Downscaling Satellite Soil Moisture using a Modular Spatial Inference Framework

Ricardo M. Llamas, Leobardo Valera, Paula Olaya, Michela Taufer and Rodrigo Vargas

Supplementary Material S1

Selection of most relevant terrain parameters used as predictors to estimate soil moisture at 1 km spatial resolution over the conterminous United States

Input topographic information was obtanied from a conus-wide 1-km digital elevation model (DEM) (Becker et al., 2009). Then, a set of 15 terrain parameters (Table S1.1) were generated using the terrain analysis module in RSAGA (Brenning et al., 2008), which is the implementation of SAGA GIS (Conrad et al., 2015) in R statistical platform (R Core Team, 2020).

Table S1.1. Terrain paramaters derived from a 1-km conus-wide digital elevation model.

Terrain Parameter	Minimun Value	Maximum Value	Mean Value
Analytical Hilldahsing	0.234	1.29	0.786
Aspect	0	6.283	3.037
Channel Network Base Level	-4.345	3013.547	672.99
Vertical Distance to Channel Network	0	3570.978	120.32
Flow Accumulation	$6.10x10^{-05}$	1176.699	32.835
Convergence Index	-100	100	-0.015
Elevation	-85.102	4148.538	521.893
Length-Slope Factor	88.051	0.599	1.227
Longitudinal Curvature	$-2.40x10^{-04}$	$2.35x10^{-04}$	$7.00x10^{-07}$
Cross Sectional Curvature	$-2.22x10^{-04}$	$2.14x10^{-04}$	$-7.00x10^{-07}$
Relative Slope Position	0	1	0.247
Slope	0.612	0.0288	0.0448
Topographic Wetness Index	6.7	0.2961	1.26
Catchment Area	1.00 x10+^{06}	3.06 x10+12	1.18 x10+^{09}
Valley Depth	2674.01	370.578	345.271

Different sets of training files were created based on the combination of different sets of the previously generated terrain parameters and latitude and longitude values for every pixel in the CONUS domain. The training files consisted of a matrix describing annual mean soil moisture values in 2010 derived from ESA-CCI soil moisture daily estimates at 0.25 degrees of spatial resolution, and coordinates values of each 1-km pixel along with the values of terrain parameters at the centroid of the ESA-CCI 0.25 degrees pixels. In total, 12,763 training points (ESA-CCI pixels centroids) were used as main input for the initial soil moisture predictions at 1-km spatial resolution. The different sets of training files that incorporated soil moisture values, different terrain parameters, and coordinates as predictors, are described in Table S1.2.

Similarly, a set of the combinations of terrain parameters and coordinates (Table S1.2) were generated based on all the 1-km pixels available in our conus domain (8,003,235 evaluation points), these files, so-called evaluation matrices (with no soil moisture values), were used to predict soil moisture at 1-km, based on models previously defined based on the training matrices.

Table S1.2. Combinations of parameters used as soil moisture predictors, derived from terrain parameters and coordinates.

	Latitude values	**Longitude Values**	**15 Terrain parameters**	**Aspect**	**Elevation**	**Slope**	**Topographic Wetness Index**	**Catchment Area**
1	▨	▨	▨					
2			▨					
3	▨	▨		▨	▨	▨		
4				▨	▨	▨		
5	▨	▨		▨		▨	▨	
6				▨		▨	▨	
7	▨	▨						
8	▨	▨		▨		▨		▨
9				▨		▨		▨
10	▨	▨		▨	▨	▨	▨	
11				▨	▨	▨	▨	

To define the best prediction parameters among the 15 terrain parameters generated and to avoid redundancy of information not directly related to soil moisture spatial distribution, we predicted soil moisture at 1-km over the conterminous United States (based on a 2010 ESA-CCI soil moisture mean annual layer), and a cross-validation analysis was performed. Different combinations of terrain parameters and the inclusion of coordinates as predictors were tested using kernel-weighted k-nearest neighbors (KKNN).

KKNN in its traditional form is a regression technique that builds many simple models from local data (Johnston et al., 2016). It is based on decision rules that classify a sampled point based on the values of the closest set of previously classified points or reference values in the sample space (Cover & Hart, 1967). This method assumes a different level of influence in the prediction space, where the k points closest to the target location have the most relevant influence, while the influence on the construction of the prediction model decreases with distance (Rorabaugh et al., 2019). To assign distance-related relevance for predicting soil moisture in this case, a weighted average of the closest soil moisture k ratios is calculated. This variant is based on the definition of kernel functions (that is, triangular, Epanechnikov, Gaussian, optimal) that serve to find the number of neighbors (k) that will be used in the prediction. The number of neighbors and the optimal kernel function are automatically selected through a 10-fold cross-validation, offering correlation and root mean square error (RMSE) values that help in evaluating the performance of each prediction (Guevara & Vargas, 2019; Rorabaugh et al., 2019).

Soil moisture predictions at 1-km over conus were obtained by ‘kknn’ package (Hechenbichler & Schliep, 2004) developed on the R-statistical platform (R Core Team, 2020). All prediction processes were performed using the University of Delaware High Performance Computing (HPC) cluster “Caviness” (UDIT Research CyberInfrastructure, 2021), a distributed-memory Linux cluster with 126 compute nodes (4,536 cores with 24.6 TiB of RAM and 200 TB of storage).

Cross validation results of soil moisture predictions using the different predefined combinations of terrain parameters and coordinates (predictors) are shown in Table S1.3 and Figure S1.1.

Table S1.3. KKNN cross-validation results derived from soil moisture predictions over conus at 1-km, based on different combination of terrain parameters and coordinates. Prediction combination number refers to predictors shown in Table S1.2; Number of points describes the number of samples in the training matrices use to construct the modes; the Kernel describes the shape found to be the one with the most effective prediction performance when defining location of k-neighbors; K number describes the optimal number of neighbors used in each sample location to construct the prediction model.

Predictors combination	**Correaltion**	**RMSE**	**Number of points**	**Kernel**	**K Number**
1	0.845	0.031	12,763	triangular	23
2	0.453	1.323	12,763	triangular	39
3	0.876	0.028	12,763	triangular	19
4	0.77	464.887	12,763	gaussian	49
5	0.846	0.031	12,763	triangular	25
6	0.795	0.027	12,763	epanechnikov	50
7	0.927	0.022	12,763	gaussian	10
8	0.859	0.03	12,763	triangular	22
9	0.259	0.043	12,763	rectangular	50
10	0.867	0.029	12,763	triangular	19
11	0.769	465.402	12,763	gaussian	47

Based on the analyses and results, geographic coordinates and 4 terrain parameters (elevation, aspect, slope and topographic wetness index) were selected as predictors for our study.

Although the highest correlation and lowest root mean square error (RMSE) between predicted and observed values in the cross-validation was obtained using only geographic coordinates as predictors, a visual analysis of the prediction output showed that rather than soil moisture predictions, the outputs show the replication of spatial patterns driven by location. The highest soil moisture values occur generally in the Northeast of CONUS, while lower values tend to occur in the Southwest. The second highest correlation and second lowest RMSE in the cross-validation analysis were obtained using geographic coordinates, elevation, aspect, and slope. However, to avoid overrepresentation of topography in the prediction outputs, we included the topographic wetness index as an additional predictor, as this index is based on topography but also describes the distribution of soil water content as is related to overland flow patterns. This last combination of prediction factors (geographic coordinates, elevation, aspect, slope and topographic wetness index) showed the third highest correlation and the third lowest RMSE, this parameters selection complies with the assumption of topography is an important factor affecting water distribution in soils since it directly affects overland flow and solar radiation rates (Hallema et al., 2016).

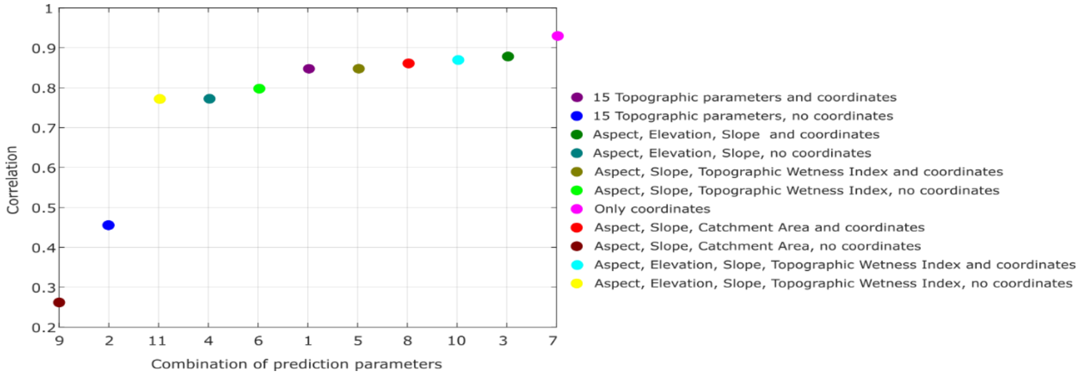

Figure S1.1. Correlation values derived from cross-validation analysis. Numbers in the X-axis refer to parameters combinations described in Table S1.3

References

Becker, J.J.; Sandwell, D.T.; Smith, W.H.F.; Braud, J.; Binder, B.; Depner, J.; Fabre, D.; Factor, J.; Ingalls, S.; Kim, S.-H.; et al. Global Bathymetry and Elevation Data at 30 Arc Seconds Resolution: SRTM30_PLUS. *Mar. Geod.* **2009**, *32*, 355–371. https://doi.org/10.1080/01490410903297766.

Brenning, A.; Bangs, D.; Becker, M. RSAGA: SAGA Geoprocessing and Terrain Analysis in R (1.3.0). 2008. Available online: https://github.com/r-spatial/RSAGA (accessed on 23 July 2021).

Conrad, O.; Bechtel, B.; Bock, M.; Dietrich, H.; Fischer, E.; Gerlitz, L.; Wehberg, J.; Wichmann, V.; Böhner, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **2015**, *8*, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.

Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Guevara, M.; Vargas, R. Downscaling satellite soil moisture using geomorphometry and machine learning. *PLoS ONE* **2019**, *14*, e0219639. https://doi.org/10.1371/journal.pone.0219639.

Hallema, D.W.; Moussa, R.; Sun, G.; Mcnulty, S.G. Surface storm flow prediction on hillslopes based on topography and hydrologic connectivity. *Ecol. Process.* **2016**, *5*, 13. https://doi.org/10.1186/s13717-016-0057-1.

Hechenbichler, K.; Schliep, K. *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*; Collaborative Research Center 386, Discussion Paper 399; Ludwig-Maximilians-Universität München: Munich, Germany, 2004. https://doi.org/10.5282/ubm/epub.1769.

Johnston, T.; Zanin, C.; Taufer, M. HYPPO: A Hybrid, Piecewise Polynomial Modeling Technique for Non-Smooth Surfaces. In Proceedings of the 2016 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), Los Angeles, CA, USA, 26–28 October 2016; pp. 26–33. https://doi.org/10.1109/SBAC-PAD.2016.12.

R Core Team. *R: A Language and Environment for Statistical Computing (4.0.3)*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.r-project.org/ (accessed on 27 August 2021).

Rorabaugh, D.; Guevara, M.; Llamas, R.; Kitson, J.; Vargas, R.; Taufer, M. SOMOSPIE: A Modular SOil MOisture SPatial Inference Engine Based on Data-Driven Decisions. In Proceedings of the 2019 15th International Conference on eScience (eScience), IEEE, San Diego, CA, USA, 24–27 September 2019; pp. 1–10. https://doi.org/10.1109/eScience.2019.00008.

UDIT Research CyberInfrastructure CAVINESS, Supporting Researchers at University Of Delaware. Available online: https://sites.udel.edu/it-rci/compute/community-cluster-program/caviness/ (accessed on 23 August 2021).