Materials, Energy, Water and Environmental Sciences   Research Articles [MEWES]

2015-08-10

Qualitative soil moisture assessment in semi-arid Africa – the role of experience and training on inter-rater reliability

Rinderer, Michael

Hydrology and
Earth System
Sciences

Open Access

Qualitative soil moisture assessment in semi-arid Africa – the role of experience and training on inter-rater reliability

M. Rinderer[1,2], H. C. Komakech[3], D. Müller[1], G. L. B. Wiesenberg[1], and J. Seibert[1,4]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[2]Nicholas School of the Environment, Duke University, 9 Circuit Drive, Durham, NC, 27708, USA
[3]Nelson Mandela African Institution of Science and Technology, P.O. Box 447 Arusha, Tanzania
[4]Department of Earth Sciences, Uppsala University, 752 36 Uppsala, Sweden

Correspondence to: M. Rinderer (michael.rinderer@duke.edu)

Abstract. Soil and water management is particularly relevant in semi-arid regions to enhance agricultural productivity. During periods of water scarcity, soil moisture differences are important indicators of the soil water deficit and are traditionally used for allocating water resources among farmers of a village community. Here we present a simple, inexpensive soil wetness classification scheme based on qualitative indicators which one can see or touch on the soil surface. It incorporates the local farmers' knowledge on the best soil moisture conditions for seeding and brick making in the semi-arid environment of the study site near Arusha, Tanzania. The scheme was tested twice in 2014 with farmers, students and experts (April: 40 persons, June: 25 persons) for inter-rater reliability, bias of individuals and functional relation between qualitative and quantitative soil moisture values. During the test in April farmers assigned the same wetness class in 46 % of all cases, while students and experts agreed on about 60 % of all cases. Students who had been trained in how to apply the method gained higher inter-rater reliability than their colleagues with only a basic introduction. When repeating the test in June, participants were given improved instructions, organized in small subgroups, which resulted in a higher inter-rater reliability among farmers. In 66 % of all classifications, farmers assigned the same wetness class and the spread of class assignments was smaller. This study demonstrates that a wetness classification scheme based on qualitative indicators is a robust tool and can be applied successfully regardless of experience in crop growing and education level when an in-depth introduction and training is provided. The use of a simple and clear layout of the assessment form is important for reliable wetness class assignments.

1 Introduction

For rainfed agriculture in semi-arid regions the soil water storage is of key importance for crop survival as it serves as the only water source during dry spells. The soil water storage is also important if water is available for irrigation. Based on differences in soil water deficits, scarce irrigation water resources can be allocated among farmers of a community in a fair manner. For farming activities like choosing the right moment to seed and for the development of crops, the moisture content in the unsaturated, shallow soil layers is of most importance.

Common techniques for measuring soil moisture are often time-consuming and/or rely on expensive equipment (e.g., time domain reflectometry, TDR) that needs electricity, maintenance and repair. Such instruments are also usually not available to farming communities in developing countries. Therefore, local irrigators in semi-arid Africa often visually assess the shallow soil wetness conditions to decide on which plots should be allocated irrigation turns. Despite their long experience in farming, for which these leaders are respected by the community members, their assessment might be disputed. A more systematic way of soil wetness assessment based on defined criteria would relieve pressure on community leaders and assure transparency in decision making and therefore avoid conflicts among farmers.

Qualitative methods have been shown to be useful complements to quantitative measurement techniques in a number of field applications in soil science (Thien, 1979), risk assessment (De Quervain, 1950; cited in Pielmeier and Schneebeli, 2003) and ecology (Metcalfe-Smith, 1994). They are based on qualitative indicators that one can identify through sight, sound or touch and that are related to quantitative properties of interest like the grain size distribution of a soil sample or the strength of a snow pack.

In hydrology qualitative indicators have been used for mapping saturated areas in some experimental studies. Dunne and Black (1970) and Dunne et al. (1975) were the first to map saturated areas with the "squishy boot" method, i.e., by walking through the catchment and mapping areas with water ponding on the soil surface. Others used this method to visually identify saturated areas (McDonnell and Taylor, 1987; Ambroise et al., 1996; Inamdar and Mitchell, 2007; Latron and Gallart, 2007; SNIFFER, 2009). Soil hydromorphic features that are visual when digging a soil profile can be useful indicators of intermittent soil saturation (Rinderer and Seibert, 2012). Local vegetation and individual plant species can also be indicators of prevailing soil moisture conditions (Ellenberg et al., 1991; Quinn et al., 1998; Kulasova et al., 2014).

The methods mentioned above do not allow for different grades of soil wetness or changes in soil wetness to be captured over time. The "spade diagnosis" method, which was originally developed in the 1930s for an applied soil texture examination in the field, is one of the earliest schemes with five qualitative wetness classes (Görbing and Sekera, 1947). The Natural Resources Conservation Service of the United States Department of Agriculture (1998) published guidelines for estimating soil moisture by feel and appearance for four different soil types and different soil moisture content. Blazkova et al. (2002) defined a qualitative classification scheme based on five wetness classes and used it for mapping moisture differences along transects and in a drainage ditch (for an application see also Kulasova et al., 2014). In their study, they did not utilize the full range of the five wetness classes, but aggregated the three wettest ones as they were interested in saturated areas. All these methods were not systematically tested in terms of correspondence between the qualitative indicators and the quantitative differences in soil water content and in terms of the reliability of the methods when applied by different people.

Rinderer et al. (2012) presented a soil wetness classification scheme based on characteristic, qualitative indicators for each wetness class to make class assignments more distinct. The indicators are based on the judgment of raters and include information such as whether their trousers would stay dry or get moist or wet when sitting on the ground, whether a squelchy noise could be heard, or whether water would squeeze out of the topsoil when stepping on the ground, or water could be seen ponding on the soil surface. The so-called "boots and trousers" method was tested in humid environmental conditions in terms of inter-rater reliability, influence of subjectivity and the relation between qualitative wetness classes and volumetric water content measured by the gravimetric and the TDR method. The definitions of the three wettest classes was subsequently applied by Ali et al. (2014) to map superficial water saturation in two nested catchments in Scotland.

Despite testing the robustness of the boots and trousers method it is still not clear if this qualitative wetness classification scheme is also applicable in drier environmental conditions with different soil types. It is also unclear whether the agreement of classifications is dependent on the prior experience, the depth of the introduction or the training of the raters. We hereby define introduction as explanation of the method (typically 5 min) and training as practical guidance in applying the method in the field (typically 10 min).

In this study we present a qualitative soil wetness classification scheme that is capable of capturing surface soil moisture differences in a semi-arid environment. It is slightly modified from the "boots and trousers" method (Rinderer et al. 2012) to incorporate every-day experiences of local people in terms of soil wetness that, in Africa, is more related to optimal seeding conditions and brick making than outdoor recreation activities that are common in Europe. The scheme is tested for its robustness and agreement between qualitative wetness classes and quantitative differences in soil water content. In particular the following questions are addressed:

1. Do the different qualitative wetness classes reflect actual differences in volumetric water content of the local soil of the study site?

2. Does the agreement of qualitative wetness classifications depend on the participants' experience in crop growing or the level of education?

3. Is the way in which the classification scheme is introduced to the participants and how they are trained important for achieving high agreement among raters?

2 Methods

2.1 Wetness classification scheme

The soil wetness classification scheme presented in this paper is based on qualitative indicators that are intuitive to local people in Tanzania from their every-day experience. In doing so, it incorporates the tacit knowledge of the perception of local people on soil wetness related to farming and brick making. It ranges from the driest class (1) called "very dry – dust dry" for which one cannot see or feel any moisture in the soil at the soil surface to the wettest class (7) for which one could see water ponding on the soil surface (Table 1). The other classes represent different grades of wetness with wetness class 2 characterizing a soil sample that feels dry but its

Table 1. Soil wetness classification scheme (Swahili version see Supplement) with the seven wetness classes based on qualitative indicators related to best conditions for seeding and brick making.

Icon	Class	Class name	Description
	1	very dry	"dust dry"
	2	dry	dry, but with some moist look
	3	below optimal	drier than optimal for seeding
	4	optimal	optimal for seeding crops
	5	above optimal	wetter than optimal – one can form a solid brick
	6	wet	when you step on the soil, water liquifies
	7	very wet	water ponding on the soil surface

darker color indicates that the water content is slightly higher than a sample of the same soil type that is classified as class 1. Wetness class 3 is characterized as moist but still too dry to be optimal for seeding plants, while wetness class 4 represents soil wetness that is optimal for seeding. A soil sample classified as class 5 would be optimal for making bricks so that they retain their shape when being dried but a soil sample of wetness class 6 would be too wet to form a stable brick. The indicators of the wetness scheme, namely, the conditions of optimal seeding and brick making, as well as the English and Swahili class definitions were developed in the course of a field workshop and interviews with a group of local farmers.

It is not intended to tie optimal seeding conditions to a specific crop but rather to reflect farmers' experience on good seeding conditions in general. The class "very dry – dusty dry" is also not necessarily related to the formation of a dust cloud when stepping on the ground, as this is strongly dependent on the soil texture. It is also not intended that raters form a brick to test its stability but it is assumed that local people have good experience in imagining these conditions from their every-day life. A vegetation cover or a litter layer as well as recent rainfall, dew or strong evaporation might affect the soil wetness conditions on the soil surface without being representative of the soil moisture at depth. This is particularly relevant as some full-grown crops can root at depths of 30 to 90 cm (Weaver and Bruner, 1927; Creswell and Martin, 1998) with average maximum rooting depths of crops up to 2 m (Canadell et al., 1996).

2.2 Field sites, data sets and test layout

The wetness classification scheme was tested in the two farming villages Mungushi and Kichangani, in the upper Pangani basin, about 20 km southeast of Arusha, Tanzania

(3°31′36″ S, 36°51′02″ W) (Fig. 1). The local soil was classified as Chromic Cambisol Colluvic Clayic (IUSS Working Group WRB, 2014), characterized by the absence of stones, low content of sand (6 %), consequently high content of silt (35 %) and very high content of clay (59 %) in the topsoil (0–16 cm depth) (see Table 2, Fig. 2a). In the underlying soil horizon (16–54 cm depth) only minor changes in the soil texture were observed, except for an intercalating sedimentary layer at 45–48 cm depth with higher sand content (40 %) and lower silt (29 %) and clay (30 %) content. The horizon at 54–58 cm depth exhibited a transition towards coarser material in the underlying horizon (58–81 cm depth) which was characterized by higher stone content (6 %), higher sand content (78 %), lower silt content (8 %) and comparatively low clay content (9 %). The fine textures in the upper two soil horizons and the layered structure of the soil profile suggest that the local soil was influenced by infrequent flooding events with delivery of predominantly fine material. Due to the high clay content in the topsoil horizons, the pore-volume derived plant available water content is comparatively low, whereas in deeper depths it might be significantly higher due to coarser particle size classes, leading to larger pore volumes. In general, the soils are fertile and heavily used for growing crops, mainly beans and corn. Due to a limited amount of rainfall (below 600 mm yr^{-1}) (Komakech and Van der Zaag, 2011) falling mainly during the rainy seasons (long rain *masika*: March–June and short rain *vuli*: October–December), agriculture in this region depends on flood irrigation during the rest of the year.

To test the wetness classification scheme we performed two experiments, one in April 2014 and another in June 2014. The first test in April was organized in the Mungushi village where 40 sampling points of different wetness were marked with flags along a 1.4 km course. The wetness of sequential

Table 2. Soil particle size distribution and soil texture of the local soil in the study area according to IUSS Working Group WRB (2014).

Soil depth [cm]	Clay ($<2\,\mu m$) [%]	Silt ($2–63\,\mu m$) [%]	Sand ($63–2000\,\mu m$) [%]	Stones ($>2000\,\mu m$) [%]	Soil texture in fine earth ($<2\,mm$)
0–16	59	35	6	0	clay (C)
16–54[a]	56	35	9	0	clay (C)
45–48[b]	30	29	40	1	clay loam (CL)
54–58[c]	11	11	76	2	sandy loam (SL)
58–81	9	8	78	6	sandy loam (SL)

[a] Matrix of the unit, whereas intercalated coarser sedimentary layers were analyzed separately. [b] Intercalated sedimentary layer in the unit at 16 to 58 cm depth. [c] Transition horizon between horizons at 16 to 58 cm and 58 to 81 cm depth.
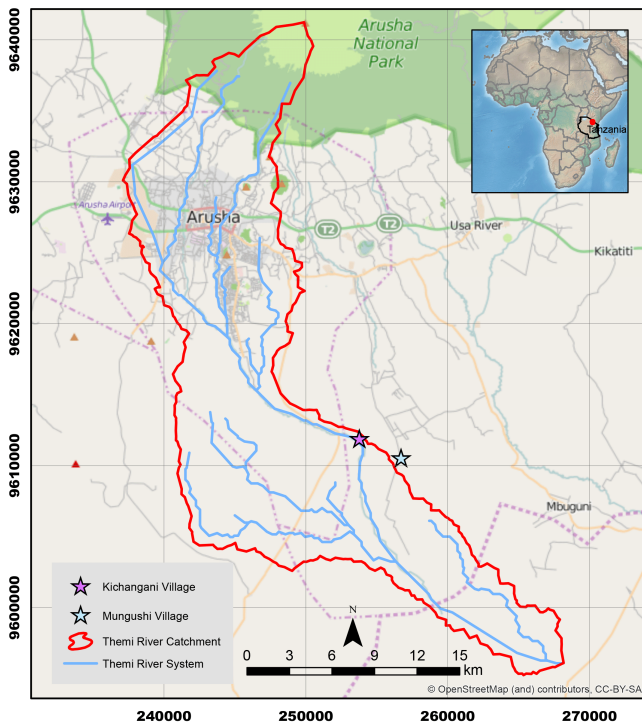

Figure 1. Themi river catchment at Arusha, Tanzania, and the two farming villages Mungushi and Kichangani were the wetness classification scheme was tested. (Background: OpenStreetMap and contributors, CC-BY-SA, insert map: Natural Earth.)
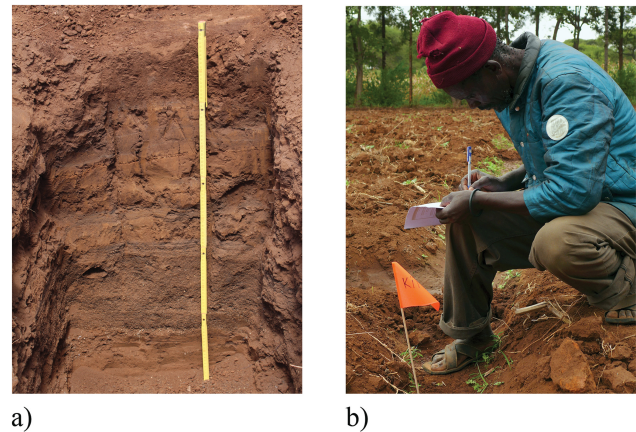

Figure 2. (a) Typical soil profile in the area where the wetness classification scheme was tested (profile depth: 1 m). **(b)** Farmer assessing the soil wetness conditions using the qualitative soil wetness scheme. (Photo: **(a)** D. Müller, **(b)** M. Rinderer.)

sampling points was chosen to be random. The test involved 40 people, namely, 14 farmers, 14 master students (called "students" hereafter), 9 PhD students and 3 professors. PhD students and professors were later combined into one group called "experts". All participants were given a brief introduction of about 5 min to the wetness classification scheme either in Swahili (farmers) or English (students, experts) and then were asked to individually classify the marked sites of different wetness along the course (Fig. 2b). Half of the farmers and students were given an additional training ($\sim 10\,\text{min}$) in which they were shown representative sites of wetness

classes 1, 4, and 7 before the test. These two groups of participants are referred to as F_{trained} and S_{trained} in the following. Farmers and students with a basic introduction are called F_{basic} and S_{basic}, respectively. When referring to all of the farmers, students and experts we use the expressions F_{all}, S_{all} and E_{all}. The assessment form used in April 2014 consisted of a matrix on an A4 paper (landscape format) with the number of the sampling sites appearing as rows and the wetness classes as columns (see Supplement 1 and 2) Participants were asked to tick the appropriate cell corresponding to their judgment of soil moisture conditions of a particular site.

In June 2014 a similar test with 18 farmers and 7 experts was organized in the neighboring village of Kichangani (42 sampling points). The second test was intended to analyze whether a more detailed and longer introduction ($\sim 20\,\text{min}$) and training ($\sim 30\,\text{min}$) organized in small subgroups of five people and an improved layout of the assessment form, would allow farmers to gain higher inter-rater reliability than during the first test in April. The new assess-

ment form consisted of an A4 portrait page with the class descriptions in the upper part and three columns for the soil wetness assessment (see Supplement 3 and 4). The first column was pre-labeled with "Site 1" to "Site 40" or "kituo 1" to "kituo 40" in Swahili, respectively. The second column was for the wetness class number and the third column was for optional comments. The flags, which indicated the sampling locations, were also labeled "kituo 1" to "kituo 40" to prevent potential conflicts between the number of the site and the number of wetness classes to assign. The wetness scheme remained the same except for some minor changes of class descriptions in the Swahili version.

Subsequently after both tests in April and in June, volumetric water content was measured by the gravimetric method taking $100 \, \text{cm}^3$ soil samples with a stainless steel cylinder (diameter: 5 cm), at 10 cm depth below the soil surface and determining the difference in weight between the original and oven-dried sample ($105 \, °\text{C}$ for 24 h). Corresponding qualitative wetness classification were made by the first author at the same time the gravimetric samples were taken to avoid the influence of a potential drying effect as sampling was slow and took longer than the qualitative test with the farmers, students and experts. A drying effect during the qualitative test was, however, considered to be small as all raters finished the course within less than an hour.

No rainfall occurred during the day of the test in April and June but in April, rainfall on the day prior to the test (no measurements taken) wetted the soil, while in June the fields were irrigated on the preceding days. A careful selection of sampling points was considered to guarantee the comparability between these two tests despite potential differences in infiltration patterns.

2.3 Statistical analysis

To evaluate the agreement between the qualitative soil wetness classes and the quantitative measurements, the distribution of gravimetrically measured volumetric soil water content was compiled for each qualitative wetness class. To assess the agreement of qualitative wetness classifications among farmers, students and experts, the frequency distribution of classification differences relative to the median of classifications of all group members, determined at each sampling point, was analyzed. First the overall agreement among group members was investigated incorporating the classification differences of all sampling points. Furthermore the frequency distribution of wetness class assignments for each sampling point was analyzed individually in order to identify which wetness classes were distinct and which ones were more difficult to identify. The median was chosen as reference as it is a robust measure of class assignments and not affected by individual outliers.

To see if individual raters had a systematic tendency to classify some wetness classes as too wet or too dry, the mean difference of classifications to the median for all sampling

points of each of the seven wetness classes was calculated for each person. Positive differences indicate a mean rater classification that was too wet and negative differences indicate a mean rater classification that was too dry compared to the reference.

Krippendorff's alpha (KA) (Krippendorff, 2004) and Cohen's kappa (CK) (Cohen, 1960) are two statistical measures to assess the degree of agreement or inter-rater reliability among raters assigning categorical values. Krippendorff's alpha is a measure to assess the degree of agreement within a group of raters (Krippendorff, 2004). If all raters agree perfectly, the observed agreement is one and so is Krippendorff's alpha. If wetness classes would be assigned randomly, Krippendorff's alpha would be equal to zero as observed and expected disagreement among all raters would be equal (Krippendorff, 2011).

Cohen's kappa was used as a measure to assess concordance between two raters, or, in our case, each individual rater and a reference (Cohen, 1960). If there is no agreement between the two rates other than what would be expected by chance, CK equals zero and if they both agree perfectly, CK would theoretically equal one. However, the maximum attainable CK value (CK_{max}) is smaller than one in cases where the codes are not equally probable and both raters do not assign all classes similarly often (Sim and Wright, 2005). As this is however normally the case, the kappa values in this paper are reported as the ratio between CK / CK_{max} and given as percentage.

3 Results

3.1 Qualitative and quantitative soil wetness

The classes of the presented, qualitative soil wetness classification scheme reflected differences in quantitative volumetric water content of the soil samples taken during the test in April and June (Fig. 3). The median volumetric water content ranged from 16 to 39 % for soil samples taken in April and from 14 to 32 % for samples taken in June. The median volumetric water content and its 25 and 75 % quantiles increased for soil samples of wetness classes 2 to 6 during the test in April and for samples of classes 1 to 5 during the test in June. However, soil samples of the following wetness classes had a similar median volumetric water content: classes 1 and 2; classes 6 and 7 (taken during the test in April); classes 5, 6, 7; and to a lesser extent, classes 3 and 4 (taken during the test in June). A Kruskal–Wallis test (Kruskal and Wallis, 1952) using an adjusted level of significance of 0.002 of Bonferroni (Dunn, 1959, 1961) indicated that the volumetric water content of the different qualitative wetness classes was not statistically significant. But it should be noted that the number of samples in each wetness class was low. A more relaxed significance test neglecting the alpha inflation and using an unadjusted significance level of 0.05 indicated, for the test in
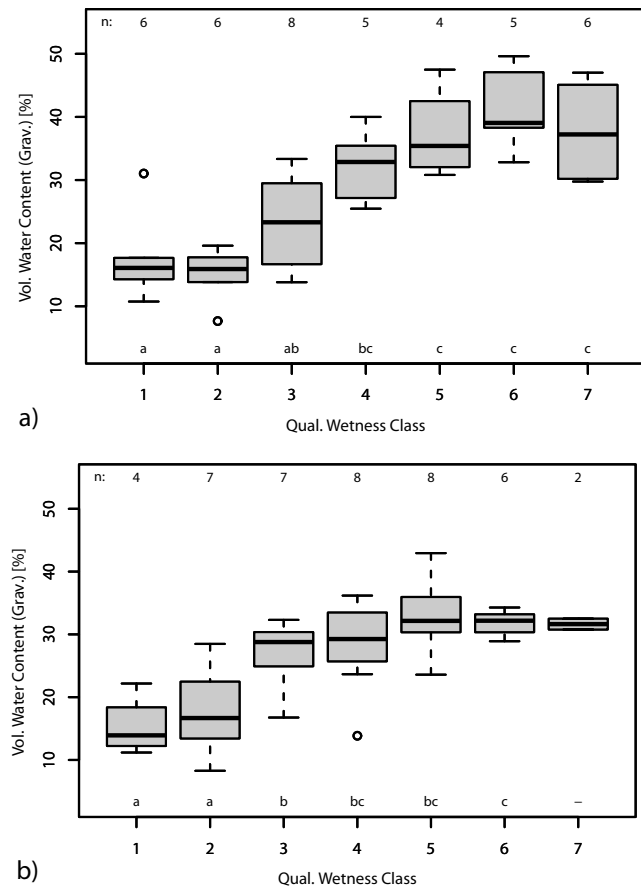
a)

b)

Figure 3. Volumetric water content for soil samples of each wetness class determined by the gravimetric method **(a)** during test in April 2014 and **(b)** during test in June 2014. (n: sample size, letters: statistically not significantly different groups.)

April, that the following classes were not significantly different from each other: classes 1, 2, 3; classes 3 and 4; and classes 4, 5, 6, 7. For the data set of the second test in June, the following classes were not significantly different from each other: classes 1 and 2; classes 3, 4, 5; and classes 4, 5 and 6. Class 7 was only represented by two samples, so could not be assessed.

3.2 Inter-rater reliability

In terms of the role of experience in crop growing and level of education on the agreement of wetness classifications, we found that during the first test in April the F_{all} showed a lower degree of agreement than S_{all} and E_{all} (Fig. 4): in about 46 % of all cases ($n = 456$) classified by F_{all} they agreed and independently assigned the same wetness class, 34 % of all classifications were off the group median by one class, 11 % by two classes, 4 % by three classes, 2 % by four classes, 2 % by five classes and 0.2 % ($= 1$ case) by six classes. The agreement of wetness classifications among S_{all} during the test in April was higher than that among F_{all}: 60 % of all
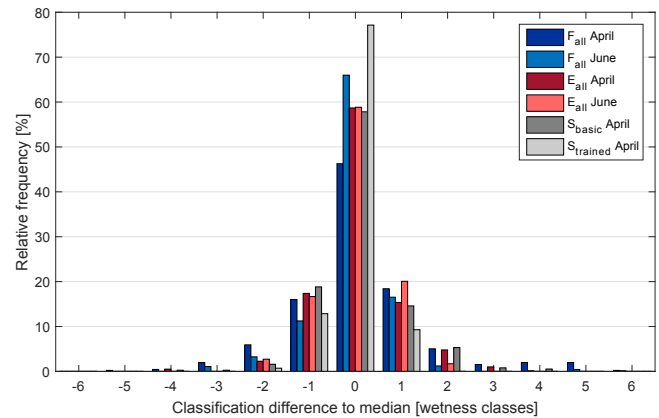

Figure 4. Deviation of wetness class assignments relative to the median of all farmers (F_{all}), and all experts (E_{all}) during the test in April and June and of all students with a basic introduction (S_{basic}) and students with training ($S_{trained}$) during the test in April.

cases ($n = 463$) classified by S_{all} were assigned to the same wetness class, 33 % of all classifications were off the group median by one class, 6 % by two classes, 1 % by three classes and 0.2 % ($= 1$ case) were off by four classes. None of S_{all} assigned a wetness class that was off by more than four classes. The agreement of wetness classifications among E_{all} during the test in April was similar to that of S_{all}: about 59 % of all cases ($n = 397$) classified by E_{all} were assigned the same wetness, 33 % of all classifications were off by one class, 7 % by two classes, 1 % by three classes and 0.5 % ($= 2$ assignments) were off by four classes. No wetness classification of the E_{all} was off the group median by more than four classes.

The difference in the degree of agreement between F_{all}, S_{all} and E_{all} during the test in April was also evident from the inter-rater reliability statistics. The Krippendorff alpha (KA) value for F_{all} (KA: 42 %) was half of KA of S_{all} (KA: 83 %) and E_{all} (KA: 82 %) during the test in April (Fig. 5 and Table 3). The median CK / CK_{max} also differed between F_{all}, S_{all} and E_{all} (43, 65 and 67 %, respectively; Fig. 5 and Table 3). The interquartile range (IQR) of CK / CK_{max} was 1.8 to 3 times larger for F_{all} than for S_{all} and E_{all} (Fig. 5 and Table 3).

During the second test in June the agreement of class assignments among F_{all} was higher and exceeded even the agreement among E_{all} (Fig. 4): in about 66 % of all cases ($n = 738$) F_{all} independently assigned the same wetness class, 28 % were off the group median by one class, 4 % by two classes, 1 % by three classes and 1 % were off by four or more classes. Only once (0.14 %) a farmer assigned a wetness class that was off by six classes. The agreement of wetness classifications among E_{all} was similar during the test in April and in June except that no expert was off the group median by more than two wetness classes during the second test (Fig. 4): 59 % of all cases ($n = 294$) classified by E_{all} during the test in June were assigned the same wetness class,
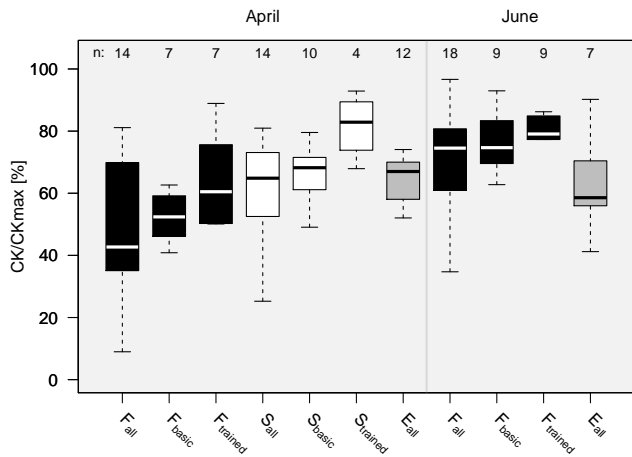
Figure 5. Inter-rater reliability among members of individual groups tested in April and June expressed as the Cohen's kappa ratio CK / CK$_{max}$ (F: farmers, black, S: students, white, E: experts, grey; "$_{basic}$" indicates the subgroup with only basic introduction, "$_{trained}$" indicates the subgroup with more detailed training, "$_{all}$" indicates that both subgroups have been considered; n: number of individuals in each group).

Table 3. Inter-rater reliability statistics for the different groups (F: farmers, S: students, E: experts, All: all participants) during tests in April and in June. ("$_{basic}$" indicates only basic introduction, "$_{trained}$" indicates more detailed training, "$_{all}$" indicated that both subgroups have been considered.) Krippendorff's alpha and the Cohen's kappa ratio CK / CK$_{max}$ can vary between 100 % (perfect agreement) and 0 % (no agreement other than that what would be expected by chance).

Test	Groups	Krippendorff alpha [%]	Median CK / CK$_{max}$ [%] (IQR)
April	F_{all}	42	43 (35–70)
	F_{basic}	49	52 (46–59)
	$F_{trained}$	41	60 (50–76)
	S_{all}	83	65 (53–73)
	S_{basic}	81	68 (61–72)
	$S_{trained}$	91	83 (74–89)
	E_{all}	82	67 (58–70)
	All	66	51 (34–62)
June	F_{all}	76	75 (61–81)
	F_{basic}	65	75 (70–83)
	$F_{trained}$	87	79 (77–85)
	E_{all}	84	59 (56–70)
	All	78	67 (59–73)

37 % of all classifications were off by one class, 4 % by two classes.

During the second test in June F_{all} achieved a similar inter-rater reliability as E_{all} (no student raters during the test in June). KA of F_{all} (KA: 76 %) was more similar to KA of E_{all} (KA: 84 %) and the median of CK / CK$_{max}$ of F_{all} (75 %) even exceeded that of E_{all} (59 %) during the second test in June (Fig. 5 and Table 3). The IQR of CK / CK$_{max}$ for F_{all} during the second test was almost half the IQR of the first test (Fig. 5 and Table 3).

In terms of the role of training on how to apply the wetness classification scheme, we found that $S_{trained}$ during the test in April and $F_{trained}$ during the test in June had a higher inter-rater reliability (KA and CK / CK$_{max}$) compared to their colleagues with only a basic introduction (Table 3). The distribution of differences in classifications relative to the median of the groups was also narrower for $S_{trained}$ during the test in April (Fig. 4) and for $F_{trained}$ during the test in June compared to their colleagues with only a basic introduction. No individual of these two groups with additional training assigned a wetness class that was off the group median by more than two classes. During the test in April the importance of additional training was not so evident among farmers. While the median CK / CK$_{max}$ was higher for $F_{trained}$ compared to F_{basic}, this was not the case for KA (Table 3) and the spread in class assignments among $F_{trained}$ and F_{basic} was for both large. In hindsight, we partly attribute this to the use of an assessment form during the test in April that seemed difficult to read/fill out.

In terms of a convergence of wetness class assignments among the raters with increasing number of rated sampling points when following the course, we found that during the first test in April the median CK / CK$_{max}$ and KA for S_{all} and E_{all} was higher but not statistically significant for the second half of the sampling points (points 21 to 40) compared to the first half (points 1 to 20). This was also true for the median CK / CK$_{max}$ for E_{all} during the second test in June (no student raters in June). F_{all} did not have a higher median CK / CK$_{max}$ and KA for the second half of the sampling points compared to the first half during both tests. The median CK / CK$_{max}$ and KA of $S_{trained}$ during the first test in April and $F_{trained}$ during the second test in June was higher for the second half of the sampling points compared to the first half but the median CK / CK$_{max}$ of their respective colleagues with only a basic introduction was not.

3.3 Identifiability of individual wetness classes

During the first test in April the spread of classification assignments by F_{all}, S_{all} and E_{all} was large for all wetness classes. F_{all} had a flat frequency distribution of class assignments for all wetness classes especially for class 2 to 5 and to a lesser extent also for class 6 (Fig. 6a). Note that during both tests, half of F_{all} did not classify any of the sampling points as class 7. S_{all} and E_{all} (graphs not shown) had narrower frequency distributions of class assignments than F_{all}. The two wettest classes, class 7 and to a lesser extent class 6, showed the smallest spread and the dry to intermediate class 2, 3 and 4 the largest spread.
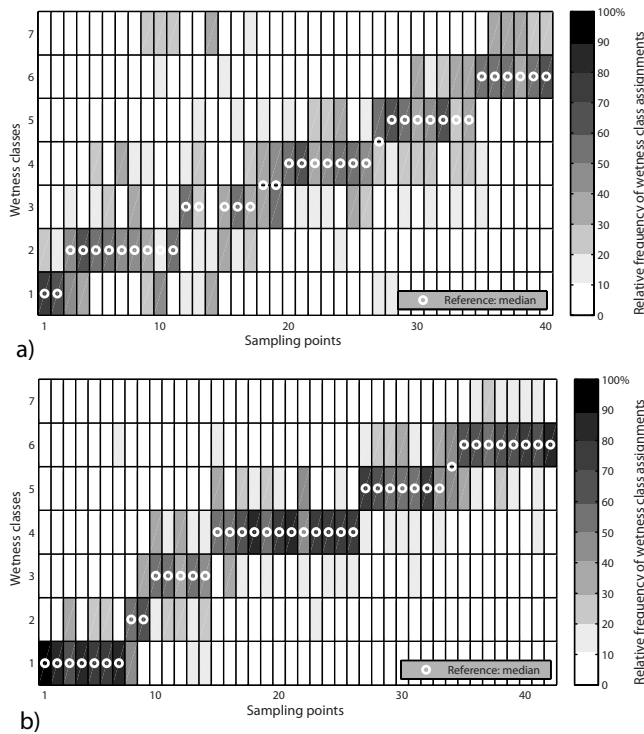
a)

b)

Figure 6. Spread of classification assignments for sampling points of individual wetness classes by **(a)** all farmers (F_{all}) in April and **(b)** all farmers (F_{all}) in June (grey shades: relative frequency of wetness class assignments for each of the sampling points, white circles: median of classifications). For example, sampling point 10 in panel **(b)**: the majority (56 %) of raters classified this point as class 3 (white circle) but 33 % of the raters assigned class 4 and 11 % of the raters assigned class 2. The difference between the two graphs shows the effect of more detailed introduction and a clear assessment form. Note that during both tests, none of the sampling points were classified as class 7 by half of F_{all}, and that the sampling points were distributed in random order of wetness classes in the field experiment, but were ordered here according to the median estimation for graphical clarity.

During the second test in June the spread in class assignments by F_{all} was smaller (Fig. 6b). The spread of class assignments by F_{all} improved especially for sample points of the dry to intermediate class 2 to 5 and also the second wettest class 6 between the first and the second test. The spread of class assignments by E_{all} was similar or only slightly smaller during the second test than during the first one (graphs not shown).

Regarding how training helped to better identify the wetness classes, we found that there was hardly any difference in spread of class assignments by F_{basic} and $F_{trained}$ for the first test in April. Both groups showed a large spread of class assignments for all wetness classes. In contrast, $S_{trained}$ had narrower frequency distributions of class assignments for almost all wetness classes compared to S_{basic}; especially for the dry to intermediate classes 2 to 5 but also for the second
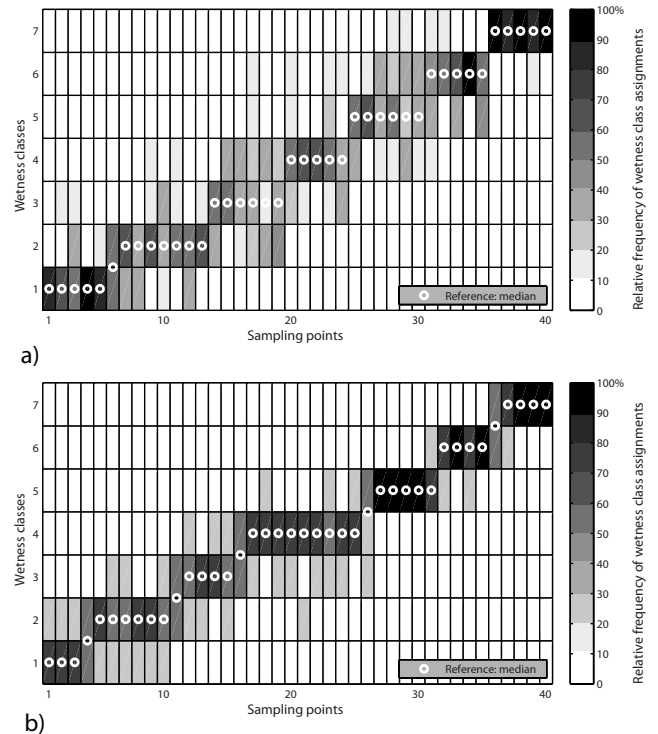

a)

b)

Figure 7. Spread of classification assignments for sampling points of individual wetness classes by **(a)** S_{basic} with basic introduction and **(b)** $S_{trained}$ with additional training during test in April (grey shades: relative frequency of wetness class assignments for each of the sampling points, white circles: median of classifications). For example, sampling point 10 in panel **(b)**: the majority (50 %) of raters classified this point as class 2 (white circle) but 25 % of the raters assigned class 1 and 25 % of the raters assigned class 3. Note that the sampling points were distributed in random order of wetness classes in the field experiment, but were ordered here according to the median estimation for graphical clarity.

wettest class 6 (Fig. 7). During the second test in June the group of $F_{trained}$ also showed less spread in class assignments compared to F_{basic} (graph not shown). The improvement was noticeable for all wetness classes.

Individual people showed a systematic tendency to rate selected wetness classes as either too dry or too wet. During the first test in April, individual farmers as well as a few students and experts on average showed a tendency to classify dry sampling sites as too wet and to a lesser extent wet sites as too dry (for F_{all} see Fig. 8a). The class 2 and 3 showed the largest mean classification differences. During the second test in June fewer individuals of farmers and experts showed a systematic bias to classify dry sites as too wet and wet sites as too dry. The mean classification difference was smaller (white and pastel colors in Fig. 8b). Note that none of the sampling points had been classified as class 7 by half of F_{all} during the test in April and in June; this is why the mean classification difference for this class is not given.
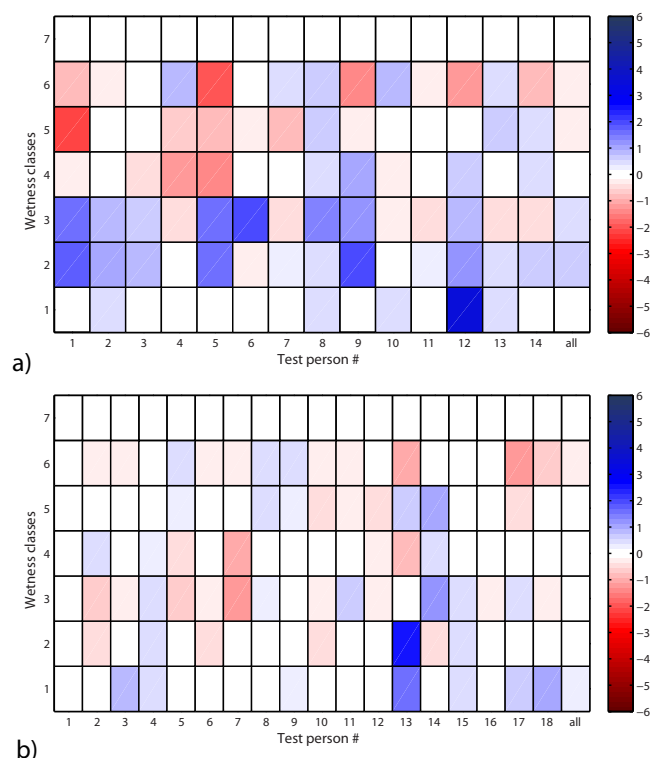
Hydrol. Earth Syst. Sci., 19, 3505–3516, 2015

www.hydrol-earth-syst-sci.net/19/3505/2015/

Figure 8. Mean classification difference for all sampling points of each wetness class per test person in group F_{all} **(a)** tested in April; **(b)** tested in June. Red colors indicate mean classification to be too dry, blue colors to be too wet compared to the median of each wetness class. For example, test person no. 5 in panel **(b)** on average classified the sampling points of wetness class 3 and 4 too dry (red colors) and the sampling points of wetness class 5 and 6 too wet (blue colors) relative to the majority of classification of the other raters. However, the bright colors show, that the average classification difference was smaller than one wetness class (class 3: −0.6, class 4: −0.4, class 5: +0.3, class 6: +0.50).

4 Discussion

The agreement in wetness class assignments among S_{all} and E_{all} during the test in April and also F_{all} during the test in June was high which shows the robustness of the method despite being based on qualitative indicators. In 93 and 91 % of all classifications the members of group S_{all} and E_{all} agreed or were off by only one wetness class during the first test in April. Despite a lower inter-rater reliability for F_{all} during the test in April, they still agreed in 81 % of all cases or were off by one wetness class. These high numbers of agreement suggest that the qualitative soil wetness classification scheme in general was intuitive to local people with different levels of education and different experience in crop production.

The within-group variability of class assignments by F_{all} could be considerably reduced by a profound basic introduction organized in small subgroups, by a redesign of the assessment form layout and by a clearer labeling of the sampling sites. In 94 % of all classifications the members of group F_{all} agreed or were off by only one wetness class. In June not only the site number was written on the flag but also the word kituo (English: station). We assume that gross misclassifications of up to six wetness classes during the first test in April might partly be due to ticking the wrong cell of the old, matrix type of assessment form. The dry to intermediate wetness classes seemed to be difficult to assign while the wettest classes were the easiest (Fig. 6). A profound basic introduction to the wetness classification scheme during the second test in June could particularly improve dry to intermediate class assignments by F_{all}. The benefit of a more detailed training was evident regardless of farming experience or education level for both F_{trained} and S_{trained}. Not only the within group agreement could be improved but also the number of gross misclassifications of more than three wetness classes could be avoided (see Table 3 and Figs. 4, 6, 7).

Compared to a test with master students in Switzerland (Rinderer et al., 2012), the agreement in this study was similar or lower. Classifications with an offset from the group median of more than two wetness classes were similarly frequent among Tanzanian students S_{all} (1 %) and experts E_{all} (2 %) compared to Swiss students (∼ 1 %), but considerably higher among Tanzanian farmers F_{all} (8 %) during the first test in April. The inter-rater reliability of F_{all} (no student raters tested) during the second test in June was however similar to that of Swiss students.

A better basic introduction, organized in small subgroups, minimized the spread of class assignments and the bias of individuals to classify wet sites as too dry and dry sites as too wet (Fig. 8). While the mean classification difference of individuals during the first test in April (see Fig. 8a) was much higher compared to the one in the study by Rinderer et al. (2012), it was similar during the second test in June (see Fig. 8b). (Note that the range of values assigned to the color ramp in Rinderer et al., 2012, is different compared to Fig. 8.)

The intermediate, qualitative wetness classes reflected actual differences in volumetric water content of the gravimetric soil samples however the median values of the two driest classes and the three wettest classes were very similar. What "looked" different was in fact similar in quantitative terms. This suggests that a classification scheme with fewer wetness classes would be sufficient to differentiate the actual range of volumetric water content. Rinderer et al. (2012) also discuss merging the two wettest classes and the three intermediate classes in their study. However, in a range of applications it still might be useful to use the seven wetness classes as qualitative differences can be very informative and in fact more important than the quantitative differences (e.g., differentiating between wet areas with shallow subsurface flow and saturated areas with overland flow in terms of flow velocities, transport processes and susceptibility to erosion (see also Blazkova et al., 2002; Ali et al., 2014; or Dunne and Black, 1970). A reduction of classes would result in a coarser

resolution of the resulting patterns and might not well capture differences in space and time. Despite being potentially less frequent, misclassification would have a larger effect on the final result when using a scheme with fewer classes.

It should be noted that the classification scheme by Rinderer et al. (2012) was developed and tested in humid environmental conditions with moor landscapes and therefore had a different range of volumetric water content assigned to the individual wetness classes. The median volumetric water content of class 1 in the Swiss study ($\sim 38\%$) was similar to the median volumetric water content of class 7 (37 %) in this study (Fig. 3a). This exemplifies that similar qualitative indicators on the soil surface can be associated with different volumetric water content and therefore the qualitative wetness classes need to be calibrated to the local soil types accounting for differences in soil textures and environmental conditions if the absolute water content is of interest. In the context of crop growing this is for instance the case in terms of assessing the plant available soil water content and the permanent wilting point of a given soil. Soils with a high content in clay minerals are also characterized by thixotrophic behavior allowing one to squeeze out water when shearing the clay mineral layers. Other limitations of this wetness classification scheme exist since only the soil surface properties are assessed, but for many full-grown crops, the soil moisture at depth is of main interest. In this case we recommend applying the classification scheme to a soil sample that is taken from a small pit, dug down to the depth with the highest root abundances with a spade (Görbing and Sekera, 1947). Other potential influencing factors are the vegetation and litter on the soil surface, wetting by dew and drizzle and drying up due to evaporation.

5   Conclusions

This study demonstrates the potential of a soil wetness classification scheme based on qualitative indicators that is capable of capturing shallow soil moisture differences in a semi-arid environment. It highlights the value of a detailed introduction and training to the method in gaining high agreement among individual raters but that neither experience in crop production nor a certain education level are a prerequisite for robust and comparable wetness classifications. The study also shows that the intermediate, qualitative wetness classes reflect quantitative differences in median volumetric water content, but that the driest and wettest classes do not.

A soil wetness classification scheme as presented here is quick to apply, needs no expert knowledge and no measuring device, but can still provide robust and reliable results on soil moisture differences. It could be exemplified that such a qualitative method can be applied successfully in a wider range of environmental conditions (Ali et al., 2014), when being calibrated/adapted to the local soil textures and cultural conditions. As farming and brick making are common

in many rural communities, we see the potential to also use this scheme in other developing countries and remote areas with limited measuring equipment and energy supply.

Collecting soft data is particularly promising for citizen science, a new approach that takes advantage of the value of distributed information captured by many local observers (Buytaert et al., 2014; Lowry et al., 2011; Turner and Richter, 2011; Peckenham and Peckenham, 2014). In such a framework the qualitative soil wetness classification scheme presented here could be used by many observers to conduct rapid spatial soil moisture assessments comprising thousands of sampling points within a catchment or an even larger extent (e.g., Open Air Laboratories (OPAL) network and Global Learning and Observations to Benefit the Environment (GLOBE) program). Trained farmers could send wetness classifications of their fields via SMS to a common decision support system. The spatial soil moisture patterns could then be used for model calibration and data assimilation to predict soil water stress and provide suggestions to local farmers on how to best use the available water resources. This vision of crowd-based collection of environmental data is currently under development in the project "iMoMo – Innovative Monitoring and Modeling of Water", funded by the Swiss Agency for Development and Cooperation (SDC) in the study area near Arusha, Tanzania. Evaluation at the end of this project will allow for an assessment of the actual impact of this qualitative method on the sustainability of crop yields and community welfare under limited water availability.

The Supplement related to this article is available online at doi:10.5194/hess-19-3505-2015-supplement.

References

Ali, G., Birkel, C., Tetzlaff, D., Soulsby, C., McDonnell, J. J., and Tarolli, P.: A comparison of wetness indices for the prediction of observed connected saturated areas under contrasting conditions, Earth Surf. Proc. Land., 39, 399–413, 2014.

Ambroise, B., Freer, J., and Beven, K.: Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France, Water Resour. Res., 32, 2147–2159, 1996.

Blazkova, S., Beven, K. J., and Kulasova, A.: On constraining TOP-MODEL hydrograph simulations using partial saturated area information, Hydrol. Process., 16, 441–458, 2002.

Buytaert, W., Zulkafli, Z., Grainger, S., Acosta, L., De Bièvre, B., Bhusal, J., Chanie, T., Clark, J., Dewulf, A., Hannah, D. M., Hergarten, C., Isaeva, A., Karpouzoglou, T., Pandey, B., Paudel, D., Sharma, K., Steenhuis, T. S., Tilahun, S., Van Hecken, G., and Zhumanova, M.: Citizen science in hydrology and water resources: opportunities for knowledge generation, ecosystem service management, and sustainable development, Frontiers in Hydrosphere, 2, 1–21, 2014.

Canadell, J., Jackson, R. B., Ehleringer, J. B., Mooney, H. A., Sala, O. E., and Schulze, E.-D.: Maximum rooting depth of vegetation types at the global scale, Oecologia, 108, 583–595, 1996.

Cohen, J.: A Coefficient of Agreement for Nominal Scales, Educ. Psychol. Meas., 20, 37–46, 1960.

Creswell, R. and Martin, F.: Dryland Farming: Crops and Techniques for Arid Regions, ECHO technical note, Fort Myers, Fl, USA, 23 pp., 1998.

De Quervain, M.: Die Festigkeitseigenschaften der Schneedecke und ihre Messung, Geofisica Pura e Applicata, 18, 179–191, 1950.

Dunn, O.: Estimation of the Medians for Dependent Variables, Ann. Math. Stat., 30, 192–197, 1959.

Dunn, O.: Multiple Comparisons Among Means, J. Am. Stat. Assoc., 56, 52–64, 1961.

Dunne, T. and Black, R. D.: Partial Area Contributions to Storm Runoff in a Small New England Watershed, Water Resour. Res., 6, 1296–1311, 1970.

Dunne, T., Moore, T. R., and Tayloer, C. H.: Recognition and prediction of runoff producing zones in humid regions, Hydrological Sciences Bulletine, 20, 305–327, 1975.

Ellenberg, H., Weber, H. E., Düll, R., Wirth, V., Werner, W., and Paulissen, D.: Zeigerwerte von Pflanzen in Mitteleuropa, Scripta geobotanica, 18, p. 248, 1991.

Görbing, J. and Sekera, F.: Die Spatendiagnose – Ziel und Grundlage der zweckmässigen Bodenbearbeitung, Verlag Sachse, Hannover, Germany, 1947.

Inamdar, S. P. and Mitchell, M. J.: Contributions of riparian and hillslope waters to storm runoff across multiple catchments and storm events in a glaciated forested watershed, J. Hydrol., 341, 116–130, 2007.

IUSS Working Group WRB: World reference base for soil resources 2014, Rome, Italy, 2014.

Komakech, H. C. and Van der Zaag, P.: Understanding the Emergence and Functioning of River Committees in a Catchment of the Pangani Basin, Tanzania, Water Alternatives, 4, 197–222, 2011.

Krippendorff, K.: Content Analysis, an Introduction to its Methodology, 2nd ed., Sage Publications, Thousand Oaks, USA, 2004.

Krippendorff, K.: Computing Krippendorff's Alpha-Reliability, University of Pennsylvania, Philadelphia, Pennsylvania, USA, 2011.

Kruskal, W. H. and Wallis, W. A.: Use of Ranks in One-Criterion Variance Analysis, J. Am. Stat. Assoc., 47, 583–621, 1952.

Kulasova, A., Beven, K. J., Blazkova, S. D., Rezacova, D., and Cajthaml, J.: Comparison of saturated areas mapping methods in the Jizera Mountains, Czech Republic, J. Hydrol. Hydromech., 62, 160–168, 2014.

Latron, J. and Gallart, F.: Seasonal dynamics of runoff-contributing areas in a small mediterranean research catchment (Vallcebre, Eastern Pyrenees), J. Hydrol., 335, 194–206, 2007.

Lowry, C. S., Loheide, S. P., Moore, C. E., and Lundquist, J. D.: Groundwater controls on vegetation composition and patterning in mountain meadows, Water Resour. Res., 47, W00J11, doi:10.1029/2010WR010086, 2011.

McDonnell, J. J. and Taylor, C. H.: Surface and subsurface water contributions during snowmelt in a small Precambrian Shield watershed, Muskoka, Ontario, Atmos. Ocean, 25, 251–266, 1987.

Metcalfe-Smith, J. L.: Biological Water-Quality Assessment of Rivers?: Use of Macroinvertebrate Communities, in: The River Handbook, edited by: Calow, P. and Petts, G. E., Blackwell Science Ltd., Oxford, UK, 144–171, 1994.

Natural Resources Conservation Serviceof the United States Department of Agriculture: Estimating Soil Moisture by Feel and Appearance, Program Aid, 1619, 1–7, 1998.

Peckenham, J. M. and Peckenham, S. K.: Assessment of Quality for Middle Level and High School Student-Generated Water Quality Data, JAWRA Journal of the American Water Resources Association, 50, 1477–1487, 2014.

Pielmeier, C. and Schneebeli, M.: Developments in the Stratigraphy of Snow, Surv. Geophys., 24, 389–416, 2003.

Quinn, P. F., Ostendorf, B., Beven, K., and Tenhunen, J.: Spatial and temporal predictions of soil moisture patterns and evaporative losses using TOPMODEL and the GASFLUX model for an Alaskan catchment, Hydrol. Earth Syst. Sci., 2, 51–64, doi:10.5194/hess-2-51-1998, 1998.

Rinderer, M. and Seibert, J.: Soil Information in Hydrologic Models: Hard Data, Soft Data , and the Dialog between Experimentalists and Modelers, in: Hydropedology – Synergetic Interaction of Soil Science and Hydrology, edited by: Lin, H., Elsevier B. V., Waltham, USA, 515–536, 2012.

Rinderer, M., Kollegger, A., Fischer, B. M. C., Stähli, M., and Seibert, J.: Sensing with boots and trousers – qualitative field observations of shallow soil moisture patterns, Hydrol. Process., 26, 4112–4120, 2012.

Sim, J. and Wright, C.: The kappa statistic in reliability studies: Use, interpretation, and sample size requirements, Phys. Ther., 85, 257–268, 2005.

SNIFFER: A Functional Wetland Typology for Scotland – Field Survey Manual, 1st ed.,Scotland and Northern Ireland Forum for Environmental Research (SNIFFER), Edinburgh, UK, 2009.

Thien, S. J.: A flow diagram for teaching texture by feel analysis, Journal of Agronomic Education, 8, 54–55, 1979.

Turner, D. S. and Richter, H. E.: Wet/dry mapping: Using citizen scientists to monitor the extent of perennial surface flow in dryland regions, Environ. Manage., 47, 497–505, 2011.

Weaver, J. E. and Bruner, W. E.: Root Development of Vegetable Crops, 1st ed., McGraw-Hill Book Company, Inc., New York, USA, 1927.