Interpolation of Irregular Soil Moisture Measurements with Machine Learning Methods

Milan Cisty 1, Frantisek Cyprich 1, David Dezericky 1
1 Faculty of Civil Engineering, Slovak University of Technology in Bratislava, Bratislava, Slovakia
milan.cisty@stuba.sk

Abstract. In this work, a method is proposed that uses machine learning techniques intending to make existing soil moisture time series complete. Authors are assuming and solving the usual situation, in which only data irregularly measured are available. Soil moisture is one of the determining variables of the stress on various ecosystems and agriculture systems and a key element of the surface water budget. A time series of this variable is useful for an evaluation of the moisture regime of soil and for decisions regarding building irrigation structures. Interpolation models proposed in this paper were verified using data from the days on which the field measurements were available. Mainly nonlinear machine learning models are proved to be suitable for a solution to this task. The Extreme Gradient Boosting Machines model and simple ensemble model provided the best results.

1. Introduction
Soil moisture is the water stored in the soil in the spaces between the soil particles. Its evaluation is necessary for several practical applications related to agricultural production, irrigation management, or evaluations of the status of an ecosystem. Soil moisture is influenced by a wide range of ecological, hydrological, geotechnical, and meteorological processes [1].

The measurement of soil moisture by standard methods is carried out by field sampling and its subsequent handling in a pedological laboratory. Soil moisture can be measured by several instruments in the field, e.g., by various types of tensiometers, neutron probes, so-called time domain reflectometry, capacitance probes or electrical resistance gypsum blocks [2]. Due to time demands, high financial costs, a lack of personnel, and weather fluctuations, such measuring of soil moisture are not usually performed on a daily basis, particularly if it is done in a location where permanent metering equipment cannot be installed. Therefore, this paper presents an interpolation of soil moisture, i.e., a supplementation of the time series of this variable measured in the field by computed values.

The sector of the remote sensing of soil moisture has greatly developed in recent years. Soil moisture measurements are based on optical, thermal, and radar satellites, and these methods provide data for convenient spatial coverage [3]. One problem can be that the coarse spatial resolution (pixel size) of satellites usually limits their application to crop field operations. To address such limitations and handle the complex nature of soil moisture dynamics, machine learning tools such as Artificial Neural Networks (ANNs) [4], Support Vector Machines (SVMs) [5], and Random Forest (RF) [6], have been tested in recent years with regard to their ability to estimate spatially distributed soil moisture.
In this work, a method is proposed that uses various machine learning techniques and different data sources to develop a soil moisture time series, assuming the usual situation, i.e., that only data irregularly measured with, for example, weekly or longer time steps between measurements, are available. In the next part of the paper (“Case Study and Data Description”), the acquisition and preparation of the data is described. The methods applied in this study are then briefly explained. In the “Results and Discussion” section, the settings of the experimental computations are described, and the results are evaluated and discussed. Finally, the “Conclusion” part of the paper summarizes the main achievements of the work.

2. Case Study and Data Description

An area of the Zahorska lowlands was selected for testing the methods described hereinafter. The Zahorska lowlands are in Slovakia and are situated in the west part of the country between the Small Carpathian Mountains and the Morava River. Three basic data sources were used in this paper:

1. manually measured data using a neutron probe
2. satellite data or data derived from satellite data (evapotranspiration and moisture from the upper level of the soil in locations close to the testing probes – Figure 1),
3. climatic data from an ECA&D dataset

The monitoring of the soil moisture at the selected stations in the Zahorska low-lands was realized using a neutron probe. The locations were selected on the basis of their hydrophysical soil characteristics (soil texture, depth to groundwater level). The monitoring of the soil moisture was carried out approximately two times a month during the growing season (April to October) and once a month in the non-vegetation period (November to March in Slovakia). The method of measuring soil moisture by a neutron probe is an indirect measurement method and is suitable for repeated measurements at the same site. The soil moisture data measured was used for the calibration and testing of the machine learning models that were used in this paper.

Remotely sensed data were used in addition to the data measured in the field, namely, data about the moisture in the upper level of the soil and evapotranspiration data. Firstly, data on the soil moisture acquired by satellite measurements of the European Space Agency (ESA) was used [7]. The data used was measured by a passive multifrequency radiometer. The soil moisture data from this source has been available since 1.11.1978 up to the present with a time step of one day. The data is available in the form of a regular grid of points with a distance of 0.25° geographical latitude x 0.25° geographical longitude. The closest nodes of this grid to the measurement locations of the soil moisture by the neutron probe are shown in Figure 1. This data on the soil moisture differs from the calculated data, as it is from a close, but not identical, location and is related to different soil depth.

Another variable based on remote sensing techniques and used in this work is evapotranspiration. Evapotranspiration represents the combined loss of soil water from the earth’s surface to the atmosphere due to the evaporation of water from the soil or plant surfaces and transpiration via the stomata of the plants [8]. In agricultural production systems, these two losses of water represent a major component of the water balance and are therefore very important processes when estimating soil moisture. Data available from the Global Land Evaporation Amsterdam Model (GLEAM) was used in this work; it is a set of algorithms that estimates the different components of evapotranspiration from remotely sensed data, i.e., transpiration, bare soil evaporation, interception loss, open water evaporation, and sublimation. It uses microwave observations, which is an advantage under cloudy conditions. The Priestley and Taylor equation [9] is used in GLEAM for the quantification of the evapotranspiration [10]. Additionally, GLEAM provides data on surface and rootzone soil moisture, which were also used as input data in this work.
The third source of the data are time series of climatic variables. The air temperature has a significantly negative correlation with the soil moisture, while the amount of precipitation has a positive correlation. Climatic data from the European Climate Assessment & Dataset (ECA&D) was used in this paper. ECA&D is currently made up of 69 participating organizations from 63 European countries. The main product of this initiative (E-OBS), which was also used in this work, is a daily gridded observational dataset for precipitation, the temperature, and sea level pressure for Europe. The climatic data in ECA&D is provided as a spatial time series in the netCDF [11] format for the period 1950 – 2018, in a spatial scope of 25°N – 75°N to 40°E -75°E, and in a spatial resolution of 0.25°x0.25°.

Figure 1. Location of the measured points, ESA and ECA&D Points

Figure 2. Overview of the regime of the basic environmental data
3. Methods

In this paper, the soil moisture in the root zone is modelled and is based on the described data. The main objective of the paper was the identification of a suitable method for this task. The methods proposed are briefly described below, together with the reasons for their selection.

Multiple linear regression was used for the comparisons, as it is a standard method used in regression tasks. A major condition for linear regression is that explanatory variables $X$ should not be mutually correlated too highly. However, such a correlation will probably occur in the addressed task; algorithms were therefore used that are more suitable in such a situation. Ridge, Lasso, and Elastic Net, which we used herein, redefine linear regression to prevent the effect of multicollinearity and help ensure a more stable model. Ridge regression reduces the magnitude of the coefficients. The Lasso regression method regulates the number of variables in the resulting model. Both regularization techniques improve the precision and interpretability of the statistical model. Elastic Net is a combination of the Lasso and Ridge methods [12].

Principal component regression (PCR) was also used to overcome the multicollinearity problem. In PCR, instead of directly using the explanatory variables, the principal components of the explanatory variables are used as independent variables. The central idea of principal component analysis is to reduce the dimensionality of a data set, which often contains a large number of interrelated variables while retaining as much of the variations present in the data set as possible. This reduction is achieved by transforming the original variables to a new set of variables, i.e., the principal components, which are uncorrelated, and which are ordered so that the first few of which retain most of the variations present in all of the original variables [13].

The previously described methods are essentially linear. Their advantage is that they are simple to apply, and their results interpret well. However, as the collection of soil moisture based on the complicated processes of a hydrological cycle has a strongly non-linear nature, machine learning (ML) regression models were used in this paper. Two models, i.e., XGBoost and Deep Learning Neural Networks, were selected, as they both represent the current state-of-the-art in ML modelling. They are characterised by a high degree of precision in comparison with other regression models. Both models additionally have built-in mechanisms that can handle the multicollinearity in the input data. Their characteristics are briefly stated below (at the level of acquiring intuition about their functioning, a more detailed description is available in the relevant literature), along with a description of their application.

XGBoost is based on a gradient boosting machine model (GBM). Boosting is a sequential technique that works on the principle of an ensemble. It combines a set of weak learners (usually shallow decision trees) to improve the accuracy of the prediction. Both GBM and XGBoost build a final model in a stage-wise manner through gradually refined estimations. Gradient boosting evaluates the precision of a model in a previous stage and then develops the next model, the inputs of which are weighted according to the previous results in such a way that the input data from the worst calculated samples will have a greater weight in the next calculation. The subsequent models are thus mainly focused on the previously inaccurately computed samples. XGBoost provides parallel tree boosting that helps to solve the problem at hand much faster. XGBoost uses a more regularized model structure to control overfitting, which gives it better performance and more precise results in comparison with GBM [14]. The price for the advantages of XGBoost is that it has many parameters to tune. A description of the parameters can be found at XGBoost WWW [15].

Deep learning neural networks (DLNN) is the second machine learning model used in this paper. DLNN is a continuation in the development of artificial neural networks that have been devised since the mid-20th century; due to its user-friendly software (SNNS, Genesis, NeuroDimension, applications in Matlab, etc.) and for other reasons, their application in the 1990s recorded a remarkable boom. Later, due to the instability of the results provided (even in the repeated solution of the same task), complicated training and other problems, other machine learning models attracted researchers more, e.g., Support Vector Machines [5], Random Forest [6], Gradient Boosting Machines [14], etc. Deep learning neural networks, which came to the foreground of the AI community at the end of the
previous decade, represent the return of artificial neural networks in their “deeper” form (a greater amount of hidden layers characterises them). They are capable of processing tasks with much larger data volumes than other algorithms. Unlike the neural networks of the past, modern deep learning provides training stability and good generalizations (with a good set of appropriate parameters). Compared to other models, the DLNN benefits are mainly manifested in problems characterised by large data volumes; however, the improvements of the artificial neural networks implanted in them are also positive for medium sized problems (regarding the amount of data processed) such as the one addressed in this article. From several software solutions for the design and training of DLNN (i.e., TensorFlow, MXNet, Caffe, Theano, Torch), the version in framework H2O [16] was selected for this paper. H2O follows the model of multilayer, feedforward neural networks for predictive modelling. H2O's deep learning functionalities include multithreaded and distributed parallel computations that can be run on a single or multinode cluster, an automatic, per neuron, adaptive learning rate for fast convergence; regularization options such as L1, L2 and/or dropout to prevent overfitting the model; grid search for hyperparameter optimization and model selection; and other advanced features.

Many researchers have shown that by combining the output of many predictors, more accurate predictions can be produced than what could be obtained from any of the individual predictors [17]. For this reason, the authors of the present paper assume that it is also important to examine ensembles, which could, in some cases, offer better results.

A grid search, combined with a repeated cross-validation methodology, was used to find the parameters of these models. In this approach, a set of model parameters from a predetermined grid is sent to the evaluating algorithm. A set of parameters was generated by the genetic algorithms (GA) (the “chromosome” in GA terminology) and sent to the repeated cross-validation mechanism, which is used for the evaluation of the parameter combinations [18] by developing and evaluating the model with them. N-times repeated k-fold cross-validation is used to find the best parameters for the final model. In the repeated cross-validation, the data set is divided into k subsets, and the training-testing evaluation is repeated k times. Each time, one of the k subsets is used as the test set, and the rest of the subsets are put together to form a training set. The subdividing of the training data into k-subsets is accomplished n times. The average error across all the k trials is then computed, which is a particular parameter’s “fitness” combination. The chromosome with the best fitness defines the best parameter’s values.

4. Results and Discussion

In this part, we report on the evaluation of the simulation of soil moisture using the models and data previously described. The prediction of soil moisture in the Zahorska Lowlands region in Slovakia (Figure 1) serves as the case study herein. The inputs were shaped into a standard form, e.g., as a table with rows and columns. Each row in the input data includes the date of the soil’s moisture measurement and the measurement value (although these are also the modelled data, they are necessary for the training and testing mechanisms). The soil moisture values from the remotely sensed nearby locations from ESA and GLEAM are in the next five columns. Remotely sensed evapotranspiration data from the GLEAM model then follows; finally, the minimal, maximal and average daily temperatures and precipitation from the ECA&D data set are provided.

Simple feature engineering was performed to improve the modelling. As a soil water regime is influenced by the current climate variables as well as by their values from previous days (e.g., the precipitation from preceding days makes soil water saturated for a longer period), we included evapotranspiration, precipitation, and the temperature from twenty days before the date of the predicted soil moisture between the inputs. In addition, with reference to the value which the soil moisture acquires, the exact time of the meteorological event is not the most crucial factor, i.e., it is not decisive whether it rained 13, 14 or 15 days ago. In each of these cases, the impact of the rain on the soil moisture after 2 weeks will be approximately the same. For this reason, aggregated values of the precipitation, evapotranspiration, and the minimum, maximum and average temperatures were also included among the input data. These may be more suitable than daily data because they can lead to a
better generalizing model. The aggregation was performed for precipitation using a summation and for the temperatures and evapotranspiration by an averaging of the values. The aggregation was made for the following intervals of days preceding the date of the prediction of the soil moisture: days 1 – 3, 1 – 5, 3 – 10 and 7 – 20.

![Correlation of Variables](image)

**Figure 3.** Mutual correlation of variables

The input data includes a total of 136 variables. The number of lines is limited by the amount of field moisture measurements performed with the neural probe from 2009 – 2018, i.e., 129.

An important feature of this data is that it has a relatively high number of columns compared to the number of lines. Many of the variables in the input data will, therefore, be highly mutually correlated, as shown in the histogram in Figure 3. These correlated variables are mostly time series of the soil moisture, temperatures and evapotranspiration of consecutive days; there is also a close correlation between the evapotranspiration and temperature values on the same days. The data file thus shows multicollinearity, which is a problem for some methods. The histogram in Figure 3 also has a distinct top around the 0.1 value, indicating that some variables show almost no correlation, e.g., the evapotranspiration and rain. The smaller top of the histogram can also be seen in this part with negative values; this indicates a negative dependency, e.g., between the values of the soil moisture measured by the satellites and temperatures.

The data was randomly divided into training and testing data sets. The training (calibrating) set was used for all the models to tune their parameters, and the testing set served for an evaluation of the precision of the resulting models.

The calculations were initially performed using multiple linear regression to illustrate the multicollinearity problem and for comparison. Using this standard method and the above described inputs, the soil moisture value was calculated and evaluated on the test data set. The comparison with the measured data revealed a correlation coefficient of 0.17, i.e., very inaccurate results. The task solved in this paper is to show various options for acquiring better results by an evaluation of the selected linear and nonlinear methods. All the calculations were performed by setting up the required computer codes in the environment of the statistical computer language R [19].

The LASSO regularisation linear method was used for the reduction of the variables and thus for the removal of the multicollinearity. The *glmnet* package was used for the calculations [20]. This R package fits LASSO and Elastic-Net regression models using a coordinate descent. For optimisation of the regularisation parameter lambda, which influences the number of variables selected for the resulting LASSO model, a search for the predefined value set (interval 0.3 – 30) was used. Variants of the model obtained with different lambda were evaluated on the training file using a 5-fold cross-validation. The identified optimum value of this regularisation parameter was 6.02. The computed soil
moisture is evaluated in Table 1. The resulting model includes 11 variables from the original 136. Elastic-Net was evaluated similarly to LASSO, with the difference that this model requires the optimisation of two parameters, i.e., in addition to lambda, the alpha parameter is also required. Alfa is Elastic-Net’s mixing parameter, with values in $(0,1)$, which determines the ratio of LASSO ($\text{alpha}=1$) and Ridge ($\text{alpha}=0$) types of regularization during the training. As shown in the table, this model does not give better results than LASSO, although this model for the resulting model selects 18 explanatory variables.

The next method tested for the calculation of the soil moisture in this paper is the principal component regression. As stated above, instead of the original variables, this method uses their linear combinations – principal components (PC) as explanatory variables. The original variables were zero centred and scaled to have unit variance before the analysis was undertaken. R language was used for the specification of the principal components, with the result that 87 principal components were indicated. The first ten, with the proportion of the variance explained, are displayed in Figure 3. The number of PCs considered in the resulting model was optimised using cross-validation. Four PCs that were identified as further PCs have a very low proportion of their variance explained (Figure 3). The evaluation of this method using statistical indices is given in Table 1.

![Figure 4. Principal components with the proportions of variance explained](image)

The calculation of the soil moisture was also performed using machine learning models. They have the potential of capturing nonlinear relations between the explanatory variables and the calculated soil moisture, which can increase the precision of the calculations. The XGBoost model was used first. The optimization of the internal parameters of XGBoost [21] was accomplished by a 10-fold 5 times repeated cross-validation. Seven parameters were tuned, so the genetic algorithms (GA) were applied in the tuning process. Genoud R version of GA was used, which combines evolutionary algorithm methods with a derivative based, quasi Newton method [22]. The program can also work without the Newton method. From empirical finding was concluded, that the probability of producing an optimum in a relatively small number of generations tends to increase with the GA population size, so we set population size to 500 and number of generation to 15. However, XGBoost offers quite stable results with differences in optimum RMSE between runs not more than 4%. The evolutionary algorithm in *genoud* uses nine operators that were set to their default values. The optimized parameters were the maximum number of iterations, the learning rate, the minimum loss reduction gamma, the maximum tree depth, the minimum child weight, the subsample ratio of the training instance, and the subsample ratio of the columns when constructing each tree. A description of the parameters and various recommendations for their settings can be found at XGBoost WWW [15]. The model is evaluated in Table 1.
Table 1. Evaluation of the models

| Model    | MAE   | RMSE  | d   | r   | $R^2$ |
|----------|-------|-------|-----|-----|-------|
| LASSO    | 3.13  | 3.45  | 0.62| 0.71| 0.51  |
| Elastic-Net | 2.73  | 3.00  | 0.79| 0.76| 0.58  |
| PCR      | 2.53  | 3.12  | 0.81| 0.70| 0.49  |
| XGBoost  | 2.27  | 2.61  | 0.89| 0.84| 0.71  |
| DLNN     | 2.81  | 3.10  | 0.75| 0.86| 0.75  |
| ensemble | 2.49  | 2.70  | 0.85| 0.86| 0.74  |

MAE - mean absolute error, RMSE – root mean square error, d – coefficient of agreement, r – correlation coefficient, $R^2$ – coefficient of determination

The second machine learning model used was the Deep Learning Neural Network. The framework h2o was utilised for the calculations. A 5-fold cross-validation was used for the search of the optimum parameters. The optimised parameters were (after the name of the parameter, its optimised value is in brackets): the number of hidden layers and the number of neurons in them (three layers – 20, 10, 20), the type of activating function (rectifier), the magnitude of the regularisation parameters (both l1 and l2 are equal to 0.01), and the parameter epsilon (1.0e-12) and rho (0.995), which influence the speed of the convergence. A total of 539 combinations of parameters were evaluated. The number of epochs of the development of the neural network was reduced to 50 during the optimisation of the parameters, due to the above-stated number of optimised neural networks (even with this reduction, it ran for 11 minutes on 20 logical cores AMD Ryzen Threadripper 1950X by 64 GB memory). The final calculations with 1000 epochs were computed after the acquisition of the parameters. This number of epochs based on the analysis still did not cause an overtraining of the model; the results are presented in Table 1.

All the models are evaluated in Table 1 using the mean absolute error, RMSE, coefficient of agreement, correlation coefficient, and coefficient of determination.

In this paper, linear and nonlinear models for simulating soil moisture were compared. The benefit of the former is their simple interpretation; the benefit of the latter is the opportunity to capture more complicated relations between the variables investigated. The results indicate that linear modelling should not be abandoned prematurely if the basic multiple linear regression does not show satisfactory results. As stated in Table 1, the more sophisticated methods leading to linear models (LASSO, Elastic-Net, PCR) demonstrated better results, with regard to the multiple linear regression, although not as good as machine learning models. Two best models (DLNN and XGBoost) acquired herein are not equal and therefore, do not provide identical results. The dissimilarity of the models can be

Figure 5. Comparison of the importance of the first 10 variables of DLNN and XGBoost

In this paper, linear and nonlinear models for simulating soil moisture were compared. The benefit of the former is their simple interpretation; the benefit of the latter is the opportunity to capture more complicated relations between the variables investigated. The results indicate that linear modelling should not be abandoned prematurely if the basic multiple linear regression does not show satisfactory results. As stated in Table 1, the more sophisticated methods leading to linear models (LASSO, Elastic-Net, PCR) demonstrated better results, with regard to the multiple linear regression, although not as good as machine learning models. Two best models (DLNN and XGBoost) acquired herein are not equal and therefore, do not provide identical results. The dissimilarity of the models can be
illustrated by Figure 4 also, where the variable importance of the DLNN and XGBoost models is compared. It can be seen that each model has different preferred variables. The diversity of the models is also shown by the fact that the statistical indicators (Table 1) are not all the best for the same model, three indicators are better for XGBoost and two for DLNN. With approximately the same degree of accuracy of the models, this means that in different conditions (e.g., weather conditions), different models can provide more precise calculations. This points to the suitability of merging models in an ensemble, which was performed by a simple averaging of the XGBoost and DLNN models. The other models were not included in the ensemble due to their lower degree of precision. As shown by the last line of the table, this model gives better balanced results for all indicators.

5. Conclusion
In this paper, the possibility of using meteorological and remotely sensed data for the calculation of a complete time series of soil moisture is assessed. Such time series are useful for the evaluation of the moisture regime of soil or for decisions regarding building irrigation or drainage structures and for some other practical and research tasks. If such data is measured in a specific location by common metering devices, such as tensiometers, neutron probes, time domain reflectometry, etc., such measurements are time demanding and require transport to the location of the measurements, the transport of the samples to the laboratory, etc. If a metering device cannot be permanently installed at a given location (e.g., if a neutron probe is used) and a data logger cannot be used, it is usual that the time series acquired will have longer time steps, which are unsuitable for several reasons.

This task is being solved using data which are described in the article to train (calibrate) models on the parts of the days on which the measurements were performed (training set of data). Several problems related to the multicollinearity of the input data for these models had to be addressed. New variables were constructed (feature engineering) and their contribution to the precision of the results were tested also. The resulting models were verified using different data from days on which the field measurements were available as those which was used for training. The final purpose of this work is to apply the model to data for which the soil moisture measurements were not originally performed.

The results show that some linear statistical models can be used (although not all of them), but mainly nonlinear machine learning models are suitable for a solution of the given task. Extreme Gradient Boosting Machines and Deep Learning Neural Network methods were used. The Extreme Gradient Boosting Machines model and simple ensemble model provided the best results. The comparison of the models is based on several statistical indicators in Table 1. It could be concluded that tested machine learning models are preferred alternative from tested models for the specification of a time series of soil moisture and for evaluating the water regime of soil. Moreover, other models and more sophisticated ensembles can be evaluated for this task in the future with the aim of even better results.

Acknowledgment(s)
This work was supported by the Slovak Research and Development Agency under Contract No. APVV-15-0489 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/0662/19.

References
[1] N. Romano, “Soil moisture at local scale: Measurements and simulations” Journal of Hydrology, vol. 516, pp. 6-20, 2014.
[2] V. Novak, H. Hlavackova, “Applied Soil Hydrology” Springer, Cham, ISBN: 978-3-030-01805-4, 2018.
[3] G. P. Petropoulos, I. Gareth, B. Brian, “Surface soil moisture retrievals from remote sensing: Current status, products & future trends” Physics and Chemistry of the Earth, Parts A/B/C, vol. 83-84, pp. 36-56, 2015.
[4] H. R. Maier, G. C. Dandy, “Neural networks for the prediction and forecasting of water
resources variables: a review of modelling issues and applications” *Environmental modelling & software*, vol. 15(1), pp. 101-124, 2000.

[5] L. Wang, “Support vector machines: theory and applications”. *Springer Science & Business Media*, vol. 177, pp. 341, ISBN: 978-3-540-24388-5, 2005.

[6] L. Breiman, “Random forests” *Machine Learning*, vol. 45(1), pp. 5–32, 2001.

[7] S. Mecklenburg, et al. “ESA's Soil Moisture and Ocean Salinity Mission: Mission Performance and Operations” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50(5), pp. 1354-1366, 2012.

[8] J. L. Hatfield, J. H. Prueger, “Spatial and temporal variation in evapotranspiration” *Evapotranspiration: from measurements to agricultural and environmental applications, InTech*, ISBN: 978-953-307-512-9, 2011.

[9] C. H. B. Priestley, R. J. Taylor, “On the assessment of surface heat flux and evaporation using large-scale parameters” *Monthly weather review*, vol. 100(2), pp. 81-92, 1972.

[10] B. Martens, et al. “GLEAM v3: Satellite-based land evaporation and root-zone soil moisture” *Geoscientific Model Development*, vol. 10(5), pp. 1903-1925, 2017.

[11] R. Rew, et al. “NetCDF User's Guide. An Interface for Data Access. 2.3” available at: <https://www.unidata.ucar.edu/software/netcdf/>, 1993.

[12] J. H. Friedman, T. Hastie, R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent” *Journal of Statistical Software*, vol. 33(1), pp. 1-22, 2010.

[13] I. Jolliffe, “Principal component analysis” *International Encyclopedia of Statistical Science, Springer, Berlin*, pp. 1094-1096, ISBN: 978-3-642-04897-5, 2014.

[14] T. Chen, C. Guestrin, “XGBoost: a scalable tree boosting system” *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York*, pp. 785–794, ISBN: 978-1-4503-4232-2, 2016.

[15] XGBoost “XGBoost Documentation” available at: <https://xgboost.readthedocs.io/en/latest/>, 2019.

[16] A. Arora, et al. “Deep Learning with H2O” pp. 47. available at: <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf>, 2016.

[17] F. Serafino, G. Pio, M. Ceci, “Ensemble Learning for Multi-Type Classification in Heterogeneous Networks” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30(12), pp. 2326-2339, 2018.

[18] T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction” *Springer series in statistics, Springer, New York*, ISBN: 978-0-387-84857-0, 2009.

[19] R Core Team: “R: a language and environment for statistical computing” *R Foundation for Statistical Computing, Vienna*, available at: https://www.R-project.org/, 2017.

[20] T. Chen, et al. “xgboost: Extreme Gradient Boosting” *R package version 0.71.2*, available at: <https://CRAN.R-project.org/package=xgboost>, 2020.

[21] W. R. Mebane Jr, J. S. Sekhon, “Genetic optimization using derivatives: the rgenoud package for R” *Journal of Statistical Software*, vol. 42(11), pp. 1-26, 2011.