

# **Evaluation of the cropland modelling component of the U.S. national scale CEAP project: estimation of soil hydraulic properties**

Attila Nemes<sup>1,2</sup> – Dennis J. Timlin<sup>2</sup> – Bruno Quebedeaux<sup>1</sup> – Vangimalla R. Reddy<sup>2</sup>

<sup>1</sup>Department of Plant Science and Landscape Architecture, University of Maryland, USA,  
<sup>2</sup>USDA-ARS Crop Systems and Global Change Lab, USA. Corresponding author: A. Nemes,  
Tel.: +1 301 504 5177; Fax: +1 301 504 5823; E-mail: [Attila.Nemes@ars.usda.gov](mailto:Attila.Nemes@ars.usda.gov)

## **1. Abstract**

The USDA-Natural Resources Conservation Service (NRCS) is partnering with the USDA Agricultural Research Service (ARS), the National Agricultural Statistics Service (NASS), Farm Service Agency (FSA) and other agencies to conduct a national assessment of environmental benefits and effects of 2002 Farm Bill programs. The resulting Conservation Effects Assessment Project (CEAP) has two components; one of which is a national-scale assessment effort using the National Resources Inventory (NRI) as a sampling base for estimating the environmental benefits of the implementation of conservation practices both on-site and off-site. The Agricultural Policy Environmental Extender (APEX) model has been proposed for use to evaluate on-site benefits of conservation practices in place in cultivated croplands. Farmer surveys have been conducted at a subset of NRI sample points about ongoing farming activities and conservation practices; and an array of databases has been generated and utilized to provide base data to the simulation model. Existing data provides baseline estimates with current practices employed at farms that use NRCS conservation programs. “No practices” alternative scenario will be run in order to estimate the benefits of those programs. An independent evaluation of the cropland component of the national-scale assessment is being performed at the USDA-ARS Crop Systems and Global Change Laboratory and the University of Maryland. As part of the systematic evaluation process, we evaluate whether the approximation of in situ soil water conditions is done properly. A pedotransfer approach is embedded into the APEX model, estimating water retention at -33 and -1500kPa matric potentials using the renowned pedotransfer function of Rawls et al. (1982). Critical review revealed that the Rawls et al. PTF delivers sub-optimal and biased estimates for US conditions – for which these equations were widely considered to be valid. An alternative solution is being suggested to estimate soil hydraulic properties.

## **2. Background**

The USDA-Natural Resources Conservation Service (NRCS) is partnering with the USDA Agricultural Research Service (ARS), the National Agricultural Statistics Service (NASS), Farm Service Agency (FSA) and other agencies to conduct a national assessment of environmental benefits and effects of 2002 Farm Bill programs. The resulting Conservation Effects Assessment Project (CEAP) has two components; one of which is a national-scale assessment effort using the National Resources Inventory (NRI) as a sampling base for estimating the environmental benefits of the implementation of conservation practices both on-site and off-site. The Agricultural Policy Environmental Extender (APEX) simulation model has been proposed for use to evaluate on-site benefits of conservation practices in place in cultivated croplands. Farmer surveys have been conducted at a subset of NRI sample points about ongoing farming activities and conservation practices; and an array of databases has been generated and utilized to provide base data to the simulation model. Existing data provides baseline estimates with current practices employed at farms that use NRCS conservation programs. “No practices” alternative scenario will be run in order to estimate the benefits of those programs. An independent evaluation of the cropland component of the national-scale assessment is being performed at the USDA-ARS Crop Systems and Global Change Laboratory and the University of Maryland which involves examination of the data and other information used, the simulation model, and abstractions in the modelling approach.

## **3. Introduction**

Simulation modeling provides a feasible alternative to field monitoring when large scale environmental concerns are to be addressed. However, various model inputs exist that cannot be collected – at least not in a

feasible manner - at large scale. Soil hydro-physical properties are among those properties. In such cases simulation models mostly rely on using information collected from small-scale (point) samples. This practice, however, raises the need for an accurate and reliable up-scaling protocol. Water quality assessments, crop simulation studies and projects/programs like CEAP typically utilize such up-scaled information. Finding the proper tool to estimate soil hydraulic properties is one of the first key elements of up-scaling. Among the applicable methods is the use of pedotransfer functions (PTFs) that establish a functional relationship between existing basic soil physical properties and the missing soil hydraulic properties. There is an abundance of methods and databases used today to develop PTFs (c.f. Pachepsky and Rawls, 2004), each yielding somewhat different – or even largely different – estimates. It is essential that great care is taken when the quality of the selected PTF is assessed. This includes e.g. ensuring that the PTF is not biased for the applicable area, ensuring sufficient complexity of the PTF development method, ensuring relevance of input variables and assessing the potential influence of differences in measurement methodology or classifications.

Accuracy in estimating soil hydraulic properties carries special importance in the modeling process. Incorrect estimates not only directly impact simulated crop growth via providing false information on water availability, but have potential indirect implications via falsely triggering certain other (e.g. chemical) simulation processes in the model that are bound to certain soil water (and/or aeration) conditions. Since the impact of such events will propagate through later time steps and potentially to yet other processes, it may result in far reaching consequences for various simulated environmental measures on-site or even off-site.

In this study we examined the appropriateness of a long established PTF that was initially suggested for use in the CEAP assessment project in the US.

#### 4. Materials and Methods

The well known PTF of Rawls et al. (1982) was suggested for use in CEAP. This set of regression equations provide point estimates on the water retention curve (WRC) - i.e. water content at particular matric potentials - using sand, silt and clay content, organic matter (OM) content, and optionally bulk density (BD) and 1 or 2 existing values on the WRC. Of the several alternatives, the selected functions for CEAP did not use such optional input variables in estimating water retention at -33 and -1500 kPa (θ<sub>33</sub> and θ<sub>1500</sub> respectively). While it has been recognized in the past that these values are not valid for all soil types and conditions, lacking a better general solution, water retention at these two matric potentials are often used to approximate field capacity (FC) and wilting point (WP) respectively, two values with special significance in crop modelling.

Initial cross testing on independent data suggested that the Rawls et al. (1982) equations return somewhat biased estimates of θ<sub>33</sub> and θ<sub>1500</sub> (and derived available water holding capacity, AWHC) for a data collection that is considered equally representative for the US, therefore we felt that an in-dept study was necessary to unveil the potential reasons for that. Differences in applied data collection methodology or biased representation of different sub areas are often the reason for such phenomenon.

Rawls et al. established their PTFs using a collection of 2541 samples, which was a subset of a larger master dataset originating from 26 sources, covering a wide variety of soil types from 32 states of the contiguous United States. Having access to the master data set, and applying limitations set forth in Rawls et al (1982), we managed to identify 2528 of the 2541 samples originally used. Their quasi-identical statistical properties (apart from one measure, OM content, which will be addressed later) allowed us to conclude that the 2528-size data set that we identified is virtually identical to the 2541-size set used originally by Rawls et al. (Table 1). The difference of 13 samples may have originated from rounding of data that later affected sample size when limitations were applied or by the removal of duplicate or errant entries.

A second data set has been used for independent testing of the suggested Rawls et al. (1982) PTFs and any alternatives generated in this study. The National Soil Survey Characterization (INSSC) database of the US Natural Resources Conservation Service, Soil Survey Laboratory was used for this purpose (Soil Survey Staff, 1997). The database contains data of some 120.000 soil horizons from all states of the US – and some beyond the US. Our data selection was limited to the top 1m of soils from the contiguous 48 states of the US, limitations in soil properties were set following guidelines in Rawls et al. (1982). Data were filtered for obvious inconsistencies and the data vector had to contain data on sand, silt and clay content, OM content, BD, and θ<sub>33</sub> and θ<sub>1500</sub> for each selected sample. This yielded 9395 samples, which were randomly split into two – 4698 ('A') and 4697 ('B') size - subsets to best facilitate an independent two-step testing of findings. Summary statistics of subset 'B' – used for testing – is also shown in Table 1. Subset 'A' was used to develop the same PTFs from a statistically identical but still independent data set and obtain baseline results.

We used the same linear regression methodology and inputs as Rawls et al. did to redevelop the original equations. We elected this approach to ensure that any differences are due to differences in the properties of the underlying data set and not due to changes in PTF methodology or using additional inputs.

However, we also wished to examine whether the simple linear regression technique originally used by Rawls et al. (1982) was (1) complex enough to unveil all the systematic relationships among inputs and outputs in this data collection, and (2) was suitable to offer resolution for the significant skew between the Rawls and the NSSC data sets in terms of different soil textures being represented (c.f. Table 1). To examine this, we utilized the k-Nearest Neighbor (k-NN) pattern recognition technique of Nemes et al. (2006) to run the same estimations. The k-NN technique does not offer fixed equations but rather looks for similarity in the specified input properties in the development database and returns the estimate as a weighted average of the output values of only a small specified number of similar samples. This technique, therefore, offers a ‘local’ solution for every query, minimizing the influence of any dissimilar samples, even if those are present in abundance.

Beyond traditionally used basic measures (RMSE, ME,  $R^2$ ) to evaluate PTFs, we correlated estimation errors with the values of input properties to reveal specific bias towards the input properties if any.

**Table 1 Statistical properties of basic data sets used to test the Rawls et al. (1982) pedotransfer function**

	Rawls et al. (1982) (N=2541)					Rawls et al. (recovered, N=2528)					NRCS NSSC subset 'B' (N=4697)				
	min	max	mean	std	med	min	max	mean	std	med	min	max	mean	std	med
Sand [%]	0.10	99.00	56.00	N/D	N/D	0.20	99.00	55.21	31.19	58.65	0.30	97.40	31.05	23.54	26.00
Silt [%]	0.10	93.00	26.00	N/D	N/D	0.00	91.00	26.76	22.92	21.20	0.40	89.50	41.21	18.70	40.70
Clay [%]	0.10	94.00	18.00	N/D	N/D	0.00	93.40	18.03	15.36	14.10	0.20	88.80	27.74	15.40	25.70
BD [g/cm <sup>3</sup> ]	0.10	2.09	1.42	N/D	N/D	0.01	2.02	1.43	0.23	1.47	0.56	2.05	1.43	0.19	1.43
OM [%]	0.10	12.50	0.66	N/D	N/D	0.00	14.99	1.10	1.91	0.34	0.00	12.43	1.24	1.58	0.67
θ <sub>33</sub> [v/v]	N/D	N/D	N/D	N/D	N/D	0.01	1.38	0.24	0.14	0.24	0.05	0.72	0.33	0.09	0.33
θ <sub>1500</sub> [v/v]	N/D	N/D	N/D	N/D	N/D	0.00	0.56	0.12	0.09	0.10	0.01	0.43	0.18	0.08	0.17

## 5. Results

While the exact same regression coefficients were not returned using the 2528 size data set – most likely due to the slight difference in the original and the regenerated data sets – the coefficients were reasonably close to those reported by Rawls et al. (1982). However, the only way we could approach all of the regression coefficients reasonably closely was that we used the existing data in the Rawls data set as if it was organic carbon (OC) content, rather than OM content. This contradicts what was reported to be used by Rawls et al. (1982), and appears to be the reason for the mismatch in OM contents between the original 2541 size set reported by Rawls et al. and the 2528 size set that we identified (c.f. Table 1). We therefore found it necessary to further examine the base data. In any further work, all OM/OC data was standardized to reflect OM content.

We managed to retrieve 23 of the 26 sources of the original data collection which were consulted for details on methodology, data availability and reported data format. We obtained indication of a number of methodological differences as well as data conversion problems. It has been identified that while many of the sources used undisturbed samples to determine θ<sub>33</sub> and θ<sub>1500</sub>, others did not. Also, while in most cases reported gravimetric water contents were correctly converted while entered into the data collection, in some cases this step was not made, yielding possible bias in the data collection. Similar conversion problems surfaced in relation to converting OC and OM content to one another, which caused the presence of a mixture of data in the data collection. In case of several data sources, however, OM/OC content =0 was indicated in the data collection for all samples obtained from the particular source – in total for over 600 samples of various horizon notations - whereas the source did not list OM/OC values at all; adding further bias/skew to the data.

After adjusting the data to reflect volumetric water content and OM content as standard for all samples, and omitting samples with no OM/OC data or no information on reported format and/or methodology, an N=1615 size subset of the originally N=2528 sized data set was identified. This subset was then used to re-develop and re-test the functional relationships between variables using only the same variables and regression technique as in the original work of Rawls et al. (1982) - as well as using the k-NN technique. The original Rawls et al. (1982) equations, the redeveloped regression equations ((a) from N=2528, adjusted for OM content; and (b) from the N=1615 subset) and estimations made using the k-NN technique ((a) from the N=1615 subset and (b) from the NSSC ‘A’ subset) were examined in terms of root-mean-squared residuals (RMSE), mean residuals (ME) (Table 2) and in terms of correlations between estimation errors and the values of input properties (Table 3). In both groups of aspects, accuracy and bias improved substantially after data were corrected. Estimations from the N=1615 size data set were near as good as those from a statistically identical (but independent) data NSSC set when the advanced k-NN technique was used for both.

**Table 2 Estimation accuracy of FC, WP and AWHC (as FC minus WP) using the various listed development data sets and the NSSC subset 'B' as test data set**

	Rawls et al. (1982) (as published) linear regression			Rawls N=2528 subset (OM/OC corrected) linear regression			Rawls N=1615 subset (all data corrected) linear regression			Rawls N=1615 subset (all data corrected) k-Nearest Neighbor			NSSC subset 'A' (N=4698) k-Nearest Neighbor		
	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC
	RMSE	0.072	0.043	0.055	0.065	0.039	0.057	0.073	0.039	0.058	0.058	0.036	0.053	0.053	0.032
ME	-0.006	-0.005	-0.001	-0.008	0.010	-0.017	-0.011	<0.001	-0.011	-0.001	<0.001	-0.001	0.002	0.001	0.001

**Table 3 Correlation coefficients (R<sup>2</sup>) between estimation errors and input variables applying various levels of data correction and 2 different estimation techniques**

	Rawls et al. (1982) (as published) linear regression			Rawls N=2528 subset (OM/OC corrected) linear regression			Rawls N=1615 subset (all data corrected) linear regression			Rawls N=1615 subset (all data corrected) k-Nearest Neighbor			NSSC subset 'A' (N=4698) k-Nearest Neighbor		
	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC
	Sand	0.067	0.004	0.085	0.068	0.002	0.079	0.116	0.020	0.097	0.027	0.021	0.006	<0.001	<0.001
Silt	0.009	0.003	0.029	0.003	0.012	0.022	0.005	0.022	0.037	0.001	0.005	0.008	<0.001	<0.001	0.002
Clay	0.079	0.028	0.056	0.108	0.038	0.062	0.188	0.155	0.058	0.086	0.018	0.051	<0.001	0.002	0.002
OM	0.319	0.319	0.089	0.111	0.132	0.018	0.083	0.038	0.043	0.001	0.003	0.005	<0.001	<0.001	<0.001

## 6. Discussion and Conclusions

Parts of the original Rawls et al. (1982) data collection were errant in representing soils, due to conversion errors, errors in reported units and misrepresentation of originally missing values. Also the data collection is skewed towards coarse textured soils – a fact that was already known to the developers and many users. Estimations using the developed PTFs therefore showed bias towards some of the input variables (e.g. OM, c.f Table 3, panel A), which could mostly be corrected after the data were corrected (panels B-C). In addition, using a more advanced estimation method further reduced dependence of estimation errors on the actual value of inputs (panel D) – using the same data – showing the inferiority of the originally applied technique – notwithstanding its simplicity. General measures of estimation accuracy also improved after using 'clean' data and an advanced technique. Those indicators came much closer to what the baseline indicators are using the randomly split other half of the NSSC data (reflecting mostly unavoidable noise in this type of data).

It is expectable that the donor database (NSSC) of the test data has certain weaknesses also (e.g. skew?). However, the Rawls data and the NSSC data should not show very significant bias compared to each other, since both were considered representative of US conditions in the past. Input specific bias between them could mostly be avoided by correcting reported data and by using a more advanced estimation technique, despite of known methodological differences in how those data were obtained for the two data collections. A significant part of the remaining correlation of estimation errors and clay content is most likely due to the different origin of the bulk of the two databases (i.e. US East Coast vs. Mid-West) and perhaps will be explained after examining the effect of clay mineralogy as the next step. This is also an early indication that for an optimal solution we may need to include an additional input variable along with using a new technique and corrected data.

Functional testing of the changed estimations is underway, and a final PTF solution for use in the CEAP modelling approach is being investigated. It will most likely involve new data, a new development technique and potentially a new input variable. One additional factor that is being investigated is the ability to account for seasonal changes of soil BD that will abruptly change following any invasive tillage operations.

## 7. References

- Nemes, A., Rawls, W.J., Pachepsky, Ya.A., 2006. Use of the non-parametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal* 70: 327-336.
- Pachepsky, Ya.A. and Rawls, W.J., 2004. [Eds.] *Development of pedotransfer functions in Soil Hydrology. Developments in Soil Science, Vol 30.* Elsevier, Amsterdam. ISBN 0-444-51705-7.
- Rawls, W.J., Brakensiek, D.L., Saxton, K.E., 1982. Estimation of soil water properties. *Trans. ASAE* 25(5): 1316-1320 & 1328.

Soil Survey Staff, 1997. National Characterization Data. Soil Survey Laboratory, National Soil Survey Center, and Natural Resources Conservation Service, Lincoln, NE.