the soil moisture response. Consequently, the variables related

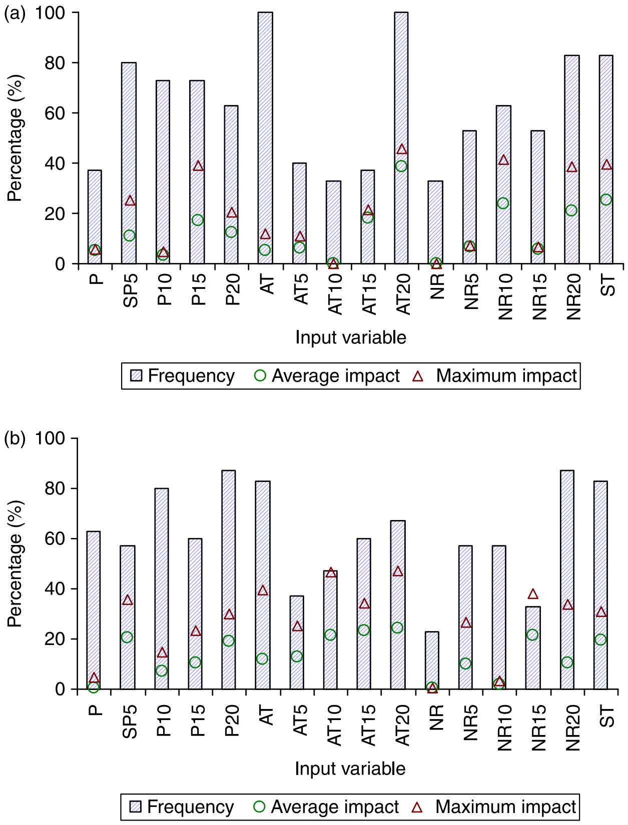

**Figure 2** | Discipulus proxy measures of different inputs for the D1 soil cover: (a) peat layer (20 cm) and (b) till layer (30 cm).

to the evapotranspiration process such as AT and NR are influential inputs. As shown in Figure 2, the precipitation inputs have a range of average impact values of 2–20% on the soil moisture of both layers, while the average impact values of the AT inputs is of 0–22% and those of NR inputs is of 5–21%. This indicates the balancing effect of both precipitation and evapotranspiration on the soil moisture dynamics of the medium thickness (50 cm) cover D1. ST has high frequency and impact values, which point out the importance of this input to the characterization of the soil moisture response in both soil layers. This close link between soil moisture and soil temperature confirms the strong relation between soil moisture and its thermal properties.

Figures 3 and 4 show the two proxy measures produced by Discipulus$^{TM}$ for the other soil covers, i.e. D2 (35 cm) and D3 (100 cm), for both layers. The storage effect on the soil moisture of the peat layer of cover D2 is less evident than cover D3. This is evident in the high frequency and impact values of $AT_{20}$ of 100% and 60%, respectively, on the moisture of the peat layer of the thin cover D2. This is a strong indication that the thin cover is controlled and dried by the evapotranspiration process, whereas the soil layers do not have sufficient water holding capacity (Figure 3).

Although the precipitation inputs have high frequency, they do not have high impact values relative to the other atmospheric conditions. Figure 4 clearly depicts the concept of the storage effect, where the impact of e.g. $P_{20}$ is strong compared to other AT and NR variables in both the peat and the till layers. This confirms the ability of the thick cover to store moisture from previous precipitation events and to release it for evapotranspiration when needed. The moisture dynamics is controlled by the precipitation history rather than the evapotranspiration. Interestingly, this phenomenon has been demonstrated by a mechanistic water balance model (Elshorbagy & Barbour 2007). Table 3 summarizes the previous discussion, where it shows the high impact input variables in each model for each soil layer. These variables are in descending order according to their impact and frequency values. As shown in the table, the increasing soil thickness results in an increasing storage effect, as clear from the dependence of the peat layer of soil cover D3 on all the precipitation inputs in characterizing its moisture response. All models presented in Table 3 include ST, which confirms the link between the soil moisture response and the soil thermal properties.

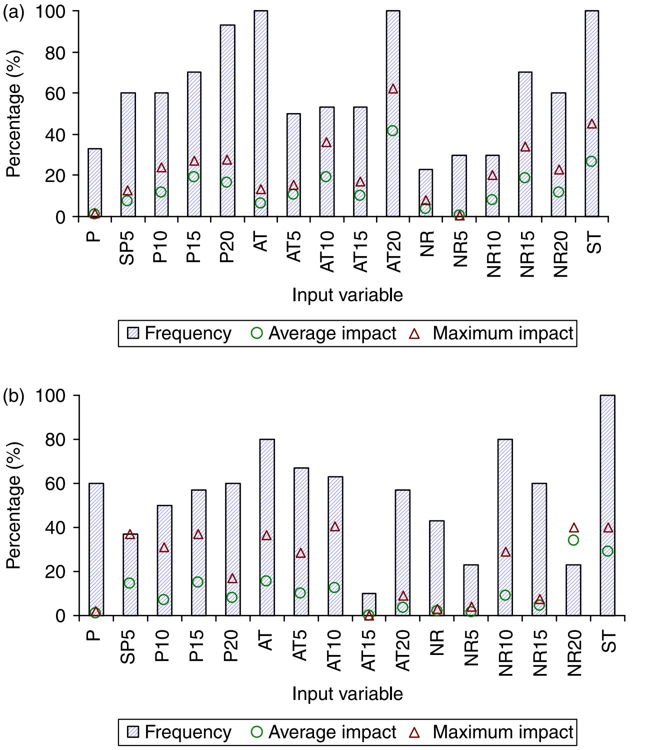

**Figure 3** | Discipulus proxy measures of different inputs for the D2 soil cover: (a) peat layer (15 cm) and (b) till layer (20 cm).

## Results of the EPR technique

Although the models provided by the EPR technique did not match the performance of the Discipulus$^{TM}$ models (Table 1), the formulae produced by the EPR tool can be a useful tool in providing more insights into the direct effect of each input variable to the characterization of the soil moisture response in a more explicit manner, and also to confirm the previous findings. The EPR-evolved models have been simplified such that terms which contribute less than 3% of the total predicted moisture values were neglected. This simplification would result in less than ±5% deterioration in the prediction accuracy. The following are the best two models, after simplification, produced by the EPR tool for the thinnest soil cover D2, where $SM_p$

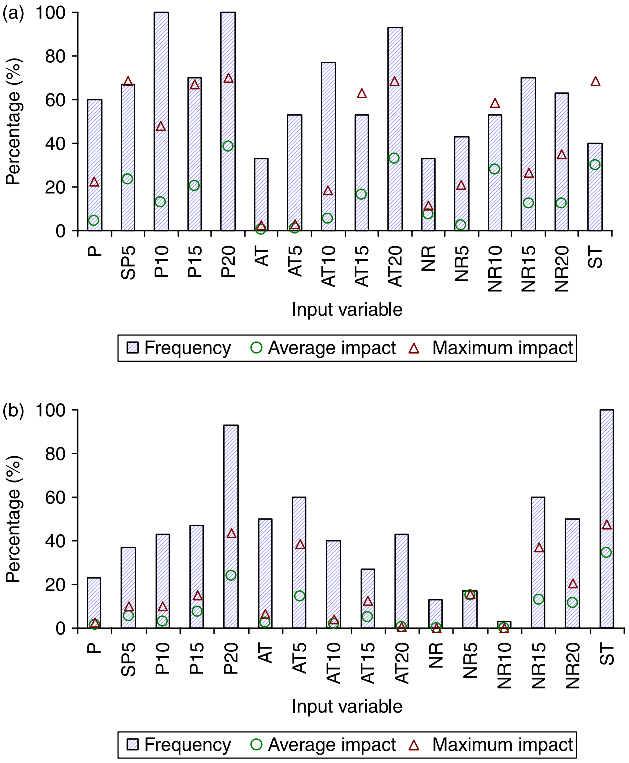

**Figure 4** | Discipulus proxy measures of different inputs for the D3 soil cover: (a) peat layer (20 cm) and (b) till layer (80 cm).

and $SM_t$ are the predicted peat and till moisture contents, respectively:

$$SM_p = -1.01\times10^{-8}\frac{(\sum AT_{20})^2\cdot(\sum NR_{15})\cdot ST_p}{(\sum AT_{10})} + 1.31\times10^{-12} AT\cdot\left(\sum NR_{10}\right)^2\cdot\left(\sum NR_{20}\right) + 0.41 \quad (3)$$

and

$$SM_t = -4.10\times10^{-12}\frac{(\sum P_{20})\cdot(\sum AT_{15})\cdot(\sum AT_{20})^3\cdot ST_t^3}{(\sum AT_{10})\cdot(\sum NR_{15})} + 2.37\times10^{-5}\frac{(\sum P_{20})\cdot AT\cdot(\sum AT_5)\cdot(\sum AT_{15})}{(\sum NR_{15})\cdot ST_t} + 0.28 \quad (4)$$

To analyze the importance of each term in characterizing the overall soil moisture response, the minimum, maximum and average contribution of each term was computed as a percentage of the overall predicted value, based on the testing dataset. This was performed by calculating the soil moisture response using the corresponding input data, then calculating the individual term percentage of the total value. This procedure was repeated for all the instances of the testing dataset, which resulted in a range of contribution values for each term. The maximum, average and minimum contribution was then calculated for each term. The whole procedure was repeated for all models developed by the EPR. For brevity, Table 4 presents only the average contribution percentage for each term in the developed six Equations (3–8), excluding terms that do not have a significant contribution percentage, i.e. less than 2%.

As shown in Table 4, the constant term (bias) has the highest contribution percentage in both equations with average values of 72.3% and 92.0% for the peat and till layers of soil cover D2, respectively. This term reflects the effect of the soil layer holding capacity on its response. The difference in the contribution of the bias between the two layers indicates the effect of the overlying peat layer in trimming the atmospheric forcing on the till layer. As a result, the peat layer is more responsive to the near surface atmospheric conditions, which minimizes the storage effect in this layer.

The till layer is not as responsive and the storage (stabilizing) effect dominates its moisture response. This difference also reflects the difference in the holding capacities between the two types of texture and thickness. In Equation (3), the second most contributing term is the first term, which includes $\sum AT_{10}$, $\sum AT_{20}$, $\sum NR_{15}$ and $ST_p$. Interestingly, the storage effect (P) is not represented

**Table 3** | List of the high impact inputs (grouped by type) of the different GP models

| Cover | Layer | High impact inputs (grouped) |
|---|---|---|
| D2 (35 cm) | Peat | $\sum AT_{20}$, $\sum AT_{10}$, $\underline{ST}$, $\sum P_{20}$,$\sum P_{15}$, $\sum NR_{15}$ |
| | Till | $\sum AT_{10}$, AT, $\underline{ST}$, $\sum P_{15}$, $\sum P_{10}$, $\sum P_5$, $\sum NR_{20}$ |
| D1 (50 cm) | Peat | $\sum AT_{20}$, $\sum AT_{15}$, $\underline{ST}$, $\sum P_{15}$, $\sum NR_{10}$, $\sum NR_{20}$ |
| | Till | $\sum AT_{20}$, $\sum AT_{15}$, $\sum AT_{10}$, AT, $\underline{ST}$, $\sum P_{15}$, $\sum NR_{15}$ |
| D3 (100 cm) | Peat | $\sum AT_{20}$, $\sum AT_{15}$, $\underline{ST}$, $\sum P_5$, $\sum P_{10}$,$\sum P_{15}$,$\sum P_{20}$ |
| | Till | $\sum AT_5$, $\underline{ST}$, $\sum P_{20}$, $\sum NR_{20}$, $\sum NR_{15}$ |

**Table 4** | Contribution of different terms to the characterization of the soil moisture response as a percentage of the total value

| Cover | Layer | Equation | Term #1 | Term #2 | Term #3 | Term #4 |
|---|---|---|---|---|---|---|
| D2 (35 cm) | Peat | 6 | 15.4 | 7.5 | 72.3 | – |
| | Till | 7 | 3.2 | 3.9 | 92.0 | – |
| D1 (50 cm) | Peat | 8 | 11.8 | 13.0 | 73.8 | – |
| | Till | 9 | 3.2 | 3.8 | 90.0 | – |
| D3 (100 cm) | Peat | 10 | 5.8 | 15.2 | 4.2 | 70.1 |
| | Till | 11 | 4.9 | 2.0 | 91.1 | – |

explicitly in this layer as the bias term included most of the storage 'signal'. This, in turn, supports the previous conclusion that the overall thickness of the soil cover affects its response and the thinner the cover the more its sensitivity to the evapotranspiration requirements. The following four equations are the simplified formulae produced by the EPR tool for both peat and till layers for D1 (50 cm) and D3 (100 cm) covers, respectively:

$$\mathrm{SM_p} = -9.06\times10^{-12}\left(\sum \mathrm{AT_{20}}\right)^4 + 4.82 \times10^{-5}\left(\sum \mathrm{AT_{10}}\right)\cdot\mathrm{ST_p} + 0.43 \tag{5}$$

$$\mathrm{SM_t} = 1.09\times10^{-10}\left(\sum \mathrm{AT_{15}}\right)^2\cdot\left(\sum \mathrm{NR_{20}}\right) - 3.20\times10^{-12}\frac{(\sum \mathrm{P_{15}})\cdot(\sum \mathrm{AT_{20}})^2\cdot(\sum \mathrm{NR_{20}})\cdot\mathrm{ST_t^2}}{(\sum \mathrm{AT_{10}})} + 0.28 \tag{6}$$

For D3 cover:

$$\mathrm{SM_p} = -1.07\times10^{-11}\frac{(\sum \mathrm{AT_{20}})^3\cdot(\sum \mathrm{NR_{15}})\cdot\mathrm{ST_p}}{(\sum \mathrm{AT_{10}})} - 3.78 \times10^{-9}\left(\sum \mathrm{AT_{20}}\right)^3 + 2.09\times10^{-6}\frac{(\sum \mathrm{P_{15}})\cdot(\sum \mathrm{AT_{15}})^2\cdot\mathrm{ST_p}}{(\sum \mathrm{NR_{15}})} + 0.403 \tag{7}$$

$$\mathrm{SM_t} = 9.88\times10^{-7}\left(\sum \mathrm{NR_{20}}\right)\cdot\mathrm{ST_t} + 1.31 \times10^{-6}\frac{(\sum \mathrm{P_{20}})\cdot(\sum \mathrm{AT_{10}})\cdot(\sum \mathrm{AT_{20}})}{(\sum \mathrm{AT_{15}})} + 0.37 \tag{8}$$

Similar analysis and conclusions could be drawn with regard to the other soil covers (Equations (5–8) and Table 4). Consistently, the bias term contributed most of the soil moisture variability and its relative contribution in the lower till layer was higher than that in the upper peat layer. This finding is important for modelling soil moisture using data-driven techniques. Unlike other hydrological variables, the variability of soil moisture content values around an average constant value is small, making the selection of the proper modelling technique/tool and the error measure of utmost importance.

## The dilemma of the modelling techniques and tools

Both EPR tool and the GP (Discipulus$^{\mathrm{TM}}$) produced several models capable of capturing the inherent complexity of the soil moisture response of various soil covers. They can both be used to identify dominant input variables to the soil moisture response, which facilitates the simulation and prediction of the behaviour of the various soil layers. The models provided by the EPR did not match the performance of the GP models, which can be attributed to the ability of the GP technique and the Discipulus tool to use logical operators, in addition to the other mathematical operators. This gives the developed GP technique superiority over some other data-driven techniques. However, the formula produced by the EPR tool can be a useful tool in providing insight into the direct effect of each input variable to the characterization of the soil moisture response, in a more explicit manner, and also to confirm the previous findings using the Discipulus tool. The attraction and convenience for decision makers of having an explicit mathematical formula cannot be overemphasized.

In this study, another GP tool–GPLAB (Silva 2004)–has been used and tested on the same experiments. Interestingly, and after so many trials, GPLAB always evolved constant values rather than mathematical formulae for the soil moisture response in the various soil covers, and for both layers. The constant values were very similar to the bias terms evolved by the EPR. This is due to the dominant relative contribution of the bias terms as discussed earlier. GPLAB, from a practical and implementation point of view, was not sensitive to the small contribution by other input variables. Minimizing an overall average error measure, e.g. squared error, may lead to an average output value for a hydrological variable such as soil moisture. Therefore, it is important to note that the conclusion drawn earlier in this study was not only regarding GP versus EPR, but tied inseparably to the EPR toolbox and the Discipulus software. Otherwise, EPR evolved more realistic formulae than GPLAB and also provided better prediction accuracy. This point is not trivial because in data-driven (black box) modelling, the adopted technique cannot be used and, therefore, should not be evaluated independent of the implementation environment (tool).

The above discussion highlights the need for awareness of the ‘tool uncertainty’ concept. This should not be confused with model structure uncertainty. Within the same technique and tool (e.g. EPR toolbox), multiple model structures could be evolved for the same application and such a multiplicity generates model structure uncertainty. However, different models and results from two different implementation tools of the same technique (e.g. both GPLAB and Discipulus represent the GP technique) is a clear case of tool uncertainty. This means that in data-driven modelling, tool uncertainty is an additional source of uncertainty to the possible predictive uncertainty. The authors advocate the idea of using multiple data-driven techniques as well as multiple implementation tools (environments) to obtain realistic prediction ensemble for hydrological processes. The authors have applied higher order neural networks (HONNs) (Elshorbagy & Parasuraman 2008) and a mechanistic system dynamics watershed model (Elshorbagy *et al.* 2005) on the same study area. Currently, the authors are in the process of presenting a comprehensive comparison of multiple data-driven and mechanistic models for predicting soil moisture content.

One last observation on GP, EPR and possibly all data-driven techniques is regarding the scientific interpretability of the evolved expressions. GP and EPR, as symbolic regression, are similar to numerical regression that produces empirical expressions, which are not dimensionally sensible in most cases. Dimensionally aware GP (Keijzer & Babovic 2002) was not investigated in this research. However, it is expected that significant deterioration of the prediction accuracy has to be tolerated to generate dimensionally correct equations for soil moisture content. Soil moisture content is a dimensionless variable that is indirectly related to the input parameters adopted in this study. The input parameters may dimensionally relate to evapotranspiration, which in turn relates to soil moisture content.

## SUMMARY AND CONCLUSION

The complex behaviour of the soil moisture response affects the reliability of most of the *in situ* methods, in addition to newly developed remote sensing techniques. Therefore, there is no single technique that is suitable for the application at the large scale and in a practical mode. Soil moisture modelling techniques are necessary to supplement soil moisture measurements.

To gain some insights into the soil moisture response to different indigenous and exogenous factors, the authors explored the utility of two evolutionary data-driven techniques: (i) Evolutionary Polynomial Regression (EPR) and (ii) Genetic Programming (GP) in modelling the soil moisture response to the net radiation, precipitation, air temperature and soil temperature. The results showed that the storage effect of the soil moisture response can be quantified using cumulation (summation of) inputs better than time-lag inputs, which can be attributed to the effect of the soil layer moisture holding capacity. This effect increases with increasing soil cover thickness. The discrepancies that exist in the sub-layers of the soil cover result from the buffering effect of the overlaying surface layer, which trims the effect of the near surface atmospheric forcing. The surface layer exhibits relatively high sensitivity to the atmospheric forcing. The insignificant differences in behaviour between the two soil layers indicate the importance of the combined effect of the two layers, as a whole, to characterize the soil moisture response. The adopted data-driven techniques were able to quantify and characterize the above-mentioned dynamics.

The overall soil thickness plays a dominant role in determining the controlling process over the soil moisture response. Relatively thin soil covers are more responsive to the evapotranspiration process and exhibit highly dynamic behaviour due to the short span memory they hold. Thick soil layers can sustain longer periods of drought as it has long span memory, and it actually responds to previous precipitation and atmospheric forcing (up to 20 preceding days). All developed models included a soil temperature term confirming the link between the soil moisture response and the soil thermal properties. Finally, the research used three different data-driven techniques and/or tools; none of them is considered superior in all aspects. This highlights the issue of tool uncertainty in data-driven modelling indicating that there is no single data-driven technique/tool that is capable of capturing the inherent variability of the complex soil moisture response processes at all times. This can be mitigated through the incorporation of more than one technique/tool for the same problem.

## ACKNOWLEDGEMENTS

The authors acknowledge the financial support of NSERC through its Discovery Grant and CRD programs. The financial and the in-kind contribution of CEMA are also appreciated.

## REFERENCES

Albertson, J. D. & Montaldo, N. 2003 Temporal dynamics of soil moisture: 1. Theoretical basis. *Water Resour. Res.* **39** (10), 1274.

Aubert, D., Loumagne, C. & Oudin, L. 2003 Sequential assimilation of soil moisture and stream flow data in a conceptual rainfall–runoff model. *J. Hydrol.* **280**, 145–161.

Babovic, V. & Abbot, M. B. 1997 The evolution of equations from hydraulic data part II: applications. *J. Hydraulic Res.* **35** (3), 411–430.

Babovic, V. & Keijzer, M. 2000 Genetic programming as model induction engine. *J. Hydroinform.* **2** (1), 35–60.

Babovic, V. & Keijzer, M. 2002 Rainfall–runoff modelling based on genetic programming. *Nordic Hydrol.* **33** (5), 331–346.

Banzhaf, W., Nordin, P., Keller, R. E. & Francone, F. D. 1998 *Genetic Programming–An Introduction: on the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann Publishers, Inc., San Francisco, California.

Brutsaert, W. 1982 *Evaporation into the Atmosphere*. D. Reidel., Dordrecht, Holland.

Carey, S. K. 2006 Energy and water exchange from a saline-sodic overburden reclamation soil cover: fort Mcmurray, Alberta. *Proceedings of CLRA Conference*, Ottawa, ON, August 20–23.

Coulibaly, P. 2004 Downscaling daily extreme temperatures with genetic programming. *Geophys. Res. Lett.* **31**, L16203.

Daly, E. & Porporato, A. 2005 A review of soil moisture dynamics: from rainfall infiltration to ecosystem response. *Environ. Eng. Sci.* **22** (1), 9–24.

Daly, E. & Porporato, A. 2006 Impact of hydroclimatic fluctuations on the soil water balance. *Water Resour. Res.* **42**, W06401.

Davidson, J. W., Savic, D. A. & Walters, G. A. 1999 Method for identification of explicit polynomial formulae for the friction in turbulent pipe flow. *J. Hydroinform.* **1** (2), 115–126.

De Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N. & Verhoest, N. E. C. 2007 State and bias estimation for soil moisture profiles by an ensemble Kalman filter: effect of assimilation depth and frequency. *Water Resour. Res.* **43**, W06401.

Detto, M., Montaldo, N., Albertson, J. D., Mancini, M. & Katul, G. 2006 Soil moisture and vegetation controls on evapotranspiration in a heterogeneous mediterranean ecosystem in Sardinia, Italy. *Water Resour. Res.* **42**, W08419.

D'Odorico, P., Ridolfi, L., Porporato, A. & Rodriguez-Iturbe, I. 2000 Preferential state of seasonal soil moisture: the impact of climate fluctuations. *Water Resour. Res.* **36** (8), 2209–2219.

Doglioni, A., Giustolisi, O., Savic, D. A. & Webb, B. W. 2008 An investigation on stream temperature analysis based on evolutionary computing. *Hydrol. Process. J.* **22**, 315–326.

Donker, N. H. W. 2001 A simple rainfall–runoff model based on hydrological units applied to the Teba catchment (south-east Spain). *Hydrol. Process. J.* **15**, 135–149.

Elshorbagy, A. & Barbour, S. L. 2007 Probabilistic approach for design and hydrologic performance assessment of reconstructed watersheds. *J. Geotech. Geoenviron. Eng.* **133** (9), 1110–1118.

Elshorbagy, A. & Parasuraman, K. 2008 On the relevance of using artificial neural networks for estimating soil moisture content. *J. Hydrol.* **362** (1–2), 1–18.

Elshorbagy, A., Julta, A., Barbour, L. & Kells, J. 2005 System dynamics approach to assess the sustainability of reclamation of distributed watersheds. *Can. J. Civil Eng.* **32**, 144–158.

Entekhabi, D., Rodriguez-Iturbe, I. & Castelli, F. 1996 Mutual interaction of soil moisture state and atmospheric processes. *J. Hydrol.* **184**, 3–17.

Fernández-Gálvez, J., Verhoef, A. & Barahona, E. 2007 Estimating soil water fluxes from soil water records obtained using dielectric sensors. *Hydrol. Process. J.* **21**, 2785–2793.

Francone, F. D. 2001 *Discipulus: Owner's Manual*. Register Machine Learning Technologies, Inc. Littleton, CO.

Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinform.* **8** (3), 207–222.

Giustolisi, O., Doglioni, A., Savic, D. A. & Webb, B. W. 2007 A multi-model approach to analysis of environmental phenomena. *Environ. Model. Softw.* **22** (5), 674–682.

Goldman, D. M., Marino, M. A. & Feldman, A. D. 1990 Runoff prediction uncertainty for ungauged agricultural watersheds. *J. Irrigation Drainage Eng.* **16** (6), 752–768.

Hong, Y.-S., White, P. A. & Scott, D. M. 2005 Automatic rainfall recharge model induction by evolutionary computational intelligence. *Water Resour. Res.* **41**, W08422.

Jayawardena, A. W., Muttil, N. & Fernando, T. M. K. G. 2005 Rainfall–runoff modelling using genetic programming. In *MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand* (ed. A. Zerger & R. M. Argent), December 2005, pp. 1841–1847.

Karl, T. R. 1986 The relationship of soil moisture parameterizations to subsequent seasonal and monthly mean temperature in the united states. *Mon. Weather Rev.* **114**, 675–686.

Keijzer, M. & Babovic, V. 2002 Declarative and preferential bias in GP-based scientific discovery. *Genet. Programming Evolvable Mach.* **3**, 41–79.

Khu, S. T., Liong, S. Y., Babovic, V., Madsen, H. & Muttil, N. 2001 Genetic programming and its application in real-time runoff forecasting. *J. Am. Water Resour. Assoc.* **37** (2), 439–451.

Koza, J. R. 1992 *Genetic Programming: on the Programming of Computers by Means of Natural Selection*. The MIT Press, Cambridge, MA.

Laucelli, D., Berardi, L. & Doglioni, A. 2005 Evolutionary polynomial regression toolbox: version 1. SA. Department of Civil and Environmental Engineering, Technical University of Bari, Bari, Italy. Available from: http://www.hydroinformatics.it/prod02.htm (accessed March 2008).

Lawford, R. G. 1992 An overview of soil moisture and its role in the climate system. In *Proceedings of the National Hydrology Research Centre Workshop* (ed. F. J. Eley, R. Granger & L. Martin), March 9–10, pp. 1–12.

Mahmood, R. 1996 Scale issues in soil moisture modeling: problems and prospects. *Prog. Phys. Geogr.* **20** (3), 273–291.

Maier, H. R. & Dandy, G. C. 2000 Application of artificial neural networks to forecasting of surface water quality variables: issues, applications and challenges. In *Artificial Neural Networks in Hydrology* (ed. R. S. Govindaraju & A. R. Rao), pp. 287–309. Kluwer, Dordrecht, The Netherlands.

Minns, T. 2000 Subsymbolic methods for data mining in hydraulic engineering. *J. Hydroinform.* **2** (1), 3–13.

Mohanty, B. P., Skaggs, T. H. & Famiglietti, J. S. 2000 Analysis and mapping of field-scale soil moisture variability using high-resolution, ground-based data during the southern Great Plains 1997 (SGP97) hydrology experiment. *Water Resour. Res.* **36** (4), 1023–1031.

Munro, R. K., Lyons, W. F., Shao, Y., Wood, M. S., Hood, L. M. & Leslie, L. M. 1998 Modelling land surface–atmosphere interactions over the Australian continent with and emphasis on the role of soil moisture. *Environ. Model. Softw.* **13**, 333–339.

Parasuraman, K., Elshorbagy, A. & Carey, S. K. 2007*a* Modelling dynamics of the evapotranspiration process using genetic programming. *Hydrol. Sci. J.* ***53** (3), 563–578.*

Parasuraman, K., Elshorbagy, A. & Si, B. C. 2007*b* Estimating saturated hydraulic conductivity using genetic programming. *Soil Sci. Soc. Am. J.* ***71**, 1676–1684.*

Rodriguez-Iturbe, I., D'Odorico, P., Porporato, A. & Ridolfi, L. 1999 On the spatial and temporal links between vegetation, climate and soil moisture. *Water Resour. Res.* **35** (12), 3709–3722.

Rollenbeck, R. & Anhuf, D. 2007 Characteristics of the water and energy balance in an Amazonian lowland rainforest in Venezuela and the impact of the ENSO-cycle. *J. Hydrol.* **337** (3–4), 377–390.

Savic, D. A., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S. & Saul, A. 2006 Sewers failure analysis using evolutionary computing. *Water Manage. J.* **159** (2), 111–118.

Segal, M. & Arritt, R. W. 1992 Nonclassical mesoscale circulations caused by surface sensible heat-flux gradients. *Bull. Am. Metrol. Soc.* **73**, 1593–1604.

Silva, S. 2004 *GPLAB; A genetic programming toolbox for MATLAB*. ECOS–Evolutionary and Complex Systems Group University of Coimbra, Portugal, Version 2.

Small, E. E. & Kurc, S. A. 2003 Tight coupling between soil moisture and the surface radiation budget in semiarid environment: implications for land-atmosphere interactions. *Water Resour. Res.* **39** (10), 1278.

Todini, E. 1996 The ARNO rainfall–runoff model. *J. Hydrol.* **175**, 339–382.

Tokar, A. & Markus, M. 2000 Precipitation-runoff modeling using artificial neural networks and conceptual models. *J. Hydrologic Eng.* **5** (2), 156–161.

Warkentin, A. A. 1992 Use of soil moisture estimates for river forecasting. In *Proceedings of the National Hydrology Research Centre workshop* (ed. F. J. Eley, R. Granger & L. Martin), March 9–10, pp. 21–33.

Warrick, A. W. 2003 *Soil Water Dynamics*. Oxford University Press, New York, NY.

Whigham, P. A. & Crapper, P. F. 2001 Modelling rainfall–runoff using genetic programming. *Math. Computer Model.* **33** (6–7), 707–721.

Wigneron, J., Olioso, A., Calvet, J. & Bertuzzi, P. 1999 Estimating root zone soil moisture from surface soil moisture data and soil-vegetation-atmosphere transfer modeling. *Water Resour. Res.* **35** (12), 3735–3745.

Williams, C. A. & Albertson, J. D. 2004 Soil moisture controls on canopy-scale water and carbon fluxes in an African savanna. *Water Resour. Res.* **40**, W09302.

Yamaguchi, Y. & Shinoda, M. 2002 Soil moisture modeling based on multiyear observations in the Sahel. *J. Appl. Meteorol.* **41**, 1140–1146.

Yoo, C., Kim, S. J. & Lee, J. 2001 Land cover change and its impact on soil-moisture-field evolution. *J. Hydrol.* **6** (5), 436–441.

First received 28 April 2008; accepted in revised form 25 November 2008