Hydrol. Earth Syst. Sci., 14, 1943–1961, 2010
www.hydrol-earth-syst-sci.net/14/1943/2010/
doi:10.5194/hess-14-1943-2010


Hydrology and Earth System Sciences

# Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application

**A. Elshorbagy**[1], **G. Corzo**[2], **S. Srinivasulu**[1], **and D. P. Solomatine**[2,3]

[1]Centre for Advanced Numerical Simulation (CANSIM), Department of Civil and Geological Engineering, University of Saskatchewan, Saskatoon, SK, S7N 5A9, Canada
[2]Department of Hydroinformatics and Knowledge Management, UNESCO-IHE Institute for Water Education, Delft, The Netherlands
[3]Water Resources Section, Delft University of Technology, Delft, The Netherlands

Received: 29 October 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 19 November 2009
Revised: 19 August 2010 – Accepted: 6 September 2010 – Published: 14 October 2010

**Abstract.** In this second part of the two-part paper, the data driven modeling (DDM) experiment, presented and explained in the first part, is implemented. Inputs for the five case studies (half-hourly actual evapotranspiration, daily peat soil moisture, daily till soil moisture, and two daily rainfall-runoff datasets) are identified, either based on previous studies or using the mutual information content. Twelve groups (realizations) were randomly generated from each dataset by randomly sampling without replacement from the original dataset. Neural networks (ANNs), genetic programming (GP), evolutionary polynomial regression (EPR), Support vector machines (SVM), M5 model trees (M5), K-nearest neighbors (K-nn), and multiple linear regression (MLR) techniques are implemented and applied to each of the 12 realizations of each case study. The predictive accuracy and uncertainties of the various techniques are assessed using multiple average overall error measures, scatter plots, frequency distribution of model residuals, and the deterioration rate of prediction performance during the testing phase. Gamma test is used as a guide to assist in selecting the appropriate modeling technique. Unlike two nonlinear soil moisture case studies, the results of the experiment conducted in this research study show that ANNs were a sub-optimal choice for the actual evapotranspiration and the two rainfall-runoff case studies. GP is the most successful technique due to its ability to adapt the model complexity to the modeled data. EPR performance could be close to GP with datasets that are more linear than nonlinear. SVM is sensitive to the kernel choice and if appropriately selected, the performance of SVM can improve. M5 performs very well with linear and semi linear data, which cover wide range of hydrological situations. In highly nonlinear case studies, ANNs, K-nn, and GP could be more successful than other modeling techniques. K-nn is also successful in linear situations, and it should not be ignored as a potential modeling technique for hydrological applications.

*Correspondence to:* A. Elshorbagy
(amin.elshorbagy@usask.ca)

## 1 Introduction

The research methodology explained in the first part of this two-companion paper was implemented in the sequence presented earlier. First, inputs of the various models were identified. A mixed approach of input selection was adopted since identification of optimum inputs was not in itself one of the objectives of this study. The next section describes the five different datasets. The two soil moisture datasets (Elshorbagy and Parasuraman, 2008) and a reduced hourly version of the evapotranspiration (AET) dataset (Parasuraman and Elshorbagy, 2008; Parasuraman et al., 2007) were used in earlier studies. This study benefited from the input structure identified in the earlier studies, and sometimes (e.g., the case of the evapotranspiration dataset) enhanced the input structure by considering more inputs identified using the mutual information content.

## 2 Datasets

### 2.1 Actual evapotranspiration

The eddy covariance (EC)-measured actual evapotranspiration data from the South West Sand Storage (SWSS) facility, located near Ft. McMurray, Alberta, Canada, is considered

in this study. The SWSS is currently the largest operational tailings dam in the world, holding approximately 435 million cubic meters of material, covering 25 $km^2$, and standing approximately 40 m high with a 20H:1V side-slope ratio. Soils consist of mine tailings sand overlain with 0.4 to 0.8 m of topsoil that is a mixture of peat and secondary mineral soil with a clay loam texture. Both vegetation species and composition vary across the SWSS, with dominant groundcover including horsetail (Equisetum arvense), fireweed (Epilobium angustifolia), sow thistle (Sonchus arvense), and white and yellow sweet clover (Melilotus alba, Melilotus officinalis). Tree and shrub species include Siberian larch (Larix siberica), hybrid poplar (Populus sp. hybrid), trembling aspen (Populus tremuloides), white spruce (Picea glauca), and willow (Salix sp.). For the SWSS facility, the ground-water table is located well below the rooting zone, at a depth between 0.8–1.0 m, and hence does not directly contribute to the evapotranspiration process. Accurate estimation of actual evapotranspiration from the reconstructed watersheds is of vital importance as it plays a major role in the water-balance of the system, which links directly to ecosystem restoration strategies. The weather station located on top of the SWSS facility measured the air temperature (AT) (°C), ground temperature (GT) (°C), net radiation (NR) ($W/m^2$), relative humidity (RH), and wind speed (WS) (m/s). Turbulent fluxes of heat and water vapor were measured using a CSAT3 sonic anemometer and thermometer (Campbell Scientific) and an LI-7500 CO2/H2O gas analyzer (Li-Cor). Ground heat flux was measured using a cm3 radiation and energy balance (REBS) ground heat flux plate placed at 0.05 m depth. In EC technique, the covariance of vertical wind speed with temperature and water vapor is used to estimate the sensible heat (H) and latent heat (LE) fluxes (Parasuraman and Elshorbagy, 2008). More information on the EC technique can be found in Drexler et al. (2004). Raw turbulence measurements were made at 10 Hz and fluxes were calculated using 30-min block averages with a 2-D coordinate rotation.

The half hourly EC-measured LE flux (the product of the latent heat of vaporization and evapotranspiration) at the SWSS facility for two growing seasons (from 3 May to 21 Sep 2005 and from 27 May 9 Sep 2006) is considered in this study. The total precipitation during the two seasons is 275 mm and 265 mm, respectively and the average day-time reference evaporation rate is 0.27 mm/h. Nevertheless for modeling purposes, the day time (08:00 h – 20:00 h) evapotranspiration alone is considered. After eliminating records of missing data, the remaining number of data instances were 5,307 data points. Since evapotranspiration is commonly perceived as being highly dependent on climatic variables, the EC-measured LE flux is modeled as a function of NR, AT, GT, RH, and WS, as well as possible combinations of these variables. The descriptive statistics of the datasets used for training, cross validation, and testing are presented in Table 1. The coefficient of variation (CV) of different variables during training, cross validation, and testing are comparable.

**Table 1.** Descriptive statistics of the AET dataset.

| | NR $W/m^2$ | AT °C | GT °C | RH | WS m/s | LE $W/m^2$ |
|---|---|---|---|---|---|---|
| Training dataset | | | | | | |
| Minimum | −189.60 | −3.40 | 4.10 | 0.14 | 0.40 | −80.20 |
| Maximum | 875.40 | 33.90 | 27.20 | 0.96 | 10.20 | 503.80 |
| Mean | 229.70 | 18.70 | 16.70 | 0.50 | 2.80 | 144.90 |
| SD | 189.40 | 5.50 | 3.80 | 0.20 | 1.70 | 90.00 |
| CV | 0.82 | 0.29 | 0.23 | 0.34 | 0.62 | 0.62 |
| Cross validation dataset | | | | | | |
| Minimum | −119.80 | −3.20 | 3.70 | 0.16 | 0.40 | −42.20 |
| Maximum | 729.50 | 33.70 | 26.40 | 0.95 | 11.00 | 405.60 |
| Mean | 224.10 | 18.70 | 16.90 | 0.50 | 2.80 | 145.90 |
| SD | 181.90 | 5.60 | 3.80 | 0.20 | 1.70 | 88.70 |
| CV | 0.81 | 0.30 | 0.23 | 0.33 | 0.60 | 0.61 |
| Testing dataset | | | | | | |
| Minimum | −414.60 | −4.30 | 3.30 | 0.15 | 0.40 | −56.30 |
| Maximum | 801.60 | 33.80 | 27.20 | 0.96 | 12.30 | 425.80 |
| Mean | 226.90 | 18.50 | 16.60 | 0.50 | 2.90 | 143.80 |
| SD | 188.90 | 5.50 | 3.70 | 0.20 | 1.80 | 89.90 |
| CV | 0.83 | 0.30 | 0.22 | 0.34 | 0.63 | 0.63 |

The AET dataset, and each dataset, was randomly sampled 100 times; creating 100 realizations of the dataset with three split samples (training, cross-validation, and testing) created from every dataset realization.

### 2.2 Soil moisture content

Over the years, several large scale soil cover (reconstructed watersheds) experiments have been conducted to assess the performance of different reclamation strategies in northern Alberta, Canada, by studying the basic mechanisms that control the moisture movement within these covers. In particular, three experimental soil covers (D1, D2, and D3) were established in the year 1999. The experimental covers were constructed over the saline-sodic overburden with thickness of 0.50 m, 0.35 m, and 1.0 m, comprising a thin layer of peat mineral mix over varying thickness of secondary (glacial/till) soil. Cover D1 consists of 20 cm of peat overlying 30 cm of till, and it is considered for this study. The soil cover has an area of 1 ha (approximately 200 m long and 50 m wide), with a 5:1 slope (5 horizontal to 1 vertical). This reconstructed watershed, compared to natural watersheds, is not stable during its initial stages, and hence evolves over time to achieve hydro-sustainability. In order to track the evolution (hydrological changes) of the watershed, intensive instrumentations were installed in the watershed. Each watershed has an individual soil station located at the middle of the slope, which measures the volumetric soil moisture content of the upper peat (SMP) and the lower till (SMT) layers, twice a day. Soil moisture is measured using TDR principles with model CS615 (Boese, 2003). The TDR sensors were

installed laterally into the soil profile. Watershed D1 has eight TDR sensors installed over a depth range of 0.05 m to 1.00 m. Hourly values of soil temperature of peat (STP) and till (STT) layers are measured using thermisters buried in the watershed at the depth ranges corresponding to the TDR sensors. Consequently, D1 has eight soil temperature sensors. A weather station located in the mid-slope measures air temperature (AT), and precipitation ($P$). Similarly, Bowen station located at the mid-slope measures net-radiation (NR) and energy fluxes. All the meteorological variables are measured in an hourly scale. More details on the field instrumentation program and the data collected can be found in Boese (2003) and Elshorbagy et al. (2007).

Average daily values of precipitation, air temperature, soil temperature (STP and STT), net radiation (NR), soil moisture (SMP and SMT) as well as possible combinations of them, are considered for modeling purposes. The ground temperature and soil moisture contents are depth averaged for each layer (upper peat and lower till). As the soil stratum is frozen during the winter, only summer (May–September) time data of years 2000 till 2006 are considered. The total number of instances available for modeling purposes was 972 data points. As the reconstructed watersheds evolve over time to achieve hydro-sustainability, the freeze-thaw cycles and decomposition of highly organic peat layer increases the porosity of the soil and consequently increase infiltration rates (Haigh, 2000). Hence, modeling the moisture dynamics of such evolving watersheds would be adding to the already challenging task of modeling soil moisture. The descriptive statistics of the datasets used for training, cross validation, and testing are presented in Table 2 for the peat and the till layer datasets, respectively. For modeling purposes, two datasets were generated from the site; one for predicting SMP and the other for SMT. The same set of inputs was used in both datasets. The coefficient of variation (CV) of different variables during training, cross validation, and testing are comparable (Table 2).

**Table 2.** Descriptive statistics of the daily peat and till moisture datasets.

| | $P$ mm | AT °C | NR W/m$^2$ | STP °C | STT °C | SMP | SMT |
|---|---|---|---|---|---|---|---|
| Training dataset | | | | | | | |
| Minimum | 0.00 | −6.30 | −10.40 | 0.50 | −0.50 | 0.304 | 0.240 |
| Maximum | 43.70 | 25.20 | 204.40 | 18.20 | 16.30 | 0.539 | 0.316 |
| Mean | 1.54 | 13.63 | 90.64 | 11.71 | 10.48 | 0.442 | 0.288 |
| SD | 4.20 | 6.10 | 50.22 | 3.79 | 3.49 | 0.055 | 0.018 |
| CV | 2.72 | 0.45 | 0.55 | 0.32 | 0.33 | 0.124 | 0.062 |
| Cross validation dataset | | | | | | | |
| Minimum | 0.00 | −3.90 | 0.00 | 0.50 | −0.70 | 0.305 | 0.241 |
| Maximum | 27.18 | 22.90 | 226.10 | 18.20 | 16.10 | 0.542 | 0.316 |
| Mean | 1.68 | 13.80 | 92.96 | 11.75 | 10.32 | 0.440 | 0.289 |
| SD | 3.99 | 4.96 | 49.98 | 4.03 | 4.17 | 0.055 | 0.018 |
| CV | 2.38 | 0.36 | 0.54 | 0.34 | 0.40 | 0.125 | 0.062 |
| Testing dataset | | | | | | | |
| Minimum | 0.00 | −6.80 | 0.00 | −0.10 | −0.60 | 0.306 | 0.241 |
| Maximum | 23.60 | 25.80 | 223.60 | 18.20 | 16.10 | 0.543 | 0.316 |
| Mean | 1.48 | 14.07 | 96.94 | 11.88 | 10.45 | 0.440 | 0.288 |
| SD | 3.32 | 5.96 | 50.91 | 3.77 | 3.56 | 0.054 | 0.018 |
| CV | 2.25 | 0.42 | 0.53 | 0.32 | 0.34 | 0.123 | 0.061 |

### 2.3 Rainfall-runoff

The rainfall-runoff dataset used in this study is taken from the Ourthe subcatchment, which is a subcatchment of River Meuse The river basin covers part of France, Belgium and The Netherlands (Fig. 1). The area analyzed in this research is approximately 22 000 km$^2$, from Borgharen-dorp (Jeker basin on the Netherlands border) to Meuse source-St Mihiel (Lorraine basin in France). This meso-scale catchment system has been widely explored with data driven and expert knowledge (de Wit, 2001; Tu et al., 2005).

The greater part of the discharge of the River Meuse is supplied by its tributaries. Groundwater, precipitation and artificial extractions constitute the discharge to a smaller extent (de Wit, 2001). The Meuse has a great number of tributaries, varying greatly in their sizes. The largest is the Ourthe, with a contributing area of 3.626 km$^2$. The Ourthe subcatchment has large discharges rising fast. Through its nature and location, close to the Dutch border, the Ourthe is also the most important Meuse tributary for flood forecasts. In its upper course, the Ourthe consists of two branches: the Ourthe Occidentale and the Ourthe Orientale, merging near Nisramont. Near Comblain-au-Pont, the Amblève joins the Ourthe and near Angleur the Ourthe also receives the Vesdre. Measuring from the source of the Ourthe Occidentale, the Ourthe is approximately 175 km long.

The average travel time from upstream to downstream is one day (Berger, 1992). More information about the hydrological properties of the basin and the data validation are referred to Berger (1992) and De Wit (2007). The daily rainfall and runoff data of the Ourthe subcatchment from 11 January, 1988 till 31 December 1998 (4.008 data points) were used for modeling purposes in this study. Two distinct datasets were created: (i) the first is a dataset where only rainfall data were used as model inputs to predict the runoff; and (ii) the second is the same dataset but the preceding time step ($t$-1) runoff, in addition to the rainfall data, were used as inputs to predict the runoff at time $t$. The descriptive statistics of the variables that are used as model outputs in this study are presented in Table 3.

Figure 2 presents the inputs identified for the AET case study using AMI method. For the two rainfall-runoff datasets, the Average Mutual Information (AMI) method was used to identify the inputs for predicting the daily runoff (Fig. 3). The inputs-output of the five case studies are presented in Table 4. One should note that in light of the focus of this study, which is the comparative analysis of various data driven techniques, the important criterion is to use the same set of inputs across all adopted models. After inputs

**Table 3.** Descriptive statistics of the output variables of all datasets.

| | Evapo-transpiration ($W/m^2$) | Peat moisture | Till moisture | Runoff ($m^3/s$) |
|---|---|---|---|---|
| Count | 5307.00 | 972.00 | 972.00 | 4008.00 |
| Minimum | −80.20 | 0.30 | 0.24 | 1.07 |
| Median | 133.09 | 0.45 | 0.29 | 11.39 |
| Average | 144.52 | 0.44 | 0.29 | 21.91 |
| Maximum | 503.77 | 0.54 | 0.32 | 370.63 |
| St. deviation | 89.79 | 0.05 | 0.02 | 29.93 |
| CV | 0.62 | 0.12 | 0.06 | 1.37 |
| Skew | 0.51 | −0.72 | −1.33 | 4.06 |

have been identified, each dataset was randomly sampled 100 times; creating 100 realizations of the dataset with three split samples (one half for training, one sixth for cross-validation, and one third for testing) created from every dataset realization. Figure 4 shows an example of this process for the peat moisture dataset. Similar process was conducted with each one of the five case studies. Based on the similarity of the statistical properties (mean and standard deviation) of the three split samples, the best 12 realizations of each dataset are identified for the modeling exercise in this study.

## 3 Model implementation

### 3.1 Artificial neural networks (ANNs)

The Levenberg-Marquardt algorithm was used for training all neural network models using the MATLAB Neural Networks toolbox. For each of the 12 dataset realizations of a case study, the ANN was executed 200 times with 200 different random weight initializations. The best model of the 200 runs was identified as the best ANN model. The cross validation sub dataset was used to stop the training process. Accordingly, 12 ANN models were developed and tested using the corresponding unseen dataset. In all optimum ANN models, the number of input nodes was equivalent to the number of inputs, and all networks had one output node. The number of hidden nodes ranged from three to 13, with an average number of seven hidden nodes in single hidden layer ANNs.

### 3.2 Genetic programming (GP)

Discipulus Software (Francone, 2001) was used to implement the program-based GP to all datasets. GP was applied to the various dataset realizations similar to the way followed with ANNs. The addition, subtraction, multiplication, comparison, conditions, division, and trigonometric operators were allowed. The program size varied from 80-512 bits, with population size of 500 and generations without improvement up to 300. The probabilities of mutation and crossover were 30% and 50%, respectively. The program was allowed to run for at least two hours. The authors experimented with the run time and observed that improvement could be almost negligible beyond two hours. Similar to the case of ANN applications, 12 non-dominated GP models were developed and tested on the corresponding testing set of each case study.

### 3.3 Evolutionary polynomial regression (EPR)

The EPR Toolbox (Laucelli et al., 2005) was used to implement the static EPR technique to all datasets, following the same experimental steps adopted with the ANNs and the GP techniques. The EPR Toolbox allows for many choices in terms of the polynomial types, functions used within the polynomial terms, and the number of terms and exponents. In this study, the default number of terms (up to five) was used whereas a comprehensive search among the possible combinations of polynomial types and functions was conducted. Accordingly, 12 non-dominated EPR models were developed and tested on the corresponding testing set of each case study. The EPR type and function developed for each case study are presented in Table 5.

### 3.4 Support vector machine (SVM)

WEKA 3.6.0 Software (Bouckaert et al., 2008) was used in this study to implement the SVM to all datasets, following the same experimental steps adopted with the previous techniques. SVM models with linear, polynomial, and radial basis function (RBF) kernels were tested on all datasets. With the exception of the Rainfall – runoff II case study, the RBF kernel was found to provide the best predictive performance. In case of the rainfall-runoff II case study, both linear and RBF kernels were almost on par. Therefore, SVM with RBF kernel was adopted in this study. The constant C (Elshorbagy et al., 2010, Part 1) and the kernel parameter $\gamma$ were optimized from an exponential range of the following values: 0.0313; 0.0625; 0.125; 0.25; 0.50; 1.00; 2.00; 4.00; 8.00; and 16.00. Non-dominated 12 SVM models were developed and tested on the corresponding testing set of each case study.

### 3.5 M5 model trees

WEKA 3.6.0 Software (Bouckaert et al., 2008) was used in this study to implement the M5 model trees to all datasets, following the same experimental steps adopted with the previous techniques. The tree pruning coefficient was optimized during the execution of the models to minimize the average squared error. A range of values from 3–30 was tested in this study. 12 M5 model tree models were developed and tested on the corresponding testing set of each case study.

**Table 4.** Inputs and outputs of all case studies.

| Case study | Inputs | Output |
|---|---|---|
| Actual evapotranspiration (half hourly) | $AT_t$ ; $GT_t$ ; $GT_{t-1}$ ; $NR_t$ ; $NR_{t-1}$ ; $Sum(NR_{-4})$ ; $RH_t$ ; $WS_t$ | AET ($W/m^2$) |
| Upper layer (peat) soil moisture content (daily) | $P_t$ ; $AT_t$ ; $NR_t$ ; $STP_t$ ; $STT_t$ ; $Sum(P_{-6})$ ; $Sum(AT_{-6})$ | $SM_P$ (dimensionless) |
| Lower layer (till) soil moisture content (daily) | $P_t$ ; $AT_t$ ; $NR_t$ ; $STP_t$; $STT_t$; $Sum(P_{-6})$ ; $Sum(AT_{-6})$ | $SM_T$ (dimensionless) |
| Rainfall-runoff I (daily) | $P_t$ ;$P_{t-1}$ ; $P_{t-2}$ ; $P_{t-3}$ ; $P_{t-4}$ | $Q_{tI}$ ($m^3/s$) |
| Rainfall-runoff II (daily) | $P_t$ ;$P_{t-1}$ ; $P_{t-2}$ ; $P_{t-3}$ ; $P_{t-4}$ ; $Q_{t-1}$ | $Q_{tII}$ ($m^3/s$) |

AT: air temperature (°C); GT: ground temperature (°C); NR: net radiation ($W/m^2$); $Sum(NR_{-4})$: the cumulative net radiation over the preceding four time steps; RH: relative humidity; WS: wind speed (m/s); $P$: precipitation (mm); STP: depth averaged soil temperature of the upper peat layer (°C); STT: depth averaged soil temperature of the lower till layer (°C); $Sum(P_{-6})$: the cumulative precipitation over the preceding six time steps (mm); $Sum(AT_{-6})$: the cumulative air temperature over the preceding six time steps (°C); $SM_P$: depth averaged soil moisture content of the upper peat layer; $SM_T$: depth averaged soil moisture content of the lower till layer; and $Q_t$: the runoff ($m^3/s$).

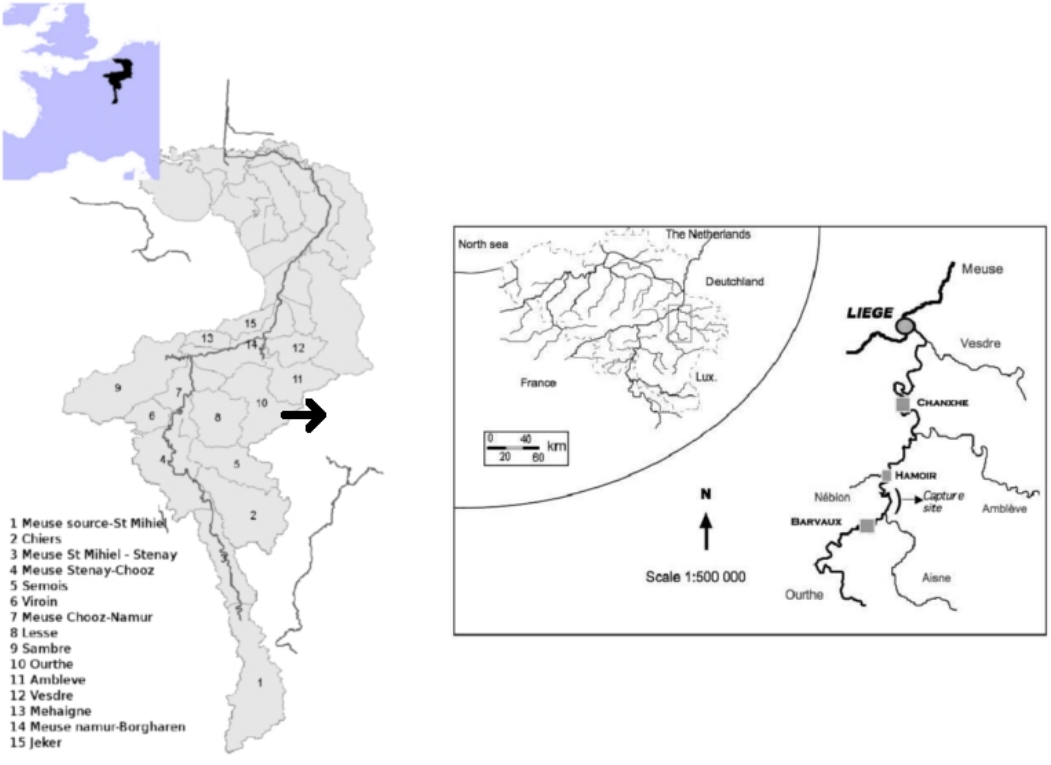


**Fig. 1.** The Meuse river basin and the sub-basins upstream of Borgharen. Sub-basin 10 (Ourthe) is used in the case study.

### 3.6 K-nearest neighbors (K-nn)

WEKA 3.6.0 Software (Bouckaert et al., 2008) was used in this study to implement the K-nn technique to all datasets, following the same experimental steps adopted with the previous techniques. The number of the nearest neighbors was optimized during the execution of the models to minimize the average squared error. A range of values from 1–50 neighbors was tested in this study. Accordingly, 12 K-nn models were developed and tested on the corresponding testing set of each case study. The ranges of the optimum numbers of nearest neighbors for each case study are presented in Table 6.

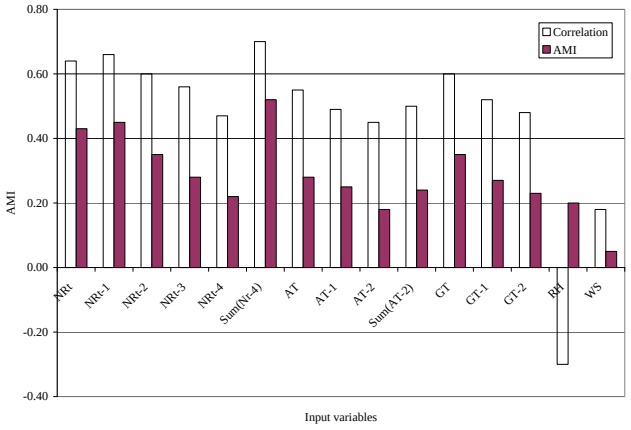


**Fig. 2.** Average mutual information and correlation of inputs-output for the evapotranspiration case study.

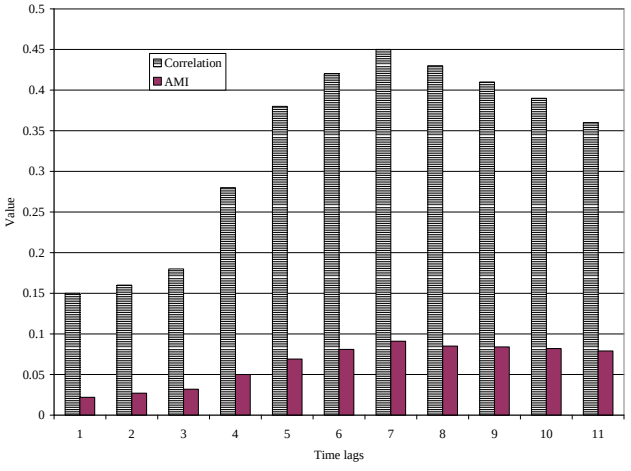


**Fig. 3.** Average mutual information and correlation of inputs-output for the rainfall-runoff case study.

## 4 Results and analysis

### 4.1 Evapotranspiration case study

The performance of the various techniques applied to the half-hourly actual evapotranspiration (AET) case study is provided in Table 7. The best, the worst, and the average of the performances of the 12 models of all techniques are shown. It is certainly useful to judge techniques based on the range of performances (difference between the best and the worst models), however, if a single value is needed, then one has to rely on the average performance. Table 7 supports the idea that in most cases, it is not possible to find a technique that dominates others with respect to all error measures. But if a technique is better than the rest with respect to two different error measures (e.g., RMSE and R), this can be a strong indication of the superiority of such a technique. In the AET case study, GP, SVM, M5 model trees, and K-nn techniques

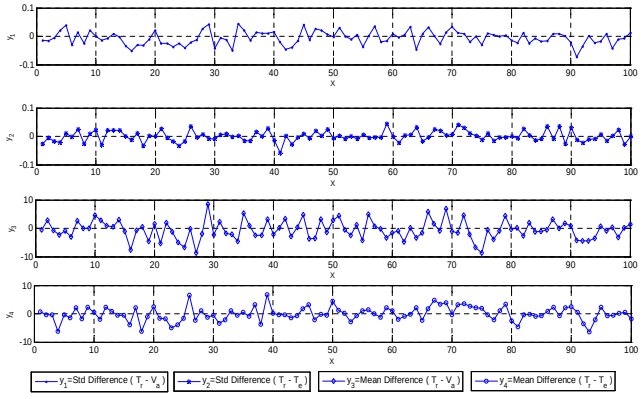


**Fig. 4.** Statistical properties of the training/cross-validation/testing subsets for 100 random realizations. $T_r$: Training; $V_a$: Validation; $T_e$: Testing.

can be identified as the best techniques, followed by EPR, in terms of the predictive accuracy. The performance of the ANNs was worse than the linear regression (MLR) technique in this particular case study. This highlights the important fact that the half-hourly AET data were captured reasonably well in a linear relationship considering the provided model inputs. Therefore, a technique that forces nonlinear structures on the input-output relationship (ANNs) may not be favorable in all cases. Certainly, the AET data are not strictly linear; that is why local and/or modular linear models (e.g., M5 and K-nn) could be optimum choices.

Since all 12 models of each technique are non-dominated models and represent possible performances of the technique under consideration, the output of all 12 models are integrated in one set and presented in Fig. 5. The figure shows the scatter plots of observed vs. predicted AET data. The scatter around the 45-degree line supports the conclusion made earlier regarding the performances of the various techniques. However, the plots allow to make two additional observations; first, all techniques were less successful in predicting high values. The tips of the data plumes were always below the 45-degree line. This might be an indication that the ideal inputs that can describe all dynamics of the process for this case study have not been optimally identified. The SVM (Fig. 5d) was more successful than other techniques in approaching the high values. The M5 model trees and MLR (Fig. 5e and g) were the least successful in this regard. Table 8 shows the ideal point error (IPE) measure calculated for all techniques. The IPE statistic, integrating all four error measures in one indicator, lends another support to the conclusions made earlier. Except the ANNs, all other techniques have close performances, with the possibility of identifying the SVM, GP, M5, and K-nn; followed by EPR as better techniques than the rest. The utility of the idea of adopting multiple models (12 in this study) based on different random realizations of the datasets to evaluate

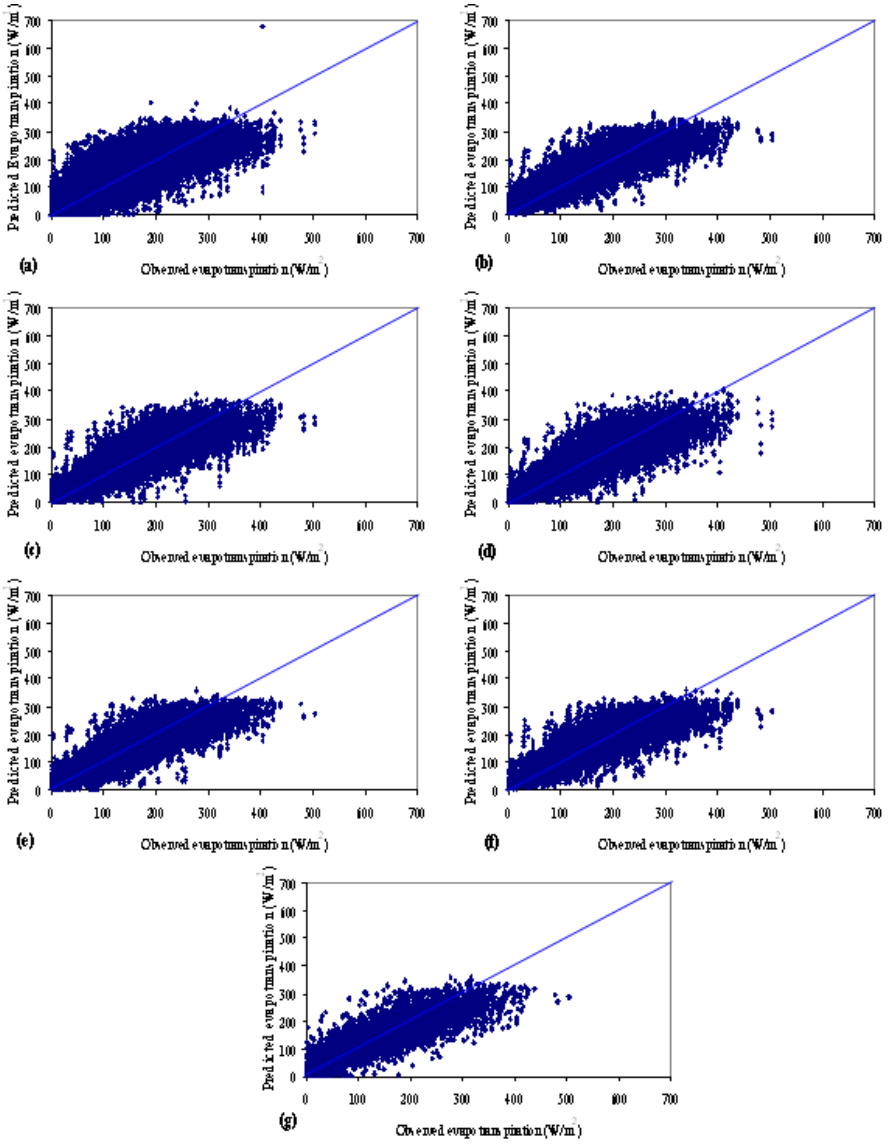


**Fig. 5.** scatter plots of observed and predicted evapotranspiration. **(a)** ANNs, **(b)** GP, **(c)** EPR, **(d)** SVM, **(e)** M5, **(f)** K-nn, and **(g)** MLR.

**Table 5.** EPR type and functions of all case studies.

| Case study | EPR type | Function ($f$) |
|---|---|---|
| Actual evapotranspiration (half hourly) | Sum [$a_i$ *X1*X2*f(X1*X2)] + $a_o$ | No function |
| Upper layer (peat) soil moisture content (daily) | Sum [$a_i$ *f(X1*X2)] + $a_o$ | Exponential |
| Lower layer (till) soil moisture content (daily) | Sum [$a_i$ *f(X1*X2)] + $a_o$ | Logarithm |
| Rainfall-runoff I (daily) | Sum [$a_i$ *X1*X2*f(X1)*f(X2)] + $a_o$ | No function |
| Rainfall-runoff II (daily) | Sum [$a_i$ *X1*X2*f(X1)*f(X2)] + $a_o$ | No function |

**Table 6.** The optimum number of nearest neighbors (K-nn) of the 12 models in each case study.

| | All 12 values | Min. | Average | Max. |
|---|---|---|---|---|
| evapotranspiration | 17–28–10–21–21–34–22–18–26–9–15–40 | 9 | 22 | 40 |
| Upper layer soil moisture | 4–4–9–5–4–3–5–5–12–7–4–4 | 3 | 6 | 12 |
| Lower layer soil moisture | 9–4–2–2–8–10–7–3–5–6–9–6 | 2 | 6 | 10 |
| Rainfall-runoff I | 19–33–9–11–3–18–8–24–44–12–6–13 | 3 | 17 | 44 |
| Rainfall-runoff II | 2–7–3–4–1–2–2–3–6–3–3–5 | 1 | 3 | 7 |

**Table 7.** Testing results of all models applied to the *evapotranspiration* dataset.

| Models | RMSE | | | MARE | | | MB | | | R | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst |
| ANNs | 46.32 | 57.08 | 85.77 | 0.52 | 1.25 | 2.25 | −1.59 | 5.87 | 57.95 | 0.87 | 0.84 | 0.74 |
| GP | 42.17 | **43.90** | 46.04 | 0.58 | 0.69 | 0.84 | −0.09 | 0.27 | 1.65 | 0.88 | **0.87** | 0.86 |
| EPR | 44.69 | 46.30 | 48.04 | 0.62 | 0.82 | 1.07 | 0.01 | 0.90 | 3.08 | 0.87 | 0.86 | 0.85 |
| SVM | 41.61 | 44.52 | 49.25 | 0.48 | **0.54** | 0.64 | −1.26 | −2.84 | −4.90 | 0.84 | **0.87** | 0.88 |
| M5 | 42.85 | **44.42** | 46.19 | 0.53 | 0.63 | 0.72 | 0.17 | **−0.03** | 1.97 | 0.86 | **0.87** | 0.88 |
| K−nn | 43.05 | 44.65 | 46.42 | 0.58 | 0.69 | 0.80 | 0.09 | −0.39 | −2.16 | 0.88 | **0.87** | 0.86 |
| MLR | 46.81 | 48.49 | 50.27 | 0.78 | 0.93 | 1.13 | −0.15 | 0.14 | 2.81 | 0.85 | 0.84 | 0.83 |
| Naïve | – | – | – | – | – | – | – | – | – | – | – | – |

various techniques presents itself through Tables 7 and 8. If the modeler picks, for example, the best model of one technique and compares it with the worst model of another technique, a different and perhaps biased conclusion might be made regarding the performance of these techniques. The best ANN model with IPE value of 0.31 is much better than the worst EPR model with IPE value of 0.37 (Table 8).

Based on the outputs of the 12 non-dominated models of each technique, the predictive uncertainty of the various techniques can be easily analyzed. The residuals (predicted value minus observed value) of the 12 models were integrated in one set to conduct probabilistic analysis. Frequency curves were constructed for the residuals of each technique. @RISK Software (Palisade Corporation, 2005) was used to fit the best probability distribution from a selection of more than 15 possible distributions. The best-found probability distributions of the residuals of the various techniques are shown in Fig. 6. The Logistic ($\alpha$, $\beta$) distribution was found to fit the residuals of all modeling techniques, with different values of location parameter $\alpha$ and scale parameter $\beta$. Ideally, the best technique is the one that has residuals represented by the narrowest, symmetrical, and tallest (has the highest probability value at zero residuals) probability distribution. Such a distribution implies the smallest level of predictive uncertainty, which could be translated to the highest level of reliability. Figure 6 reveals that, not only in terms of the predictive accuracy, but also the predictive uncertainty SVM is the best, flollowed by GP, K-nn, M5 and EPR. Clearly, the ANN technique leads to the most uncertain results with the widest range of residuals, whereas the MLR is occupying the middle position.

The Kolmogorov-Smirnov (KS) nonparametric test was conducted on the model residuals of all techniques to test the null hypothesis that the model residuals of any two techniques are sampled from the same distribution. The test was conducted at the default significance level of $p = 0.05$. The matrix of the $p$-values is given as Table 9. With the exception of K-nn and M5 techniques, there is strong statistical evidence that the residuals of the various techniques are stemming from different distributions. Even though the visual assessment of Fig. 6 shows that the SVM, M5, and K-nn are very similar, the KS test indicates that the SVM performs differently.

### 4.2 Peat (upper layer) soil moisture case study

The performance of the various techniques applied to the daily soil moisture data of the upper peat layer (SMP) case study is provided in Table 10. Unlike the evapotranspiration case study, Table 10 shows that both ANNs and GP techniques can be considered superior to other modeling techniques due to their domination with respect to the four error measures. It has to be noted that in case of soil moisture content, low values of the RMSE and the MARE might be misleading because the entire dataset is limited to a narrow

**Table 8.** IPE testing results of all models applied to all datasets.

| | Evapotranspiration (AET) | | | Peat moisture (SMP) | | | Till moisture (SMT) | | | Rainfall-runoff I (R-R II) | | | Rainfall-runoff II (R-R I) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst |
| ANNs | 0.31 | 0.51 | 0.79 | 0.58 | *0.65* | 0.71 | 0.49 | 0.57 | 0.92 | 0.51 | 0.57 | 0.69 | 0.24 | 0.47 | 0.78 |
| GP | 0.29 | *0.30* | 0.33 | 0.56 | **0.63** | 0.72 | 0.43 | *0.53* | 0.67 | 0.50 | 0.55 | 0.58 | 0.17 | **0.20** | *0.22* |
| EPR | 0.31 | 0.33 | 0.37 | *0.65* | 0.68 | 0.72 | 0.55 | 0.58 | 0.63 | 0.52 | 0.56 | 0.68 | 0.19 | 0.22 | 0.28 |
| SVM | 0.28 | **0.29** | 0.32 | 0.65 | 0.80 | 0.90 | 0.55 | 0.60 | 0.69 | 0.52 | 0.57 | 0.62 | 0.24 | 0.37 | 0.54 |
| M5 | 0.29 | *0.30* | 0.31 | 0.57 | 0.64 | 0.74 | 0.49 | 0.56 | 0.63 | 0.50 | **0.52** | *0.53* | 0.18 | **0.20** | 0.22 |
| K-nn | 0.29 | *0.31* | 0.32 | 0.57 | 0.65 | 0.71 | 0.44 | **0.51** | 0.52 | 0.52 | 0.54 | 0.57 | 0.55 | 0.59 | 0.67 |
| MLR | 0.34 | 0.36 | 0.39 | 0.72 | 0.74 | 0.78 | 0.57 | 0.60 | 0.63 | 0.51 | 0.53 | 0.55 | 0.44 | 0.48 | 0.52 |
| Naïve | – | – | – | – | – | – | – | – | – | – | – | – | 0.32 | 0.35 | 0.42 |

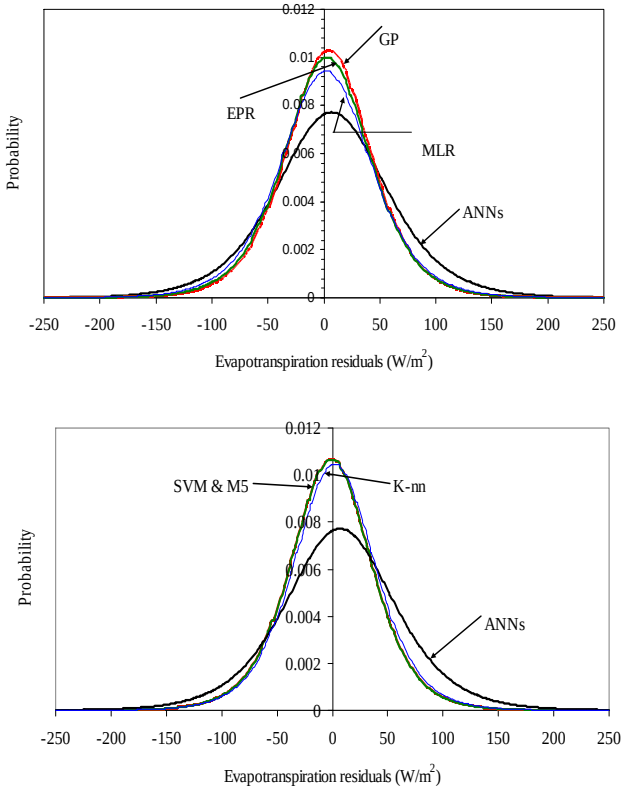


**Fig. 6.** Probability distribution of the 12 model residuals of all techniques (evapotranspiration case study).

range (0.30–0.55) of values (Table 3). In this case, the R statistic becomes the most important indicator (Elshorbagy and Parasuraman, 2008). For example, if an average-all model is constructed just by assuming that the best predictor is the average soil moisture value of all observations in the training dataset, the predicted value will be always 0.442. In this case, the RMSE and the MARE values are 0.05 and 0.10, respectively, but the R statistic value is almost zero; indicating an extremely poor model. Accordingly, ANNs and

**Table 9.** The *p*-values of the two samples K-S test on the model residuals (evapotranspiration).

| | ANNs | GP | EPR | SVM | M5 | K-nn | MLR |
|---|---|---|---|---|---|---|---|
| ANNs | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| GP | | 1 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| EPR | | | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| SVM | | | | 1 | 0.0000 | 0.0000 | 0.0000 |
| M5 | | | | | 1 | **0.051** | 0.0000 |
| K-nn | | | | | | 1 | 0.0000 |
| MLR | | | | | | | 1 |

GP are the best modeling techniques for this case study (producing the *R* values of 0.60 and 0.61, respectively), followed by the K-nn and the M5 techniques. The MLR is clearly inferior to other techniques, which points to the possibility that the SMP dataset is a highly nonlinear dataset. The authors believe that this is a major reason for the relative success of ANNs in this case study compared to the previous (AET) case study. The moisture storage effect (Elshorbagy and El-Baroudy, 2009; Elshorbagy and Parasuraman, 2008) attributes to the nonlinearity of the process. Techniques that can handle highly nonlinear data (ANNs and GP) were quite successful, followed closely by local/modular models (M5 and K-nn). Even though the EPR technique was relatively close to the K-nn and M5, the performance of the SVM technique was the poorest with an *R* value of 0.44; slightly higher than the MLR.

The scatter plots (Fig. 7) show clearly that the error measures, including the IPE (Table 8), reflect only the average overall performance of the models, and favored models that produce scatter with less dispersion (e.g., GP and EPR). However, the plots reveal that ANNs outperforms other techniques where, at least, the trend of the higher range of peat moisture values was captured better than the other techniques could do. Similar to the AET case study, frequency curves were constructed for the residuals of each technique (Fig. 8). Interestingly, the best-found probability distributions of the residuals of the various techniques differed. The LogLogistic

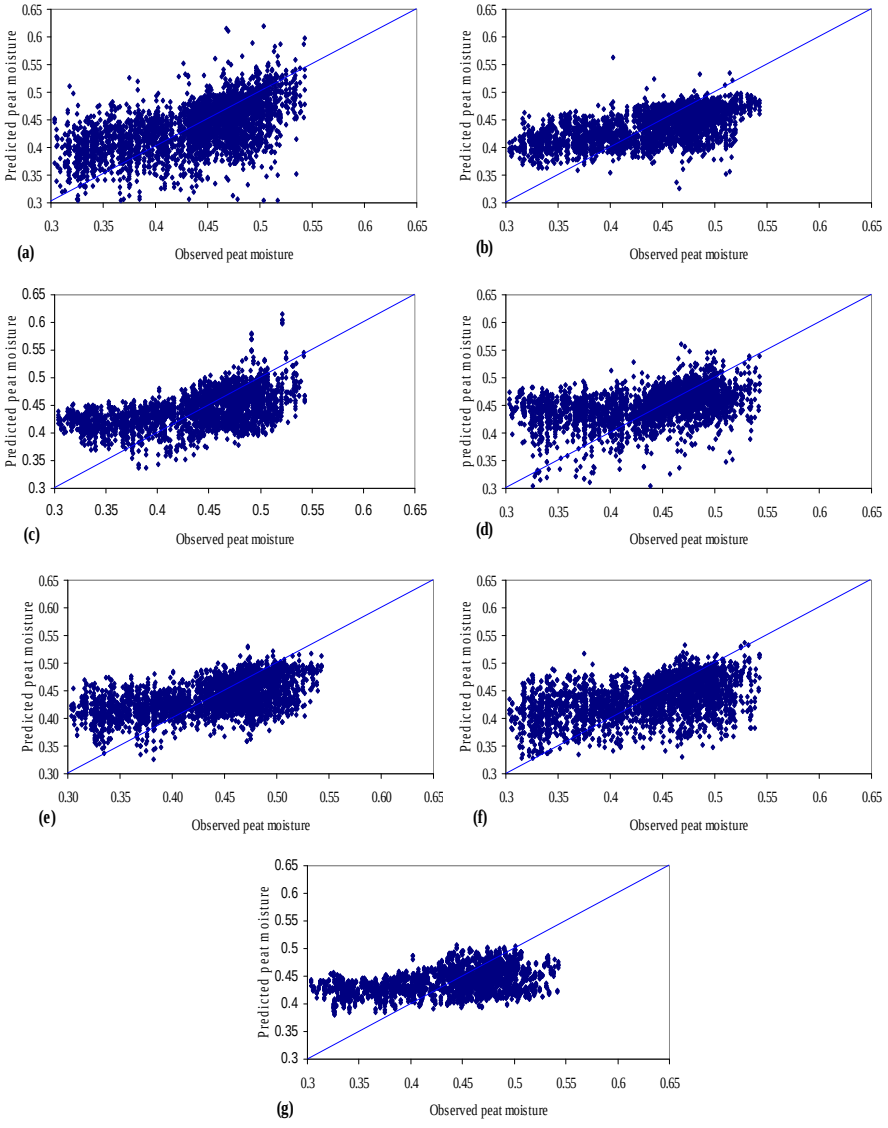


**Fig. 7.** Scatter plots of observed and predicted peat moisture content, **(a)** ANNs, **(b)** GP, **(c)** EPR, **(d)** SVM, **(e)** M5, **(f)** K-nn, and **(g)** MLR.

**Table 10.** Testing results of all models applied to the *Peat moisture* dataset.

| Models | RMSE | | | MARE | | | MB | | | R | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst |
| ANNs | 0.041 | 0.046 | 0.051 | 0.076 | **0.083** | 0.090 | 0.000 | **−.001** | −.009 | 0.66 | **0.60** | 0.53 |
| **GP** | 0.040 | **0.044** | 0.050 | 0.076 | **0.084** | 0.091 | 0.000 | **−.001** | −.007 | 0.70 | **0.61** | 0.47 |
| **EPR** | 0.045 | 0.047 | 0.050 | 0.087 | 0.091 | 0.097 | 0.000 | **0.001** | 0.006 | 0.56 | 0.52 | 0.46 |
| **SVM** | 0.048 | 0.051 | 0.053 | 0.081 | 0.092 | 0.098 | −.004 | 0.011 | 0.016 | 0.57 | 0.44 | 0.35 |
| M5 | 0.041 | **0.045** | 0.050 | 0.075 | **0.084** | 0.098 | 0.001 | **0.000** | 0.004 | 0.66 | 0.57 | 0.37 |
| K-nn | 0.042 | 0.047 | 0.051 | 0.073 | **0.083** | 0.090 | 0.000 | **0.000** | 0.005 | 0.62 | 0.53 | 0.43 |
| MLR | 0.049 | 0.050 | 0.052 | 0.96 | 0.099 | 0.104 | 0.000 | 0.001 | 0.004 | 0.43 | 0.40 | 0.33 |
| Naïve | – | – | – | – | – | – | – | – | – | – | – | – |

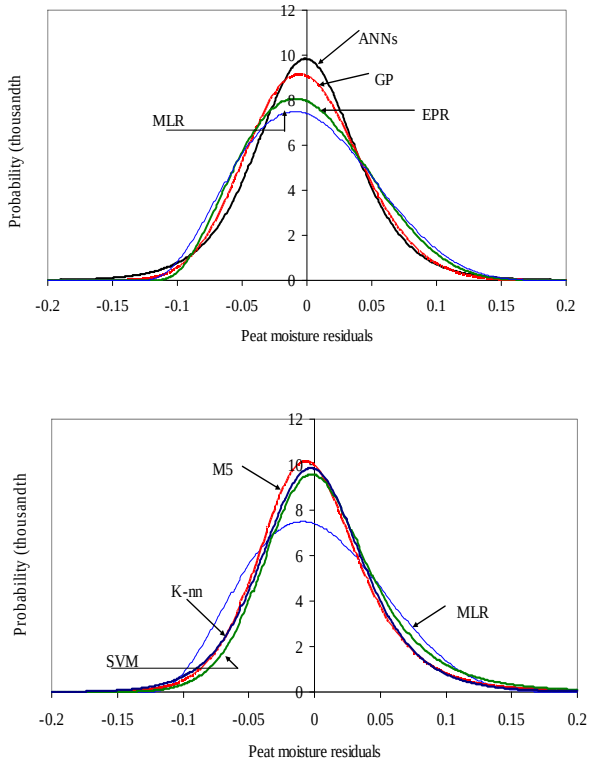


**Fig. 8.** Probability distribution of the 12 model residuals of all techniques (peat moisture case study).

($\gamma$, $\beta$, $\alpha$) probability distribution was found to fit the residuals of SVM and K-nn, and M5 modeling techniques, Logistic ($\alpha$, $\beta$) for ANNs, Lognormal ($\mu$, $\sigma$) for GP, Beta ($\alpha 1$, $\alpha 2$) for EPR and MLR techniques. This reflects the fact that the adopted modeling techniques are different in the way that they predict the output and minimize the errors, even if their average overall error values are close. The frequency curves reflect the considerable outperformance of the ANNs, K-nn, M5, and SVM over other more uncertain and biased techniques, such as MLR and the EPR techniques. An important observation here is the lower uncertainty of the SVM technique. The small uncertainty of the SVM technique reflected by the probability distribution is affected by the narrow range of residuals and small overall RMSE, however, the SVM models are poor in capturing the trend of the SMP data – this is indicated by the lower $R$ value.

The Kolmogorov-Smirnov nonparametric test was conducted on the model residuals of all techniques to test the null hypothesis that the model residuals of any two techniques are sampled from the same distribution. The matrix of the $p$-values is given in Table 11. There is strong statistical evidence that the residuals of the various techniques are stemming from different populations.

**Table 11.** The $p$-values of the two samples K-S test on the model residuals (peat moisture).

| | ANNs | GP | EPR | SVM | M5 | K-nn | MLR |
|---|---|---|---|---|---|---|---|
| ANNs | 1 | 0.0000 | 0.0021 | 0.0000 | 0.0000 | 0.0156 | 0.0000 |
| GP | | 1 | 0.0001 | 0.0000 | 0.0015 | 0.0000 | 0.0000 |
| EPR | | | 1 | 0.0000 | 0.0003 | 0.0000 | 0.0069 |
| SVM | | | | 1 | 0.0000 | 0.0000 | 0.0000 |
| M5 | | | | | 1 | 0.0000 | 0.0000 |
| K-nn | | | | | | 1 | 0.0000 |
| MLR | | | | | | | 1 |

### 4.3 Till (lower layer) moisture case study

The till moisture case study (SMT) is similar to the previous case study with regard to the small variability in the dataset, and the nonlinear response to the climatic variables due to the large storage effect. Table 3 shows that the variability (CV) of the till moisture data is half of that of the peat moisture data, whereas the skew in the till moisture dataset is nearly double that of the peat moisture. The error measures shown in Table 12 (and in particular the R statistic) reveal that K-nn, GP, and ANNs are better candidates than other modeling techniques based on the same argument mentioned earlier regarding the $R$ statistic. Similar to the previous case study, SVM and MLR techniques were the lowest in the rank with regard to the prediction accuracy. The small variability, combined with the high nonlinearity, of the SMT dataset contributed to the relative success of the K-nn technique in this particular case study. The failure of the MLR is an indicator of the potential utility of the ANNs for modeling the SMT.

Frequency curves were constructed for the residuals of each technique (Fig. 9) to investigate the predictive uncertainty. The graph in this case provides useful and more insightful view of the predictive reliability of the various techniques. The K-nn, GP, ANNs, and the SVM are clearly less uncertain and less skewed than EPR and other linear techniques (M5 and MLR) in this case study. The best-found probability distributions of the residuals of the various techniques differed across techniques. The LogLogistic ($\gamma$, $\beta$, $\alpha$) distribution was found to fit the residuals of SVM and K-nn, and ANNs modeling techniques, Logistic ($\alpha$, $\beta$) for GP, Lognormal ($\mu$, $\sigma$) for EPR and MLR, and ExtremeValue (a, b) for M5. This reflects the fact that some of the adopted modeling techniques are really different in the way that they predict the output and minimize the errors, whereas some similarity is identified among the ANNs, K-nn, and SVM techniques. This similarity is only in terms of approaching the optimum solution, and leaving model residuals to be similarly distributed, but not necessarily in the distribution parameters. Similar to the SMP case study, less uncertainty with the use of the SVM is due to model residuals that stay around the mean, and thus, reduce the variability and the average error. This should not be confused with the poor

**Table 12.** Testing results of all models applied to the Till moisture dataset.

| Models | RMSE | | | MARE | | | MB | | | R | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave. | Worst | Best | Ave. | Worst | Best | Ave. | Worst | Best | Ave. | Worst |
| ANNs | 0.014 | 0.015 | 0.020 | 0.037 | **0.041** | 0.058 | 0.000 | −.002 | −.006 | 0.63 | *0.55* | 0.21 |
| GP | 0.012 | **0.015** | 0.020 | 0.034 | **0.040** | 0.046 | 0.000 | −.001 | 0.002 | 0.72 | **0.57** | 0.38 |
| EPR | 0.015 | 0.016 | 0.017 | 0.042 | **0.044** | 0.047 | 0.000 | **0.000** | 0.002 | 0.52 | 0.44 | 0.32 |
| SVM | 0.015 | 0.016 | 0.017 | 0.038 | **0.040** | 0.043 | 0.001 | 0.003 | 0.005 | 0.57 | 0.48 | 0.32 |
| M5 | 0.014 | 0.016 | 0.017 | 0.037 | **0.042** | 0.047 | 0.000 | **0.000** | 0.002 | 0.59 | 0.46 | 0.30 |
| K-nn | 0.013 | **0.015** | 0.017 | 0.034 | **0.038** | 0.040 | 0.000 | **0.000** | 0.002 | 0.70 | **0.57** | 0.49 |
| MLR | 0.016 | 0.016 | 0.017 | 0.043 | **0.045** | 0.047 | 0.000 | **0.000** | 0.002 | 0.50 | 0.41 | 0.32 |
| Naïve | – | – | – | – | – | – | – | – | – | – | – | – |

**Table 13.** The *p*-values of the two samples K-S test on the model residuals (till moisture).

| | ANNs | GP | EPR | SVM | M5 | K-nn | MLR |
|---|---|---|---|---|---|---|---|
| ANNs | 1 | 0.0040 | 0.0043 | 0.0000 | 0.0094 | 0.0000 | 0.0000 |
| GP | | 1 | 0.0400 | 0.0000 | **0.1843** | 0.0000 | 0.0006 |
| EPR | | | 1 | 0.0000 | 0.1667 | 0.0000 | 0.0101 |
| SVM | | | | 1 | 0.0000 | 0.0001 | 0.0000 |
| M5 | | | | | 1 | 0.0000 | 0.0007 |
| K-nn | | | | | | 1 | 0.0000 |
| MLR | | | | | | | 1 |

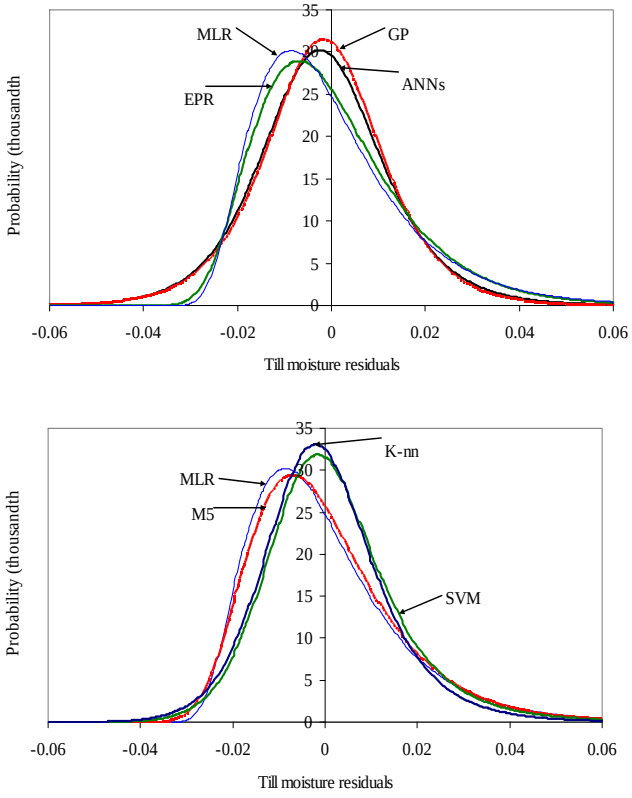


**Fig. 9.** Probability distribution of the 12 model residuals of all techniques (till moisture case study).

accuracy of capturing trends in the data (low *R* value in Table 12 and even high IPE value in Table 8).

The Kolmogorov-Smirnov nonparametric test was conducted on the model residuals (raw data; not fitted distribution) of all techniques to test the null hypothesis that the model residuals of any two techniques are sampled from the same population. The matrix of the *p*-values is given as Table 13. The K-S test reveals that there is no evidence to reject the hypothesis in the case of the EPR and M5, and also GP and M5. The visual analysis of Fig. 9 confirms the finding regarding EPR and M5; however, M5 and GP are visually different. The reason is that the graph presents the best-fit distributions that should be used to make conclusions regarding the potential of the techniques and their possible performance on untested cases in the future. The K-S is a nonparametric test that relies on the cumulative frequency of the sample itself. For the rest of the adopted techniques, there is strong statistical evidence that the residuals of the various techniques are stemming from different populations.

### 4.4 Rainfall-runoff case study I

The performance of the various techniques applied to the daily rainfall-runoff I (R-R I) case study is provided in Table 14. In this case study, the preceding runoff was not used as an input for the models, therefore, the information content can be considered limited (only rainfall of the current and the three preceding days were used). The performances of all techniques were almost on par as shown by close values of average RMSE and *R* (Table 14) as well as close values of the IPE indicator (Table 8). Nonetheless, one can observe that M5, GP, and MLR were slightly better and less biased (lower MB values) than the other techniques. In a situation like this R-R I case study, where the information content itself is limited; it may not be possible to differentiate among the various modeling techniques. The limiting factor for the prediction accuracy becomes the information content rather

**Table 14.** Testing results of all models applied to the Rainfall-Runoff I .

| Models | RMSE | | | MARE | | | MB | | | R | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst |
| ANNs | 24.94 | 26.25 | 28.24 | 1.04 | 1.47 | 2.03 | 0.59 | −2.25 | −8.78 | 0.59 | 0.53 | 0.40 |
| GP | 22.93 | **24.91** | 27.49 | 1.61 | 1.71 | 1.83 | 0.52 | 1.05 | 1.84 | 0.66 | 0.57 | 0.52 |
| EPR | 24.32 | 27.05 | 39.93 | 1.55 | 1.69 | 1.81 | −0.05 | **0.05** | 1.66 | 0.61 | 0.54 | 0.49 |
| SVM | 25.22 | 26.16 | 26.83 | 1.01 | **1.11** | 1.18 | −4.99 | −6.06 | −7.75 | 0.60 | 0.54 | 0.47 |
| M5 | 24.11 | **24.64** | 25.57 | 1.48 | 1.60 | 1.65 | 0.05 | −0.47 | −1.92 | 0.62 | **0.58** | 0.54 |
| K−nn | 25.13 | 25.98 | 27.39 | 1.45 | 1.58 | 1.70 | −0.74 | −1.55 | −3.28 | 0.58 | 0.52 | 0.44 |
| MLR | 24.20 | **24.93** | 25.77 | 1.5 | 1.61 | 1.71 | 0.01 | 0.12 | −1.55 | 0.60 | 0.56 | 0.53 |
| Naïve | – | – | – | – | – | – | – | – | – | – | – | – |

**Table 15.** Testing results of all models applied to the Rainfall-Runoff II.

| Models | RMSE | | | MARE | | | MB | | | R | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave. | Worst | Best | Ave | Worst |
| ANNs | 5.61 | 9.13 | 14.77 | 0.10 | 0.21 | 0.43 | −0.27 | −0.69 | 7.54 | 0.99 | 0.97 | 0.91 |
| GP | 4.28 | **4.92** | 6.03 | 0.09 | 0.11 | 0.14 | 0.03 | 0.06 | 0.62 | 0.99 | **0.99** | 0.98 |
| EPR | 4.69 | 5.55 | 6.95 | 0.10 | 0.11 | 0.15 | 0.02 | 0.01 | −0.34 | 0.99 | 0.98 | 0.97 |
| SVM | 6.47 | 10.12 | 15.62 | 0.09 | 0.12 | 0.15 | −0.02 | −0.59 | −1.53 | 0.98 | 0.94 | 0.87 |
| M5 | 4.4 | 5.2 | 6.0 | 0.09 | **0.09** | 0.10 | 0.00 | **0.00** | 0.44 | 0.99 | **0.99** | 0.98 |
| K−nn | 10.4 | 11.8 | 13.8 | 0.33 | 0.37 | 0.42 | −1.26 | −1.86 | −2.63 | 0.96 | 0.93 | 0.89 |
| MLR | 6.8 | 7.8 | 9.4 | 0.31 | 0.35 | 0.41 | −0.06 | 0.07 | 0.48 | 0.97 | 0.97 | 0.95 |
| Naïve | 8.8 | 10.1 | 12.1 | 0.12 | 0.12 | 0.12 | −0.01 | 0.04 | −0.44 | 0.96 | 0.94 | 0.92 |

than the predictive capability of the various techniques. A linear (e.g., MLR) or a modular linear (M5) technique is sufficient for such dataset.

The best-found probability distributions of the residuals of the various techniques did not differ. The Logistic ($\alpha$, $\beta$) probability distribution, with different parameter values for each technique, was found to fit the residuals of all modeling techniques. This reflects the fact that the adopted modeling techniques produce residuals that have similar nature, and that all techniques were similar in the way that they predict the output and minimize the errors (Fig. 10). Even though the visual analysis of Fig. 10 shows almost no practical differences among the various probability distributions, the $p$-values of the K-S test (Table not shown because all values are zeros) indicate that there is strong evidence to reject the null hypothesis. Based on the K-S test, the model residuals of the various techniques could represent different distributions. There is no contradiction between the K-S test results and the visual test because a slight shift on the graph might be translated to a statistically significant difference.

### 4.5 Rainfall-runoff case study II

This rainfall-runoff II (R-R II) case study is the same as the previous R-R I dataset with one difference; that is the preceding runoff was used as an additional input. In such a strongly autocorrelated series as the daily runoff, providing the preceding runoff as an input to predict the current runoff make strong information content at the disposal of the predictive models. Even though the MLR technique may not be suitable for this case study because one of the inputs (preceding runoff) is autocorrelated, it is used to show how much information can be a captured by a global linear model. In addition to this, a naïve model for predicting the daily runoff was developed just by using the preceding runoff value as an estimate of the current runoff. The performance of the various techniques applied to the daily R-R II case study is provided in Table 15. GP, M5, and EPR, followed by the MLR, techniques are better choices than the other techniques for this case studies. They provide the lowest RMSE, MARE, MB, and the highest R values. The IPE indicator in Table 8 also mostly supports this finding. Expectedly, the presence of the preceding runoff as an input in this case study makes the input-output relationship more globally linear than nonlinear. The superiority of the MLR over the ANNs supports this idea. Instance-based leaning techniques that use simple average of the nearest neighbors (K-nn) may not be a good choice. K-nn found almost most of the information within a range of very small number of neighbors (average of 3 neighbors, Table 6), but the failure to regress the information weakens the input-output relationship. The information capture in linear models could be even enhanced by local/modular techniques, such as the M5 model trees.

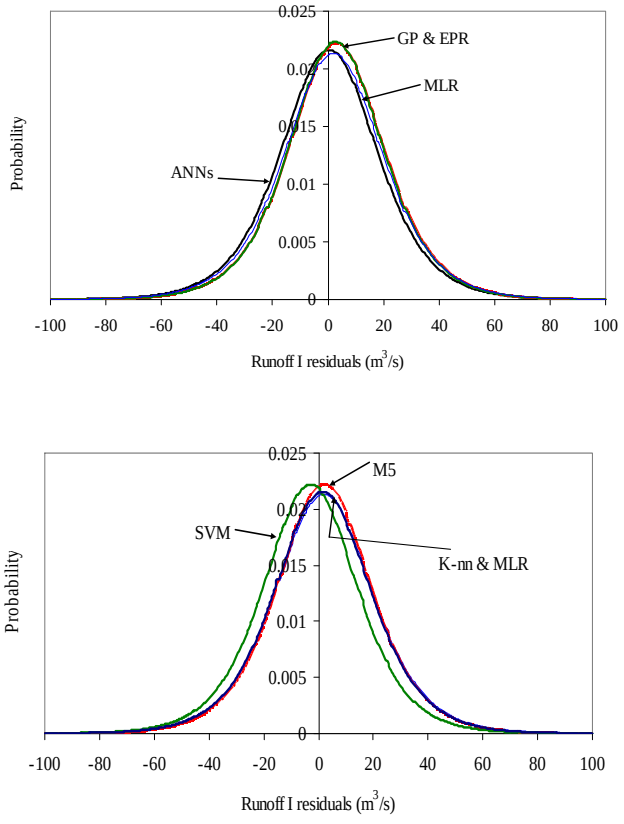


**Fig. 10.** Probability distribution of the 12 model residuals of all techniques (rainfall-runoff I case study).

Figure 11 shows the scatter plots of observed vs. predicted runoff II data. The scatter around the 45-degree line supports the conclusion made earlier regarding the superiority of the GP, M5, and EPR, and the inferiority of K-nn, ANNs, and SVM techniques. The success of GP, EPR, and M5 across all ranges of the dataset is noticeable (Fig. 11b, c, e). With the exception of the SVM and naïve models, the best-found probability distributions of the residuals of the various techniques did not differ. The Logistic ($\alpha$, $\beta$) probability distribution, with different parameter values for each technique, was found to fit the residuals of ANNs, GP, EPR, M5, K-nn, and MLR techniques, whereas Normal ($\mu$, $\sigma$) was found to fit the residuals of the SVM and the naïve models. In spite of the similarity in the best-fit distribution, the parameters were completely different even visually (Fig. 12). All modeling techniques produced symmetrical distributions of model residuals, but GP, EPR, and M5 possess the smallest predictive uncertainty. The $p$-values of the K-S test (Table 16) indicate that there is strong evidence to reject the null hypothesis. Based on the K-S test, the model residuals of the various techniques could represent different distributions.

**Table 16.** The $p$-values of the two samples K-S test on the model residuals (rainfall-runoff II).

| | ANNs | GP | EPR | SVM | M5 | K-nn | MLR |
|---|---|---|---|---|---|---|---|
| ANNs | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| GP | | 1 | 0.0003 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| EPR | | | 1 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| SVM | | | | 1 | 0.0000 | 0.0000 | 0.0000 |
| M5 | | | | | 1 | 0.0000 | 0.0000 |
| K-nn | | | | | | 1 | 0.0000 |
| MLR | | | | | | | 1 |

## 5 Discussion

After evaluating the various data driven modeling (DDM) techniques from both perspectives of prediction accuracy and uncertainty, one of the means to gain further insight into their modeling capabilities is to compare the performance deterioration in the testing phase to that in the training phase. Less deterioration may indicate a higher level of reliability and less uncertainty about the technique's performance in future and untested applications. The percent deterioration is calculated for each technique by dividing the difference between training and testing performance by the training performance. A negative percent means that the performance of the technique during the testing phase was better than that during the training phase. Table 17 presents the percent deterioration in both RMSE and MARE for all techniques and case studies. For each technique, the average values of RMSE and MARE of the 12 models were used. A few observations can be noted from Table 17: (i) ANNs had the highest level of performance deterioration in all case studies, which is an intricate characteristic of the technique and perhaps any highly nonlinear technique. ANNs seem to go after some individual and local patterns even when training is stopped by cross-validation; (ii) similar to ANNs, SVM suffered from similar phenomenon in four out of the five case studies. This might be counter intuitive and requires further investigation because a technique that employs the concept of error tolerance and flatness of the approximation function should do better in this regard. Users of SVM are encouraged to study further the effect of the error tolerance and the flatness coefficient C on the technique performance; (iii) in nonlinear case studies (e.g., peat and till soil moisture), the compromise between improving the prediction accuracy while reducing the deterioration might be difficult. The deterioration of the K-nn technique in both case studies was the highest, while performing relatively better than other techniques in terms of the prediction accuracy and uncertainty; (iv) EPR, almost similar to MLR, was excellent in its generalization ability. The deterioration of performance during the testing phase was very small in all case studies; highlighting a great potential of this technique; and (v) in most cases GP and M5 model trees were not far from the EPR regarding the performance deterioration. Therefore, whenever EPR, GP, and

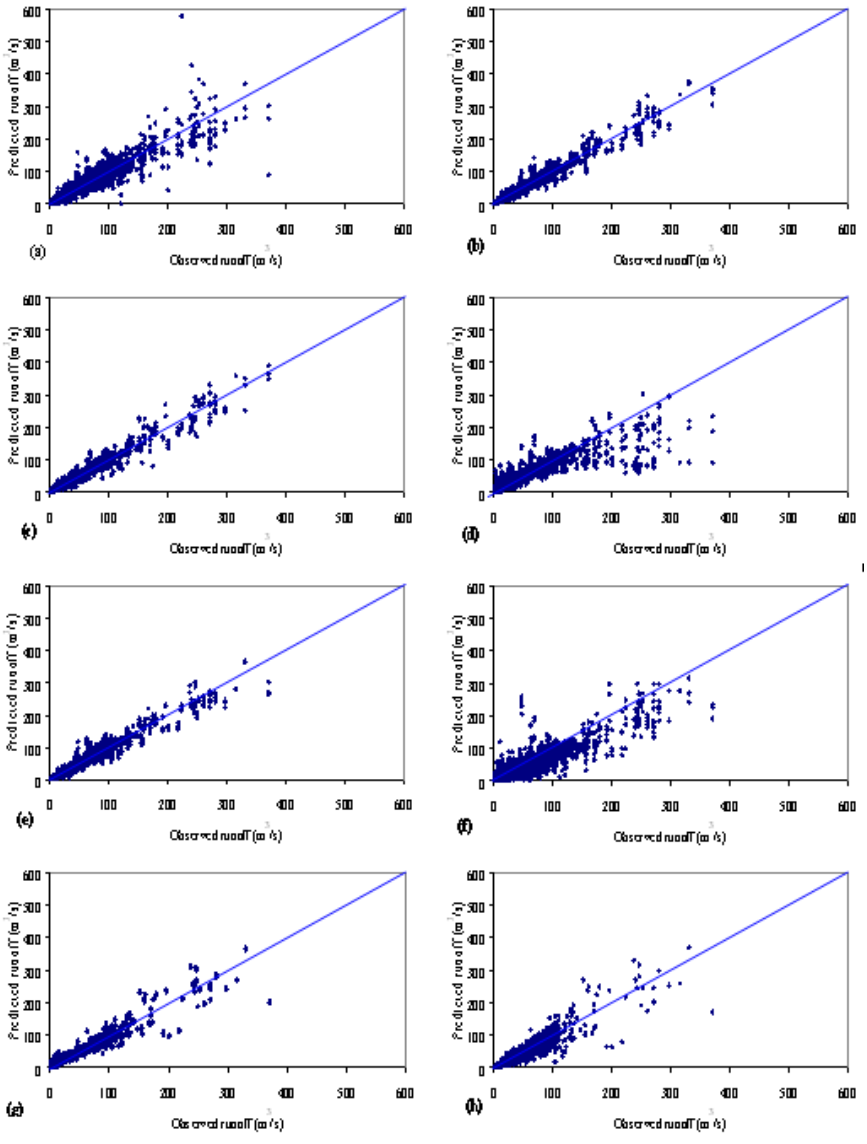

**Fig. 11.** scatter plots of observed and predicted runoff II, **(a)** ANNs, **(b)** GP, **(c)** EPR, **(d)** SVM, **(e)** M5, **(f)** K-nn, **(g)** MLR, and **(h)** naive.

**Table 17.** The percent deterioration of model performance during testing compared to training.

| | AET | | SMP | | SMT | | R-R I | | R-R II | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | MARE | RMSE | MARE | RMSE | MARE | RMSE | MARE | RMSE | MARE |
| ANNs | **29** | **118** | **27** | **26** | **23** | **26** | **18** | −8 | **127** | **65** |
| GP | *0* | *10* | 11 | 10 | 12 | 9 | 13 | 0 | 11 | 2 |
| EPR | *1* | *12* | *4* | *4* | 2 | 2 | 16 | *−1* | 7 | *−1* |
| SVM | **22** | **47** | 11 | 19 | **17** | **29** | **20** | **12** | **140** | **73** |
| M5 | 1 | *12* | 12 | 12 | 8 | 7 | 8 | 1 | 15 | 6 |
| K-nn | 4 | 16 | **26** | **29** | **24** | **26** | 12 | 9 | 45 | 46 |
| MLR | *−1* | *11* | 1 | 2 | 0 | 1 | 1 | 0 | 0 | *−1* |

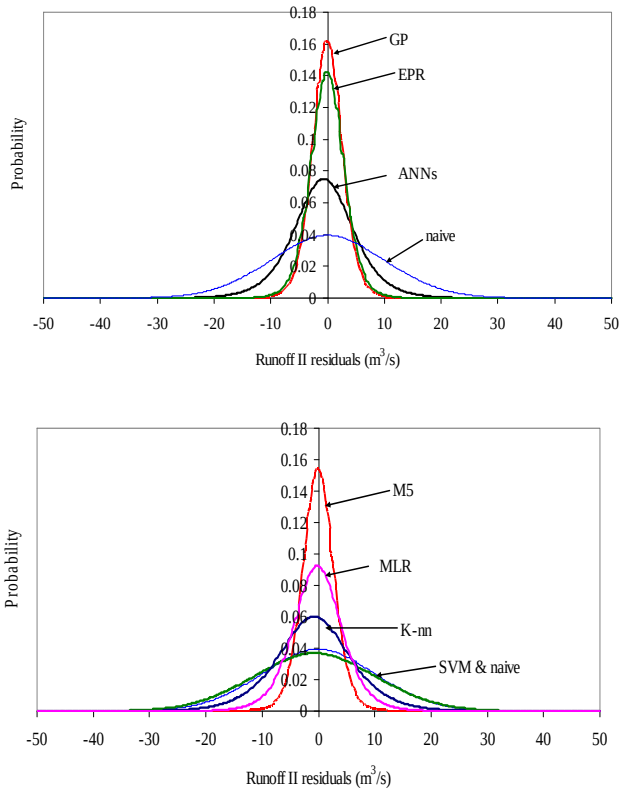


**Fig. 12.** Probability distribution of the 12 model residuals of all techniques (rainfall-runoff II case study).

M5 are comparable to other techniques in terms of prediction accuracy and uncertainty, they deserve to be given preference as candidate modeling techniques.

One of the fundamental questions of this research study is whether there are real differences among the techniques under consideration with regard to their predictive capabilities. The results and analysis show that serious evaluation of the various techniques has to rely on multiple ways, such as the average overall error represented by multiple error measures, scatter plots of the observed vs. predicted outputs, probabilistic analysis of the model residuals, and statistical tests of the significance of the differences among the residuals of various models/techniques. As an example, the SVM technique performed well on the peat moisture case study in terms of the overall average error measures and the probability distribution of the residuals, however, the scatter plots reveal that the models were not behavioral; i.e., could not capture the trend of the phenomenon at all. On the other hand, the superiority of the ANNs over other techniques on the same dataset was revealed by the scatter plots. The analysis presented in the previous section shows that SVM, M5, K-nn, and GP techniques were the best candidates for modeling the evapotranspiration case study. In the peat moisture case study, ANNs, GP, and followed by K-nn, M5, and EPR provided the best performances, whereas ANNs, GP, and K-nn were the best for modeling the till moisture dataset. Even though the K-S test show that the difference between the residuals of GP and M5 was insignificant, this should be treated with caution. The test compares the residuals but fail to assess the difference in the $R$ statistic, which is the key indicator in this particular case study. M5 was not successful in this nonlinear dataset. For the rainfall-runoff I dataset, all techniques were on par, and perhaps there is no need for a sophisticated nonlinear model. In the last case study (rainfall-runoff II), that has an autoregressive term and hence can be described by less non-linear mappings, GP, M5, and EPR were obviously better than the other techniques.

Neural networks could be one of the optimum modeling choices for highly nonlinear case studies (e.g., peat and till soil moisture), but could be completely dominated by other techniques as it was the case for the AET and the rainfall-runoff II case study, depending on the level of linearity in the dataset. M5 is an excellent choice for linear and some nonlinear dataset; it performed poorly only in the till moisture dataset. EPR, though it was not a top choice except in the rainfall-runoff II case study, was never completely dominated by other methods, and sometimes it was among the best techniques. The excellent generalization ability (minimum performance deterioration during the testing phase) of the EPR adds to its potential for hydrological applications. However, in nonlinear datasets, EPR was always less successful than GP. GP was the only technique that was always either the top model or, at least, among the best models regarding both prediction accuracy and uncertainty. The ability of GP to adapt the structural complexity of the generated model/program to the dataset could be one of the main reasons of its superb predictive capability. The SVM seems to be significantly affected by the choice of kernels. In this study, the RBF kernel was chosen based on its performance on the cross validation sample of most case studies (four out of five cases). In the linear rainfall-runoff II case study, when a linear kernel was tested, the prediction accuracy, represented by RMSE, MARE, and $R$, improved by 20–25%.

Two limitations of this study have to be noted. First, the effect of the model inputs on the predictive capabilities was not investigated. Adding more important inputs, or removing some of them, affects the degree of linearity/nonlinearity of the input-output relationship, and thus, the model performance. Such an effect may differ from one technique to the other. Second, some capabilities of the various techniques and tools were not, and perhaps cannot be, thoroughly covered. The Discipulus software for GP was run for almost two hours each time. It was observed that allowing from 24–48 h of run could slightly improve the results. The EPR tool allows for multiobjective optimization, rather than just minimizing the squared error, but it was not tried in this study. Also instance-based techniques (K-nn) could be further improved using weighted average or regression of the nearest neighbors. ANNs could be trained using Bayesian regularization algorithm (Demuth and Beale,

**Table 18.** The gamma test results on all case studies.

| Case study | Error variance MLR technique | Γ statistic | Error variance best technique | V-ratio | Gradient | M statistic |
|---|---|---|---|---|---|---|
| AET | 2302 | 2778 | 1928 (GP) | 0.207 | 0.0414 | 1200 |
| SMP | 0.0025 | 0.0018 | 0.002 (ANNs) | 0.410 | 0.2970 | 500 |
| SMT | 0.0003 | 0.0002 | 0.0002 (K-nn) | 0.273 | 0.4140 | 520 |
| R-R I | 459 | 495 | 397 (M5) | 0.560 | 0.1040 | 1300 |
| R-R II | 26 | 27 | 7 (GP) | 0.013 | 0.1100 | 1100 |

2001), which could improve the generalization ability. In this study, multiobjective cost functions were avoided as much as possible. However, future research by the authors and/or other researchers could add to this experiment and build on it.

In this experiment, the main objective was to investigate the predictive capabilities of the various data driven techniques. However, a brief ensemble prediction analysis was conducted in this study. For every case study (e.g., AET), ensemble prediction was calculated by averaging the predicted output values from the six modeling techniques. This process was repeated for each of the 12 dataset realizations. A summary of the ensemble prediction accuracy is provided in Table 19. In the cases of the AET and the P-R I dataset, ensemble predictions were not different from the results of the best individual technique. However, ensemble predictions were better than the best individual techniques in the cases of SMP and SMT. Interestingly, in the case of the P-R II case study, the best individual technique (GP) performed better than the ensemble. Apparently, when GP performed notably better than the other techniques, the results of the ensemble (averages of all techniques) will not improve the prediction accuracy (Table 19).

The non-parametric Gamma test (Γ-test) (Chuzhanova et al., 1998; Evans and Jones, 2002, and recently applied in hydrology by Remesan et al., 2008) was conducted to gain insight into the predictability and the complexities of the modeled processes, and possible leads into selection of suitable modeling techniques. The Γ statistic was calculated for every dataset using the original training and cross-validation subsets as one integrated subset (all unique points). The V-ratio, gradient, and the M-test were all calculated using the scaled data (zero mean and 0.5 standard deviation as specified by WinGamma software). The Γ statistic was calculated using the unscaled data to facilitate the comparison with the error variance of the various modeling techniques (Table 18). The following observations can be made: (i) for the AET case study, the error variance of the linear regression technique (2302) was already lower than the  statistic; indicating that complex nonlinear model (e.g., ANNs) may not be necessary. The low gradient value of 0.041 shows that a noncomplex smooth function can be used for modeling the AET process, whereas the reasonably low V-ratio indicates that there is high predictability in the output variable. GP, shown to perform well on all case studies, achieved the lowest error variance. Even though it is lower than the estimated Γ, but when it is divided by the AET variance (Table 1), the ratio is 0.23; similar to the V-ratio.; (ii) for the R-R I case study, similar to the AET, there is no need for nonlinear complex model, especially in light of the high V-ratio that indicates low level of predictability. The low level of predictability is attributed to the lack of appropriate inputs, which was rectified in the R-R II case study. All techniques were found to perform on par. The slight superiority of the M5 (ratio of error variance to output variance is 0.44), which is a modular linear technique can be attributed to the fact that it does not produce a smooth function. This is something that the Γ-test may not capture well; (iii) similar conclusions can be made for the R-R II case study. Nonlinear techniques, such as ANNs, will not perform well. The very low V-ratio that indicates very high predictability might be achieved by techniques that can outperform MLR, yet have the ability to adapt to linear situations. As expected GP, EPR, and M5 performed extremely well in this case; (iv) both SMP and SMT case studies, the MLR technique failed to achieve the estimated Γ value, and actually produced ratios of error variance to output variance of 1.0 and 0.8, respectively. This finding points to the possibility that more complex nonlinear models are needed. As the results of this study show, in addition to GP, the ANNs and K-nn were relatively more successful in the SMP and SMT case studies. However, it should be noted that Γ-test relates well to the model performance with regard to the squared error, but in cases where the criterion of performance is the *R* statistic, the test may not be the optimum tool; (v) the M-test indicates the number of data points that is perhaps needed to achieve the accuracy indicated by the V-ratio. It can be noticed from Table 18 that the size of the datasets needed for developing nonlinear models for the peat and till soil moisture are slightly more than what was used in this study. For the other three case studies, the size of the training datasets exceeded the M-test.

The Γ-test may assist in the selection of the appropriate modeling techniques by applying first multiple linear regression models and evaluating the residuals against the Γ-test values. Decision can be made regarding the need for a complex nonlinear technique. If there is a need for

**Table 19.** Accuracy results of ensemble prediction based on average output values of six techniques.

| | RMSE | | | MARE | | | BIAS | | | Correlation (R) | | | IPE | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst | Best | Ave | Worst |
| AET | 42.246 | **43.694** | 47.848 | 0.519 | **0.738** | 1.037 | 0.170 | **0.570** | 9.296 | 0.89 | **0.88** | 0.86 | 0.28 | **0.31** | 0.37 |
| SMP | 0.040 | **0.042** | 0.044 | 0.076 | **0.080** | 0.087 | 0.000 | **0.002** | 0.004 | 0.71 | **0.67** | 0.61 | 0.57 | **0.60** | 0.64 |
| SMT | 0.013 | **0.014** | 0.015 | 0.036 | **0.038** | 0.042 | 0.000 | **0.000** | 0.001 | 0.71 | **0.63** | 0.51 | 0.46 | **0.50** | 0.56 |
| P−R I | 23.318 | **24.342** | 25.126 | 1.395 | **1.507** | 1.613 | 0.030 | **−1.498** | −3.080 | 0.64 | **0.59** | 0.55 | 0.49 | **0.51** | 0.53 |
| P−R II | 5.201 | **6.204** | 9.475 | 0.108 | **0.135** | 0.310 | −0.300 | **−0.531** | −1.510 | 0.99 | **0.98** | 0.96 | 0.22 | **0.26** | 0.48 |

such technique, then ANNs and K-nn (in addition to GP, for example) should be seriously considered. If it is concluded that complex nonlinear techniques are not needed, then improvement of results can be sought using GP, EPR, and M5. When complex nonlinear techniques are not needed, and the predictability is low (i.e., high V-ratio) significant improvement may not be at all possible.

## 6 Conclusions

Neural networks (ANNs) that have hidden nodes with nonlinear transfer functions may impose on the data a model with complexity level that is higher than that needed by many hydrological data. The results of the experiment conducted in this research study show that ANNs were a sub-optimal choice for the actual evapotranspiration (AET) and the two rainfall-runoff case studies. In the nonlinear case studies (peat and till soil moisture), ANN models were the most successful ones. In general, genetic programming (GP) was the most successful technique due to its ability to adapt the model complexity to the modeled data. Evolutionary polynomial regression (EPR) performance could be close to the GP with datasets that are more linear than nonlinear. Support vector machines (SVM) are sensitive to the kernel choice and if appropriately selected, the performance of SVM can improve. M5 model trees performs very well with linear and semi linear data, which cover wide range of hydrological situations. In nonlinear case studies, ANNs, K-nearest neighbors (K-nn), and GP could be more successful than other modeling techniques. K-nn was also successful in linear situations, and it deserves more attention as a potential modeling technique for hydrological applications.

The results of this study show that a winner modeling technique cannot be easily declared. DDM techniques should be applied in ensemble fashion. Multiple groups (realizations) of each dataset should be randomly generated, by sampling without replacement, and should be divided into three split samples of training, cross-validation for stopping the training phase, and testing for applying the model once. Developing multiple non-dominated models of each technique, based on the multiple realizations of the dataset, allows for evaluating the predictive accuracy and uncertainty in a comprehensive way. Multiple overall average error measures, frequency distributions of model residuals, and scatter plots of observed vs. predicted data should be all used as one package to evaluate the predictive capabilities of the modeling techniques. Gamma test can be used as a guide to assist in the selection of the appropriate modeling technique for a particular dataset. Further studies can be built on the experiment presented in this research to evaluate other data driven techniques and to study the impact of input selection and input pre-processing on the relative predictive capabilities of the techniques.

Edited by: R. Merz

## References

Berger, H. E. J.: Flow Forecasting for the River Meuse, PhD Thesis, Technische Universiteit Delft, 1992.

Boese, K.: The design and installation of a field instrumentation program for the evaluation of soil-atmosphere water fluxes in a vegetated cover over saline/sodic shale overburden, M.Sc. thesis, University of Saskatchewan, Saskatoon, Sask., 2003.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., and Scuse, D.: WEKA Manual for version 3.6.0. University of Waikato, Hamilton, New Zealand, 2008.

Chuzhanova, N. A., Jones, A. J., and Margett, S.: Feature selection for genetic sequence classification, Bioinformatics, 14(2), 139–143, 1998.

Demuth, H. and Beale, M.: Neural Network Toolbox Learning For Use with MATLAB, The Math Works Inc, Natick, Mass, 2001.

De Wit, M. J. M.: Effect of Climate Change on the Hydrology of the River Meuse. RIVM, National Institute op Public Health and the Environment, 2001.

Drexler, J. Z., Snyder, R. L., Spano, D., and Paw, K. T.: A review of models and micrometeorological methods used to estimate wetland evapotranspiration, Hydrol. Processes, 18, 2071–2101, 2004.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology, Hydrol. Earth Syst. Sci., 14, 1931–1941, doi:10.5194/hess-14-1-2010, 2010.

Elshorbagy, A. and El-Baroudy, I.: Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content, J. Hydroinfo., 11(3–4), 237–251, 2009.

Elshorbagy, A. and Parasuraman, K.: On the relevance of using artificial neural networks for estimating soil moisture content, J.

Hydrol., 362(1–2), 1–18, 2008.
Elshorbagy, A., Jutla, A., and Kells, J.: Simulation of the hydrological processes on reconstructed watersheds using system dynamics, Hydrol. Sci. J., 52, 538–562, 2007.
Evans, D. and Jones, A. J.: A proof of the gamma test, Proceedings of Royal Society. Series A, 458, 2759–2799, 2002.
Francone, F. D.: Discipulus: Owner's Manual. Register Machine Learning Technologies, Inc., Littleton, CO, 2001.
Haigh, M. J.: The aims of land reclamation, Land Reconstruction and Management, A. A. Balkema Publishers, Rotterdam, The Netherlands, 1, 1–20, 2000.
Laucelli, D., Berardi, L., and Doglioni, A.: Evolutionary polynomial regression toolbox: version 1.SA., Department of Civil and Environmental Engineering, Technical University of Bari, Bari, Italy, available at: http://www.hydroinformatics.it/prod02.htm, last access: March 2008, 2005.
Palisade Corporation Inc. Guide to using @RISK. Advanced risk analysis for spreadsheets, Palisade Corporation, NY, USA, 2005.
Parasuraman, K. and Elshorbagy, A.: Model Structure Uncertainty and its Quantification Using Ensemble-Based Genetic Programming Framework, Water Resour. Res., 44, W12406, doi:10.1029/2007WR006451, 2008.
Parasuraman, K., Elshorbagy, A., and Carey, S. K.: Modelling dynamics of the evapotranspiration process using genetic programming, Hydrological Science J., 53(3), 563–578, 2007.
Remesan, R., Shamim, M. A., and Han, D.: Model data selection using gamma test for daily solar radiation estimation, Hydrol. Proc., 22, 4301–4309, 2008.