Conversion of the Time Series of Measured Soil Moisture Data to a Daily Time Step – A Case Study Utilizing the Random Forests Algorithm

Milan Cisty*[*1], Lubomir Celar[2]
[1]Department of Land and Water Resource Management, Faculty of Civil Engineering, STU, Radlinskeho 11, Bratislava, Slovakia
e-mail: milan.cisty@stuba.sk
[2]Department of Land and Water Resource Management, Faculty of Civil Engineering, STU, Radlinskeho 11, Bratislava, Slovakia
e-mail: lubomir.celar@stuba.sk

Cite as: Cisty, M., Celar, L., Conversion of the Time Series of Measured Soil Moisture Data to a Daily Time Step – A Case Study Utilizing the Random Forests Algorithm, J. sustain. dev. energy water environ. syst., 4(2), pp 183-192, 2016, DOI: http://dx.doi.org/10.13044/j.sdewes.2016.04.0015

ABSTRACT

Modeling the water content in soil is important for the development of agricultural information systems. Various data are necessary for such modelling. In this paper the authors are proposing a methodology for a frequent situation, i.e., when the modeler is facing a problem due to the lack of available data. Soil water prediction, e.g., for irrigation planning, should be performed with a daily time step. Unfortunately, past measurements of soil moisture, which are necessary for the calibration of a model, are often not available at such a frequency. In the case study presented the soil moisture data were acquired every two weeks. The authors have tested a model utilizing the Random Forests (RF) algorithm, which was used for the conversion of the original data to data with a daily time step. The accuracy of the application of RF to this task is compared with a neural network-based model. The testing accomplished shows that the RF algorithm performs with a higher degree of accuracy and is more suitable for this task.

KEYWORDS

Irrigation, Soil moisture, Data-driven model, Random forests, Temporal downscaling.

INTRODUCTION

The spatial and temporal distribution of soil water is a critical part of many disciplines, including agriculture, forest ecology, hydro-climatology, civil engineering, water resources modelling, etc. The hydrological processes involved in creating a soil water regime particularly include precipitation, evapotranspiration, the infiltration into the soil of surface water, the movement of groundwater, the infiltration of moisture from groundwater, etc. A soil moisture regime, especially in the root zone of plants, is important according to different aspects, e.g., when a decrease in soil moisture below a certain value occurs, the soil water becomes less available to plants [1]. The long-term monitoring of a soil water regime is useful for understanding the conditions of a region's ecology or food production. With the increasing recognition of the importance of soil moisture as a geophysical variable [2], soil moisture monitoring networks began to be established [3] and large databases of soil moisture information made available [4]. In [5], the influence of land use on soil moisture dynamics is investigated for monitoring

* Corresponding author

purposes. However, such monitoring consumes a lot of time, equipment, staff and funds. The monitoring of soil moisture is an important tool, but in many cases it can be complemented by other methods, including mathematical modeling. The accuracy of any modeling, besides the selection of a suitable model, largely depends on the available input data. However, it often happens that the available data are not representative or detailed enough for the process to be modelled.

While one is modelling natural processes, both state variables and time series data are needed. This study is focused on solving a situation when no appropriate time series is available; more specifically, when the past measurement of soil moisture at the required frequency (a daily step) is not available. Only measurements which were accomplished with a larger gap between each other are at the researcher's disposal.

Infilling missing data has traditionally been done using different statistical methods, although various problems with their application are known. E.g., Dumedah *et al*. [6] examines 5 statistical methods and 9 artificial neural networks to assess their suitability to infill missing soil moisture data. In particular, multiple regression has problems of multicollinearity, heteroscedasticity, and data normality assumptions [7]. In most meteorological time series, nonlinearity is another problem that may hamper time series analysis using linear methods. In particular, soil moisture data suffer from nonlinearity in addition to the problem of missing values. Dumedah and Coulibaly [8] evaluated statistical infilling methods for soil moisture and found that even simple methods such as monthly average replacement and rank stability methods outperformed regression-based techniques. For this reason, many of the techniques presented in this paper will focus on infilling data on the basis of time series methods; we would especially like to show the benefits of multivariate machine learning techniques, because they can better account for the stochastic components [9], higher order interactions, and hysteresis in the data [10].

The novelty of this paper consists in the application of prediction tools for interpolation tasks. Various methods used for infilling missing data could be applied [11]. There are various types of models for hydrological predictions: physically-based, conceptual and data-driven models are among the most well-known. While physically-based models mainly depend on our knowledge of physical laws, data-driven models extract knowledge only from the monitored data describing the inputs and outputs of the process modelled, i.e., they are better suited for solving a problem. Artificial Neural Networks (ANN) [8] and other data-driven models such as RF [12], Support Vector Machines [13], etc., can, under certain conditions, enter into gaps of a mathematical description and replace them with the knowledge stored in the data. Their usage is based on the principle that from the known inputs and outputs (e.g., measured), they learn how to generate the correct output from the input [14]. Then in the application phase, the unknown outputs can be generated from the known inputs. The advantages of ANN and RF data-driven models are their ability to learn from the model and generalize knowledge from them, often without a detailed knowledge of the various state variables of the process. That makes them a suitable alternative tool to address complex processes.

The work submitted compares soil water content models based on a data-driven methodology with the aim of accomplishing an interpolation task and obtaining a time series of soil moisture with a daily time step. The particular focus of this paper is the use of the alternative data-driven method RF, which was firstly used as a classifier. It's functionality was then extended to regression, which makes it suitable for a soil moisture interpolation task. In the following part of the paper "Methods", the methods of the specific machine learning algorithms involved in this study are briefly explained. Then in the "Case study description" the data acquisition and preparation is presented. In the "Results" part, the settings of the experimental computations are described and the results

evaluated. Finally, the "Conclusion" part of the paper summarizes the main achievements and conclusions of the work and proposes ideas for future work in this area.

METHODS

From a methodology point of view, the novelty of this paper consists in the application of a regression type of data-driven algorithm to an interpolation task. As can be seen hereinafter, this is somewhat similar to commonly used applications of these algorithms to make predictions. A brief introduction of the methods which were used now follows.

Random forests

RF is a fully non-parametric data mining method requiring no distributional assumptions of the covariate relation to the response. This methodology was already successfully tested in soil mapping, e.g. [15]. RF are robust and optimize predictive accuracy by fitting an ensemble of trees to stabilize model estimates. RF consists of a set of regression trees (if we are addressing a regression problem as in this work). The resulting RF prediction is an average of the values of these many tree outputs, each one of which is grown on a bootstrap sample of the training data. The user chooses the number of trees that will be in the RF ensemble. A bootstrap sample means that each tree is trained using a sample obtained by randomly drawing N cases with replacements from the original dataset, where N is the number of variables in that dataset. With each of these bootstrapped training sets, a different tree is obtained. For the regression, the values predicted by each tree are averaged to obtain the RF prediction. More details and more mathematically founded explanations can be found in [12]; the modeller or user of the RF job is more focused on setting the proper parameters of this algorithm, e.g., the optimisation of the model.

Optimization of the model

RF has three tuneable parameters:

- Ntree − the number of trees to grow;
- Mtry − the number of variables randomly sampled as candidates at each tree split;
- Nodesize − the minimum size of the terminal nodes, which has the main effect on the final precision of the model.

Two concepts are applied in this work as the means used for optimally setting these parameters: grid search and repeated cross-validation. These two concepts run together, but for the sake of a simpler explanation, they are separately described in the next two paragraphs.

The grid search is designed in the optimization process to choose the values for each parameter of the model from a grid of predefined values. The grid search involves running the model with the parameters actually chosen in the current iteration, in which the model tries to learn the dependencies between the inputs and outputs. The evaluation of the results is accomplished with a statistical coefficient (Root Mean Square Error (RMSE) was used in this study). Then the best combination of the parameters is finally chosen from that iteration in which the highest degree of precision of the model was achieved.

This precision is evaluated as the average value from more runs of the so-called cross-validation process. In each run of the grid search of the parameters of the model, this process is accomplished. A so-called "repeated cross-validation" is used in the present paper [16], which consists of randomly dividing the training data into several approximately equal-sized data sets called "folds." The training process uses all the folds except one as the inputs to the model, and the one unused fold is used as the validating

data. This process runs as many times as the number of folds that were created. Each fold is used as the validating data in this procedure. "Repeated cross-validation" means that the initial random splitting of the training data into folds is repeated more than once.

The abovementioned precision of the model in each iteration of the grid search is now the average value of the assessed statistic (e.g., RMSE) from all the runs of the model, e.g., if there are two repetitions and five folds, the resulting statistic is the average value from ten particular values.

The use of cross-validation in the optimization process improves the selection of the parameters and is a necessity for small amounts of available training data, which is our case in the case study presented.

Multilayer perceptron

ANN are inspired by biological processes in the human brain and are applied to various technical problems for which sufficient, representative data are available. Generally an ANN is defined as a computing system that has the ability to learn and retain information (and the relationships between them) and allows their further use. The most commonly used ANN is a Multilayer Perceptron (MLP). It is a feed forward network with a controlled type of learning. The input signals pass through this type of network in a forward direction, from input layer to output layer. The basic element of an MLP is a neuron, which generally has more inputs and one output. The neurons in the network are linked to each other, and these connections transform the signal coming from the previous neurons by the connection's weights. The sum of these weighted signals is then transformed by the activation function of the neuron (nonlinear), which affects the output to the next neuron. An MLP uses three or more layers of neurons – the input layer, one or more hidden layers, and the output layer, all with a nonlinear activation function. The nonlinearity included in this flow of the input signal (the activation function, hidden layer, etc.) allows the network to learn complex nonlinear tasks.

The application of an ANN model is divided into three separate parts. The first is called the "learning phase" and is about training the model with the training input data. The actual output of the network must be known for this type of ANN in the learning stage. The learning of the MLP is accomplished by the error back propagation method. An error in this sense means the difference between the expected and actual output of the MLP. The signal transmitted between the neurons is changed depending on adjustable parameters called "weights", as was mentioned in the previous paragraph. The main goal of the learning process is to define these weights. Finding the appropriate network parameters is repeated until the error between the desired and actual output from the ANN is minimal.

More details of the back propagation learning method are described in the general literature on this subject [14]. Therefore, we will not deal with this in more detail here.

In the next "verification" phase of the ANN application, the trained network is verified with the test data (the actual output of the network must be known at this stage too); if this is accomplished with satisfactory results, the model is ready to use an actual application (where the output data are unknown).

DESCRIPTION OF THE CASE STUDY

For the testing purposes of the methods described in the previous section, data were taken from a probe installed in the village of Bac on the Danubian Lowlands (Slovakia) (Figure 1). In this and other nearby locations moisture in the unsaturated soil zone and ground water levels are monitored. In some of these localities continuous monitoring has been performed since 1999, but samples are taken at two-week intervals from the Bac

probe. The interpolation of these measured values to a daily interval should be accomplished by the data-driven models presented in this study to obtain homogeneous data in the area with the aim of obtaining the spatial interpolation of the soil moisture and a prediction of the irrigation needs. The soil profile at the Bac area has a complicated layered structure. There is loam on the surface that passes into sandy loam; at a depth of about 90-100 cm, sand is present, and under it gravel soil is found. This is the most important reason why the researchers decided to use data-driven modelling in this task instead of the more complicated, additional data requiring physically-based models. The moisture content of the soil profile was monitored using neutron probes at a distance of 10 cm from each other. The measurements were made at a 2-week frequency. At each site, calibration curves were taken at different seasons. They served for the refinement of the computational relationships recommended by the manufacturer of the neutron probe.

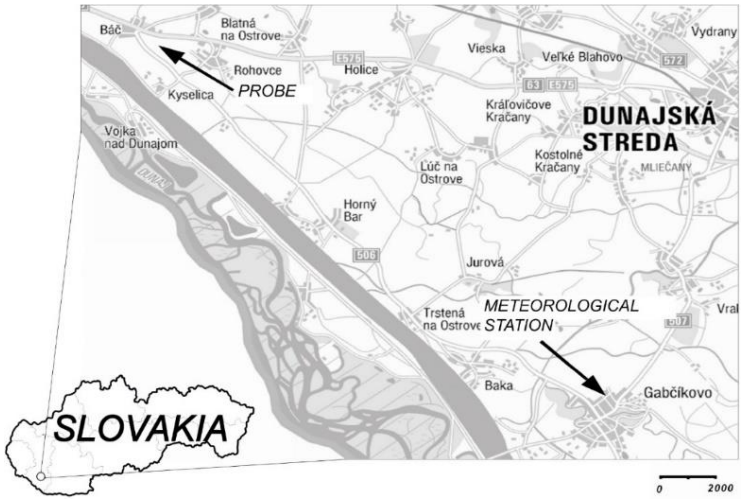

Figure 1. Map of the area studied with the soil probes and a meteorological gauge station

The measured moisture content of the soil from five horizons at the Bac probe (0, 20, 30, 40, 50 cm) was used for the modelling purposes. These data were collected at the mentioned 2-week intervals during the period April 1999-December 2009. A total of 189 data vectors is available, but data from November to February were excluded, as there is no need to model soil moisture in winter. Thus 117 data vectors were used. These data about soil moisture were used in further calculations as dependent variables that will be calculated on the basis of data taken from the nearby Gabcikovo meteorological station. As can be seen in Figure 2, the soil moisture between the various layers is highly correlated, and this fact will be used in the proposed model.

The following data were available from the Gabcikovo climatic station: the average daily temperature, relative humidity, wind speed, sunshine and daily precipitation (Figure 3). After the results of correlation analysis and other issues were considered, only the average daily temperature and daily precipitation totals were used. These two variables are used as inputs. They are taken from a various time interval before the day in which the value of the soil moisture was computed (e.g. $T_{t\text{-}3}$, is the temperature 3 days before the day on which the prediction of the soil moisture is computed). In such a way in every model every variable from more than one day preceding day in which prediction is accomplished is taken. This structure of the input data is introducing into the calculations time dynamics, since the RF model itself is basically static. As the input, the total precipitation amounts for the previous 20 days (U_{20}) and the average temperature for the same interval (T_{20}) are also used in some of the models. These two variables are intended to represent the past

meteorological conditions in the study area, which affect the consequences of the rainfall and temperatures from previous days on the soil moisture value.

For the data-driven modelling two data sets were used: the training set (1999-2005) to build the model (determine its parameters) and the test set (2006-2007) to measure its performance.

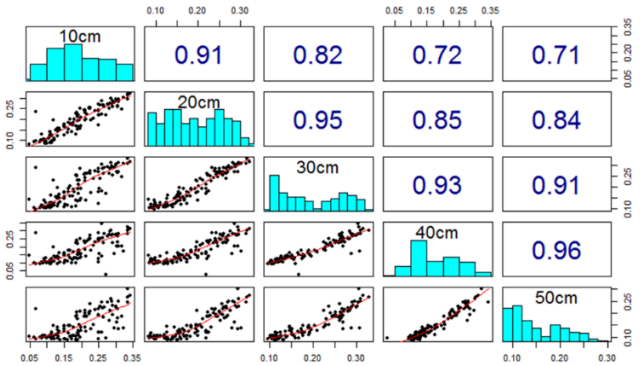

Figure 2. Correlogram of the soil moisture values at different depths. Texts "10cm", "20cm" etc., in the diagonal are variable names for soil moisture in depth 10 cm, 20 cm etc. On the right side of diagonal are correlation coefficients, on the left side are scatter graphs smoothed by Lowes line

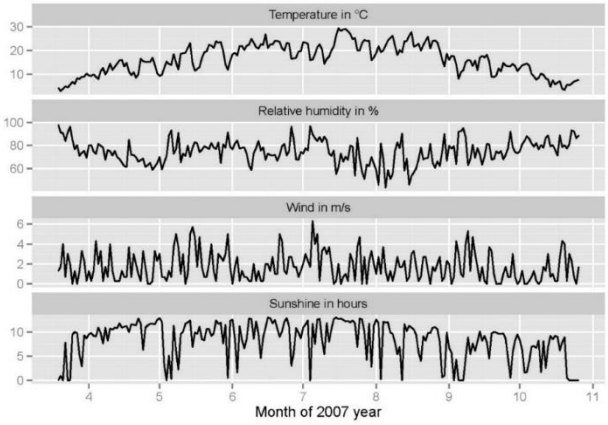

Figure 3. Climatic data for Gabcikovo in year 2007

RESULTS

The following data model structures for evaluating soil moisture θ_t were evaluated using RF and MLP. T are temperatures and Z precipitations from previous days:

- Model 1: $\theta_t = f(T_{t-1}, T_{t-2}, T_{t-3}, T_{t-4}, T_{t-5}, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5}, U_{20}, T_{20})$;
- Model 2: $\theta_t = f(T_{t-1}, T_{t-2}, T_{t-3}, T_{t-4}, T_{t-5}, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5})$;
- Model 3: $\theta_t = f(T_{t-1}, T_{t-2}, T_{t-3}, T_{t-4}, T_{t-5}, T_{t-6}, T_{t-7}, T_{t-8}, T_{t-9}, T_{t-10}, Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5}, Z_{t-6}, Z_{t-7}, Z_{t-8}, Z_{t-9}, Z_{t-10})$;
- Model 4: $\theta_t = f(T_{t-1}, T_{t-2}, T_{t-3}, Z_{t-1}, Z_{t-2}, Z_{t-3}, U_{20}, T_{20})$;
- Model 5: $\theta_t = f(Z_{t-1}, Z_{t-2}, Z_{t-3}, Z_{t-4}, Z_{t-5}, U_{20}, T_{20})$.

These data were used for the computation of the soil moisture in the first layer at a depth of 10 centimetres. Moreover, for determining the soil moisture in the deeper layers (20, 30, 40 and 50 cm below the soil surface), the previously mentioned correlation of the soil moisture between the adjacent layers was used (Figure 2). This was accomplished in that the layers were successively evaluated from top to bottom and, starting from the

second layer, the previously computed soil moisture from the upper layer was also used as an input.

To increase the accuracy of the simulation, a normalization of the input data was accomplished. A normalization should be made to eliminate the possibility that some variables will have a greater impact on the learning than the others. In the normalization process, all the variables were transformed into the range (-1, 1), which guarantees that they will have equal importance in the resulting model.

Tuning the ANN was made by the trial-and-error process with various settings (different amounts of neurons in the hidden layer; the learning rule was either momentum or Levenberg-Marquardt and other parameters). In this way a suitable architecture of the ANN was found (with one hidden layer containing 5 neurons, an activation function hyperbolic tangent in the hidden layer, and the linear function was selected in the output layer). The Levenberg-Marquardt algorithm was chosen as a learning rule, and the calculations were carried out by the NeuroSolutions neural network simulator.

The results were statistically evaluated by the Mean Square Error (*MSE*) and correlation coefficient (*R*).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta_i^m)^2 \tag{1}$$

$$R = \frac{\sum_{i=1}^{n}(\theta_i - \bar{\theta})(\theta_i^m - \bar{\theta})}{(n-1)\sigma.\sigma_m} \tag{2}$$

where:

- θ is the measured value;
- θ^m is the predicted value;
- $\bar{\theta}$ is the average value;
- σ and σ_y are the standard deviations of the measured and modeled data;
- n is the amount of the data information.

The results for the testing data and five data models mentioned using MLP are evaluated in Table 1 and for RF in Table 2. The total of *R* and *MSE* in the last column of these tables is evaluated too; it evaluates the overall precision in all five layers.

Table 1. MLP model evaluation by *R* and *MSE* (test data 2006-2007)

Model		10 cm	20 cm	30 cm	40 cm	50 cm	Sum
M1	*R*	0.889	0.842	0.801	0.718	0.716	3.967
	MSE	0.0012	0.00117	0.00164	0.00237	0.00184	0.00823
M2	*R*	0.671	0.700	0.715	0.613	0.597	3.296
	MSE	0.00314	0.0025	0.00224	0.00295	0.00237	0.0132
M3	*R*	0.713	0.668	0.609	0.556	0.576	3.122
	MSE	0.00321	0.0032	0.00376	0.00426	0.00264	0.01708
M4	*R*	0.878	0.870	0.840	0.839	0.851	4.279
	MSE	0.00123	0.00106	0.00131	0.00152	0.00121	0.00633
M5	*R*	0.666	0.739	0.751	0.776	0.783	3.715
	MSE	0.00683	0.00329	0.00221	0.0018	0.00127	0.0154

Table 2. RF model evaluation by *R* and *MSE* (test data 2006-2007)

Model		10 cm	20 cm	30 cm	40 cm	50 cm	Sum
M1	*R*	0.868	0.886	0.854	0.831	0.831	4.270
	MSE	0.00186	0.00101	0.00121	0.00377	0.00128	0.00912
M2	*R*	0.779	0.795	0.760	0.739	0.742	3.815
	MSE	0.00326	0.00154	0.0038	0.00305	0.00306	0.01471
M3	*R*	0.816	0.841	0.812	0.802	0.792	4.063
	MSE	0.00188	0.00127	0.00283	0.00181	0.00192	0.00971
M4	*R*	0.865	0.878	0.856	0.839	0.828	4.266
	MSE	0.00152	0.0011	0.0012	0.00133	0.00123	0.00637
M5	*R*	0.859	0.885	0.862	0.839	0.843	4.286
	MSE	0.00464	0.00303	0.00361	0.00133	0.00304	0.01566

The correlation coefficients for the calculations using MLP ranged from 0.556 to 0.889, with the average value of 0.735. For RF the correlation coefficients ranged from 0.739 to 0.886 with an average value of 0.828. This means that RF offers more stable results. For this reason, the RF model was selected as more suitable for the final interpolation and the input data into the final model was selected according to the first M1 model, which has the best performance. In addition, the authors also consider the RF model as preferable because it does not suffer from the difficulties of accidentally falling into the local minima, which frequently happens with a neural network model. Figure 4 illustrates the daily values of moisture evaluated at a depth of 20 cm for the growing season in the year 2007 (which was selected for the testing purposes).

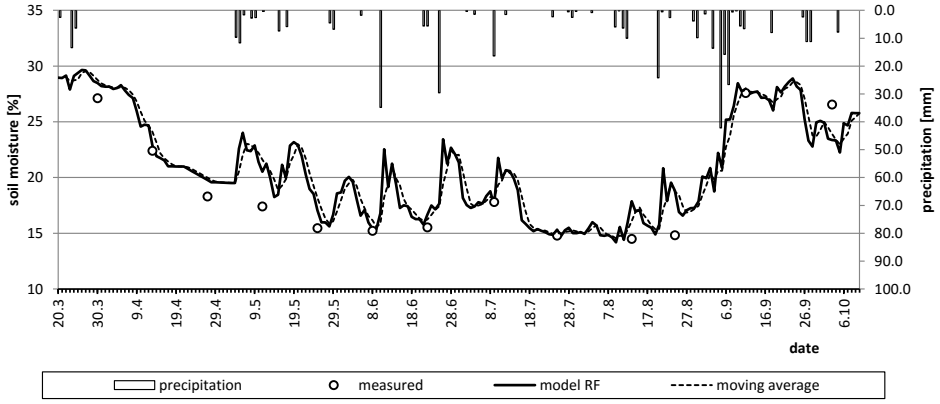

Figure 4. Interpolation of the soil moisture to the daily values in year 2007

CONCLUSION

The development of machine learning has made significant progress in the last two decades, and these systems are also finding their applications in water management and hydrology. In the work presented the authors evaluated the model for a soil water content simulation and interpolation, the aim of which was to obtain a daily time series of the soil moisture data. The methodology was based on the neural network methodology (MLP) together with the newer type of data-driven model RF. Both models were based on temperature and precipitation data from the Gabcikovo climatic station and on the soil moisture data measured in the past at approximately two-week time intervals at the Bac probe (Slovakia). Both methods have been evaluated using the same training and testing data sets, with the aim of comparing them. The results of the models are herein presented

graphically (Figure 4) and by using statistical measures (Tables 1, 2). From this comparison it can be seen that the results of the RF method are more accurate with respect to the measured data. This work also confirms the better potential of RF for application from another point of view, i.e., its results are quite stable and similar in repeated calculations, unlike the MLP, where the results did not show such stability and were often changing in the recalculation process.

In view of the fact that the proposed model predicted soil moisture using only data about temperature and precipitation, the authors considers the accuracy achieved sufficient. Presented statistical model can be used in situations, when more extensive and detailed data that would be needed for physical-based modeling (e.g., by Hydrus) are not available. Some inaccuracies in the calculation of soil moisture, especially in the values following the greater rainfalls, are resulting from a smaller quantity of data for model calibration during such events. In the future, it would be useful to develop methods that allow to deal with this (usual) lack. It would also be necessary to have climatic measurements closer to the area of measuring soil moisture. Bigger rains (after which some inacuraces in soil moisture computations ocuur) are often local in nature and amount of precipitation fallen in the weather station may not exactly match the amount of rainfall at the soil moisture measurement probe (Figure 1). This could be another source of error in soil moisture model during its calibration and application.

Both methods could be used as an alternative method to standard measurements, and the simulation of the soil moisture and particularly RF suitability for the interpolation task were confirmed.

ACKNOWLEDGMENTS

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0665/15 and 1/0625/15.

REFERENCES

1. Seneviratne, S. I. et al., Investigating Soil Moisture-climate Interactions in a Changing Climate: A Review, *Earth Sci. Rev*, Vol. 99, No. 3-4, pp 125-61, 2010, http://dx.doi.org/10.1016/j.earscirev.2010.02.004
2. Legates, D. R. et al., Soil Moisture: A Central and Unifying Theme in Physical Geography, *Prog. Phys. Geog.*, Vol. 35, No. 1, pp 65-86, 2011, http://dx.doi.org/10.1177/0309133310386514
3. Ochsner, T. E. et al., State of the Art in Large-scale Soil Moisture Monitoring, *Soil Sci. Soc. Am. J.*, Vol. 77, No. 6, pp 1888-1919, 2013, http://dx.doi.org/10.2136/sssaj2013.03.0093
4. Dorigo, W. A. et al., The International Soil Moisture Network: A Data Hosting Facility for Global in Situ Soil Moisture Measurements, Hydrol, *Earth Syst. Sci*., Vol. 15, No. 5, pp 1675-1698, 2011, http://dx.doi.org/10.5194/hess-15-1675-2011
5. Zucco, A. et al., Influence of Land use on Soil Moisture Spatial–temporal Variability and Monitoring, *Journal of Hydrology*, Vol. 516, pp 193-199, 2014, http://dx.doi.org/10.1016/j.jhydrol.2014.01.043
6. Dumedah, B. et al., Assessing Artificial Neural Networks and Statistical Methods for infilling missing Soil Moisture Records, *Journal of Hydrology*, Vol. 515, pp 330-344, 2014, http://dx.doi.org/10.1016/j.jhydrol.2014.04.068
7. Mair, A. and Fares, A., Assessing Rainfall Data Homogeneity and Estimating Missing Records in Mākaha Valley, O'ahu, Hawaii, *J. Hydrol. Eng.*, Vol. 15, No. 1, pp 61-66, 2010, http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000145

8. Dumedah, G. and Coulibaly, P., Evaluation of Statistical Methods for infilling missing Values in High-resolution Soil Moisture Data, *J. Hydrol.*, Vol. 400, No. 1-2, pp 95-102, 2011, http://dx.doi.org/10.1016/j.jhydrol.2011.01.028
9. Hong, Z., *A data-driven approach to soil moisture collection and prediction using a wireless sensor network and machine learning techniques*, Thesis at University of Illinois at Urbana-Champaign, 2015.
10. Gheyas, I. A. and Smith, L. S., A Neural Network-based Framework for the Reconstruction of Incomplete Data Sets, *Neurocomputing*, Vol. 73, No. 16-18, pp 3039-3065, 2010, http://dx.doi.org/10.1016/j.neucom.2010.06.021
11. García, C. et al., Dealing with Missing Values, *Data Preprocessing in Data Mining*, Vol. 72, pp 59-105, 2015, http://dx.doi.org/10.1007/978-3-319-10247-4_4
12. Liu, Y., et al., New Machine Learning Algorithm: Random Forest, *Information, Computing and Applications*, Vol. 7473, pp 246-252, 2012, http://dx.doi.org/10.1007/978-3-642-34062-8_32
13. Gill, M. K., Asefa, T., Kemblowski, M. W. and McKee, M., Soil Moisture Prediction using Support Vector Machines, *Journal of the American Water Resources Association*, Vol. 42, No. 4, pp 1033, 2006, http://dx.doi.org/10.1111/j.1752-1688.2006.tb04512.x
14. Solomatine, D. P., *Data-Driven Modeling and Computational Intelligence Methods in Hydrology*, Encyclopedia of Hydrological Sciences, Vol. 1, John Wiley & Sons, 2005.
15. Guo, D. et al., Digital Mapping of Soil Organic Matter for Rubber Plantation at Regional Scale: An Application of Random Forest Plus Residuals Kriging Approach, *Geoderma*, Vol. 237, pp 49-59, 2015, http://dx.doi.org/10.1016/j.geoderma.2014.08.009
16. James, G. et al., *An Introduction to Statistical Learning: With Applications in R*, Springer-Verlag New York, 2013, http://dx.doi.org/10.1007/978-1-4614-7138-7

Paper submitted: 22.08.2015
Paper revised: 02.11.2015
Paper accepted: 03.11.2015