Vol. 41, No. 1, pp 25-29, 2013
Indian Journal of Soil Conservation
Online URL:http://indianjournals.com/ijor.aspx?target=ijor:ijsc&type=home

Comparative evaluation of nearest neighbor and neural networks approach to estimate soil water retention at field capacity and permanent wilting point

N. G. Patil[1], C. Mandal and D. K. Mandal

National Bureau of Soil Survey and Land Use Planning, Amravati Road, Shankarnagar, P.O. Nagpur, India.
[1]E-mail nitpat03@yahoo.co.uk

ARTICLE INFO

Article history :
Received : March, 2011
Revised : July, 2012
Accepted : November, 2012

Key words :
Field capacity,
K-nearest neighbor Algorithm ,
Neural regression,
Pedotransfer function,
Permanent Wilting point,
Vertisols

ABSTRACT

Evaluation of neural and k nearest neighbor (kNN) techniques of developing pedotransfer functions (PTF) to predict soil water held at -33 kPa (Field Capacity FC) and -1500 kPa (Permanent Wilting Point PWP) of Vertisols of India is presented. Soil profile information of 26 representative sites comprising 157 soil samples was used for PTF development. Four levels of input information were used, (1) Textural data (data on sand, silt, and clay fraction-SSC), (2) Level 1+bulk density data (SSCBD), (3) Level 2+organic matter (SSCBDOM), and (4) Level 1+organic matter (SSCOM), kNN PTFs predicted FC with greater accuracy evidenced by lower root mean square error -RMSE (0.0695) compared to neural PTFs (0.0775). Performance of neural PTFs exhibited improvement in RMSE (from 0.076 to 0.0672) as the input variables increased. The performance of kNN PTF was better (RMSE, 0.0315) than neural PTF using input level 1 (RMSE, 0.0402) to estimate PWP. At highest level of input, neural and kNN PTFs were almost at par (RMSE, 0.0353 and 0.0358) in terms of prediction error. Better prediction by kNN PTFs (FC/ PWP) with lowest input level (SSC) was significant as accurate predictions were possible without more input. In general, kNN PTFs showed advantage over neural PTFs.

1. INTRODUCTION

Modeling soil water dynamics constitutes a core part of many simulations pertaining to hydrological process, irrigation planning, soil-plant-water relationship, crop modeling, *etc*. However, data on soil hydraulic properties are not usually available because conventional methods of measurement are arduous, time intensive and expensive. Therefore soil hydraulic properties are routinely generated employing indirect estimation techniques. Use of Pedo Transfer Functions (PTF) is one of the widely used techniques. Most of the PTFs reported in the literature are derived using regression approach. Neural regression is considered effective tool for developing PTFs and a vast array of neural PTFs (Jain *et al*., 2004; Minasny *et al.,* 1999; Minasny and Mc Bratney 2002; Patil *et al*., 2010; Scahaap *et al.,*1998) are available.

ANN and kNN techniques: Literature survey shows that neural regression technique is favoured by researchers for developing PTFs. ANN can mimic the behavior of complex systems by varying the strength of network components (basic soil properties) on each other as well as its range of choice of structures of interconnections among components. The neural network typically consists of '*j*' input neurons, '*k*' hidden neurons, and 'l' output neurons. The input and output neurons are related through a network of neurons. Advantage of using neural networks (non-parametric approach) to develop PTFs lies in the fact that they do not require *a prior* regression model, which relates input and output data (Schaap *et al.,* 1998). Analogue approach like k Nearest Neighbor (kNN) based on similarity functions is another alternative preferred by researchers (Lall and Sharma, 1996; Rajagopolan and Lall, 1999) when *a prior* information on relationship is unknown.

Being essentially empirical, PTFs are location specific and their spatial application is always prone to errors. Thus, database used in development of PTFs must have sufficient spread to represent the variations in the soilscape of the area. Obviously, the development database used in PTF development is critical to the accuracy of predictive ability of derived PTF. Since acquisition of the data is a continued process, it is essential that the PTFs are also improved by adding to the development database. Unfortunately, regression PTFs do not provide such flexibility. Thus,

whenever data are added, PTFs must be developed again repeating the process of calibration, validation and testing.

Alternatively, pattern recognition algorithms can be used to replace equation fitting techniques. Recently Amir Lakzian *et al*. (2010) evaluated different techniques including statistical, k nearest neighbour, (kNN), Artificial Neural Network (ANN) and PTFs were calibrated to predict the soil water content. Their results showed that kNN PTFs performed better than other PTFs in prediction of FC and PWP. Another study by Nemes *et al.* (2009) recommended that the statistical PTFs developed by Rawls *et al.* (1982) not be used in the context of the national scale. They suggested alternative technique *i.e.* more advanced PTF development k-Nearest Neighbor as a desirable technique. Similarly Patil *et al.* (2011, 2012a, 2012b) have argued that pattern recognition algorithms like kNN could replace neural regression resulting in PTFs that overcome the constraint faced by neural PTFs because reference database can be easily appended and the additional data can be used to improve accuracy of developed. They have reported superior performance of kNN over ANN as a tool of PTF calibration. kNN technique is one of the easiest machine learning technique because classification is achieved by identifying the nearest neighbours to a query example and using those neighbours to determine the class of the query. This study was aimed at development of PTFs to estimate field capacity (FC) and permanent wilting point (PWP) of Vertisols and their intergrades in India. The neural and kNN techniques were evaluated for their efficacy in developing PTFs.

2. MATERIALS AND METHODS

Data reported by Pal *et al.* (2003) was used for the study which included basic soil information and soil water retention properties. Salient features of the development database used in the study are presented (Table 1). Except sand content, all the basic soil properties exhibited relatively lower coefficient of variation. The database contains information on twenty six profiles collected from the Indian states of Madhya Pradesh, Maharashtra, Karnataka, Andhra Pradesh, Tamil Nadu, Gujarat and Rajasthan (Fig. 1).

They represent sub-humid (moist), sub-humid (dry), semi-arid (moist), semi-arid (dry), arid climatic regions

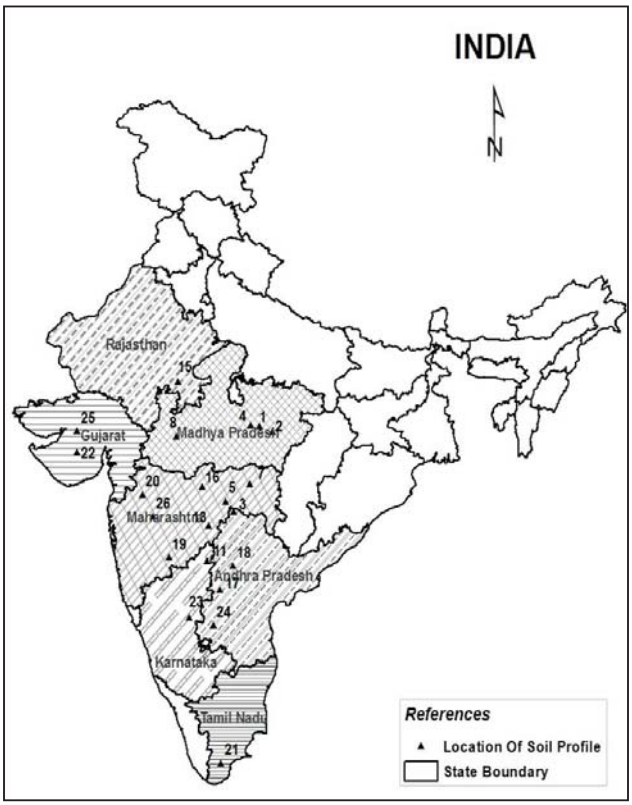

Fig.1. Location of Vertisol profiles in different states of India.

with Mean Annual Rainfall (MAR) of 1127- 1448 mm, 1011-1084 mm, 924-977 mm, 583-842 mm and ≤533 mm, respectively. The majority of these soils are developed in the alluvium of weathered Deccan basalt. Two techniques were used to build PTF namely artificial neural networks (ANN) based regression and kNN. For developing ANN based PTFs, software 'Neurointelligence' (Alyuda Research Company, USA) was used. Based on the earlier experience, (Patil *et al*., 2010), feed forward neural network model with three hidden nodes was preferred. The data set were partitioned into 'training' (95 samples), validation (22), and test (22) sets (4 samples were discarded because of discrepancy). Upon finding an appropriate network model, the PTF was calibrated. For network training, Levenberg-Marquardt (L-M) algorithm was chosen due to the fact that the data is small. Software developed by Nemes *et al.* (2008) was used to build PTFs for estimating FC and PWP from

Table: 1
Statistical summary of soil properties of 143 soil samples

	Sand (%)	Silt (%)	Clay (%)	Bulk Density ($Mg\ m^{-3}$)	Organic matter (%)	FC (m^3m^{-3})	PWP (m^3m^{-3})
Mean	9.534	32.800	57.664	1.46	0.52	0.38	0.20
S.E.	0.928	0.730	1.012	0.01	0.02	0.01	0.00
Variance	123.377	76.213	146.613	0.02	0.06	0.01	0.00
Coef. Var.	1.164	0.266	0.209	0.09	0.46	0.21	0.24
Minimum	0.200	16.400	12.200	1.10	0.08	0.21	0.08
Maximum	48.520	52.410	79.210	1.80	1.55	0.58	0.32

basic soil properties like textural distribution, bulk density and organic matter in hierarchical order. The software/tool combines kNN algorithm with the bootstrap data-subset selection technique to allow the development of model ensembles; that can be used to estimate the uncertainty of the final model output. They have reported that the PTFs developed using kNN were as efficient as the PTF developed using most advanced neural computing techniques.

Four levels of input information were used to avoid possible bias towards one set of inputs and dependencies between basic soil properties and FC/PWP were established.

- Input level 1 Textural data (data on sand, silt, and clay fraction-SSC)
- Input level 2 Level 1+bulk density data (SSCBD)
- Input level 3 Level 2+organic matter (SSCBDOM)
- Input level 4 Level 1+organic matter (SSCOM)

Performance Evaluation

Performance of the k nearest (kNN) algorithm was evaluated against estimations made by neural network models, developed using the same data and input soil attributes. Performances of the developed PTFs was evaluated based on (i) root mean square error (RMSE), (ii) index of agreement (d), (iii) maximum absolute error (ME) iv) mean absolute error (MAE) and v) coefficient of determination (R^2). RMSE, d, ME, and MAE statistics were calculated using following equations respectively, where *n* represents the number of data used for modeling and E_i and M_i represent measured and computed value respectively. The unit of errors is m^3m^{-3}.

Root Mean Square Error (Fox 1981)

$$RMSE \quad \sqrt{\frac{\sum_{i=1}^{n} (Ei - Mi)^2}{n}} \qquad (1)$$

Index of Agreement (Willmott and Wicks. 1980)

$$d \quad 1 \quad \frac{\sum_{i=1}^{n} (Ei - Mi)^2}{\sum_{i=1}^{n} \left|Ei - \overline{M}\right| \left|Mi - \overline{M}\right|^2} \qquad (2)$$

Maximum Absolute Error (Loague and Green. 1991)

$$ME = Max \ |Ei-Mi| \qquad (3)$$

Mean Absolute Error (Schaeffer 1980).

$$MAE \quad \sum_{i=1}^{n} \frac{|Ei - Mi|}{n} \qquad (4)$$

Linear correlation coefficient (Pearson 1900)

$$r \quad \frac{1}{n-1} \sum_{i=1}^{n} \frac{(M_i - \overline{M})(E_i - \overline{E})}{S_M S_E} \qquad (5)$$

Where S_M and S_E represent sum of measured and computed values respectively. It is used here as a coefficient of determination (R^2) by squaring 'r'.The RMSE statistic indicates the model's ability to predict away from the mean. RMSE imparts more weight to high values because it involves square of the difference between observed and predicted values. Ideally the model should have the smallest MAE and smallest overall dispersion (RMSE). Degree of agreement d is dimensionless index that assists in understanding the closeness of measured and estimated variables, r indicates strength of dependence of two variables on each other.

3. RESULTS AND DISCUSSION

The performance of PTFs developed using kNN and neural networks could be judged from the statistical indices (Table 2).

It could be observed that at lowest input level (SSC), the performance of kNN PTF was relatively better (Fig. 2) as indicated by lower RMSE (0.0639) than RMSE of 0.076 at the same input level in neural PTF. Other indices (d, ME, MAE, R^2) also confirmed better ability of kNN PTF. Incremental addition of bulk density data as input variable did not improve performance of kNN PTF as evidenced by increased RMSE (0.0712) in predicting FC. The RMSE

Table : 2
Statistical indices to evaluate performance of hierarchical kNN and Neural PTFs developed

	RMSE	d	ME	MAE	R^2
Input kNN PTF to estimate FC					
SSC	6.390	0.793	16.321	4.213	0.480
SSCBD	7.121	0.710	16.485	5.012	0.341
SSCBDOM	7.431	0.662	17.023	5.321	0.286
SSCOM	6.842	0.750	17.650	4.236	0.411
Mean	6.946	0.730	16.870	4.696	0.380
Neural PTF to estimate FC					
SSC	33.323	0.186	41.752	32.847	0.043
SSCBD	9.032	0.570	19.962	7.398	0.094
SSCBDOM	7.651	0.690	16.184	6.235	0.287
SSCOM	6.724	0.760	15.592	5.132	0.450
Mean	14.183	0.558	23.373	12.903	0.219
kNN PTF to estimate PWP					
SSC	3.152	0.810	7.412	2.312	0.519
SSCBD	3.532	0.741	7.624	2.745	0.393
SSCBDOM	3.531	0.720	8.365	2.784	0.402
SSCOM	3.424	0.773	8.258	2.543	0.439
Mean	3.410	0.760	7.915	2.596	0.438
Neural PTF to estimate PWP					
SSC	4.022	0.605	8.687	3.275	0.457
SSCBD	3.481	0.795	7.415	2.836	0.527
SSCBDOM	3.584	0.770	9.425	2.825	0.562
SSCOM	3.666	0.722	8.247	3.056	0.581
Mean	3.688	0.720	8.444	2.998	0.532

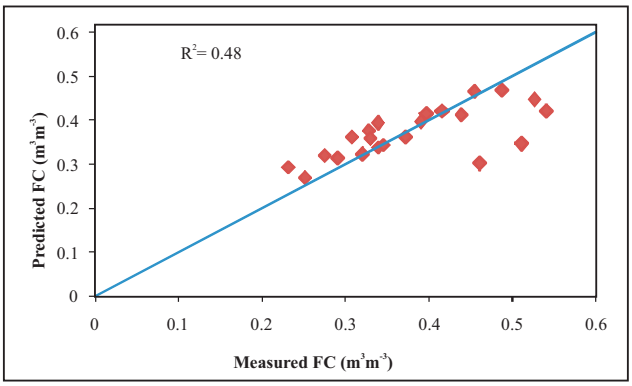

Fig.2. Measured and predicted FC using kNN PTF with input level 1 (SSC)

continued to increase with addition of organic matter as input with addition of bulk density. However, addition of organic matter alone (without BD) as additional input with textural composition exhibited lower RMSE (0.0684). In general, kNN PTFs had lower mean RMSE (0.0695) compared to neural PTFs (0.1418). Other statistical indicators also indicated that kNN PTFs predicted FC with greater accuracy irrespective of input/predictor variable level. Performance of neural PTFs exhibited improvement in RMSE (from 0.3332 to 0.0672) as the input variables increased. These results were expected as neural networks (or any predictive method) are known to show better predictive ability with increase in number of input variables. However, the lowest RMSE (0.0672) was recorded at input of texture and OM (Fig. 3). The magnitude of ME and MAE were also lower for this PTF. Thus, information on bulk density alone or in combination with organic matter could not enhance ability of neural networks to mimic the relationship between input and predicted variable-FC. These soils are known to be poor in OM status (< 2 %). But, OM status can only partly explain poor performance of predictive models. On the other hand, lower level input of texture and OM lowered the RMSE, ME and MAE while improving d (0.76). Only R^2 value suggested better performance by neural PTFs using highest hierarchical level (SSCBDOM) as compared to other levels, but

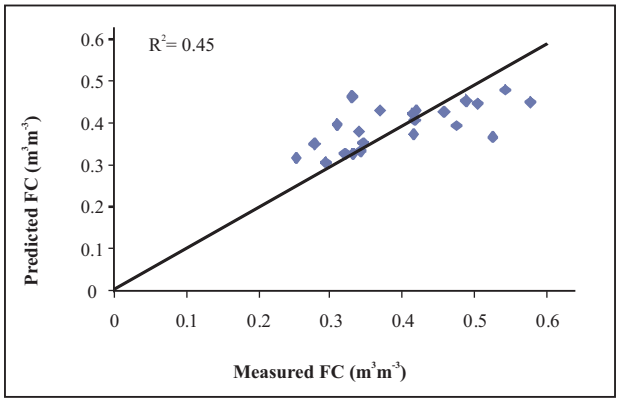

Fig.3. Measured and predicted FC using neural PTF with input level 4 (SSCOM)

performance evaluation by other indices confirmed the status of PTF using SSCOM input as the best neural PTF to predict FC. The important finding was that FC could be predicted with greatest accuracy using kNN PTF that used textural composition (SSC) as an input (lowest input level).

The performance of kNN PTF was better (RMSE 0.0315) than neural PTF using the same input level (0.0402) to estimate PWP. Degree of agreement, ME, MAE, R^2 values also confirmed these findings. The difference between magnitude of RMSE was however much lower as compared to the difference in predicting FC. With additional input of bulk density and OM, the performance of kNN PTFs declined as suggested by all statistical indices. Thus PWP was predicted with greater accuracy with input of texture (SSC) data alone than (Fig. 4) the addition of other input variables. Inclusion of bulk density resulted in marginally better neural PTF,-RMSE 0.0348 as against 0.0353 in kNN. The difference was not statistically viable for conclusive argument. However, mean absolute error in neural prediction was relatively higher. Identical results were observed when OM replaced BD as an input in addition to SSC. At highest level of input, neural and kNN PTFs were almost at par (RMSE 0.0353 and 0.0358) in terms of prediction error. Among the neural PTFs, the best performance was observed at input level-SSCBD (Fig. 5) followed by SSCOM. The PTFs did not exhibit improvement trend with increased input.

In general, kNN PTFs showed marginal advantage over neural PTFs. Better prediction of FC and PWP at lowest input level (SSC) followed by PTFs using SSC and OM was significant as accurate predictions were possible without more input. It was evident that as a tool, kNN performed better than neural networks. Though bulk density and organic matter/carbon are known to influence soil water retention, the underlying relationship between FC/PWP and BD and/or OM could not be captured by kNN as effectively as neural networks. However, this opinion is an interpretation that needs to be substantiated. The kNN technique however proved to be competitive alternative to

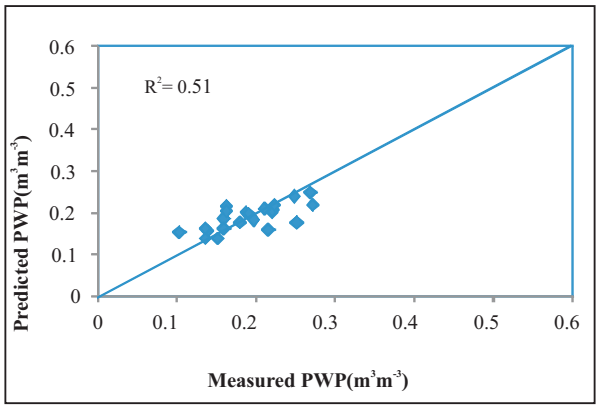

Fig.4. Measured and predicted PWP using kNN PTF with input level 1 (SSC)

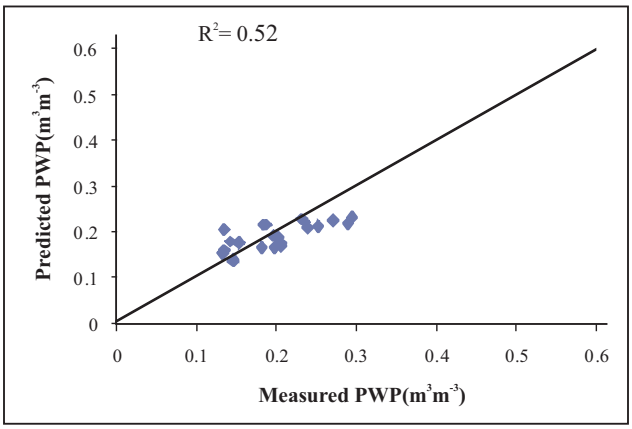

Fig.5. Measured and predicted PWP using neural PTF with input level 2 (SSCBD)

neural networks to develop PTFs, especially since re-development of this PTF is needed as new data become available.

4. CONCLUSIONS

Neural regression and kNN techniques of PTF development were evaluated. kNN and PTFs were recommended for estimating AWC of Vertisols. It was concluded that kNN technique of calibrating PTFs can be as competitive as widely used neural regression with additional benefit of appending the development data as and when desired. These findings will facilitate refinement of PTFs with acquisition of more data.

REFERENCES

Amir Lakzian, Mohammad Bannayan Aval and Nasrin Gorbanzadeh 2010. Comparison of Pattern Recognition, Artificial Neural Network and Pedotransfer Functions for Estimation of Soil Water Parameters. *Sci. Biol.*, 2 (3) : 114-120

Fox, D.G. 1981. Judging air quality model performance: a summary of the AMS workshop on dispersion models performance. *Bull. Am. Meteorol. Soc.*, 62:599-609.

Jain, S. K., Singh, V. P. and Van Genuchten., M. Th. 2004. Analysis of soil water retention data using artificial neural networks. *J. Hydro. Engg.*, 9: 415-420.

Lall, U. and Sharma A., 1996. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resour. Res.,* 32:679-693.

Loague, K. and Green R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. *J. Contam. Hydrol.*, 7: 51-73.

Minasny, B. and McBratney, A.B. 2002. The neuro-m method for fitting neural network parametric pedotransfer functions. *Soil Sci. Soc. Am. J.*, 66: 352-361.

Minasny, B., McBratney, A.B. and Bristow, K.L. 1999. Comparison of different approaches to the development of pedotransfer functions for water retention curves. *Geoderma*, 93: 225-253.

Nemes A., R.T. Roberts, W.J. Rawls, Ya.A. Pachepsky and M.Th. van Genuchten 2008. Software to estimate −33 and −1500 kPa soil water retention using the non-parametric k-Nearest Neighbor technique. *Environ. Modeliling Soft.*, 23(2): 254-255.

Nemes, D.J. Timlinb, Ya. A. Pachepskyc and J. Rawls 2009. Evaluation of the Pedotransfer Functions for their Applicability at the U.S. National Scale. *Soil Sci. Soc. Am. J.*, 73(5) : 1638-1645, doi: 10.2136/sssaj2008.0298.

Nemes, W.J. Rawls, Ya. A. Pachepsky and M. Th. van Genuchten 2006. Sensitivity Analysis of the Nonparametric Nearest Neighbor Technique to Estimate Soil Water Retention. *Vadose Zone J.*, 5: 1222-1235.

Nemes, Walter J. Rawls and Yakov A. Pachepsky 2006. Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties. *Soil Sci. Soc. Am. J.*, 70: 327-336.

Pal, D.K., Bhattacharya T., Ray, S.K., and Bhuse, S.R. 2003. Developing a model on the formation and resilience of naturally degraded black soils of the peninsular India as a decision support system for better land use planning. Unpublished report, NBSS and LUP, Nagpur, India.

Patil, N.G. and Chaturvedi, A. 2012. Pedotransfer functions based on nearest neighbour and neural networks approach to estimate available water capacity of shrink-swell soils. *Ind. J. Agri. Sci.*, 82 (1): 35-38.

Patil, N.G., Pal, D.K., Manda, C. and Mandal, D.K. 2011. On Describing Soil Water Retention Characteristics of Vertisols and Pedotransfer Functions Based on Nearest Neighbor and Neural Networks Approach to Estimate AWC. *J. Irri. Drain. Engg.,* doi: http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000375. Available Online 30 April 2011 Print (2012)138(2): 177-184.

Patil, N.G., Tiwary, P., Pal, D., Bhattacharya, T., Sarkar, D., Mandal, C., Mandal, D., Chandran, P., Ray, S., Prasad, J., Lokhande, M. and Dongre, V. 2012b. Soil Water Retention Characteristics of Black Soils of India and Pedotransfer Functions Using Different Approaches *J. Irri. Drain. Engg.,* doi: http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000527. Available Online 15 Aug 2012.

Patil, N.G., Rajput, G.S., Nema, R.K. and Singh, R.B. 2010. Predicting hydraulic properties of seasonally impounded soils. *J. Agri. Sci.*,148: 159-170. doi:10.1017/S002185960999030X.

Pearson, K. 1900. On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling. *Philosophical Magazine Series 5*, 50 (302): 157-175. doi:10.1080/14786440009463897.

Rajagopolan, B. and Lall U., 1999. A k-nearest-neighbor simulator for daily precipitation and other variables. *Water Resour. Res*., 35:3089-3101.

Rawls, W.J., Brakensiek, D.L. and Saxton, K.E. 1982. Estimation of soil water properties. Trans. *Amer. Soc. of Agric. Engg.*, 25(5):1316-1328.

Schaap, M. G., Leij, F.L. and Van Genuchten, M.Th. 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Sci. Soc. Ame. J.*, 62: 847-855.

Schaeffer, D.L. 1980. A model evaluation methodology applicable to environmental assessment models. *Ecol. Model,* 8:275-295.

Willmott, C.J., and Wicks,D.E. 1980. An empirical method for the spatial interpolation of monthly precipitation within California. *Phys. Geogr.*, 1:59-73.