A FRAMEWORK FOR ESTIMATING REPRESENTATIVE AREA OF A GROUND SAMPLE USING REMOTE SENSING

P. J. Deshpande[1,*], A. Sure[1], O. Dikshit[1], S. Tripathi[1]
[1] Department of Civil Engineering, Indian Institute of Technology Kanpur, India
(prasadj, anudeep, onkar, shiva)@iitk.ac.in

KEY WORDS: Hydro-meteorological variable, Regionalisation, Image fusion, Heterogeneity, Remote sensing, Google Earth Engine

ABSTRACT:

Modelling hydro-meteorological variables over land and atmosphere comprise of ground sampling at selected locations and predicting over the other locations. Remote sensing data can be effectively used to improve predictions by prudently choosing sampling locations of variables co-dependent on the prediction variable. This paper presents a framework for estimating the representative area of a ground sample and thereby determining the number of samples required for prediction with a given level of uncertainty and spatial resolution. Application of the proposed framework for soil moisture as the prediction variable is presented using Google Earth Engine and Scikit-learn libraries implemented in Python 3 programming language.

1. INTRODUCTION

Modelling of hydro-meteorological variables over land and atmosphere comprises sampling at selected locations followed by prediction at other unmeasured locations in study area. Since the hydro-meteorological variables are spatially and temporally varying, precise and frequent measurement is not feasible at every location. The auxiliary variables, available at the unsampled locations through remote sensing are used for making the predictions better. Although ground-based sampling is expensive as compared to remote sensing, it is more accurate. On the other hand, remote sensing covers a large spatial extent and hence, unlike ground-based sampling it can be useful to predict the required hydro-meteorological variable throughout the study area.

The ground-based samples are characterized by their accuracy and spatial extent represented by them. Heterogeneity present over the land-atmosphere interface governs the area represented by each sample. Each ground sample delivers information for an area in which the prediction variable has the same value. The spatial extent of this homogeneous area is termed as 'representative area' for the sample. In addition to spatial heterogeneity, each sample location has an inherent variability because of its stochastic nature.

For the given study area and conditions, deciding the number of ground samples is challenging. The prior knowledge in the form of physical properties like land use and land cover, elevation, soil type is necessary in this scenario. The area represented by a ground sample depends upon the surrounding heterogeneity at the sampling location. Lesser the heterogeneity, smaller is the number of samples required for that area and vice-versa. Since the prediction variable is not measured at every location, there is no direct knowledge of its heterogeneity. The use of the auxiliary or predictor variables is made to check the heterogeneity of the prediction variable. These predictor variables are remotely sensed variables which are co-dependent on the prediction variable and hence they can be used for the estimation of the prediction variable. The prediction variable considered in this work is soil moisture and various remotely sensed predictor variables are taken as input datasets. The fusion of predictor variables and the ground sample is employed to obtain information about the heterogeneity of the prediction variable in the study area.

Traditionally remote sensing has been used to estimate soil moisture at varying spatial and temporal resolution (Woodhouse, 2006). Passive and active microwave remote sensing has been used effectively using brightness temperature and backscattering coefficients (Petropoulos, 2017). The principle and wavelength in deriving the surface soil moisture content differ in the passive and active domain (Ulaby et al., 1986). Various algorithms are derived based on radiative transfer theory, water balance model, land-surface models, integral equation models, polarimetry based (Chandrasekhar, 1960; Dubois et al., 1995; Schlenz et al., 2008; Lu et al., 2012). Remote sensing has delivered an efficient way to estimate hydro-meteorological variables with the least workforce.

For making predictions, based on the heterogeneity present in the study area, a suitable resolution (grid cell size) can be decided. The ground samples collected act as the training data for the representative area of the sample. Therefore, the grid cell size can be chosen depending upon the representative area of the ground sample. The smallest representative area of the ground samples is chosen as the grid cell for prediction.

The objective of this work is to derive the heterogeneity map of soil moisture, for a Critical Zone Observatory (CZO), in the Ganga basin, India. For this paper, the modelling is carried out by temporal averaging the predictor variables, i.e. an average is taken by considering the complete time period to eliminate the temporal variations in the heterogeneity. This objective also includes the determination of the improved sampling locations and their representative area. Depending upon the determined grid cell size, a decision to acquire more predictor datasets with a finer spatial resolution is made.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-687-2019 | © Authors 2019. CC BY 4.0 License.

2. STUDY AREA AND DATASET DESCRIPTION

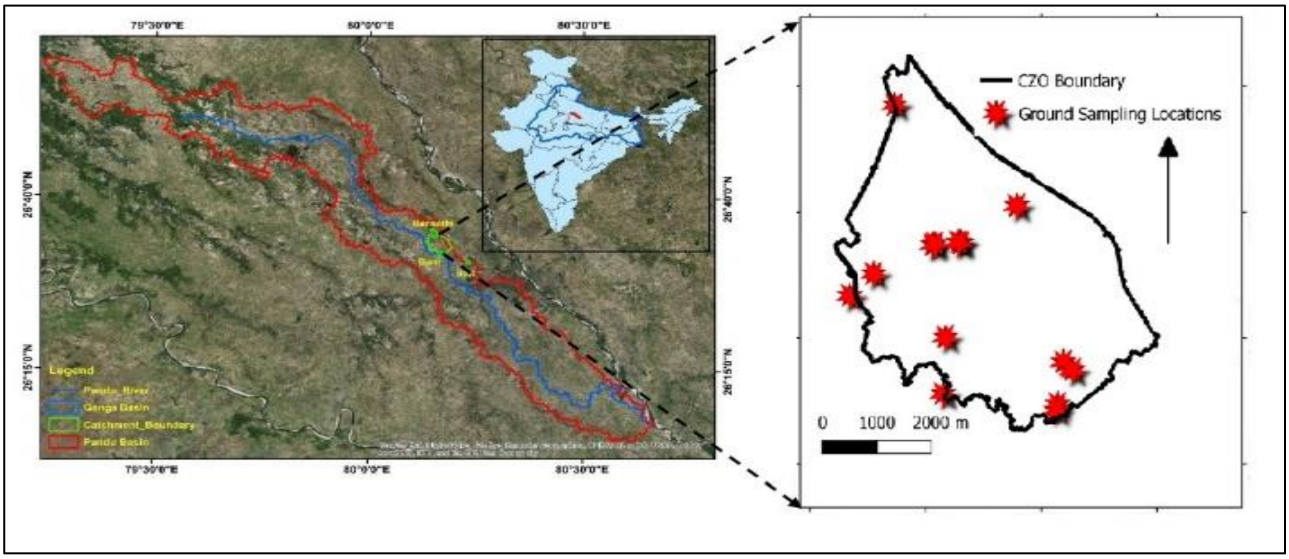

Figure 1. The study area and the locations of stations in soil moisture sensor network (Gupta et al., 2018)

2.1 STUDY AREA

The proposed framework was applied on the bounding box enclosing the study area which is known as HEART (Heterogeneous Ecosystem of an Agro Rural Terrain) - CZO. The CZO is a small watershed (21 km^2; 80°8'0" E - 80°11'0" E and 26°31'43.93" N 26°36'14.85" N) of the Pandu river basin, a tributary of the River Ganges, India. This study area is selected, because it has a network of 15 in-situ soil moisture measurement is sites (Gupta et al., 2019). The in-situ data is available from Aug 2017 to Oct 2018 i.e. for 14 months, and hence the remote sensing input data are also averaged over this time period, which enables a comparison with the ground measurements.

2.2 Datasets Used

The literature suggests that soil moisture being an integral element in the hydrological cycle, is primarily dependent on meteorological parameters (temperature, precipitation, evapotranspiration), crop characteristics, soil properties, land use and land cover, elevation. These interdependent variables define the soil moisture profile spatially and temporally. Thus, the remote sensing datasets chosen for this study are direct or indirect indicators of soil moisture variable (Huffman et al., 2001; Huffman et al., 2007; Entekhabi et al., 2010; Mu et al., 2013; Entekhabi et al., 2014). Table 1 lists the remote sensed products used in the study and Figures 2.1 to 2.4 show their spatial variations.

Sr	Product	Spatial Resolution	Temporal Resolution	Min Value in study area	Max Value in study area	Unit
1	SRTM DEM	30 m	Static	120	143	m
2	LAI MODIS	500 m	4 day	0.0000	1.3118	m² / m²
3	NDVI MODIS	250 m	16 day	0.2527	0.5515	
4	Evapotrans piration	1000 m	8 day	0.0000	16.0167	kg/m²
5	Soil Moisture SMAP	0.25 arc degrees	3 day	11.3784	11.3784	%
6	Rainfal TRMM	0.25 arc degrees	1 Month	0.1040	0.1040	mm/hr

Table 1. List of input datasets

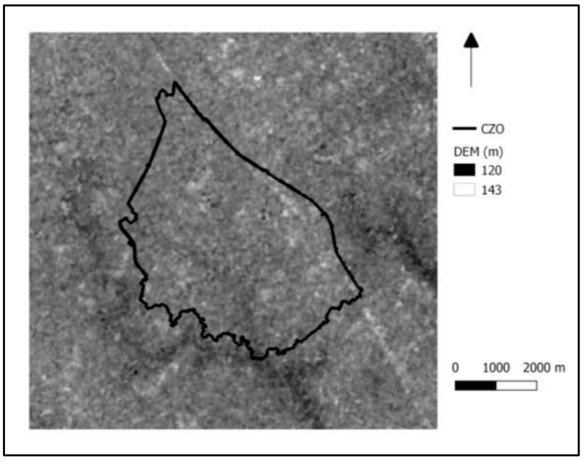

Figure 2.1. Digital Elevation Model (DEM)

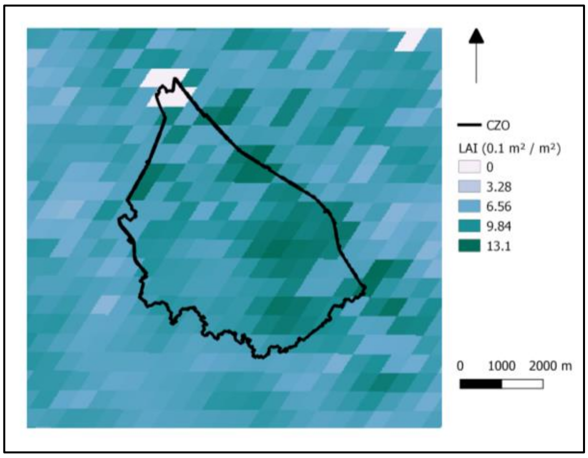

Figure 2.2. Leaf Area Index (LAI)

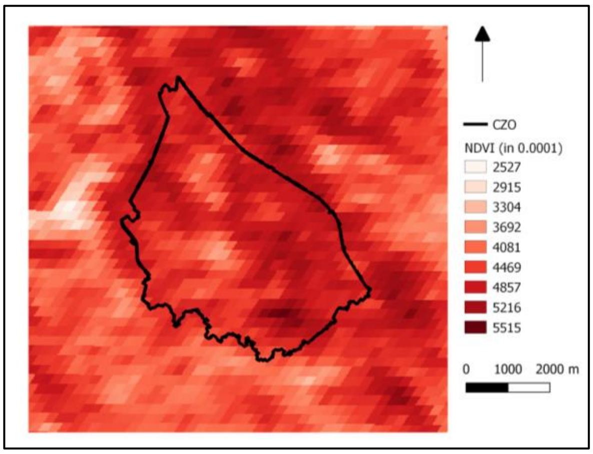

Figure 2.3. Normalised Differenced Vegetation Index (NDVI)

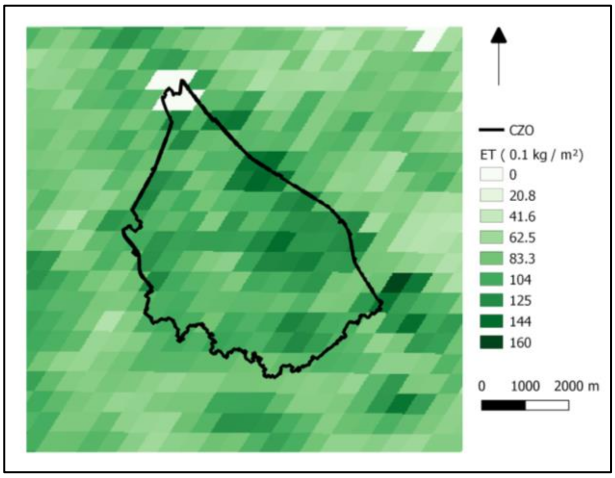

Figure 2.4. Evapotranspiration (ET)

3. METHODOLOGY

The methodology for the proposed framework is implemented in Python 3 programming language with the Google Earth Engine (GEE) and Scikit-learn libraries (Pedregosa et al., 2011; Gorelick et al., 2017).

3.1 Pre-processing of the Datasets

All the remote sensing input datasets were extracted for the study area. Since the input data are collected from different satellites having different spatial and temporal resolutions; all the datasets were brought to the common platform by temporal averaging. Also, since the spatial resolution of input datasets varied from 30 m (SRTM DEM) to approx. 25 km (TRMM Rainfall), the datasets were resampled to the finest resolution available, i.e. 30 m. Thus, 30 m was the minimum cell size considered for prediction. Nearest neighbour technique was employed for resampling. Pre-processing resulted in a temporally lumped input datasets having the same spatial resolution of 30 m.

3.2 Discretisation of Input Datasets

In this step, discretisation or binning of the continuous input datasets with a given level of information loss was carried out. Entropy which a measure of the information content was monitored while discretising the input dataset (Shannon, 1948; Meurer 2015). As the number of bins reduces, the information loss increases, but the computational resources needed for further processing reduces. Hence, to decide the number of bins an iterative procedure was applied for each input dataset. Starting from 2 bins, the number of bins were increased while the entropy of binned inputs was compared with the entropy of the original data. The process was terminated when the entropy gain because of the addition of a new bin was within the pre-specified threshold limits.

The given limit for entropy loss is user determined and depends upon the level of accuracy required for the analysis. For this work, the analysis was carried out for different percentage of allowable entropy loss threshold for each input dataset. This step resulted in the discretised input datasets with the finite number of bins corresponding to every input dataset. The discretisation or binning was carried out using Scikit-learn library in Python programming language.

3.3 Discretised Input Combinations

The discretised datasets act as the inputs for determining the prediction variable. For a given set of discretised input values, the expected value of the prediction variable should be the same irrespective of the location of the grid cell under consideration. Hence, in this step, all possible combinations of the discretised input values were obtained. There is a limit for the possible number of combinations since the inputs are discretized in a finite number of bins. This step resulted into a dictionary in which each unique combination of discretised input data values was given a unique identification (id).

3.4 Heterogeneity Map

Regionalisation of the homogeneous grid cells is carried in this step. Once all the possible combinations of input datasets were given the unique id's, the input datasets were processed cell by cell. Depending upon the values present in the grid cell under consideration, the equivalent unique identifier was assigned to the corresponding output grid cell. Once, all the grid cells were processed, an image containing the output values was formed. Regions with the same identities (having the same cell value representing the prediction variable) were created based on the positional adjacency of the cells with the same identity or the output value. The patches in which the adjacent grid cells had similar identities (unique values) represent homogeneous regions. The smallest single cluster thus formed represents the suitable grid cell size for prediction.

4. RESULTS AND DISCUSSION

Different sets of results (binned input datasets, and corresponding heterogeneity maps) were obtained as the algorithm was applied for different levels of information loss. The results thus obtained are presented and discussed in this section.

4.1 Discretised Input Datasets

Table 2 shows the results for different values of thresholds and the corresponding discretised inputs and their combinations.

Case	Threshold for entropy loss	No. of bins						Possible Combinations	Actual Combinations
		DEM	LAI	NDVI	Rainfall	ET	Soil Moisture		
1	10	2	2	2	1	2	1	16	16
2	1	2	2	2	1	2	1	16	16
3	0.1	2	3	2	1	3	1	36	26
4	0.01	2	3	3	1	3	1	54	40
5	0.001	2	3	3	1	3	1	54	40
6	0.0001	3	3	3	1	3	1	81	60

Table 2. Results for different thresholds on entropy loss

It can be seen from Table 2 that, the possible number of combinations depends upon the choice of threshold, which ultimately affects the heterogeneity map. Here the threshold on entropy is a crucial parameter to selected. The threshold is related to entropy loss, i.e. for lower threshold the resulting map will be more heterogeneous (showing more variation) and vice versa. Hence, the choice of threshold depends upon the required accuracy for the modelling as well as the variability available in the input datasets. For example, Figs. 2.1 & 4 of original and discretised DEM to show the effect of discretisation.

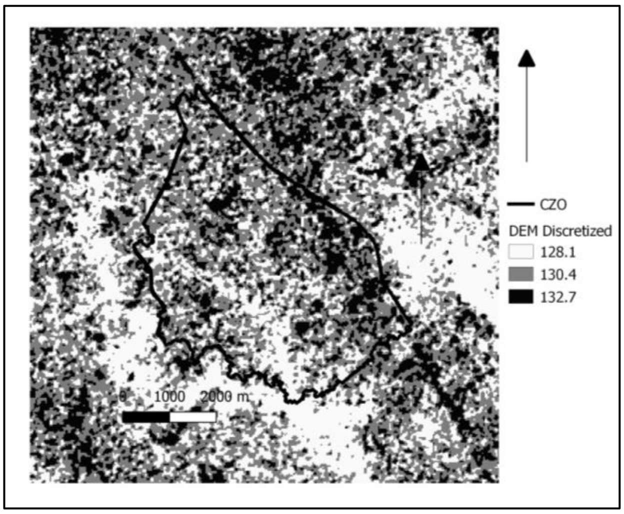

Figure 4. Discretized DEM for threshold = 0.0001

Figures 5.1 to 5.3 present the heterogeneity map of soil moisture derived for different levels of entropy threshold. As expected, the variability of the heterogeneity map increases with the reduction in the threshold value.

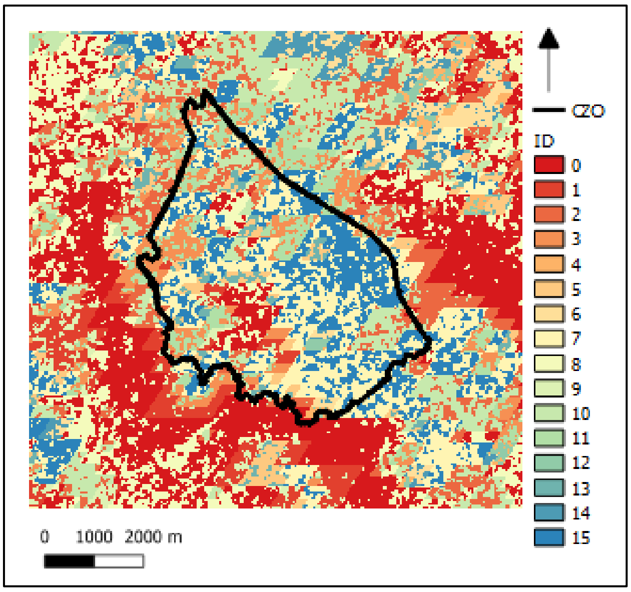

Figure 5.1. Threshold Value = 10

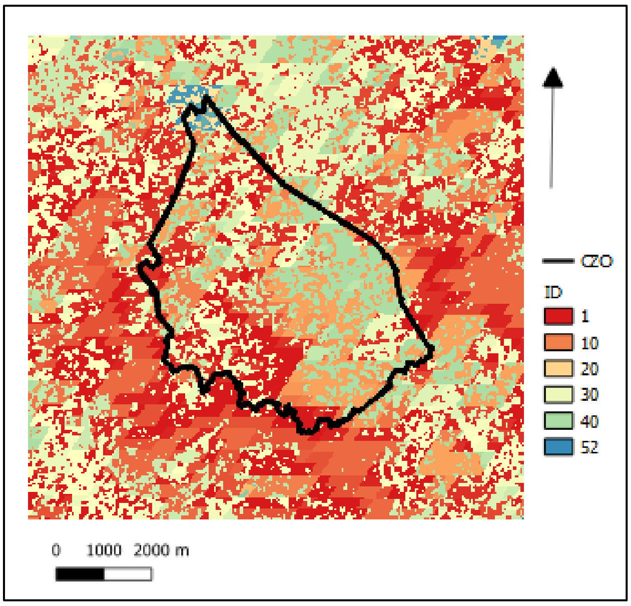

Figure 5.2. Threshold Value = 0.01

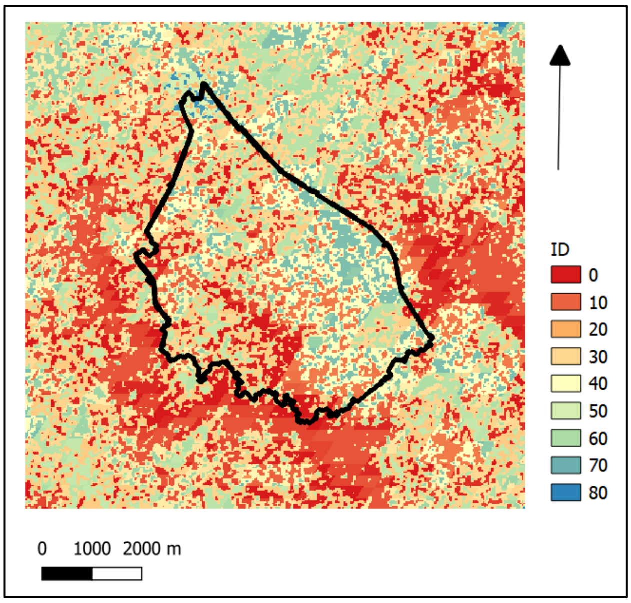

Figure 5.3. Threshold Value = 0.0001

4.2 Summary and Conclusion

The paper proposes a framework for estimating hydro-meteorological variables at unmeasured locations based on limited ground samples and remote sensed data of auxiliary variables. Although the research scope is broad and can be applied for any hydro-meteorological variable, we have tested it for soil moisture estimation. This method is flexible in terms of the spatiotemporal resolution of the input datasets. If some of the input datasets have coarse resolution and/or having less variability over the study area, then the algorithm automatically marks them as 'less significant' in the heterogeneity map. The variables having finer resolution and/ or having more variability contributes more to the heterogeneity map.

The proposed framework provides an overview of prevailing heterogeneity among the predictor variables present throughout the study area. The heterogeneity maps derived for the study area provides an alternate method to select sampling ground locations. Sampling in the same homogeneous region can be avoided to minimize redundancy. Hence, cost-effectiveness in ground sampling can be achieved. Further a representative area of each sample can be determined if the location of the sample is already known. A grid can be formed having cell size equal to smallest homogeneous patch and henceforth it can be used for prediction applications. Here, this smallest homogeneous patch may be larger than the finest resolution dataset available amongst the input datasets. This means that in spite of the availability of finer resolution data, the prediction can be efficiently carried out to a larger scale to achieve computational efficiency.

4.3 Future Scope

The proposed framework does not consider the temporal dynamics of the homogeneous patches. Since the input variables are dynamic and are further dependent on other hydro-meteorological variables, the heterogeneity map varies with time.

Heterogeneity map depends directly upon the discretisation or binning of the continuous input variables. Binning can be done according to the natural breaks, equal intervals, minimum message length or any method of choice (Liu et al., 2002). Each of the discretisation methodologies has its advantages and disadvantages. It is important to note that the discretisation method should try to keep the number of bins finite.

Heterogeneity map obtained from the above framework can contain a large number of small patches of homogeneous regions. From Fig 5.3, (threshold value = 0.0001)) it is evident that the spatial aggregation of various small patches having heterogeneous areas can result into large homogeneous regions (Marceau et al., 1994). Thus, the representative area of each sample, as well as the grid size for prediction can be increased. This step can be carried out if it is permissible to lose some level of information during the process of spatial aggregation. Discretisation is a tuning parameter for heterogeneity map behaviour in the radiometric domain while spatial aggregation is a tuning parameter in the spatial domain.

The proposed framework can be validated by comparing the in-situ measurement data with the heterogeneity map produced by the remote sensing data. The inter-patch (heterogeneous region) and intra-patch (homogeneous region) can be validated is sufficient in-situ samples are available for the study region.

ACKNOWLEDGEMENTS

Authors would like to express sincere thanks to Ministry of Human Resource Development (MHRD), Department of Science and Technology (DST), Government of India and Critical Zone Observatory for providing necessary datasets and information and funds.

REFERENCES

Chandrasekhar, S., 1960. *Radiative Transfer*, 1st ed. Dover Publications, INC, New York.

Dubois, P.C., Van Zyl, J., Engman, T., 1995. Measuring soil moisture with imaging radars. *Geoscience and Remote Sensing, IEEE Transactions* on 33, 915–926.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., Kimball, J., Piepmeier, J.R., Koster, R.D., Martin, N., McDonald, K.C., Moghaddam, M., Moran, S., Reichle, R., Shi, J.C., Spencer, M.W., Thurman, S.W., Tsang, L., Van Zyl, J., 2010. The Soil Moisture Active Passive (SMAP) Mission. *Proceedings of the IEEE* 98, 704–716. doi.org/10.1109/JPROC.2010.2043918

Entekhabi, D., Yueh, S., O'Neill, P.E., Kellogg, K.H., 2014. SMAP Handbook. (2 April 2019)

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202, 18–27. doi.org/10.1016/j.rse.2017.06.031

Gupta, S., Karumanchi, S., Dash, S., Adla, S., Tripathi, S., Sinha, R., Paul, D., Sen, I., 2019. Monitoring Ecosystem

Health in India's Food Basket. *Eos* 100. doi.org/10.1029/2019EO117683

Gupta, S., Tripathi, S., Sinha, R., Karumanchi, S.H., Paul, D., Tripathi, S.N., Sen, I.S., Dash, S., 2018. Setting Up a New CZO in the Ganga Basin: Instrumentation, Stakeholder Engagement and Preliminary Observations. *American Geophysical Union Fall Meeting 2017*, AGU, New Orleans, USA.

Huffman, G.J., Adler, R.F., Morrissey, M.M., Bolvin, D.T., Curtis, S., Joyce, R., McGavock, B., Susskind, J., 2001. Global Precipitation at One-Degree Daily Resolution from Multisatellite Observations. *Journal of Hydrometeorology* 2, 36–50. doi.org/10.1175/1525-7541(2001)002<0036: GPAODD>2.0.CO;2

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Wolff, D.B., Adler, R.F., Gu, G., Hong, Y., Bowman, K.P., Stocker, E.F., 2007. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *Journal of Hydrometeorology* 8, 38–55. doi.org/10.1175/JHM560.1

Liu, H., Hussain, F., Tan, C.L., Dash, M., 2002. Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery 6*, 393–423.

Lu, H., Koike, T., Yang, K., Hu, Z., Xu, X., Rasmy, M., Kuria, D., Tamagawa, K., 2012. Improving land surface soil moisture and energy flux simulations over the Tibetan plateau by the assimilation of the microwave remote sensing data and the GCM output into a land surface model. *International Journal of Applied Earth Observation and Geoinformation* 17, 43–54. doi.org/10.1016/j.jag.2011.09.006

Marceau, D.J., Howarth, P.J., Gratton, D.J., 1994. Remote sensing and the measurement of geographical entities in a forested environment. 1. The scale and spatial aggregation problem. *Remote Sensing of Environment* 49, 93–104. doi.org/10.1016/0034-4257(94)90046-9

Meurer, K., A Simple Guide to Binning Data Using an Entropy Measure, 2015. kevinmeurer.com/a-simple-guide-to-entropy-based-discretization (2 April 2019).

Mu, Q., Zhao, M., Running, S.W., 2013. MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3) Algorithm Theoretical Basis Document Collection 5 (2 April 2019).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

Petropoulos, G.P., 2017. *Remote Sensing of Energy Fluxes and Soil Moisture Conten*t, 1st ed. CRC Press.

Schlenz, F., Loew, A., Mauser, W., 2008. Soil Moisture Retrieval from Passive Microwave Data: A Sensitivity Study Using a Coupled Svat-Radiative Transfer Model at the Upper Danube Anchor Site, in: *IGARSS 2008. IEEE International*. pp. II–680.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 379–423, 623–656.

Ulaby, F.T., Moore, R.K., Fung, A.K., 1986. *Microwave remote sensing: Active and passive, From Theory to Applications, Microwave Remote Sensing #3*. Addison-Wesley Publishing Company, United States.

Woodhouse, I.H., 2006. *Introduction to Microwave Remote Sensing, 1st ed*. CRC Press.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-687-2019 | © Authors 2019. CC BY 4.0 License.