Catena 167 (2018) 363–372

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

Identifying representative sites to simultaneously predict hillslope surface and subsurface mean soil water contents

Xiaoming Lai[a,b], Zhiwen Zhou[a,b], Qing Zhu[a,b,c,*], Kaihua Liao[a]

[a] Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Jiangsu Collaborative Innovation Center of Regional Modern Agriculture & Environmental Protection, Huaiyin Normal University, Huaian 223001, China

ARTICLE INFO

Keywords:
Temporal stability analysis
k-means clustering
Up-scaling

ABSTRACT

Many approaches have been proposed to identify the representative sampling sites for estimating the spatial mean soil water contents. However, comparisons on these approaches have seldom been conducted to simultaneously predict the surface and subsurface mean soil water contents. In this study, five approaches were evaluated in identifying representative sites to estimate the surface and subsurface mean soil water contents on a typical hillslope in Taihu Lake Basin, China. They were temporal stability analysis (TSA), *k*-means clustering with environmental factors as inputs (EFs), combinations of TSA and EFs (EFs + TSA), *k*-means clustering with surface soil water contents as inputs (Theta), and combinations of TSA and Theta (Theta + TSA). The correlation coefficient (*r*) and root mean squared error (RMSE) between estimated and measured mean soil water contents were used to evaluate the accuracies during the calibration period (the first 25 dates) and validation period (the last 18 dates). Results showed the optimal number of representative sites on this hillslope was six. When > 6 representative sites were selected, the TSA had the lowest accuracies for estimating both surface and subsurface mean soil water contents during validation period (mean RMSE ≥ 0.011 $m^3 m^{-3}$). The Theta and Theta + TSA had better accuracies in estimating surface mean soil water contents during both calibration and validation periods (mean RMSE < 0.007 $m^3 m^{-3}$). However, to estimate surface and subsurface mean soil water contents simultaneously, the EFs and EFs + TSA were more promising (mean RMSE < 0.011 $m^3 m^{-3}$ during validation period), especially the EFs which only required one-time collection of environmental factors. These findings will be beneficial for choosing proper approach to calibrate and validate the remote sensed soil water contents.

1. Introduction

In-situ monitoring and remote sensing are two most common techniques to measure soil water contents at different spatial scales (Robinson et al. 2008; Brocca et al. 2010; Zhu et al. 2012; Vereecken et al. 2014). Relative to in-situ monitoring, remote sensing technique is more promising in fast and cost-efficient measurements of surface soil water contents at large spatial scales. However, remotely sensed soil water contents require the ground-based in-situ observations at the corresponding pixel for the calibration and validation (Mohanty and Skaggs 2001; Martínez-Fernández and Ceballos 2005; Mohanty et al. 2017), which are highly costly and time-consuming (Brocca et al. 2010; Faridani et al. 2016). Therefore, identifying the representative sites of in-situ observations to predict the spatial mean soil water contents by balancing the predicting accuracy and sampling costs is of great significance in hydrological studies.

Temporal stability analysis (TSA) has been commonly used to identify the representative sites for predicting mean soil water contents in previous studies (e.g., Grayson and Western 1998; Cosh et al. 2004; Martínez-Fernández and Ceballos 2005; Vivoni et al. 2008; Ran et al. 2017). The TSA was proposed by Vachaud et al. (1985) and defined as the time invariant associations between spatial locations and classical statistical parameters of soil water contents (e.g. spatial average or relative ranking). Traditionally, the location with the temporal mean relative difference (MRD) closest to zero or the lowest value of the standard deviation of relative difference (SDRD) was recognized as the representative site (e.g., Grayson and Western 1998; Martínez-Fernández and Ceballos 2005; Brocca et al. 2010). However, the performances in identifying the representative site to estimate the mean soil water contents by these two parameters were challenged in previous studies (Hu et al. 2012; Liao et al. 2017). Some other indicators, such as the index of temporal stability (ITS) (Jacobs et al. 2004), the mean absolute bias error (Hu et al., 2010), were proposed by previous studies and found have better performances in identifying the

* Corresponding author at: Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China.
E-mail address: qzhu@niglas.ac.cn (Q. Zhu).

https://doi.org/10.1016/j.catena.2018.05.016
Received 21 October 2017; Received in revised form 15 May 2018; Accepted 16 May 2018
0341-8162/

representative sites.

Contradictory prediction accuracies in estimating the mean soil water contents were derived by using TSA to identify the representative site. For example, Zhao et al. (2010) and Penna et al. (2013) found that the determination coefficients between predicted and actual mean soil water contents were higher than 0.90, while Cosh et al. (2004) and Vivoni et al. (2008) summarized that the determination coefficients were below 0.85. This may be attributed to that only one representative site was generally determined in previous studies, which resulted in large uncertainty. To reduce the representative error, both in the studies by Van Arkel and Kaleita (2014), She et al. (2015) and Ran et al. (2017), multiple representative sites were identified by TSA and prediction accuracies were sensitive to the number of representative sites. To improve the prediction accuracy, representative sites were also identified by combining TSA with other prior information (e.g., soil properties, landscape heterogeneity) (Vereecken et al. 2008). For example, Zhou et al. (2007) integrated soil information into the TSA and provided more accurate representative locations for capturing mean soil water contents. Ran et al. (2017) considered the village groups as a stratification layer and applied TSA region by region to identify multiple representative sites and achieved good accuracy. Thereby, combining the TSA with the stratification of the study area based on prior information is an alternative in identifying the optimal number of representative sites and may perform better than the traditional TSA.

The *k*-means clustering algorithm is a new approach in identifying the representative sites for predicting mean soil water contents. It is the most popular and simplest clustering method that separates the multivariate data into *k* clusters so that the squared error between the empirical mean of the cluster and the points in the cluster is minimized (Jain 2010). Van Arkel and Kaleita (2014) firstly applied soil properties and terrain attributes as inputs into the *k*-means clustering to determine the representative sites and better accuracies were achieved than TSA. Liao et al. (2017) evaluated the performances of TSA, *k*-means clustering and random sampling strategy in identifying representative sites, and found the *k*-means clustering performed better than other approaches. However, only several limited number of clusters were taken into account in these studies, thus the optimal number of representative sites could not be determined. For example, in the study by Van Arkel and Kaleita (2014), only schemes of 1, 2, 3, 4 clusters were considered, and Liao et al. (2017) only considered 2, 4, 6, 8 clusters. In addition, the combination of TSA with stratifications of the study area by the *k*-means clustering to identify the representative sites has rarely be investigated in previous studies.

Previous studies were mostly focused on identifying the representative sites for predicting surface mean soil water contents and subsurface mean soil water content estimation had been less addressed (e.g., Vivoni et al. 2008; Zhao et al. 2010; Mohanty et al. 2017; Liao et al. 2017). Efforts have been made to estimate subsurface soil water contents by integrating remote sensed surface soil water contents with soil hydrologic models and assimilation schemes in previous (e.g., Albergel et al. 2008; Faridani et al. 2016; González-Zamora et al. 2016). However, validation of the estimated subsurface soil water contents by in-situ measurements is still an inevitable issue (Teuling et al. 2006). Whether the representative sites identified by the surface soil and terrain information can be used to estimate the subsurface mean soil water contents have not been fully revealed. Several studies have been conducted on this issue but mixed conclusions have been received. For example, Penna et al. (2013) and Wang et al. (2013) indicated that one representative site recognized at surface soil layer was adequate to estimate the mean soil water contents at other depths. However, Gao and Shao (2012) and Heathman et al. (2012) concluded that none point was sufficient to represent the mean soil water contents for different soil layers. Previous studies were mostly relied on the surface soil water content data to identify the representative sites to simultaneously predict the surface and subsurface mean soil water contents. However, the soil water contents have high spatial variability in both horizontal and vertical directions. Therefore, new approaches should be tested to identify the representative sites to simultaneously estimate the surface and subsurface mean soil water contents.

Therefore, based on the surface (10 cm) and subsurface (30 cm) in-situ soil water content observations from 77 sampling sites on a hillslope, the objectives of this study were to: (i) find the optimal number of representative sites by methods of TSA, *k*-means clustering algorithm and the combinations of TSA and *k*-means clustering algorithm; (ii) evaluate the accuracies in predicting the surface mean soil water contents by multiple representative sites identified by these methods; (iii) investigate the feasibility of estimating the subsurface mean soil water contents by representative sites identified by surface soil and terrain information.

2. Materials and methods

2.1. Study hillslope

Two adjacent hillslopes which are separated by ditch and respectively covered by green tea (*Camellia sinensis* (L.) *O. Kuntze*) and moso bamboo (*Phyllostachysedulis (Carr.) H. de Lehaie*), were selected as the study region in the hilly area of Taihu Lake Basin, China (Fig. 1). The weather of this region belongs to the north subtropical-middle subtropical transition monsoon climate, with annual mean temperature and mean precipitation of 15.9 °C and 1157 mm, respectively. The soil type is categorized as shallow lithosols according to FAO soil classification and the parent material is quartz sandstone. Soil texture of this study region is classified as silt loam, and the thickness of soil varies from < 0.3 m at the summit slope position to about 1.0 m at the foot slope position. Detailed descriptions of this study area can be found in the study by Lai et al. (2017).

2.2. Soil water content measurement

For monitoring volumetric soil water contents, access polyvinyl chloride tubes were installed at 77 sites on this hillslope (Fig. 1). A portable time-domain reflectrometry TRIME-PICO-IPH soil moisture probe (IMKO, Ettlingen, Germany) was employed to measure the soil water contents of these 77 sites on 43 dates from January 2013 to March 2016. Considering the non-saline and non-viscous soils on the study hillslope, the factory-set calibration curve was adopted to infer the volumetric soil water contents from the measured dielectric constant. However, the uncertainty of measurements could still exist, which provided by the factory was $\pm$ 0.03 $m^3\,m^{-3}$, and provided by Cosh et al. (2016) was $\pm$ 0.05 $m^3\,m^{-3}$. Before the campaign on each survey date, the TRIME-PICO-IPH probe was adjusted in buckets with dry and saturated beads following the standard procedure in the user manual. Soil water contents were measured at the depth interval of 0- to 20-cm and 20- to 40-cm (representing the soil water contents at the depths of 10- and 30-cm, respectively). More detailed information in soil water content measurements is presented in Lai et al. (2017).

2.3. Soil properties and terrain attributes

Soil samples at 0- to 20-cm and 20- to 40-cm depth intervals were collected around each soil water content access tube. After air dried, weighted, ground and sieved, these soil samples were used to determine the rock fragment contents (RFC), soil clay content (clay), silt content (silt) and sand content (sand), and the organic matter (OM) (Lai et al. 2017). In addition, the depths to bedrock (DB) of all 77 sites were also determined when installing the access tubes and taking soil samples using a hand auger.

A high spatial resolution (1 m × 1 m) digital elevation model (DEM) of the study hillslope was obtained based on the 1:1000 contour map. Terrain attributes including elevation, slope, plane curvature (PLC), profile curvature (PRC), and topographic wetness index (TWI) were