SPATIAL AND TEMPORAL PATTERNS OF SOIL MOISTURE: A STUDY ON SOIL MOISTURE OBSERVATION AND MODELING

By

Jingnuo Dong

B.S.
China Agricultural University
Beijing, China
2010

M.S.
Oklahoma State University
Stillwater, OK, USA
2013

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
Doctor of Philosophy
May 2020

SPATIAL AND TEMPORAL PATTERNS OF SOIL MOISTURE:
A STUDY ON SOIL MOISTURE OBSERVATION AND MODELING

Thesis Approved:

Tyson Ochsner (Dissertation Advisor)

Phillip Alderman

H.K. Dai

Ye Liang

ACKNOWLEDGMENTS

I would love to express my gratitude to my advisor Dr. Tyson E. Ochsner for his research advice, great patience, and keeping open-minded in the selection of research methods and topics.

I would like to thank my committee members Dr. H.K. Dai, Dr. Ye Liang, and Dr. Phillip Alderman for their comments and suggestions on my dissertation.

I am grateful for all the support provided by soil physics group members - Philip Pope, William Wedge, Dr. Jason Patton, Dr. Andres Patrignani, Destiny Kerr, Dr. Briana Wyatt, and Haley VanVleet.

Special thanks to my 2017 CSSS "memory" project group members, especially the group leader Dr. Yao Liu.

I would like to thank my friends Dr. Weiqi Cui, Yang Jiao, Henry Han, and Yu Zhong for their special support in the past few years.

This acknowledgment is limited to the help that is directly related to this dissertation. I would also thank the people that may have contributed to this dissertation indirectly but I didn't list above. It is too brief and my gratitude is always hard to articulate in these short passages.

Acknowledgments reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: Jingnuo Dong

Date of Degree: May 2020

Title of Study: Spatial and temporal patterns of soil moisture: a study on soil moisture observation and modeling

Major Field: SOIL SCIENCE

Abstract: This dissertation addressed three research questions related to observing and modeling soil moisture spatial and temporal patterns. 1) What is the spatial pattern like for soil moisture at the mesoscale? A mobile device was used to measure soil moisture along a 150-km transect 18 times over 13 months. Spatial structures of soil moisture, sand content, and precipitation were characterized and compared. Soil texture turned out to exert a stronger influence than precipitation on mesoscale soil moisture patterns. 2) How can we effectively upscale in situ soil moisture measurements to a larger scale? A phase space analysis method was applied to upscale soil moisture from the point-scale to field-scale, and its performance was evaluated and compared to traditional scaling methods - linear regression and CDF matching. The phase space method was not able to improve the prediction accuracy compared to the two traditional scaling methods. 3) How can we discover and model hidden structures in soil moisture temporal dynamics? An approach from computational mechanics, called ϵ-machines, was applied to symbolized soil moisture time series and their first and second order derivatives. Based on the reconstructed ϵ-machines, soil moisture time series exhibit a degree of complexity hidden structures, and unpredictability. Statistical complexities tend to increase with the orders of derivatives for soil moisture processes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

General Introduction

Soil moisture is a key variable in the hydrological cycle, strongly influencing water and energy fluxes at the land surface. Due to technology developed in the past few decades, soil moisture can now be measured nondestructively with various instruments at multiple spatial scales. However, there exists a problematic scale gap between large-scale soil moisture measurements from satellite remote sensing and small-scale measurements from typical in situ sensors. Soil moisture satellites provide information with global coverage but coarse resolution (e.g. 36×36 km), and typical in situ monitoring stations, like those of the Oklahoma Mesonet, provide measurements representing only a few hundred cm^3 of soil. And, this scale gap profoundly limits soil moisture spatial estimation, drought monitoring, wildfire forecasts, and fundamental understanding of soil moisture spatial structure. There are few sources of soil moisture information at the field or watershed scale, a critical scale for land and water management decisions. To improve our understandings of these mesoscale soil moisture spatial patterns, we propose

(1) to directly observe soil moisture spatial patterns at the mesoscale, i.e. 1-100 km, using a mobile cosmic-ray neutron detector and to relate those patterns to land surface characteristics and atmospheric processes;

(2) to determine an effective upscaling process for typical in situ soil moisture sensors; and

(3) to discover and model temporal structures of soil moisture dynamics.

Each of the next three chapters is devoted to one of these research objectives.

CHAPTER 2

Soil texture often exerts a stronger influence than precipitation on mesoscale soil moisture patterns

Dong, J. & Ochsner, T. E. (2018). Soil texture often exerts a stronger influence than precipitation on mesoscale soil moisture patterns. Water Resources Research, 54, 2199-2211. https://doi.org/10.1002/2017WR021692

Abstract

Soil moisture patterns are commonly thought to be dominated by land surface characteristics, such as soil texture, at small scales and by atmospheric processes, such as precipitation, at larger scales. However, a growing body of evidence challenges this conceptual model. We investigated the structural similarity and spatial correlations between mesoscale (∼1-100 km) soil moisture patterns and land surface and atmospheric factors along a 150-km transect using 4-km multisensor precipitation data and a cosmic-ray neutron rover, with a 400-m diameter footprint. The rover was used to measure soil moisture along the transect 18 times over 13 months. Spatial structures of soil moisture, soil texture (sand content), and antecedent precipitation index (API) were characterized using autocorrelation functions and fitted with exponential models. Relative importance of land surface characteristics and atmospheric processes were compared using correlation coefficients (r) between soil moisture and sand content or API. The correlation lengths of soil moisture, sand content, and API ranged from 12-32 km, 13-20 km, and 14-45 km, respectively. Soil moisture was more strongly correlated with sand content (r = -0.536 to -0.704) than with API for all but

one date. Thus, land surface characteristics exhibit coherent spatial patterns at scales up to 20 km, and those patterns often exert a stronger influence than do precipitation patterns on mesoscale spatial patterns of soil moisture.

2.1 Introduction

It has often been argued that soil moisture patterns are dominated by land surface characteristics, such as soil series (Seyfried, 1998), soil texture (Li et al., 2014; Manns et al., 2014), and topographical attributes (Brocca et al., 2007) at small scales (roughly extents <100 m), and by atmospheric processes at larger scales (roughly extents of 0.1-1000 km) (Vinnikov et al., 1999). However, some studies have provided evidence that land surface characteristics may control soil moisture patterns at much larger scales ($\sim 10^2$ km) (Oldak et al., 2002) and that atmospheric processes could have stronger influence on a smaller scale ($\sim$10 km) than on a regional scale ($\sim$100 km) (Joshi and Mohanty, 2010). In addition, the spatial patterns of mean relative difference of soil moisture at the regional scale may be strongly correlated with soil hydraulic properties (Wang and Franz, 2015). One reason for the continuing controversy about the scales at which land surface versus atmospheric processes dominate is the fact that most of these prior studies have relied on aircraft-based remotely sensed datasets of limited duration, e.g. a few days, or on inferences from sparse in situ observations. The scarcity of suitable mesoscale observations, i.e. the scale gap, has long hindered understanding of the main drivers of spatial patterns of soil moisture at the mesoscale.

The well-known scale gap between large-scale soil moisture measurements from satellite remote sensing and small-scale measurements from in situ sensors profoundly limits soil moisture spatial estimation, soil moisture modeling, and fundamental understanding of soil moisture spatial structure and scaling (Robinson et al., 2008; Western et al., 2002). To fully understand these limitations and to develop strategies for closing the scale gap requires, among other things, a robust definition of spatial

scale. The spatial scale of a measurement set can be defined by the scale triplet, which consists of support, spacing, and extent (Bloschl and Sivapalan, 1995). The support is the footprint or sensing volume of the measurement device, the spacing is the average distance between measurement locations, and the extent is the maximum distance between any two measurement locations.

The scale gap has made it particularly difficult to accurately perceive the mesoscale spatial structure of soil moisture. In meteorology, mesoscale usually refers to atmospheric patterns with horizontal extents from a few kilometers to several hundred kilometers (Pielke, 2002). Here mesoscale refers to observations with a support of $\sim$1 km and extent of $\sim$100 km. The development of the cosmic-ray neutron method has created new opportunities for observing and understanding soil moisture at the mesoscale (Zreda et al., 2008). The count rate of fast cosmic-ray neutrons observed by an above-ground detector near the land surface is inversely related to the soil moisture in a circular footprint centered on the detector with a diameter ranging from 260-480 m (Kohli et al., 2015). The penetration depth of these neutrons in the soil decreases as the soil moisture increases, with penetration depths ranging from 15-55 cm for volumetric water content values ranging from 0.50-0.05 cm^3 cm^{-3} (Kohli et al., 2015).

Stationary fast neutron detectors have been used to monitor soil moisture in the COSMOS (COsmic-ray Soil Moisture Observing System) and CosmOz networks (Hawdon et al., 2014; Zreda et al., 2012) and in a variety of field-scale soil moisture studies, i.e. studies with extents of roughly 0.5-1 km (Baroni and Oswald, 2015; Franz et al., 2012; Lv et al., 2014; Zhu et al., 2015). But, to observe mesoscale soil moisture patterns, larger, mobile fast neutron detectors called rovers must be utilized. The first published roving survey was conducted by (Desilets et al., 2010) with a 37-km transect in Hawaii. Subsequent roving surveys were conducted to examine temporal stability of spatial patterns (Chrisman and Zreda, 2013), reconstruct spatial soil

moisture fields (Chrisman and Zreda, 2013; Dong et al., 2014; Franz et al., 2015), and compare with stationary probes (Franz et al., 2015). Recently, globally available datasets of soil properties were developed and tested to support roving surveys (Avery et al., 2016).

These prior studies show that cosmic-ray neutron rovers are effective tools for observing mesoscale soil moisture patterns. Rovers allow spatially continuous measurements of soil moisture along the travel path, which can prevent some uncertainties that could arise with other measurement techniques. Typically, with in situ soil moisture sensors the support is too small, the spacing is too large, or the extent is too small to accurately perceive mesoscale patterns, while with satellite remote sensing of soil moisture, the support is typically too large, e.g. 25-40 km (Mohanty et al., 2017). The on-the-go rover measurements have a spacing which is less than or equal to the support (0.25-1 km), depending on the travel speed and neutron count integration time. The extent of rover measurements is limited only by availability of roads and driving time.

In this study, we apply the unique cosmic-ray neutron rover technology to re-examine a persistent question in hydrology: At what scales do atmospheric versus land surface factors dominate the spatial pattern of soil moisture? In a previous rover study (Dong et al., 2014), distinctive mesoscale soil moisture patterns were observed and those patterns appeared to be related to soil texture. Therefore, soil texture and precipitation were selected for this study as examples of important land surface characteristics and atmospheric processes influencing soil moisture patterns. Clearly other factors such as vegetation type, solar radiation, and topography, to name a few, also influence spatial patterns of soil moisture, but completing an exhaustive analysis of all such factors was not the objective of this study. Rather, our objectives were: 1) to observe an extended time series of mesoscale spatial patterns in soil moisture, 2) to compare the spatial structure of soil moisture with that of land surface characteristics

(i.e. soil texture in this case) and atmospheric processes (i.e. precipitation), and 3) to determine the relative importance of land surface characteristics and atmospheric processes in defining the mesoscale spatial patterns of soil moisture.

2.2 Materials and Methods

2.2.1 Study area

The study area is in the Central Great Plains ecoregion in north central Oklahoma, USA, with a small part of the study area extending to the east into the Cross Timbers ecoregion. Data were collected from east to west spanning a 150-km long transect along public roads passing by four long-term monitoring stations of the Oklahoma Mesonet (McPherson et al., 2007), the Perkins, Marena, Marshall, and Lahoma stations (Fig.1). The roads are unpaved along most of the transect. The transect is located within the Cimarron River watershed with the average annual precipitation (2006-2015) ranging from 880 mm at the Perkins station on the eastern end to 732 mm at the Lahoma station on the western end. Land cover in this region consists primarily of pasture and rangeland dominated by warm-season grasses and cropland dominated by rainfed winter wheat, with small areas of primarily deciduous forest (Fig. 1). Soil texture ranges from sand to clay with the finest-textured soils near the middle of the transect and the coarsest-textured soils formed in alluvial sand deposits along the north side of the Cimarron River, a pattern which is reflected in the sand content map of the study area (Fig. 1). The most common soil orders are Mollisols, Alfisols, and Inceptisols in this area.

2.2.2 Rover transects

The cosmic-ray neutron rover (Hydroinnova LLC, Albuquerque, NM) used in this study consists of two pairs of 1.15-m long by 8.0-cm diameter, cylindrical metal, ^{3}He gas-filled fast neutron detectors. Each pair of detectors is shielded by 2.6-cm

of polyethylene and housed in a sealed enclosure. A control module contained in a separate case integrates a data logger (QI-DL-2100, Quaesta Instruments), a GPS receiver and a barometric pressure sensor. A neutron pulse monitor (QI NPM-2000, Quaesta Instruments) is connected to each neutron detector and sends neutron count totals to the data logger, where data are stored with a removable SD card. Mobile devices can connect to the control module via Bluetooth or USB cable to monitor the status of the rover while collecting data.

The rover was mounted in a pick-up truck and transported with an average speed of $\sim$48 km h^{-1}. This speed was slow enough to allow time to accumulate adequate neutron counts for acceptably-low uncertainty at the desired spatial resolution and fast enough that data collection for each pass along the transect was logistically manageable in a typical working day. The neutron counts were logged continuously and accumulated neutron counts were recorded at the end of each minute. Instantaneous GPS coordinates and atmospheric pressure were measured and recorded at the end of each minute as well. The stationary footprint of the rover is approximately 400-m diameter (Kohli et al., 2015). The typical travel distance associated with each one minute neutron count was 800 m given the speed of 48 km h^{-1}. So for this transect, the footprint or support of each measurement, i.e. a pixel, is about 800 m $\times$ 400 m.

Neutron counts along the transect were measured 18 times starting May 2015 and ending June 2016. All the transects followed approximately the same routes (Fig. 2.1), but the locations of the measurements on the different dates were not exactly the same because of occasional small detours due to road closures or road conditions. Moreover, the spacings between measurements within a transect were not exactly the same due to small variances in driving speed. In order to facilitate autocorrelation analysis of the resulting volumetric water content data, we preprocessed the raw data. To make the measurements of each transect evenly spaced and aligned, all the corners and the start and end of the transect were set as endpoints of a series of line segments.

Pixel centers with 800-m spacing were calculated along the segments between each pair of adjacent endpoints. The neutron counts of any original data points that had their footprints overlap with the newly generated pixel were averaged to estimate the neutron counts for that pixel on that date. The subsequent autocorrelation analysis and correlation analysis used these evenly-spaced data.

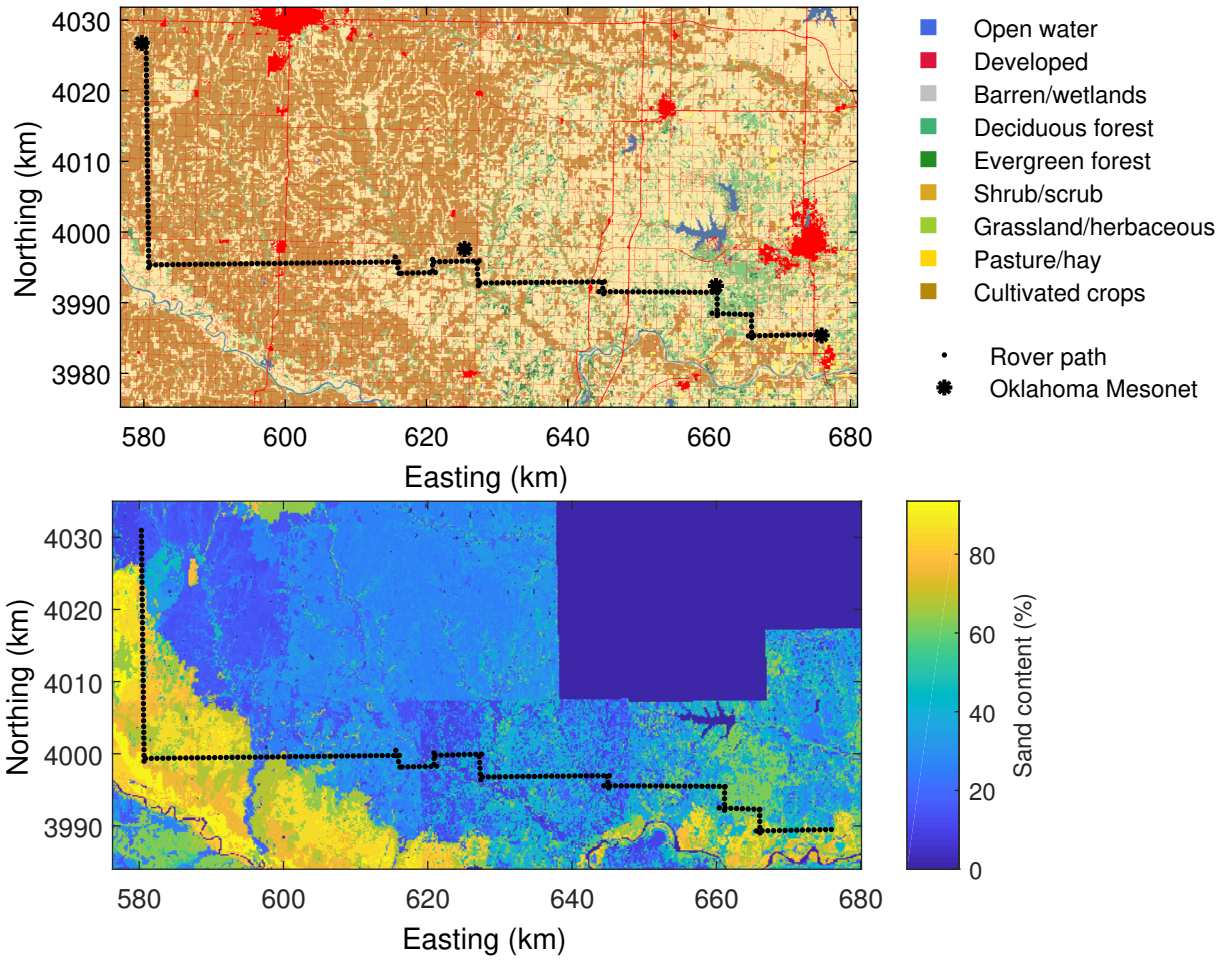


Figure 2.1: Land cover map of the (top) study area based on National Land Cover Database 2011 (NLCD 2011) and (bottom) sand content map based on SSURGO. Both maps are in the UTM coordinate system. Black dots represent the typical path of the rover transects. Locations of four calibration sites are marked in the land cover map with black stars, which from east to west are Mesonet sites at Perkins, Marena, Marshall, and Lahoma.

In order to facilitate plotting of all the transects in one figure, an additional stan-

dardization process was applied. The transect on 7 Aug. 2015 was selected as the standard transect. In theory, any date could be selected as the standard, but for convenience, we selected one of the dates with significant correlation lengths for antecedent precipitation index (API). This will be discussed in detail in the subsequent sections. The raw data from all the transects were then aligned to the standard transect using the averaging procedures described above. The data from these standardized transects were only used for visualization (Fig. 2.5) and not for any of the statistical analyses.

2.2.3 Neutron intensity correction

Neutron count rates for each transect were normalized to a reference atmospheric pressure of 98.0 kPa, which was the mean atmospheric pressure for the standard transect. Correction factors were calculated using an exponential model (Desilets and Zreda, 2003). The neutron count rates were also normalized for variability of incoming neutron flux intensity using the date of the standard transect as the reference date (Hawdon et al., 2014). Incoming neutron flux intensity data (http://www.nmdb.eu/) from the Dourbes station, which is the station with the most similar geomagnetic condition and altitude to the study area, were used in calculating the correction factors for all survey dates.

2.2.4 Rover calibration and neutron counts conversion

Four calibration campaigns were conducted in summer of 2015, one at each of the four Oklahoma Mesonet stations along the transect. During calibration, the rover was parked at the west side of each Mesonet station and neutron counts were logged for approximately 2 hours. During that time, 14 soil cores (0 – 40 cm soil layer) were taken at three radial distances around the rover, 5 m, 50 m, and 100 m. Six cores were taken within 5 m of the rover, and 4 cores at each of the other two distances.

The 3.81-cm diameter soil cores were taken using a handheld tube sampler (Giddings Machine Company, Windsor, CO) with a slide hammer on top of it. Each soil core was divided into three samples: 0-5 cm, 5-10 cm, and 10-40 cm. Gravimetric water content and bulk density for each sample was measured by oven-drying at 105∘C. The field average soil moisture was determined by weighting all the samples according to the weighting function given by (Kohli et al., 2015).

The shape-defining function (Eq.2.1) was used for calibrating the rover and converting neutron counts into gravimetric water content

$$\theta_g = \frac{a_0}{N/N_0 - a_1} - a_2 - w_{lat} \tag{2.1}$$

where θ_g (g g^{-1}) is the soil gravimetric water content, w_{lat} (g g^{-1}) is the soil lattice water content, $N$is the fast neutron count rate corrected as described above (counts per minute, cpm), $a_0 = 0.0808$, $a_1 = 0.372$, and $a_2 = 0.115$ (Desilets et al., 2010). All the calibration measurements were used with Eq. 1 to estimateN_0, which in theory is a constant of the rover and represents the neutron intensity over dry soil when all hydrogen sources within the footprint are taken into account.

Lattice water exists in the crystal lattice of minerals and is defined as the amount of water released from the soil between 105 °C and 1000 °C (Zreda et al., 2012). Lattice water can considerably affect fast neutron counts, and lattice water varies substantially in space (Zreda et al., 2012). In order to estimate lattice water content along the transect, we established a linear regression between clay content and lattice water. A total of 17 data points were used in the regression, six of which (Bondville, Neb Field 3, Brookings, Rosemount, Fort Peck, and Freeman Ranch) were from the COSMOS network Central Plains locations (Avery et al., 2016), four were from the calibration sites, and seven were from cosmic-ray neutron rover calibration campaigns conducted near El Reno and Marena, OK in 2014. Clay content for six of the points were measured from field samples, and the rest were from the SSURGO. All soil lattice

water content measurements were performed by Activation Laboratories, Ontario, Canada.

2.2.5 Land surface characteristics and atmospheric data

Soil properties, such as clay content, sand content, and bulk density, for the surface layer along the transect were retrieved from the Soil Survey Geographic Database (SSURGO). Spatial data files and the corresponding tabular data files were downloaded from the Web Soil Survey (http://websoilsurvey.sc.egov.usda.gov/). The basic component of the SSURGO spatial soil maps is called the map unit, which is a set of polygons associated with a series of soil characteristics. For each map unit along the transect, weighted averages of clay content, sand content, and bulk density were calculated based on the percent composition of all the soil series in that map unit. To calculate area-averaged values for the soil properties, raster maps with a resolution of 200 m were created based on the polygon maps. Soil properties for the standard transect were then established using the same approach described in Section 2.2. The resulting clay content for each point on the transect was used to estimate lattice water content for that point. Gravimetric water content for each point along transect on each measurement date was then determined using Eq. 1 and was converted to volumetric water content using the point-specific bulk density. The same procedures were applied in determining clay content for the regression with lattice water content, except the 6 sites of the COSMOS network, for which clay content were measured from field samples. Sand content was chosen to represent the influence of land surface characteristics on soil moisture because it strongly influences soil water retention (Minasny and McBratney, 2007).

Spatial variability in precipitation was chosen to represent the influence of atmospheric processes on soil moisture because precipitation typically shows stronger spatial variability than atmospheric variables like air temperature or humidity (Brotzge

and Richardson, 2003). Hourly multisensor (radar and rain gauge) precipitation estimates with a spatial resolution of 4 km were retrieved from National Weather Service (NWS) Arkansas-Red Basin River Forecast Center via the Oklahoma Mesonet. This multisensor product is produced using the "P3" algorithm which merges data from rain gauges with gridded radar-based rainfall estimates (Kitzmiller et al., 2013). Daily accumulated precipitation was calculated based on the hourly precipitation extracted from the images (NetCDF files). API was calculated based on the daily rainfall for the analysis of rainfall effects on soil moisture spatial patterns (Saxton et al., 1967). In this index, the influence of a prior daily precipitation total decreases exponentially as the number of subsequent days increases. The API value for each transect pixel on each transect date was calculated using precipitation data for the prior 90 d (Saxton et al., 1967) following the method of (Crow and Zhan, 2007), which accounts for the seasonality of atmospheric demand.

To observe the strength and stability of the correlations between soil moisture and sand content/API, Pearson correlation coefficients were calculated for each transect date. For each rover transect pixel, the nearest API pixel was identified and the API value was assigned to that transect pixel to construct an API transect with the "same" resolution of water content. Critical values for correlation coefficients were calculated for each pair of variables and each date to determine statistical significance.

2.2.6 Autocorrelation function

The autocorrelation function is often used to understand patterns in time series and spatial fields. The definition of an autocorrelation function is usually derived from the theoretical spatial covariance function $C_Z(h)$, which is defined as

$$C_Z(h) = E[(Z(x) - \mu)(Z(x+h) - \mu)] \tag{2.2}$$

where x denotes the location of a sample, h is the lag distance, Z is a second-

order stationary spatial process, and μ is the expectation of Z. In this study, Z could be soil moisture, sand content, or API. The theoretical autocorrelation function (correlogram) is defined as

$$\rho_Z(h) = \frac{C_Z(h)}{C_Z(0)} \tag{2.3}$$

where $C_Z(0)$ is the covariance at $h = 0$, which is actually the variance of Z. The empirical autocorrelation function is used in estimating ρ_Z from observations. This function can be written as:

$$\hat{\rho}_Z(h) = \frac{\frac{1}{|N(h)|}\sum_{(i,j)\in N(h)}(Z(x_i) - \overline{Z})(Z(x_j) - \overline{Z})}{Var(Z)} \tag{2.4}$$

where $\overline{Z}$ is the mean of Z, (i, j) denotes the pairs of samples such that $|x_i - x_j| = |h|$, $|h|$ is usually a range of lag distances, and $|N(h)|$ is the number of pairs in each set.

In (Vinnikov et al., 1999), a linear combination of exponential functions was used to characterize the spatial covariance in soil moisture attributable to land surface characteristics and atmospheric processes. As is shown in Eq. 3, the autocorrelation function is proportional to the covariance function, so an exponential model (Eq. 5) was selected to fit the empirical autocorrelation function for key spatial variables in this study,

$$\rho(h) = a\exp(-h/L) \tag{2.5}$$

where a is a constant, and L is often called the correlation length, which has a similar physical meaning to the range in variogram models (Western et al., 2004). When h is large relative to L, autocorrelation $\rho(h)$ tends to zero (uncorrelated).

The empirical autocorrelation $\hat{\rho}_Z$ for real data can often be negative for some lag distances, which prevents fitting of an exponential model. We used significance bands

to determine at what lag distance the autocorrelation is not significantly different from zero, and only the part of the autocorrelation function that was significantly greater than zero was fitted with the exponential model. A linear regression between $\ln \rho(h)$ and h was used in fitting the model. A generalized Bartlett's formula (Francq and Zakoian, 2009), which does not assume the spatial process is linear, was chosen to calculate 95% significance bands,

$$\hat{\rho}_Z(h) \pm 1.96\sqrt{\frac{1 + C_{Z^2}(h)/C_Z^2(0)}{n}} \tag{2.6}$$

where C_{Z^2} is the autocorrelation function for Z^2, and 1.96 is the 0.975 quantile of normal distribution. Prior to autocorrelation analyses, all the data (volumetric water content, sand content, and API) were normalized to the range of [0,1] with the feature scaling approach, $Z' = (Z - Z_{\min})/(Z_{\max} - Z_{\min})$.

2.3 Results and Discussion

2.3.1 Lattice water correction and N_0 calibration

The regression of clay content and lattice water is shown in Fig 2.2. The clay content varied from 14.6 % to 34.2 % with minimum values located at Marena, OK (Cross Timbers Experimental Range) and maximum values located at Freeman Ranch, TX. The corresponding lattice water content values varied from 0.020 to 0.052 $\mathrm{cm}^3\ \mathrm{cm}^{-3}$. Clay content and lattice water content were linearly related with $r^2 = 0.699$, which is greater than the r^2 (0.539) of a similar regression analysis for 24 samples from Mollisols in the continental U.S. in (Avery et al., 2016). The lattice water can strongly affect the cosmic-ray neutron intensity (Zreda et al., 2012) and would be difficult to measure for all locations along the transect. In contrast, clay content can be estimated from existing databases.

Using this linear relationship, lattice water content was approximated from clay

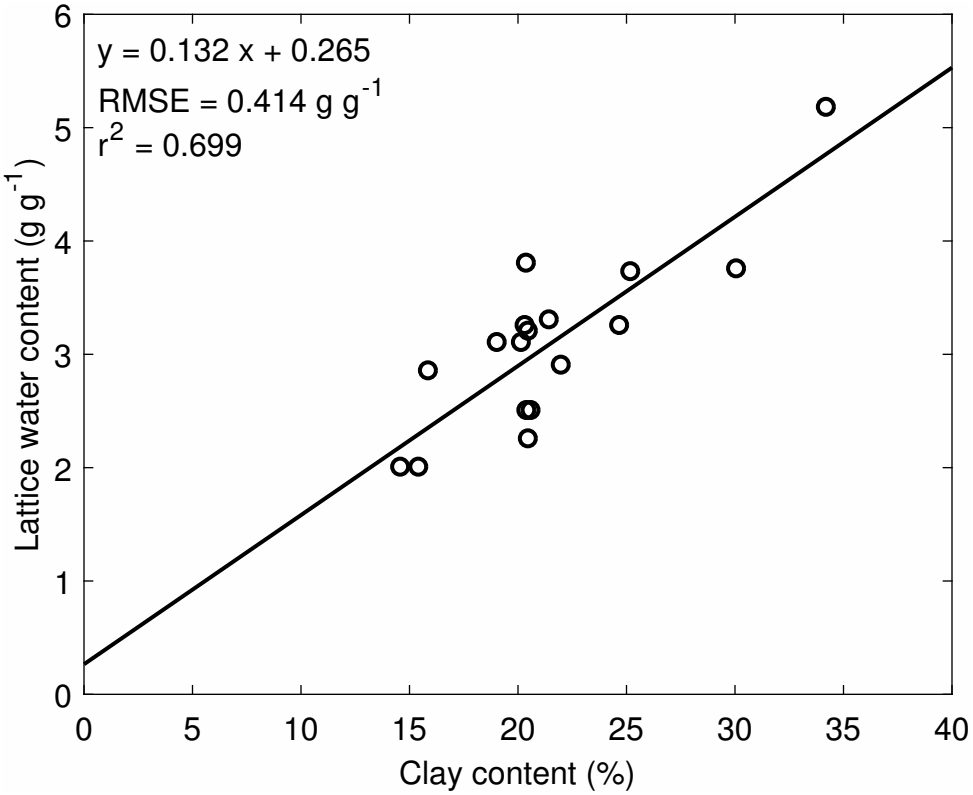


Figure 2.2: Linear regression between clay content and lattice water content including six samples from the COSMOS network Central Plains locations, four from the calibration sites, and seven from prior rover calibration campaigns conducted near El Reno and Marena, OK

content and treated as a variable along the transect (Fig. 2.3). The spatially-variable lattice water content correction is designed specifically to minimize the effect of soil texture on the rover-based soil moisture observations. The estimated gravimetric lattice water content varies between 1-4% along the transect (Fig. 2.3). For a bulk density of 1.4 g cm^{-3}, these lattice water levels correspond to a volumetric water content correction from 1.4-5.6%, a range of 4.2% (0.042 cm^3 cm^{-3}). If we had used one average lattice water value for the transect, then there would have been a soil texture influence equal to half of that range, i.e. $\pm$ 0.021 cm^3 cm^{-3}.

The calibrated shape defining function and the calibration data are plotted in Fig 2.4. The circles represent the weighted field-average soil moisture measurements at the four calibration sites. The sampling protocol produced relatively precise estimates

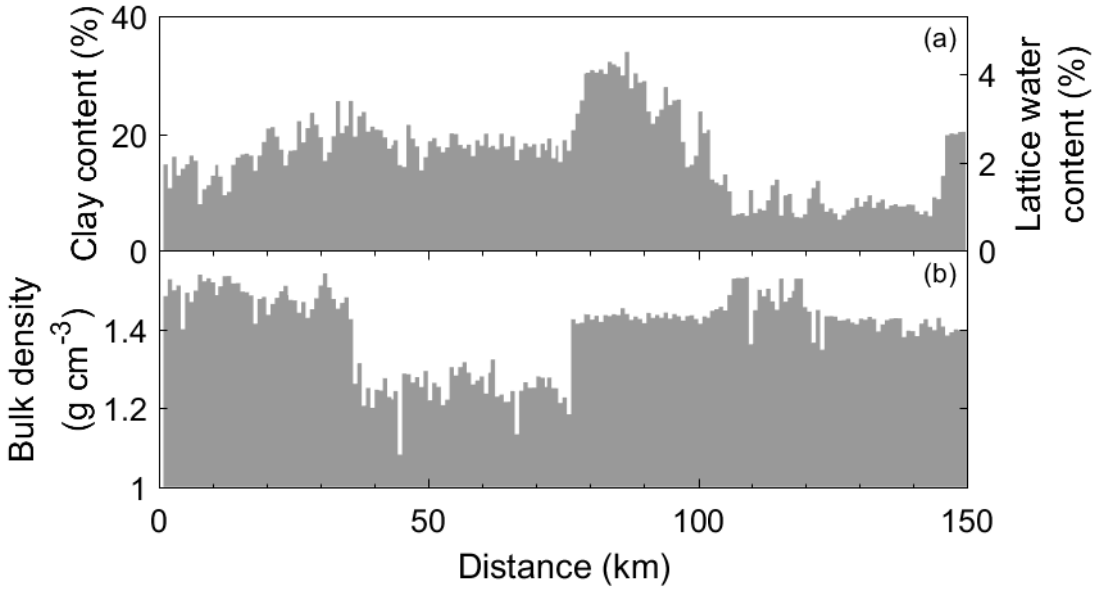


Figure 2.3: Variation of (a) clay content (left y axis), lattice water (right y axis), and (b) bulk density for the surface layer of soil along the transect at 800 m resolution.

of the field-average soil moisture at each site, with the standard error of the mean $<$ 0.011 g g^{-1} at all four locations. The shape defining function provided an excellent fit to the calibration data with $r^2 = 0.994$, which is slightly better than the calibration $r^2 = 0.966$ in the study of Dong et al. (2014). The calibrated value of $N_0$is 556 cpm, which is reasonable given that the maximum neutron count value we have observed with this rover is >500 cpm. The calibrated N_0 represents the neutron intensity over dry soil with vegetation water content similar to that of the calibration sites on the calibration dates. The effects of spatial or temporal variability in vegetation were neglected in this study, and the low RMSE of the calibration (0.006 g g^{-1}) shows that any differences in vegetation among the calibration sites were not large enough to introduce substantial errors.

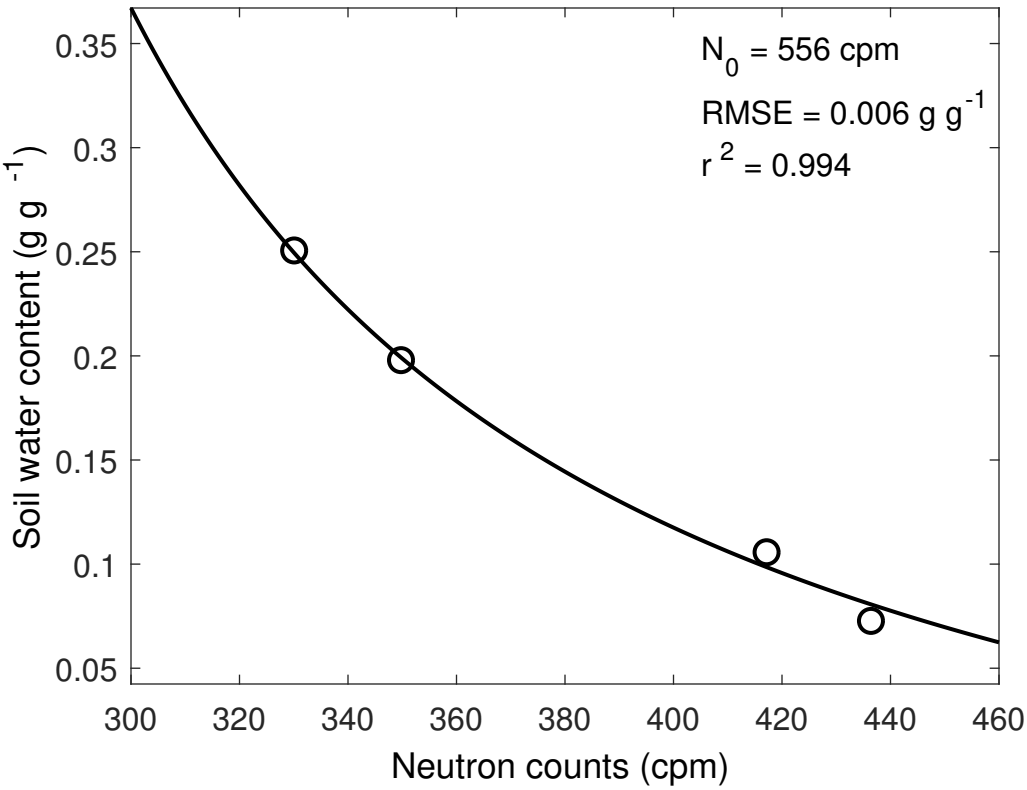


Figure 2.4: Calibration curve for neutron intensity versus gravimetric water content. Circles represent weighted averages of gravimetric water content for each calibration site.

2.3.2 Corrected neutron counts and volumetric water contents

Corrected neutron counts and the corresponding volumetric water contents for all the transects are shown in Fig. 2.5. Relatively high neutron counts and low water contents were consistently observed at distances between 100-150 km along the transect (Fig. 2.5). The maximum, mean and minimum neutron counts observed across the whole study were 521, 373, and 275 cpm. The corresponding minimum, mean, and maximum soil water content are 0.0113 cm^3 cm^{-3}, 0.106 cm^3 cm^{-3}, and 0.437 cm^3 cm^{-3}. The uncertainty in the volumetric water content data can be assessed, in part, by considering the uncertainty in the underlying neutron counts. The counting uncertainty (coefficient of variation) depends on the count number (N) by Poisson statistics and is given by $N^{-0.5}$, which results in an uncertainty of $\sim$5% at the ob-

served mean neutron count rate of 373 cpm (Zreda et al., 2012). This is similar to the uncertainty reported by (Chrisman and Zreda, 2013) using a different rover. The corresponding uncertainty in soil water content is approximately $\pm$ 0.03 g g^{-1} at the observed mean neutron count rate.

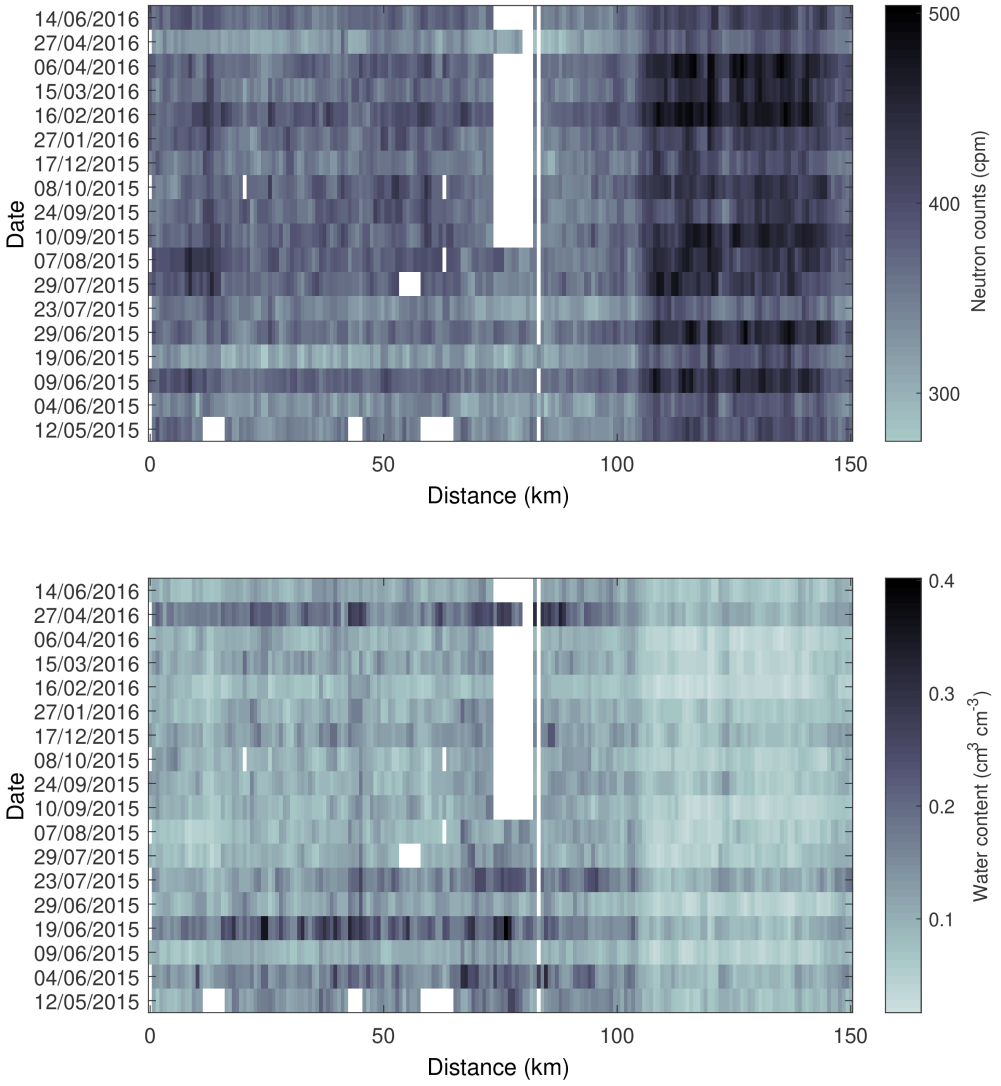


Figure 2.5: (top) Corrected neutron count rates and (bottom) volumetric water content along the transect for all dates. All transects were standardized to represent the same locations as the standard transect (7 August 2015). White patches show the locations and dates that no data were collected, primarily due to road closures.

2.3.3 Spatial patterns

The standard transect (7 Aug. 2015) was selected as an example to show how volumetric water content, sand content, and API vary along the transect (Fig. 2.6). There were no visually obvious and consistent similarities between patterns of API and soil water content along the transect across the measurement dates. Since the resolution of the precipitation data is 4 km, the API transect is coarser than the other two variables. For 7 Aug. 2015, API generally increases from the start of the transect to the end, but the soil water content shows no such east to west trend. Rather the highest soil water contents occurred at transect distances from 25-100 km, with lower water contents at each end of the transect. The general trends of sand content and volumetric water content were roughly inverse. This was most obvious at the distance of 100-150 km, where sand content was consistently high and water content was low across all transect dates.

The autocorrelation functions for volumetric water content, sand content, and API with 95% significance bands are plotted for 7 Aug. 2015 in Fig. 7. The autocorrelation values between the two significance bands are not significantly different from zero. For this transect date, significant positive autocorrelations existed for volumetric water content for lag distances < 15 km (Fig. 2.7a). As the lag distances increased, the autocorrelation decreased and became negative, suggesting a relatively large scale trend or cycle. Similarly, sand content displayed positive autocorrelation at lags $<$ 25 km and became negatively autocorrelated at larger lag distances (Fig. 2.7b). Since the number of samples of API was less than that of the other variables (spatial resolution: 4 km versus 800 m), the significance bands were wider than those for the other variables and fewer significant autocorrelations were observed (Fig. 2.7c). On this transect date, significant positive autocorrelation existed for API at lags < 30 km, but, unlike for volumetric water content and sand content, significant negative autocorrelations were not observed for API on this date.

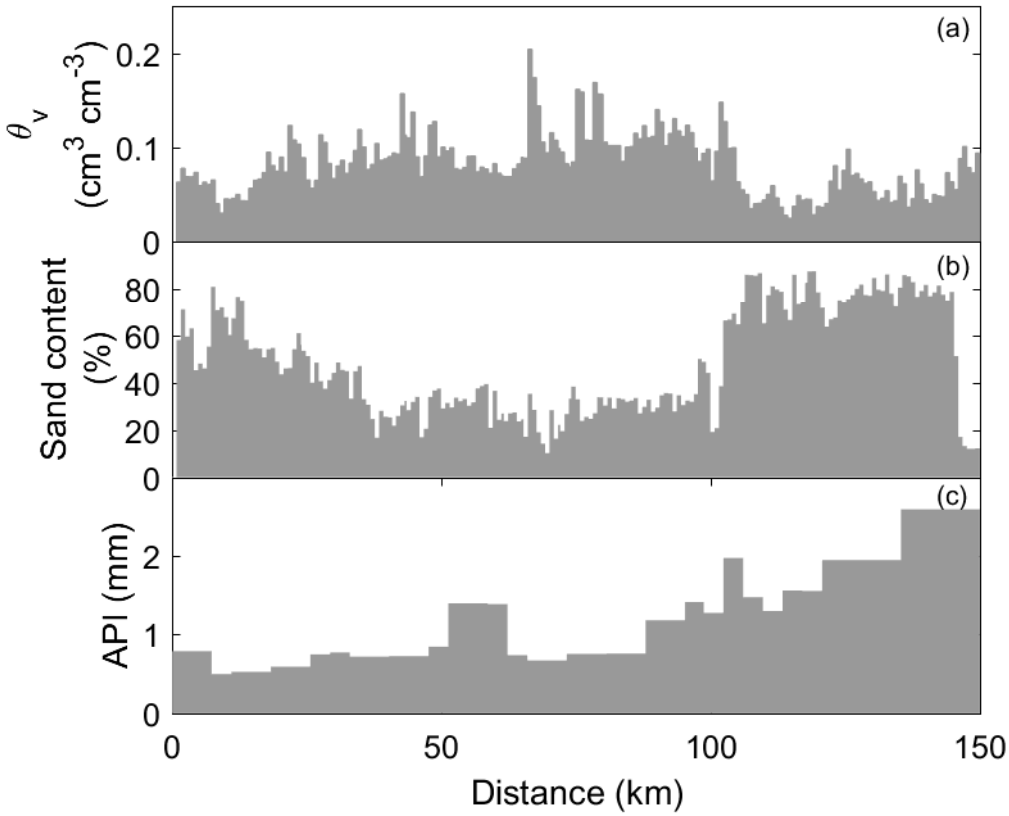


Figure 2.6: Variation of (a) volumetric water content and (b) sand content for the surface layer of soil along the transect on 7 August 2015 at 800 m resolution. Variation of the antecedent precipitation index, API, along the transect at 4 km resolution (c).

In order to estimate the correlation lengths, exponential models were applied for each transect date to fit the part of each autocorrelation function that was significantly greater than zero (Fig. 2.8). For sand content and API, the r^2 values of the fitted models were > 0.9 for all dates. The range of $r^2$for volumetric water content was 0.693-0.973. This example from the transect on 7 Aug. 2015 shows the general relationship of correlation lengths between soil water content, sand content, and API. The correlation lengths are reflected as the reciprocal of the slopes in Fig. 2.8. For the dates when API exhibited positive autocorrelations, it usually had the largest correlation length. Soil moisture typically had the smallest correlation length, which was often close to the correlation length for sand content (Fig. S3).

The y-intercept of the autocorrelation function may be interpreted as indicating

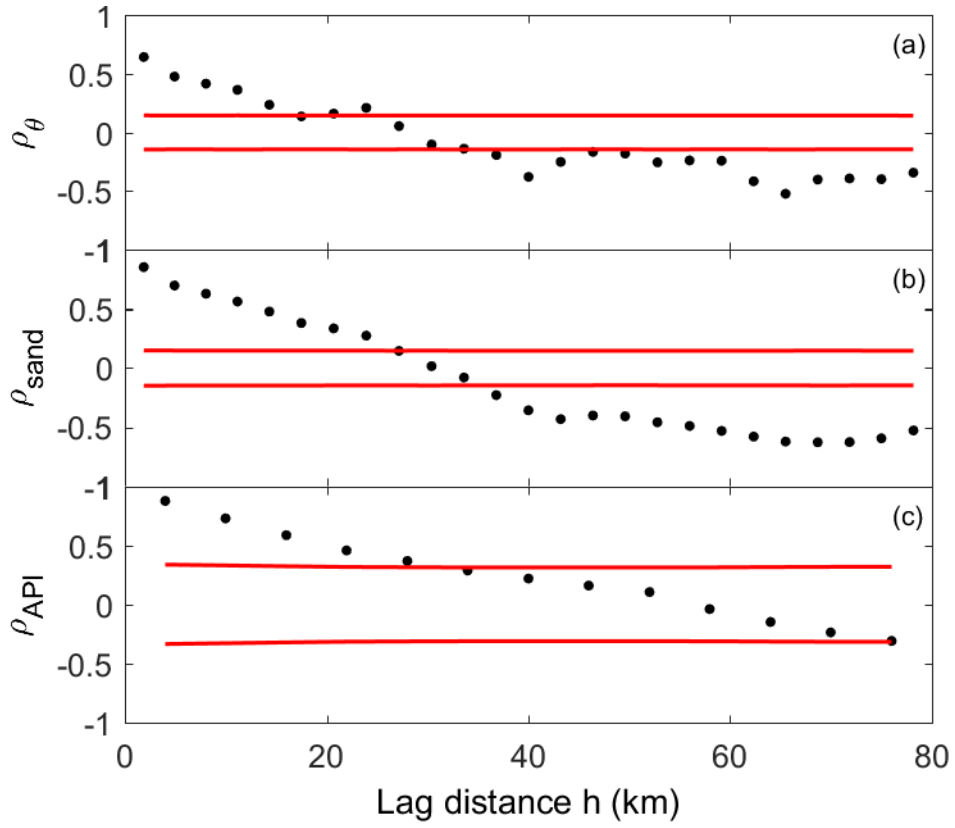


Figure 2.7: Spatial autocorrelation functions for (a) volumetric water content and (b) sand content of the surface layer of soil and for API (c) along the transect on 7 August 2015. Data points between the two significance bands (red lines) are not significantly different from zero.

the proportion of the variance associated with the random errors in the measurements (Vinnikov et al., 1999). For example, Fig. 2.8 shows an intercept of about -0.3 in log space for the soil moisture data. The proportion of the variance associated with random errors in the measurement is then approximately 1-$e^{-0.3}$ = 0.26. The soil moisture variance for that particular transect date was ∼0.002 (cm^3 cm^{-3})^2. Thus, the random error in the soil moisture measurements was approximately $\sqrt{0.26 \cdot 0.002}$ = 0.02 cm^3 cm^{-3}. The majority of this random error is likely due to the uncertainty in the neutron counts as described in Section 3.2.

A summary of the estimated correlation lengths is given in Table 2.1. The estimation is based on the dates with the fitted exponential models having $r^2 \geq 0.95$, since when poorer model fit occurs the error of the estimated correlation length is

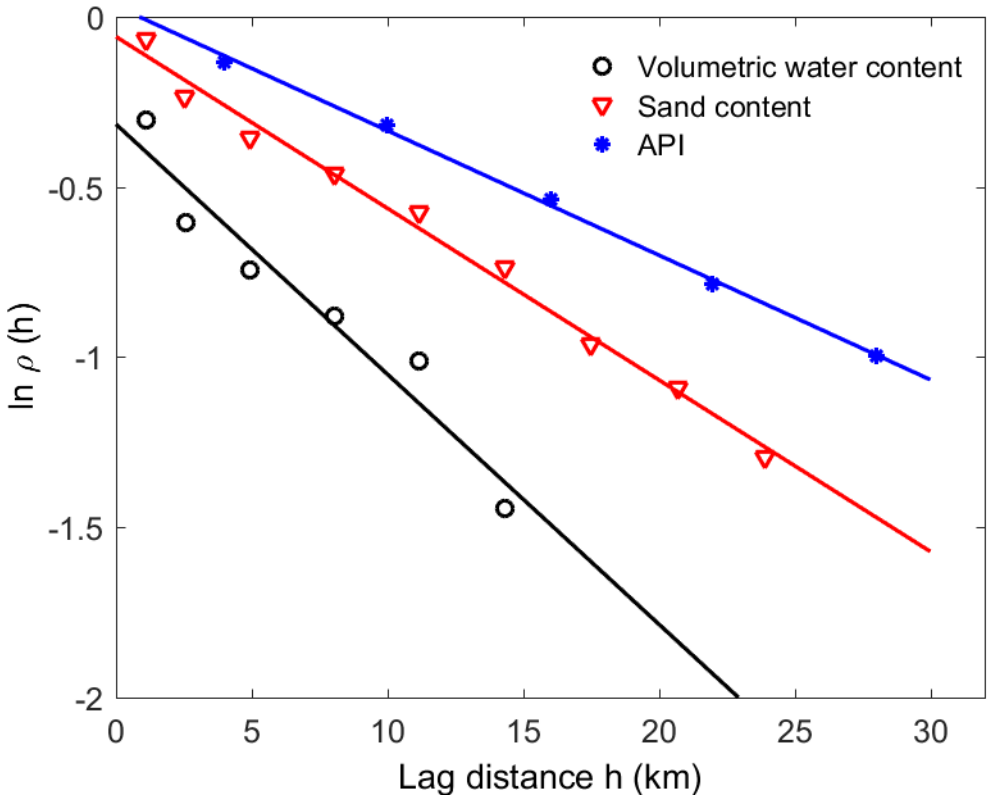


Figure 2.8: Natural-log of the autocorrelation functions for volumetric water content and sand content of the soil surface layer and for API on 7 August 2015. Only correlations significantly greater than zero are included. The solid lines are the best fit exponential functions.

amplified after transforming back from log-space. Volumetric water content displayed significant positive autocorrelation for all transect dates. Because the transect routes were not exactly the same each time, the correlation lengths for sand content for each date were different. The minimum and median correlation lengths for volumetric water content are essentially equal to those for sand content, indicating similarity in the spatial structures of land surface characteristics and soil moisture patterns at spatial scales up to 20 km.

Table 2.1: Percentage of transect dates for which significant autocorrelations were observed based on the 95% significance bands; percentage of transect dates with acceptable fit for the exponential model ($r^2 \geqslant 0.95$); and minimum, median, and maximum correlation length for volumetric water content, sand content, and antecedent precipitation index (API)

	Dates with significant autocorrelation	Dates with acceptable fit ($r^2 \geqslant 0.95$)	Minimum	Median	Maximum
	%	%		km	
Volumetric water content	100	50	12	17	32
Sand content	100	89	13	18	20
API	89	72	14	25	45

The maximum correlation length for volumetric water content was 32 km, which is greater than that of sand content (20 km) and smaller than that of API (45 km). This result is consistent with the hypothesis that the spatial autocorrelation of volumetric water content is also influenced by atmospheric processes. The median correlation length for API was 25 km and the maximum was 45 km, indicating that the spatial scale of variability in atmospheric processes was typically larger than that for soil water content or sand content. However, the minimum correlation length for API, 14 km, was similar to the minimum correlation lengths for water content and sand content, proving that atmospheric processes reflected in the API data do not always vary at larger spatial scales than land surface characteristics or soil moisture, but rather, in some cases, the scales of these spatial processes are intermingled. We also analyzed the data with water content and sand content resampled at the spatial scale of API, and none of the main findings were affected (data not shown). Coarsening the soil moisture and sand content data to 4-km had no obvious effects on the correlation lengths, but it did decrease the precision with which the correlation lengths could be estimated.

Prior studies have promoted the conceptual model that the spatial patterns of soil moisture are controlled primarily by land surface characteristics, such as soil texture,

at relatively small spatial scales and by atmospheric processes, such as precipitation, at relatively large spatial scales (Kim and Barros, 2002; Ryu and Famiglietti, 2006; Vinnikov et al., 1999). For example, (Vinnikov et al., 1999) proposed an ideal spatial covariance function $C(h) = \sigma_s^2 \exp(-h/L_s) + \sigma_a^2 \exp(-h/L_a)$, suggesting that the spatial variabilities of soil moisture should consist of a land-surface-related component (correlation lengths of 10-20 m) and an atmospheric-forcing-related component (correlation lengths approximately 400-500 km). The results in Table 2.1 suggest that this conceptual model needs to be revised to reflect the existence of spatial patterns in land surface characteristics, particularly soil texture, at the mesoscale.

The soil moisture correlation lengths found here overlap substantially with the 10-30 km range of the variograms reported for 0-5 cm soil moisture in the Southern Great Plains 1997 Hydrology Experiment (Ryu and Famiglietti, 2006). That experiment resulted in 16 days of soil moisture data generated by aircraft remote sensing using the Electronically Scanned Thinned Array Radiometer (ESTAR) at 800-m resolution over a 30-d period in the summer. The SGP97 experiment covered west central Oklahoma and included some of our study area. The spatial correlations in the 10-30 km length scale observed in SGP97 were attributed to spatial patterns of soil texture, just as in our study. Longer scale soil moisture correlations represented by variogram ranges from 60-100 km were also observed during SGP97 and were attributed to large-scale rainfall events. In our study, the correlation length of API varied from 14-45 km and was not a dominant influence on the correlation length of soil moisture.

2.3.4 Relative importance of land surface characteristics and atmospheric processes

Correlation coefficients between sand content and soil water content, and between API and soil water content illustrate the relative importance of land surface characteristics and atmospheric processes in controlling the mesoscale patterns of soil moisture

(Fig. 2.9). For better visual comparisons, the absolute values of the correlation coefficients were plotted. For all the survey dates, sand content was negatively correlated with soil water content, with correlation coefficients ranging from -0.536 to -0.704. The correlation coefficients between API and volumetric water content show more irregular patterns with positive correlations on most dates, but unexpected negative correlations on five dates. All the correlation coefficients were statistically significant ($\alpha = 0.05$). Soil volumetric water content was more strongly correlated with sand content than with API for all but one survey date.

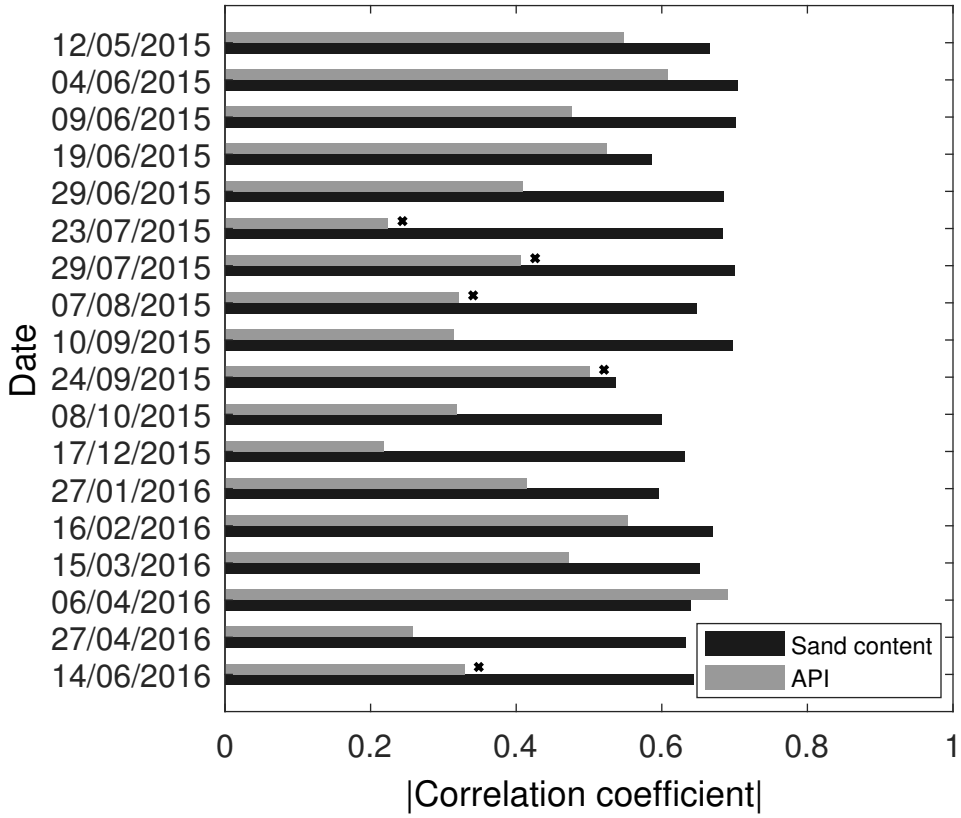


Figure 2.9: Absolute values of correlation coefficients between soil water content and sand content (black), and between soil water content and API (grey) for all survey dates. All the correlation coefficients of sand content were negative. Most of the correlation coefficients between soil water content and API were positive, but some were negative, and these were marked with crosses.

The scale mismatch between API and the other variables does affect the correlation coefficients to some degree. The water content-sand content correlations were slightly

stronger at 4-km resolution than at 800-m resolution (data not shown). Also, the statistical significance of the correlation coefficients is affected because the degrees of freedom for sand content and soil water content are ∼180 (number of samples - 1), but ∼40 for API. If we specify 40 degrees of freedom, the three correlation coefficients of API with the smallest absolute values would become insignificant. Clearly, there could be value in gridded precipitation data with higher spatial resolution than 4 km, but the scale mismatch between API and the other variables does not alter the main findings of this study.

In previous field scale studies with extents <600 m, significant negative correlations have been observed between soil water content and sand content with correlation coefficients ranging from -0.586 to -0.795 (Hu and Si, 2013) and from -0.62 to -0.66 (Gomez-Plaza et al., 2001). Likewise, analysis of ESTAR data from the SGP'97 experiment, which encompassed our study area, revealed scale-dependent correlations between sand content and soil water content with correlation coefficients ranging from -0.08 to -0.68 (Kim and Barros, 2002). The strength of these correlations generally increased as the spatial support of the data was increased from 0.7 km^2 to 408 km^2. Our results strengthen the growing body of evidence that soil texture strongly influences spatial patterns of soil moisture not only at the field scale and below, but also at the mesoscale. Furthermore, our results indicate that this influence persists and is relatively stable throughout the year.

API has been widely used to represent soil water content in satellite remote sensing studies (Jackson and LeVine, 1996) based on time series comparisons between API and soil water content measured at a point (Teng et al., 1993). The irregular and sometimes illogical correlations between soil water content and API found in this study indicate that the spatial influence of precipitation on soil water content patterns is less stable and more complex compared to that of soil texture. Underappreciated uncertainties may arise when applying the time-domain based API in

the spatial domain. Our findings are consistent with the results of recent analyses of in situ soil moisture data from Nebraska, Utah, Michigan, Oklahoma, and the southeastern US, all of which have shown that soil texture more strongly influences regional soil moisture patterns than does precipitation (Wang et al., 2017a,b). The correlations observed here are further evidence that soil texture exerts a stronger and more consistent influence on mesoscale soil moisture patterns than does precipitation.

2.3.5 Limitations of the study

Several limitations of this study are worth noting. First, it is important to realize that all the correlation lengths estimated here, and in any study with real data, are influenced by the support, spacing, and extent of the observations. All other things being equal, the apparent correlation lengths tend to increase as the spacing, support, or extent increase, with the effect of extent being most important (Western and Bloschl, 1999). Relatively unbiased estimates of correlation length should be obtained if the spacing is less than two times the true correlation length, the support is less than 20% of the true correlation length, and the extent is larger than five times the "true" correlation length (Western and Bloschl, 1999). Unfortunately, we have no independent way of knowing the true correlation length, so these guidelines cannot easily be applied.

The correlation lengths for sand content and API reported in this study and the correlation coefficients relating those variable to soil water content are also dependent upon the accuracy of the underlying sand content and precipitation data. We could find no published study on the accuracy of the precipitation data from the P3 algorithm, but one study reported a +5% bias of an earlier P1 algorithm in Oklahoma based on hourly data (Young et al., 2000) and another study reported a -5 to -10% bias for radar-based, long-term mean areal precipitation in Oklahoma (Johnson et al., 1999). There has also been limited published research to evaluate the accuracy of the

SSURGO soil texture data. Drohan et al. (2003) reported that field measurements of soil texture were within the ranges of the SSURGO estimates for 25 out of 30 forested plots in Pennsylvania, but provided no information on the width of those ranges. Thus, the available evidence suggests that the sand content and precipitation data employed in this study are relatively reliable.

Another potential limitation of this study is the possibility of uncorrected influences of extraneous factors on the neutron counts recorded by the rover. For example, qualitative inspection of prior rover data in Nebraska suggested a bias due to the influence of dry gravel roads (Franz et al., 2015), and thus the transect data in this survey could also have a similar dry bias. Likewise, spatial or temporal variations in vegetation water content along the transect could introduce small calibration errors in the rover-based soil water content estimates (Avery et al., 2016). The cosmic-ray neutron rover method is still relatively new and clearly has scope to be further refined with use, but at present there is no evidence to suggest that measurement errors arising from the above factors could change the main conclusions of the study.

2.4 Conclusion

Understanding of soil moisture spatial variability at the mesoscale, i.e. ∼1-100 km, has long been hindered by the scale gap between in situ soil moisture sensors and satellite soil moisture products. In this study, the scale gap was overcome by using a cosmic-ray neutron rover to observe mesoscale spatial patterns of soil moisture along a 150-km transect over a period of 13 months. These data allowed us to re-examine the scales at which atmospheric versus land surface factors dominate the spatial pattern of soil moisture. The land surface characteristics, as reflected in soil texture, exhibited correlation lengths ranging from 13 to 20 km and significant negative autocorrelation at lags > 40 km, patterns which were also reflected in the soil moisture observations. Atmospheric processes, as represented by API, showed spatial structure that was less

stable over time, with correlation lengths ranging from 14 to 45 km, and sometimes showed no significant spatial autocorrelation at the mesoscale. Furthermore, soil texture (i.e. sand content) was more strongly correlated with soil moisture than was API for 17 out of 18 dates in this study.

The mesoscale spatial patterns that the rover captured in this study were generally consistent, in terms of the spatial autocorrelation structure, with patterns previously observed using airborne remote sensing in the US Great Plains. Further research is needed to evaluate the scales at which atmospheric versus land surface factors dominate the spatial pattern of soil moisture in other regions. This will likely depend on the scales and degrees of spatial variability in the key atmospheric and land surface factors, such as precipitation and sand content. For example, API may exhibit smaller correlation lengths in mountainous regions where orographic precipitation dominates (Daly et al., 1994) than in the Great Plains, and thus, the correlations between API and soil moisture may differ. Likewise, sand content may exhibit less spatial variability in regions such as the US Midwest (e.g. Illinois, Indiana, Ohio) (Miller and White, 1998), and as a result, different relationships between soil moisture and sand content may exist. Although such regional differences are plausible, the strong influence of soil texture on soil moisture patterns discovered here is consistent with the results of a growing number of regional and national scale analyses [e.g. (Kim and Barros, 2002; Wang et al., 2017a,b)].

The mesoscale spatial patterns of soil moisture were apparently largely controlled by soil texture along this transect, and the patterns in soil texture exhibited a larger spatial scale than has been indicated in some previous studies, a scale that substantially overlapped with that of mesoscale atmospheric processes. The original dataset (18 rover transect files) and the corresponding soil moisture transect files are included in a public project (https://osf.io/59j6c/) under the Open Science Framework (OSF). In addition, that project includes a 200-m resolution sand content map for the study

area and scripts for preprocessing rover transect files. Detailed data file descriptions and algorithm descriptions are included in the OSF project page. Further analyses of these data are expected to yield additional insights into mesoscale soil moisture patterns and their controls and to provide a unique validation data set for soil moisture estimates from land surface and hydrologic models.

Acknowledgments

This work was funded by the U.S. National Science Foundation EPSCoR program (Grant no. 131789) and by the USDA National Institute of Food and Agriculture Hatch Project and the Division of Agricultural Sciences and Natural Resources at Oklahoma State University. The rover data and supporting information are available through the Open Science Framework at https://osf.io/59j6c/.

References

Avery, W. A., Finkenbiner, C., Franz, T. E., Wang, T. J., Nguy-Robertson, A. L., Suyker, A., Arkebauer, T., and Munoz-Arriola, F. (2016). Incorporation of globally available datasets into the roving cosmic-ray neutron probe method for estimating field-scale soil water content. *Hydrology and Earth System Sciences*, 20(9):3859–3872.

Baroni, G. and Oswald, S. E. (2015). A scaling approach for the assessment of biomass changes and rainfall interception using cosmic-ray neutron sensing. *Journal of Hydrology*, 525:264–276.

Bloschl, G. and Sivapalan, M. (1995). Scale issues in hydrological modeling - a review. *Hydrological Processes*, 9(3-4):251–290.

Brocca, L., Morbidelli, R., Melone, F., and Moramarco, T. (2007). Soil moisture

spatial variability in experimental areas of central italy. *Journal of Hydrology*, 333(2-4):356–373.

Brotzge, J. A. and Richardson, S. J. (2003). Spatial and temporal correlation among oklahoma mesonet and oasis surface-layer measurements. *Journal of Applied Meteorology*, 42(1):5–19.

Chrisman, B. and Zreda, M. (2013). Quantifying mesoscale soil moisture with the cosmic-ray rover. *Hydrology and Earth System Sciences*, 17(12):5097–5108.

Crow, W. T. and Zhan, X. W. (2007). Continental-scale evaluation of remotely sensed soil moisture products. *Ieee Geoscience and Remote Sensing Letters*, 4(3):451–455.

Daly, C., Neilson, R. P., and Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of applied meteorology*, 33(2):140–158.

Desilets, D. and Zreda, M. (2003). Spatial and temporal distribution of secondary cosmic-ray nucleon intensities and applications to in situ cosmogenic dating. *Earth and Planetary Science Letters*, 206(1-2):21–42.

Desilets, D., Zreda, M., and Ferre, T. P. A. (2010). Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resources Research*, 46.

Dong, J., Ochsner, T. E., Zreda, M., Cosh, M. H., and Zou, C. B. (2014). Calibration and validation of the cosmos rover for surface soil moisture measurement. *Vadose Zone Journal*, 13(4).

Drohan, P., Ciolkosz, E., and Petersen, G. (2003). Soil survey mapping unit accuracy in forested field plots in northern pennsylvania. *Soil Science Society of America Journal*, 67(1):208–214.

Francq, C. and Zakoian, J. M. (2009). Bartlett's formula for a general class of non-linear processes. *Journal of Time Series Analysis*, 30(4):449–465.

Franz, T. E., Wang, T. J., Avery, W., Finkenbiner, C., and Brocca, L. (2015). Combined analysis of soil moisture measurements from roving and fixed cosmic ray neutron probes for multiscale real-time monitoring. *Geophysical Research Letters*, 42(9):3389–3396.

Franz, T. E., Zreda, M., Rosolem, R., and Ferre, T. P. A. (2012). Field validation of a cosmic-ray neutron sensor using a distributed sensor network. *Vadose Zone Journal*, 11(4).

Gomez-Plaza, A., Martinez-Mena, M., Albaladejo, J., and Castillo, V. M. (2001). Factors regulating spatial distribution of soil water content in small semiarid catchments. *Journal of Hydrology*, 253(1-4):211–226.

Hawdon, A., McJannet, D., and Wallace, J. (2014). Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across australia. *Water Resources Research*, 50(6):5029–5043.

Hu, W. and Si, B. C. (2013). Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition. *Geoderma*, 193:180–188.

Jackson, T. J. and LeVine, D. E. (1996). Mapping surface soil moisture using an aircraft-based passive microwave instrument: Algorithm and example. *Journal of Hydrology*, 184(1-2):85–99.

Johnson, D., Smith, M., Koren, V., and Finnerty, B. (1999). Comparing mean areal precipitation estimates from nexrad and rain gauge networks. *Journal of Hydrologic Engineering*, 4(2):117–124.

Joshi, C. and Mohanty, B. P. (2010). Physical controls of near-surface soil moisture across varying spatial scales in an agricultural landscape during smex02. *Water Resources Research*, 46.

Kim, G. and Barros, A. P. (2002). Space-time characterization of soil moisture from passive microwave remotely sensed imagery and ancillary data. *Remote Sensing of Environment*, 81(2-3):393–403.

Kitzmiller, D., Miller, D., Fulton, R., and Ding, F. (2013). Radar and multisensor precipitation estimation techniques in national weather service hydrologic operations. *Journal of Hydrologic Engineering*, 18(2):133–142.

Kohli, M., Schron, M., Zreda, M., Schmidt, U., Dietrich, P., and Zacharias, S. (2015). Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. *Water Resources Research*, 51(7):5772–5790.

Li, T., Hao, X. M., and Kang, S. Z. (2014). Spatiotemporal variability of soil moisture as affected by soil properties during irrigation cycles. *Soil Science Society of America Journal*, 78(2):598–608.

Lv, L., Franz, T. E., Robinson, D. A., and Jones, S. B. (2014). Measured and modeled soil moisture compared with cosmic-ray neutron probe estimates in a mixed forest. *Vadose Zone Journal*, 13(12).

Manns, H. R., Berg, A. A., Bullock, P. R., and McNairn, H. (2014). Impact of soil surface characteristics on soil water content variability in agricultural fields. *Hydrological Processes*, 28(14):4340–4351.

McPherson, R. A., Fiebrich, C. A., Crawford, K. C., Elliott, R. L., Kilby, J. R., Grimsley, D. L., Martinez, J. E., Basara, J. B., Illston, B. G., Morris, D. A., Kloesel, K. A., Stadler, S. J., Melvin, A. D., Sutherland, A. J., Shrivastava, H., Carlson, J. D., Wolfinbarger, J. M., Bostic, J. P., and Demko, D. B. (2007). Statewide

monitoring of the mesoscale environment: A technical update on the oklahoma mesonet. *Journal of Atmospheric and Oceanic Technology*, 24(3):301–321.

Miller, D. A. and White, R. A. (1998). A conterminous united states multilayer soil characteristics dataset for regional climate and hydrology modeling. *Earth interactions*, 2(2):1–26.

Minasny, B. and McBratney, A. B. (2007). Estimating the water retention shape parameter from sand and clay content. *Soil Science Society of America Journal*, 71(4):1105–1110.

Mohanty, B. P., Cosh, M. H., Lakshmi, V., and Montzka, C. (2017). Soil moisture remote sensing: State-of-the-science. *Vadose Zone Journal*, 16(1).

Oldak, A., Jackson, T. J., and Pachepsky, Y. (2002). Using gis in passive microwave soil moisture mapping and geostatistical analysis. *International Journal of Geographical Information Science*, 16(7):681–698.

Pielke, R. (2002). *Mesoscale Meteorological Modeling*. Academic Press.

Robinson, D. A., Campbell, C. S., Hopmans, J. W., Hornbuckle, B. K., Jones, S. B., Knight, R., Ogden, F., Selker, J., and Wendroth, O. (2008). Soil moisture measurement for ecological and hydrological watershed-scale observatories: A review. *Vadose Zone Journal*, 7(1):358–389.

Ryu, D. and Famiglietti, J. S. (2006). Multi-scale spatial correlation and scaling behavior of surface soil moisture. *Geophysical Research Letters*, 33(8).

Saxton, K., Asce, A., Lenz, A., and Asce, F. (1967). Antecedent retention indexes predict soil moisture. *Journal of the Hydraulics Division, ASCE*, 93(HY4):223–241.

Seyfried, M. (1998). Spatial variability constraints to modeling soil water at different scales. *Geoderma*, 85(2-3):231–254.

Teng, W., Wang, J., and Doraiswamy, P. (1993). Relationship between satellite microwave radiometric data, antecedent precipitation index, and regional soil moisture. *International Journal of Remote Sensing*, 14(13):2483–2500.

Vinnikov, K. Y., Robock, A., Qiu, S., and Entin, J. K. (1999). Optimal design of surface networks for observation of soil moisture. *Journal of Geophysical Research-Atmospheres*, 104(D16):19743–19749.

Wang, T. and Franz, T. E. (2015). Field observations of regional controls of soil hydraulic properties on soil moisture spatial variability in different climate zones. *Vadose Zone Journal*, 14(8).

Wang, T. J., Franz, T. E., Li, R. P., You, J. S., Shulski, M. D., and Ray, C. (2017a). Evaluating climate and soil effects on regional soil moisture spatial variability using eofs. *Water Resources Research*, 53(5):4022–4035.

Wang, T. J., Liu, Q., Franz, T. E., Li, R. P., Lang, Y. C., and Fiebrich, C. A. (2017b). Spatial patterns of soil moisture from two regional monitoring networks in the united states. *Journal of Hydrology*, 552:578–585.

Western, A. W. and Bloschl, G. (1999). On the spatial scaling of soil moisture. *Journal of Hydrology*, 217(3-4):203–224.

Western, A. W., Grayson, R. B., and Bloschl, G. (2002). Scaling of soil moisture: A hydrologic perspective. *Annual Review of Earth and Planetary Sciences*, 30:149–180.

Western, A. W., Zhou, S. L., Grayson, R. B., McMahon, T. A., Bloschl, G., and Wilson, D. J. (2004). Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes. *Journal of Hydrology*, 286(1-4):113–134.

Young, C. B., Bradley, A. A., Krajewski, W. F., Kruger, A., and Morrissey, M. L. (2000). Evaluating nexrad multisensor precipitation estimates for operational hydrologic forecasting. *Journal of Hydrometeorology*, 1(3):241–254.

Zhu, Z. L., Tan, L., Gao, S. G., and Jiao, Q. S. (2015). Observation on soil moisture of irrigation cropland by cosmic-ray probe. *Ieee Geoscience and Remote Sensing Letters*, 12(3):472–476.

Zreda, M., Desilets, D., Ferre, T. P. A., and Scott, R. L. (2008). Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophysical Research Letters*, 35(21).

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T. E., and Rosolem, R. (2012). Cosmos: the cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences*, 16(11):4079–4099.

Supplemental materials

This file includes additional figures showing the empirical spatial autocorrelation functions for volumetric water content and antecedent precipitation index (API) for each transect date, as well as the best fit exponential models for the autocorrelation of volumetric water content, sand content, and API for each date. All the detailed information about the data and the processing steps were described in the paper.

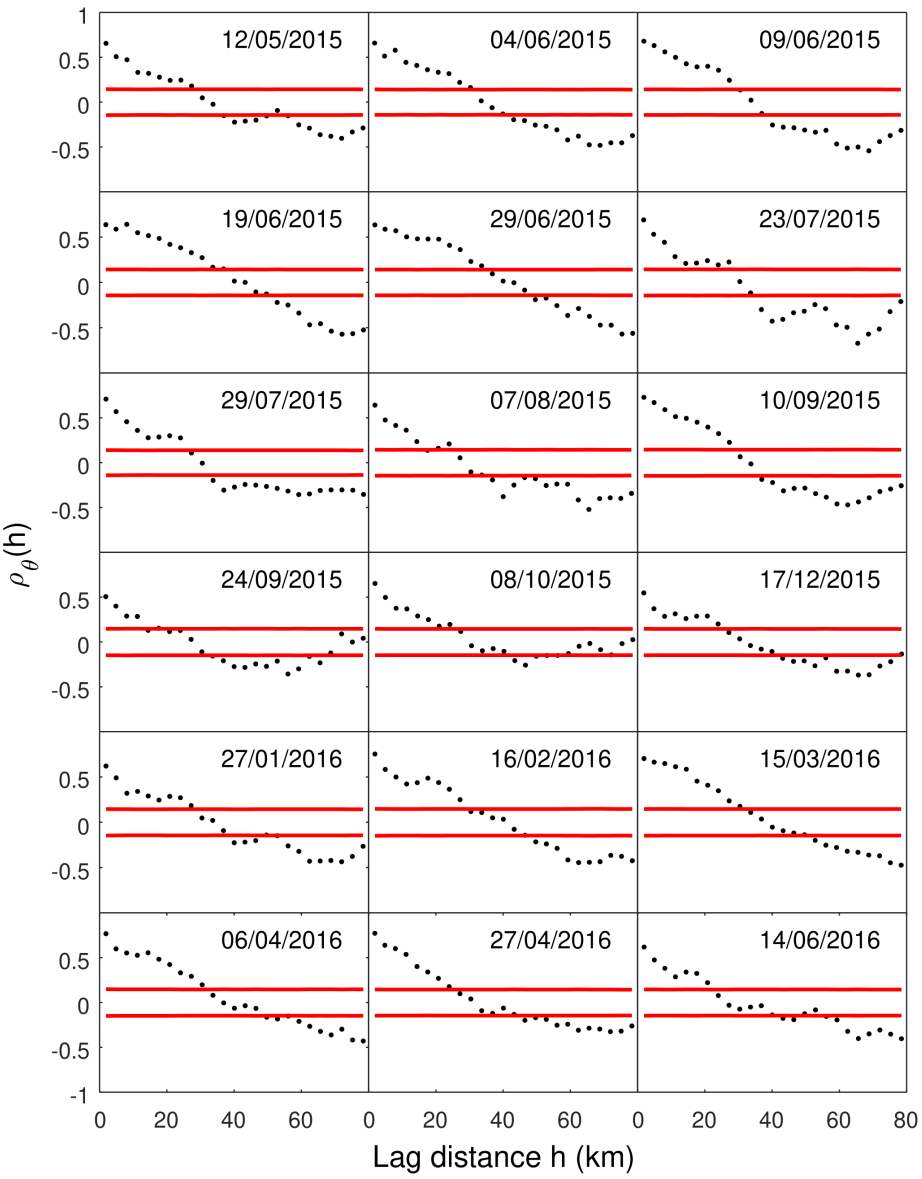


Figure S1: Spatial autocorrelation functions for volumetric water content for all dates. Data points between the two significance bands (red lines) are not significantly different from zero.

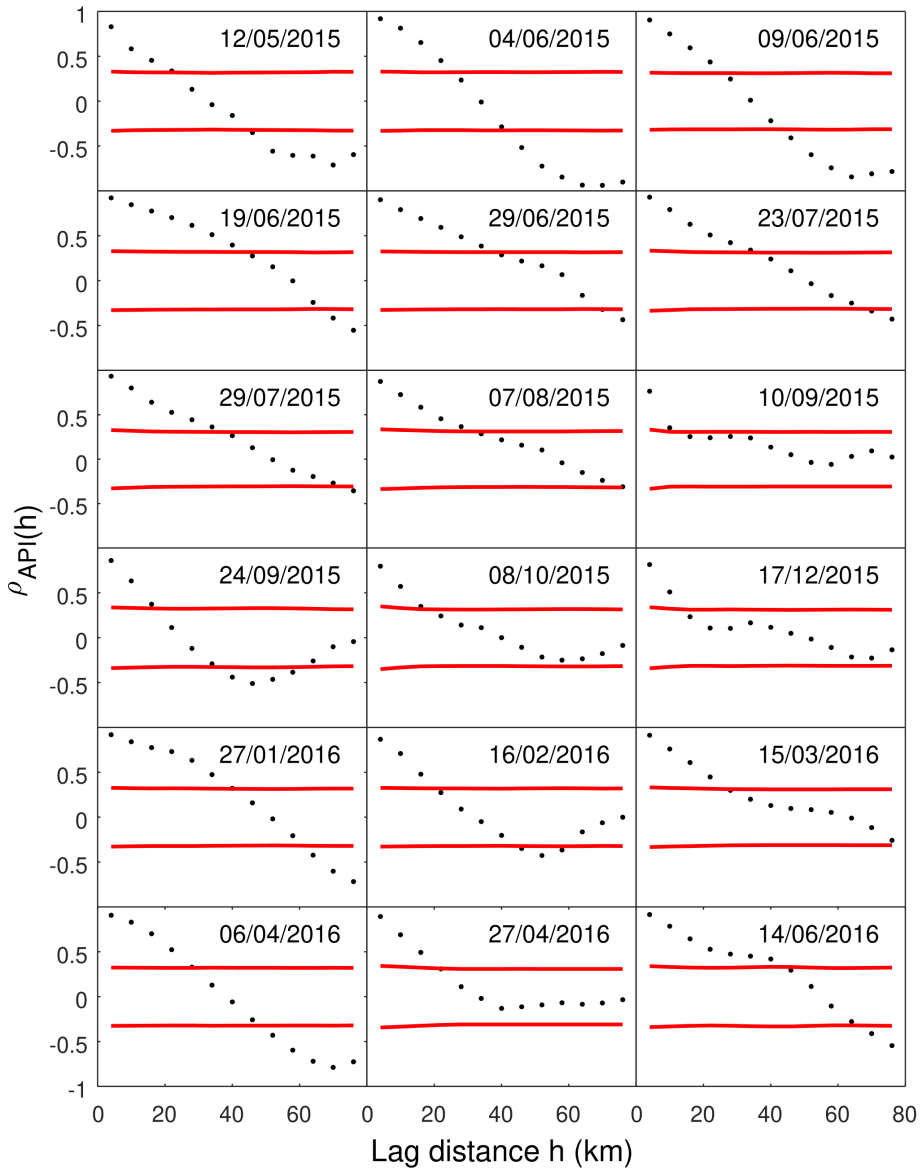


Figure S2: Spatial autocorrelation functions for antecedent precipitation index (API) for all dates. Data points between the two significance bands (red lines) are not significantly different from zero.

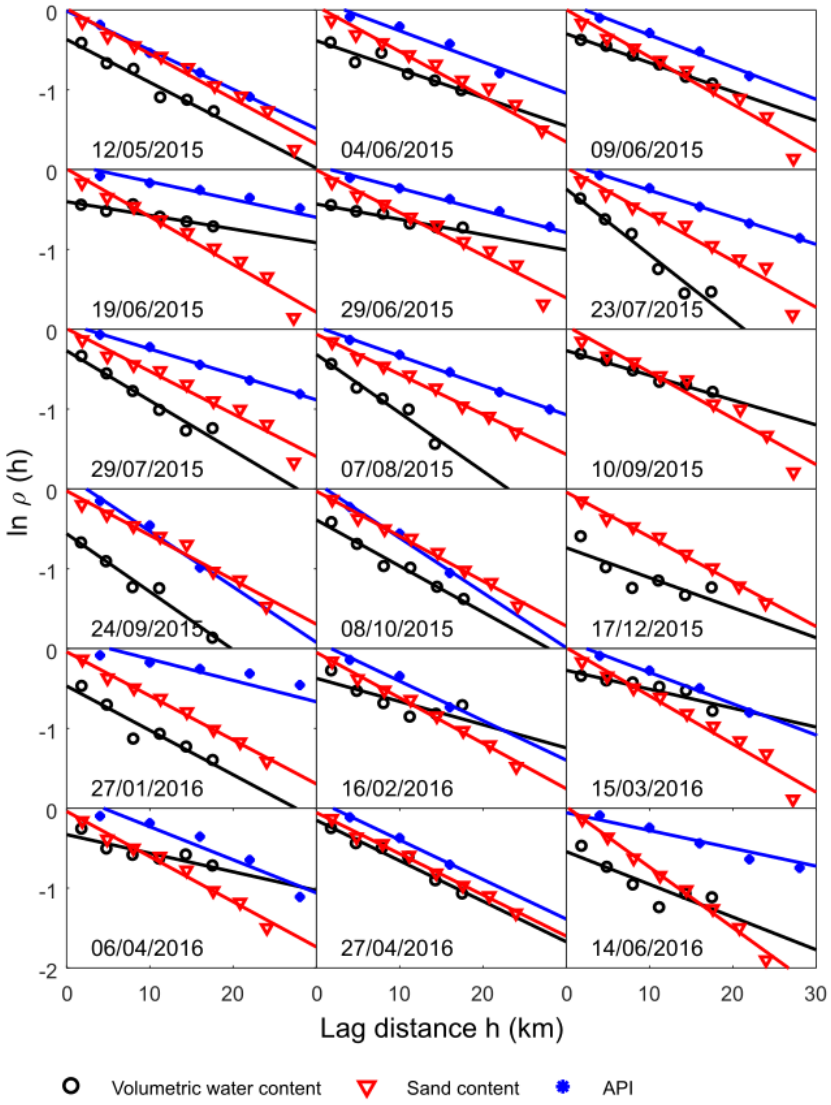


Figure S3: Natural-log of the autocorrelation functions for volumetric water content and sand content of the soil surface layer and for API on all dates. Only correlations significantly greater than zero are included. The solid lines are the best fit exponential functions.

CHAPTER 3

Upscaling of in situ soil moisture measurements based on phase space analysis

3.1 Abstract

In studies of soil moisture, upscaling and point-scale measurements is crucial for understanding spatial structures of soil moisture, yet the existing upscaling methods have various limitations. This research aims to develop a new upscaling method for point-scale soil moisture measurements based on phase space analysis. Soil moisture data at the two different spatial scales were obtained from in situ monitoring stations and co-located cosmic-ray neutron sensors. The soil moisture phase space was reconstructed from the observed data, and the relationship between the field-scale soil moisture and the phase space representation of the point-scale soil moisture was represented by a local polynomial map, which was applied to upscale soil moisture from the point-scale to field-scale. The performance of this new method was evaluated and compared to traditional scaling methods - linear regression and CDF matching. Upscaling soil moisture using the local polynomial map proved to be possible, but this new method failed to improve the prediction accuracy compared to linear regression and CDF matching.

3.2 Introduction

Soil moisture strongly influences mass and energy exchange at the land surface. At the mesoscale, this is specifically reflected in affecting atmospheric circulation (Ookouchi et al., 1984) and storm initiation (Taylor et al., 2011). In the past few decades,

observation of soil moisture and other meteorological variables at the mesocale (∼1-100 km) has often relied on in situ monitoring networks (McPherson et al., 2007). However, the relatively low spatial density of sites within existing in situ networks has greatly hindered our ability to monitor mesoscale spatial patterns of soil moisture. For example, the Oklahoma Mesonet, which is one of the most dense mesoscale in situ networks, has an average distance of 32 km between sites (Elliott et al., 1994). This is much larger than the footprint ($\sim$ 10 cm in length) of a typical in situ soil moisture sensor. This large discrepancy between the spacing and support of in situ observations, coupled with the relatively large small-scale variability in soil moisture, make it difficult to find spatial correlations and make spatial predictions, i.e. maps, based on the data from existing in situ networks (Ochsner et al., 2019). One necessary step toward solving this problem is to upscale the point measurements of the in situ networks to represent the spatial mean of the field in which they are located, thereby reducing the influence of small-scale spatial variability.

Current soil moisture upscaling methods can generally be classified into two categories. The first category assumes a linear relationship exists between the point-scale measurements of soil moisture within a particular field or watershed and the field-scale soil moisture for that field or watershed, i.e.

$$\theta_{(f)} = \sum_{i=1}^{n} a_i \cdot \theta_{(p)i} + b. \tag{3.1}$$

where $\theta_{(p)i}$ represents point-scale measurements for each of i locations within a field having a spatial mean, $\theta_{(f)}$, a_i and b are constants, and $\sum_i a_i = 1$. Because one often cannot measure the areal soil moisture, the arithmetic or weighted average of $\theta_{(p)i}$ is often used to estimate the areal soil moisture $\theta_{(f)}$. This means that Eq. (3.1) reduces to $\theta_{(f)} = \sum_i a_i \theta_{(p)i}$, and $\sum_i a_i = 1$. With intensive sampling, a linear relationship between two scales can be established successfully in some cases (De Rosnay et al., 2009).

One of the most widely used linear upscaling methods is the time stability analysis, which can be considered as a special case of this category. Representative sites are selected to replace the field-scale measurement, which means that Eq. (3.1) reduces to $\theta_{(f)} = \theta_{(p)rep}$ (Vachaud et al., 1985). This selection process implicitly assumes stationarity or even ergodicity existing in both spatial and temporal soil moisture processes (Appx. A).

Another major category of upscaling methods assumes the probability distributions of time series of a point-scale measurement and of the field-scale measurement are similar. The basic idea for scaling is thus to "stretch" the point-scale probability distribution so that the transformed distribution has a similar shape as that of the field-scale. One example of this approach is the cumulative distribution function (CDF) matching method (Eq. 3.2).

$$cdf(\theta_{(f)}) = cdf(\theta_{(p)}) \tag{3.2}$$

CDF matching has been widely used in both upscaling and downscaling. In soil moisture studies, CDF matching has been used for reducing systematic differences between two data sets errors (Drusch et al., 2005; Reichle and Koster, 2004).

These upscaling methods consider soil moisture in a one-dimensional phase-space, which means that any two numerically identical values of soil moisture in a time series are indistinguishable. However, the phase-space of soil moisture may have higher dimensions because the dynamics of soil moisture are theoretically nonlinear (Rodriguez-Iturbe et al., 1991). For nonlinear time series, scalar measurements, such as the value of soil moisture at a given time, are projections of some unobserved variables onto the real axis (Kantz and Schreiber, 2004). We hypothesize that upscaling relationships based on approximations of the higher-dimensional variables will be more accurate than directly mapping between the original one-dimensional time series. This hypothesis follows the logic that more sophisticated structures revealed

in higher-dimensional phase space may lead to more sophisticated and accurate upscaling relations.

In this study, the areal soil moisture data were measured using cosmic-ray neutron probes (CRNPs). CRNPs are unique upscaling tools with an ideal footprint size suitable to provide field-scale measurements. CRNPs are also non-destructive instruments that can be easily installed nearby existing point-scale soil moisture stations. The weaknesses of CRNPs are that their footprint and measurement depth are functions of soil moisture and they don't measure homogeneously within the footprint. These facts can make it difficult to interpret the data and may create inaccuracies in the upscaling process (Kohli et al., 2015). Furthermore, hourly CRNP soil moisture time series are usually noisy so temporal smoothing or aggregating is often needed.

We propose to develop and evaluate a phase space approach to soil moisture upscaling with the aid of CRNPs using concepts from the field of nonlinear dynamics. Previously, the approaches of nonlinear dynamics have been applied to prediction of ocean water levels (Frison et al., 1999a,b) and hydrologic systems (Doscher, 1997), but not yet soil moisture. CRNPs have been used for observing meso-scale soil moisture patterns (Franz et al., 2016; Hawdon et al., 2014; Zhu et al., 2014) but not yet for upscaling in situ soil moisture measurements. The objectives of this study are to (1) develop an upscaling method for in situ soil moisture measurements using a phase space approach; and (2) compare its performance to two common upscaling approaches: linear regression and CDF matching.

3.3 Materials and Methods

3.3.1 Phase space reconstruction

We assume that the soil moisture spatial field is a continuous stochastic field. Measuring soil moisture in space with a finite support volume means to "coarse-grain" the continuous field into different discrete levels, i.e. different resolutions. The sup-

port/footprint of the measurement device defines the resolution, which we call *scale* in this research. For example, a soil moisture sensor installed in an in situ network usually has a footprint of several tens or hundreds of square centimeters. This soil moisture sensor coarse-grains the continuous soil moisture field into a discrete field with a resolution equal to its footprint. Since its footprint is usually small compared to the size of the study area that we are interested in, the scale of its measurements can be called "point-scale". In this research, we focus on two scales - point-scale (resolution: $10^1 - 10^2$ cm^2) and field-scale (resolution: $10^4 - 10^6$ m^2), and try to find the relationship between these two scales.

A *phase-space* is the set of all possible states of a dynamical system. In this study, the system refers to the water contained in a certain soil matrix. A *state* in the system's phase space can be defined by the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$, $\boldsymbol{\theta} \in R^n$, where n is the number of dimensions of the state space and $\boldsymbol{\theta}$ can be any variable of the system, e.g. soil moisture. As the dynamical system evolves in time, the vector $\boldsymbol{\theta}$ leaves a *trajectory* representing a state's dynamic. If trajectories with any number of initial conditions converge to the same sub-region in the phase space then an attractor exists in the dynamic (Doscher, 1997). One of the main themes of nonlinear dynamics approaches is to determine phase-space structures, like attractors, from observations. One strategy is to reconstruct the phase-space directly from time series observations using Takens embedding theorem (Takens, 1981). The basic idea is to construct an d-dimensional vector $\boldsymbol{\theta}$ for each time, t, using a series of equally spaced samples in the past. This can be written as

$$\boldsymbol{\theta}(t) = [\theta(t), \theta(t-\tau), \theta(t-2\tau), ..., \theta(t-(d-1)\tau)] \tag{3.3}$$

where τ is the delay time or lag (Bradley and Kantz, 2015).

The optimal time lag τ is usually determined by finding the first minimum of the average mutual information (Appx. B) of the time series. Mutual information,

I, characterizes the mutual dependence of two random variables. In this research, we applied the method introduced by Fraser and Swinney (1986) of using mutual information to determine the time lag τ. The mutual information of a time series is defined as a function of τ

$$I(\tau) = \sum_{\theta(t),\theta(t-\tau)} P[\theta(t), \theta(t-\tau)] \cdot \log_2 \left[\frac{P[\theta(t), \theta(t-\tau)]}{P[\theta(t)]P[\theta(t-\tau)]} \right] \tag{3.4}$$

where $P[\theta(\,\cdot\,)]$ is the probability of θ , and $P[\theta(t), \theta(t-\tau)]$ is the joint probability of θ with time lag τ. $I(\tau)$ was calculated with a Matlab function in MATS-Toolkit which was developed by the EEG Analysis group, AUTh.

To estimate the embedding dimension, d, an algorithm called false nearest neighbor (FNN) is often used (Kennel et al., 1992). FNN assumes that two points far away in the actual state-space can appear close together when an embedding dimension lower than the true system dimension is used. Those two points are then called false neighbors. The aim of FNN is to find the smallest d so that if two points are neighbors in the d-embedding, they are also neighbors in $(d+1)$-embedding. Specifically, the percentage of false nearest neighbors were calculated for each possible embedding dimension d. Two criteria were devised by Kennel et al. (1992) to examine if nearest neighbors are false. The first one is defined as

$$\sqrt{\frac{R_{d+1}^2(t) - R_d^2(t)}{R_d^2(t)}} > R_{tol} \tag{3.5}$$

where $R_d^2(t)$ is the square of the Euclidian distance between a point $\boldsymbol{\theta}(t)$ and its nearest neighbor in the d-embedding dimensional space. When there is a big difference between $R_{d+1}^2(t)$ and $R_d^2(t)$, i.e. the criterion is larger than R_{tol}, the point $\boldsymbol{\theta}(t)$ and its nearest neighbor are determined to be false nearest neighbors. In this research, R_{tol} was set to 10, which was numerically tested by Kennel et al. (1992) .

In practice, when the number of data points is limited, real neighbors could be

distant. Thus, Eq. 3.5 may not be adequate for distinguishing between real and false neighbors when the data points are sparse. A second criterion was defined to avoid this kind of misjudgement:

$$\frac{R_{d+1}(t)}{R_A} > A_{tol} \tag{3.6}$$

where R_A is the size of the attractor in the d-embedding dimensional space, which is approximated by the value of the standard deviation of the data set. The threshold A_{tol} is set to be 2 with the following reason. If the two neighbors are distant ($R_{d+1}(t) \approx R_A$), increasing the embedding dimension from d to $d+1$ could approximately double the distance $R_{d+1}(t)$(Kennel et al., 1992). If either of these two criteria are met, the nearest neighbor is determined to be false. All data points and their nearest neighbors were tested , and the percentage of pairs of neighbors that meet both criteria were thus calculated to determine the appropriate embedding dimension.

3.3.2 The upscaling for soil moisture

In this study, we define scaling as a mapping of states from one scale to another. It can be expressed as a transformation $\boldsymbol{\phi}$

$$\boldsymbol{\phi} : R^{d_p} \rightarrow R^{d_f}, \quad \boldsymbol{\theta_{(f)}} = \boldsymbol{\phi}(\boldsymbol{\theta_{(p)}}) \tag{3.7}$$

where $\boldsymbol{\theta}_{(\cdot)}$ is the state at one scale, $\boldsymbol{p}$ represents point-scale with dimension m, and $\boldsymbol{f}$ represents field-scale with dimension n, which may or may not equal m. The following three upscaling methods simplified this transformation to varying degrees.

If the driving system and the response system pass the MFNN test, a function ϕ relating the two systems is implied to exist. This also suggests predicting the response system is possible.

In this study, we used local polynomial maps to predict the field-scale soil moisture $\theta_{(f)}$ using the point-scale soil moisture $\boldsymbol{\theta_{(p)}}$ phase space embedded vector with

dimension m (Eq. 3.8) using the algorithm as in Abarbanel et al. (1994).

$$\boldsymbol{\phi} : R^{d_p} \rightarrow R, \quad \theta_{(f)} = \boldsymbol{\phi}(\boldsymbol{\theta_{(p)}}) \tag{3.8}$$

A local neighborhood is formed by finding N_B nearest neighbors of a point $\theta_{(p)}$ in the d-embedding dimensional space. If the polynomial map is linear, the neighborhood size is estimated as $N_B = 2(d+1)$. For a quadratic map, $N_B = (d+1)(d+2)$ (Abarbanel et al., 1994). A local polynomial map was constructed between the field-scale measurement $\theta_{(f)}$ and the N_B nearest neighbors of the corresponding point-scale measurement $\theta_{(p)}$ in the m-dimension phase space. A least-squares fit was applied to construct the local maps $\boldsymbol{\phi}$, i.e. Eq. 3.9 was minimized to estimate parameters for each local quadratic map ϕ_j.

$$\sum_{k=1}^{N_B} |\theta_{(f)}^{(k)}(j) - \phi_j(\theta_{(p)}^{(k)}(j))|^2 \tag{3.9}$$

where $\theta_{(p)}^{(k)}(j)$ indicates the kth nearest neighbor of $\theta_{(p)}$ at time j from the training data set and $\theta_{(f)}^{(k)}(j)$ represents the value of $\theta_{(f)}$ measured simultaneously with $\theta_{(p)}^{(k)}(j)$. This local polynomial map prediction method is independent of the MFNN test, which means if the two systems fail to pass the MFNN test, this prediction can still be performed.

To achieve the best modeling performance and compensate for the effect of insufficiently populated phase space, we enlarged the neighborhood size N_B to approximately 20% of the data set, which is 10^2 times larger than the recommended N_B Abarbanel et al. (1994). However, the neighborhood size can theoretically be as large as the whole dateset, which is also called global model (Abarbanel, 2012).

For sake of comparison with the phase space method, linear regression and cdf matching upscaling relationships were also established between the field-scale and

point-scale measurements. The transformation ϕ can be expressed as

$$\phi : R \rightarrow R, \quad \theta_{(f)} = a\theta_{(p)} + b \tag{3.10}$$

where the parameters a and b were estimated using least squares fit. The field-scale soil moisture values were then predicted and compared with the measured values. For the CDF matching method, the soil moisture time series for both scales were first ranked and paired. The differences δ were calculated between each pair of the two ranked data sets. A 4th order polynomial function was fit to the ranked point-scale soil moisture $\theta(p)$ and the corresponding differences $\delta(\theta(p))$(Drusch et al., 2005).

$$\phi : R \rightarrow R, \quad \theta_{(f)} = \theta_{(p)} + \delta(\theta_{(p)}) \tag{3.11}$$

3.3.3 Data and statistical evaluation

The point-scale soil moisture data were obtained from the Oklahoma Mesonet, which is a network of long-term automated environmental monitoring stations (McPherson et al., 2007). With more than one hundred stations located around the state, the Mesonet was designed to monitor processes at the meso-scale. The Oklahoma Mesonet can provide long-term time series of soil moisture with a maximum time resolution of 30 min at three depths (5 cm, 25 cm, and 60 cm). Soil moisture data are measured by heat dissipation sensors (CS-229, Campbell Scientific, Logan, UT), which determine soil matric potential by sending heat pulses and detecting the resulting temperature change before and after a pulse introduced (Zhang et al., 2019). Soil moisture is then determined from the matric potentials using site- and depth- specific water retention curves (Scott et al., 2013). These data were used as point-scale soil moisture to reconstruct $\boldsymbol{\theta_{(p)}}$.

The field-scale soil moisture data were monitored by cosmic-ray neutron probes (abbr. CRNPs, Hydroinnova LLC, Albuquerque, NM). CRNPs measure ambient fast

neutron counts in the environment, which are inversely correlated to the amount of water in the soil (Desilets et al., 2010). CRNPs can be calibrated to measure soil moisture using the following equation:

$$\theta_g = \frac{a_0}{N/N_0 - a_1} - a_2 - w_{latt} \tag{3.12}$$

where θ_g (g g^{-1}) is the soil gravimetric water content, N is the fast neutron count rate (counts per minute, cpm), N_0 represents the constant neutron count rate when all hydrogen sources within the footprint are excluded, w_{latt} is the lattice water, $a_0 = 0.0808$, $a_1 = 0.372$, and $a_2 = 0.115$ (Desilets et al., 2010; Zreda et al., 2012). Due to the high speed and long travel paths of fast neutrons, the footprint of CRNPs is about 400-m diameter (Kohli et al., 2015). With this large footprint and 60-min time resolution, CRNPs can provide appropriate data to reconstruct $\boldsymbol{\theta_{(f)}}$.

In order to compare soil moisture at these two different scales, CRNPs were installed at three Mesonet stations - Stillwater (36.12093, -97.09527), Marena (36.06434, -97.21271), and Lake Carl Blackwell (36.14730, -97.28585). These three stations are located in grassland in north central Oklahoma. Soil textures of the surface soil at the stations are loam for Marena and Lake Carl Blackwell, and silty clay loam for Stillwater. The average annual precipitation values of the three stations are similar ranging from 831 mm (Lake Carl Blackwell) to 884 mm (Marena). The time periods of CRNP data collection are 05/2017 - 09/2018 (Marena), 03/2017 - 10/2018 (Stillwater), and 06/2017 - 06/2018 (Lake Carl Blackwell) respectively (Fig. 3.1). The lengths of time series are listed in Table 3.3.

A calibration campaign for the CRNPs was conducted at the three Oklahoma Mesonet sites in the the summer of 2017. Based on the calibration, N_0 values for each site were determined. By using Eq. 3.12 with site-specific N_0, neutron counts N were converted to gravimetric water content θ_g. The three times series of θ_g were converted to volumetric water content θ_v for each site with bulk density assumed to

be 1.40 (g cm^{-3}).

In order to obtain CRNP time series with relatively low noise, hourly CRNP data were smoothed by a Savitzky-Golay filter with a Matlab built-in function "smoothdata.m". To match the measurement depths of the CRNPs, depth-weighted averages of soil moisture for the Mesonent stations were calculated using the method introduced by Kohli et al. (2015). Linear gap filling was conducted for missing data for both scales.

For each pair of time series for the three sites, the first 70% of the data were used as the training sets, and the remaining 30% were used for validation. To evaluate and compare the performances for the three methods, bias and root mean square error RMSE were calculated for each model and each site respectively.

3.4 Results and discussion

3.4.1 Nonlinear properties of soil moisture time series

As shown in Fig. 3.1, soil moisture time series at the point-scale and field-scale reflect roughly similar dynamics but differ in many details. The point-scale soil moisture time series is generally smoother, and some "plateaus" often occur at high soil moisture levels. That is a result of the different sensing devices employed. The heat dissipation sensors' output ranges from -852 kPa to -8.5 kPa (Illston et al., 2008), although the lower limit has recently been decreased by a recalibration that was not used in this study (Zhang et al., 2019). Wet conditions or inaccurate calibrations can cause measurements to exceed the upper limit of the sensors. Thus soil moisture dynamics at high soil moisture levels (near saturation) may not be discernible in the point-scale. The field-scale soil moisture time series are generally noisier than the point-scale ones. To reduce the noise, data are often aggregated by applying moving windows, which ranged from 6 h - 24 h in some previous studies (Evans et al., 2016; Heidbüchel et al., 2015). The difference of the soil moisture values between the two scales at Stillwater

is relatively large, which implies some sensor bias could exist. This could be due to inaccuracy in the calibration parameter N_0 (Table 3.1), since the field-scale soil moisture values all depend on N_0.

The ideal delay time, τ was chosen to best unfold the geometric structures of the soil moisture process in the phase space. Fig. 3.2 shows the mutual information of the point- and field-scale time series with different time lags. All data sets follow the same pattern of steadily decreasing mutual information for the first 200-400 time lags afterwards reaching a low baseline level with no distinct first minima, like some prior studies (Frison et al., 1999a). One possible reason could be that the scale for the temporal patterns of hourly soil moisture time series is large, which requires longer time series to capture more cycles. This is probably more obvious for the field-scale, for which the mutual information slightly increases around 650 hrs and 1900 hrs for all three sites leaving a wide window (200-500 hrs) for τ selection. Another possible reason for the lack of distinct first minima could be the noise existing in the time series. This is reflected in the point-scale mutual information, which fluctuates at a low level and lacks similar patterns for the three sites. To avoid underestimating τ by this kind noise effect, i.e. accidentally selecting a local minimum that is created mainly by noise before the mutual information stabilizes, an empirical noise-tolerance level of 10^{-4} was used for seeking the first minima. Using this threshold, τ ranges from 5.8 to 16 days for the point-scale and 6.8 to 9.5 days for the field-scale Table 3.1. The mutual information at the selected values of τ reaches a level < 0.2 for all data sets.

The appropriate embedding dimension for each time series was determined based on the FNN criteria in Eqs. 3.5 and 3.6. The percentages of false nearest neighbors as embedding dimension increases are shown in Fig. 3.3. Similar patterns are shown for all sites and scales. When $d = 3$, the percentage of false nearest neighbors dropped to nearly zero, which implies the behavior of the system can be sufficiently captured

by embedding in a 3-dimensional space. The value of 3 for the embedding dimension has also been used for soil water potential data (Doscher, 1997) and soil surface temperature (Koçak et al., 2004).

For each time series, Fig. 3.4 shows the phase portraits, which are the reconstructed 3-dimensional phase spaces represented in a two dimensions. Similar patterns can be found across sites for both scales. The point-scale and field-scale soil moisture have distinct patterns, in which some structure can be perceived. The trajectories of the soil moisture are more clustered at the field-scale than at the point-scale which can be attributed to the distinct drying processes at the two scales Fig. 3.1. The differences in the phase portraits between the two scales may imply that soil moisture at different scales has different trajectory structures and dynamical behaviors. The differences of the trajectory structure among locations are not as apparent as the differences between scales. It clearly shows similar shape of boundaries and distinct internal structures across locations. Similar differences in the phase portraits among different locations can be found in (di2).

3.4.2 Upscaling results

The upscaling results for the three different methods are shown in Fig. 3.5 to 3.7. Basic statistics of the linear regression functions are summarized in Table 3.3. To illustrate and compare the performance of the three methods, results for both the calibration and validation portions of the data were plotted with a 1:1 line. The calibration results (left column in each figure) show the ability of each method to fit the calibration data set. The validation results are shown in the right column of each figure, indicating the performance of each method in prediction mode.

For the linear regression and CDF matching, points are not uniformly distributed around the 1:1 line (Fig. 3.5 and Fig. 3.6). The maximum of the predicted θ always corresponds to a wide range of the observed θ. This is because the point-

scale soil moisture remains close to saturated for longer time than the field-scale soil moisture Fig. 3.1. The point-scale observations apparently omit some details of the beginning of drying processes. The RMSE values for the validation data sets are slightly smaller than those for the calibration stage for all three sites, which indicates that the linear regression and CDF matching methods are well calibrated and do not suffer appreciably from over-fitting at these sites.

The RMSEs for the CDF matching method are comparable to previous studies which range from 0.03 cm^3 cm^{-3} to 0.044 cm^3 cm^{-3} (Scipal et al., 2008; Mittelbach et al., 2012). For linear regression method, RMSEs for all three sites are larger than the reported 0.022 cm^3 cm^{-3} reported by De Rosnay et al. (2009). They are comparable to the RMSEs for one-point (0.060 cm^3 cm^{-3}) and 12-point (0.027 cm^3 cm^{-3}) upscaling study done by Crow et al. (2005). This may also be attributed to the inaccuracy near saturation of the point-scale sensors, and the linear regression method was affected more than the CDF matching method.

The RMSE values for the phase-space method were lower than those for the linear and CDF matching methods during the calibration stage (Fig. 3.7). This is consistent with the hypothesis of this study that the structures reflected in phase space reconstruction can provide more detailed and better mapping between the two scales. However, unlike the results for linear regression and CDF matching, the RMSEs for the local polynomial maps method are larger at the validation stage than at the calibration stage, which indicates that this method may suffer from over-fitting or other problems during the calibration stage (Fig. 3.7). Compared to the linear regression and CDF matching results, the validation RMSEs for the local polynomial map method are larger (Fig. 3.5 to Fig. 3.7). This is probably because the phase-space representation was not adequately constructed, even though the local polynomial maps had better fits with $R^2 > 0.7$ for two of the three sites. The phase space neighborhood may not be adequately populated with observations, in which case the

prediction power of this method would be weakened. The R^2 for CDF matching and local polynomial map for Stillwater are both negative (Table. 3.3). This occurs when the method is not linear and the fitting with the model is worse than simply estimating with the mean of the dataset. One possible reason could be the period of 10/2018-03/2019 when water content is persistently in a high level, which is suspected to be abnormal, since the detailed wetting and processes were barely reflected in the time series.

The validation results of the last 30% of each time series are shown in Fig. 3.8. The linear regression and CDF matching validation generally follow the observed data and display smoother temporal patterns than the predictions of local polynomial maps. The temporal patterns for the linear upscaling are similar to the point-scale observations in Fig. 3.1 due to the nature of the method. For the linear method, each prediction only depends on the upscaling equations (Table. 3.2) and the corresponding point-scale soil moisture observation. In contrast, the CDF matching method is able to more accurately predict the peaks in field-scale soil moisture associated with rainfall events than is the linear upscaling method. The advantage for the CDF matching method likely arises due to its allowance for a nonlinear relationship between the point- and field-scale time series. As a result, the CDF matching method exhibits a slightly smaller bias and RMSE than the linear upscaling method for the validation data sets Table 3.3.

3.5 Conclusion

We demonstrated that upscaling of soil moisture using the method of phase space analysis is possible, but the proposed method was unable to improve the prediction accuracy compared to linear regression and CDF matching. The phase space method usually requires a large number of observations in order to reconstruct a well populated phase space representation that reflects the dynamical behavior of the system.

In this study, with the 1-hour temporal resolution, a soil moisture time series of 1-2 years long is likely not enough. In this and many similar circumstances, the applicability of using the local polynomial map is probably limited.

For upscaling of in situ soil moisture, CRNPs and the cosmic-ray neutron sensing techniques are well-suited because of their unique footprint size. Although the noise level is usually high in CRNP time series, noise can be effectively reduced through smoothing methods, which could make data qualified for upscaling analysis. The unsynchronized behavior of the two systems implies that some errors in soil moisture upscaling are unavoidable due to not only randomness but also some deterministic reasons. This gives some insights into the next step toward upscaling the thousands of existing in situ soil moisture stations. For site-specific empirical scalings, linear regression and CDF matching methods can easily be applied. However, if a widely applicable or even universal upscaling relationship were to be developed, the phase space structures and nonlinear behaviors may need to be considered.

References

Deterministic chaotic dynamics in soil moisture across Nebraska, author=Di, Chongli and Wang, Tiejun and Istanbulluoglu, Erkan and Jayawardena, AW and Li, Siliang and Chen, Xi, journal=Journal of Hydrology, volume=578, pages=124048, year=2019, publisher=Elsevier.

Abarbanel, H. (2012). *Analysis of observed chaotic data*. Springer Science & Business Media.

Abarbanel, H. D., Carroll, T., Pecora, L., Sidorowich, J., and Tsimring, L. (1994). Predicting physical variables in time-delay embedding. *Physical Review E*, 49(3):1840.

Bradley, E. and Kantz, H. (2015). Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097610.

Crow, W. T., Ryu, D., and Famiglietti, J. S. (2005). Upscaling of field-scale soil moisture measurements using distributed land surface modeling. *Advances in Water Resources*, 28(1):1–14.

De Rosnay, P., Gruhier, C., Timouk, F., Baup, F., Mougin, E., Hiernaux, P., Kergoat, L., and LeDantec, V. (2009). Multi-scale soil moisture measurements at the gourma meso-scale site in mali. *Journal of Hydrology*, 375(1-2):241–252.

Desilets, D., Zreda, M., and Ferre, T. P. A. (2010). Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resources Research*, 46.

Doscher, C. (1997). Applying nonlinear dynamics to hydrologic systems: a review. *Applied Engineering in Agriculture*, 13(2):199–207.

Drusch, M., Wood, E., and Gao, H. (2005). Observation operators for the direct assimilation of TRMM microwave imager retrieved soil moisture. *Geophysical Research Letters*, 32(15).

Eckmann, J. P. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656.

Elliott, R., Brock, F., Stone, M., and Harp, S. (1994). Configuration decisions for an automated weather station network. *Applied Engineering in Agriculture*, 10(1):45–51.

Evans, J., Ward, H., Blake, J., Hewitt, E., Morrison, R., Fry, M., Ball, L., Doughty, L., Libre, J., Hitt, O., et al. (2016). Soil water content in southern england derived from a cosmic-ray soil moisture observing system–cosmos-uk. *Hydrological Processes*, 30(26):4987–4999.

Franz, T. E., Wahbi, A., Vreugdenhil, M., Weltin, G., Heng, L., Oismueller, M., Strauss, P., Dercon, G., and Desilets, D. (2016). Using cosmic-ray neutron probes to monitor landscape scale soil water content in mixed land use agricultural systems. *Applied and Environmental Soil Science*, 2016.

Fraser, A. M. and Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134.

Frison, T. W., Abarbanel, H. D., Earle, M. D., Schultz, J. R., and Scherer, W. D. (1999a). Chaos and predictability in ocean water levels. *Journal of Geophysical Research: Oceans*, 104(C4):7935–7951.

Frison, T. W., Earle, M. D., Abanel, H. D., and Scherer, W. D. (1999b). Interstation prediction of ocean water levels using methods of nonlinear dynamics. *Journal of Geophysical Research: Oceans*, 104(C6):13653–13666.

Hawdon, A., McJannet, D., and Wallace, J. (2014). Calibration and correction procedures for cosmic-ray neutron soil moisture probes located across australia. *Water Resources Research*, 50(6):5029–5043.

Heidbüchel, I., Güntner, A., and Blume, T. (2015). Use of cosmic ray neutron sensors for soil moisture monitoring in forests. *Hydrology & Earth System Sciences Discussions*, 12(9).

Illston, B. G., Basara, J. B., Fiebrich, C. A., Crawford, K. C., Hunt, E., Fisher, D. K., Elliott, R., and Humes, K. (2008). Mesoscale monitoring of soil moisture across a statewide network. *Journal of Atmospheric and Oceanic Technology*, 25(2):167–182.

Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*, volume 7. Cambridge university press.

Kennel, M. B., Brown, R., and Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403.

Koçak, K., Şaylan, L., and Eitzinger, J. (2004). Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding. *Ecological Modelling*, 173(1):1–7.

Kohli, M., Schron, M., Zreda, M., Schmidt, U., Dietrich, P., and Zacharias, S. (2015). Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. *Water Resources Research*, 51(7):5772–5790.

McPherson, R. A., Fiebrich, C. A., Crawford, K. C., Kilby, J. R., Grimsley, D. L., Martinez, J. E., Basara, J. B., Illston, B. G., Morris, D. A., Kloesel, K. A., et al. (2007). Statewide monitoring of the mesoscale environment: A technical update on the Oklahoma Mesonet. *Journal of Atmospheric and Oceanic Technology*, 24(3):301–321.

Mittelbach, H., Lehner, I., and Seneviratne, S. I. (2012). Comparison of four soil moisture sensor types under field conditions in switzerland. *Journal of Hydrology*, 430:39–49.

Ochsner, T. E., Linde, E., Haffner, M., and Dong, J. (2019). Mesoscale soil moisture patterns revealed using a sparse in situ network and regression kriging. *Water Resources Research*, 55(6):4785–4800.

Ookouchi, Y., Segal, M., Kessler, R., and Pielke, R. (1984). Evaluation of soil moisture effects on the generation and modification of mesoscale circulations. *Monthly weather review*, 112(11):2281–2292.

Pecora, L. M. and Carroll, T. L. (2015). Synchronization of chaotic systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(9):097611.

Reichle, R. H. and Koster, R. D. (2004). Bias reduction in short records of satellite soil moisture. *Geophysical Research Letters*, 31(19):n/a–n/a. L19501.

Rodriguez-Iturbe, I., Entekhabi, D., and Bras, R. L. (1991). Nonlinear dynamics of soil moisture at climate scales: 1. stochastic analysis. *Water resources research*, 27(8):1899–1906.

Rosenstein, M. T., Collins, J. J., and De Luca, C. J. (1993). A practical method for calculating largest lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, 65(1-2):117–134.

Rulkov, N. F., Sushchik, M. M., Tsimring, L. S., and Abarbanel, H. D. (1995). Generalized synchronization of chaos in directionally coupled chaotic systems. *Physical Review E*, 51(2):980.

Scipal, K., Drusch, M., and Wagner, W. (2008). Assimilation of a ers scatterometer derived soil moisture index in the ecmwf numerical weather prediction system. *Advances in water resources*, 31(8):1101–1112.

Scott, B. L., Ochsner, T. E., Illston, B. G., Fiebrich, C. A., Basara, J. B., and Sutherland, A. J. (2013). New soil property database improves oklahoma mesonet soil moisture estimates. *Journal of Atmospheric and Oceanic Technology*, 30(11):2585–2595.

Strogatz, S. H. (2001). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 2 edition.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

Taylor, C. M., Gounou, A., Guichard, F., Harris, P. P., Ellis, R. J., Couvreux, F., and De Kauwe, M. (2011). Frequency of sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nature Geoscience*, 4(7):430.

Vachaud, G., Passerat de Silans, A., Balabanis, P., and Vauclin, M. (1985). Temporal stability of spatially measured soil water probability density function1. *Soil Science Society of America Journal*, 49(4):822–828.

Zhang, Y., Ochsner, T. E., Fiebrich, C. A., and Illston, B. G. (2019). Recalibration of sensors in one of the worlds longest running automated soil moisture monitoring networks. *Soil Science Society of America Journal*, 83(4):1003–1011.

Zhu, Z., Tan, L., Gao, S., and Jiao, Q. (2014). Observation on soil moisture of irrigation cropland by cosmic-ray probe. *IEEE Geoscience and Remote Sensing Letters*, 12(3):472–476.

Zreda, M., Shuttleworth, W. J., Zeng, X., Zweck, C., Desilets, D., Franz, T. E., and Rosolem, R. (2012). Cosmos: the cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences*, 16(11):4079–4099.

Table 3.1: Summary of the dynamical characteristics of the time series. τ is the time lag with resolution of 6 hours, d is the number of embedding dimensions, and λ_1 is the maximal Lyapunov exponent.

Station	Point-scale			Field-scale		
	τ (day)	d	λ_1	τ (day)	d	λ_1
Marena	16	3	0.029	9.5	3	0.088
Stillwater	12	3	0.040	13	3	0.133
Lake Carl Blackwell	5.8	3	0.050	6.8	3	0.087

Table 3.2: Regression equations for the upscaling methods of linear regression and CDF matching

Station	Linear regression	CDF matching
Marena	$y = 0.691x + 0.029$	$y = -494x^4 + 405x^3 - 119x^2 + 15.5x - 0.745$
Stillwater	$y = 0.485x + 0.017$	$y = -254x^4 + 261x^3 - 95.8x^2 + 15.3x - 0.814$
Lake Carl Blackwell	$y = 0.578x + 0.053$	$y = -66.6x^4 + 47.0x^3 - 13.6x^2 + 2.35x - 0.145$

Table 3.3: Basic statistics for the three upscaling methods.

Station	data length		Linear regression		CDF matching		Local polynomial map	
	hrs	days	Bias	R^2	Bias	R^2	Bias	R^2
Marena	12497	521	0.011	0.408	0.018	0.329	0.030	0.747
Stillwater	18595	775	0.011	0.484	0.026	-9.96	0.019	-10.3
Lake Carl Blackwell	9128	380	0.00086	0.445	0.0029	0.639	0.018	0.737

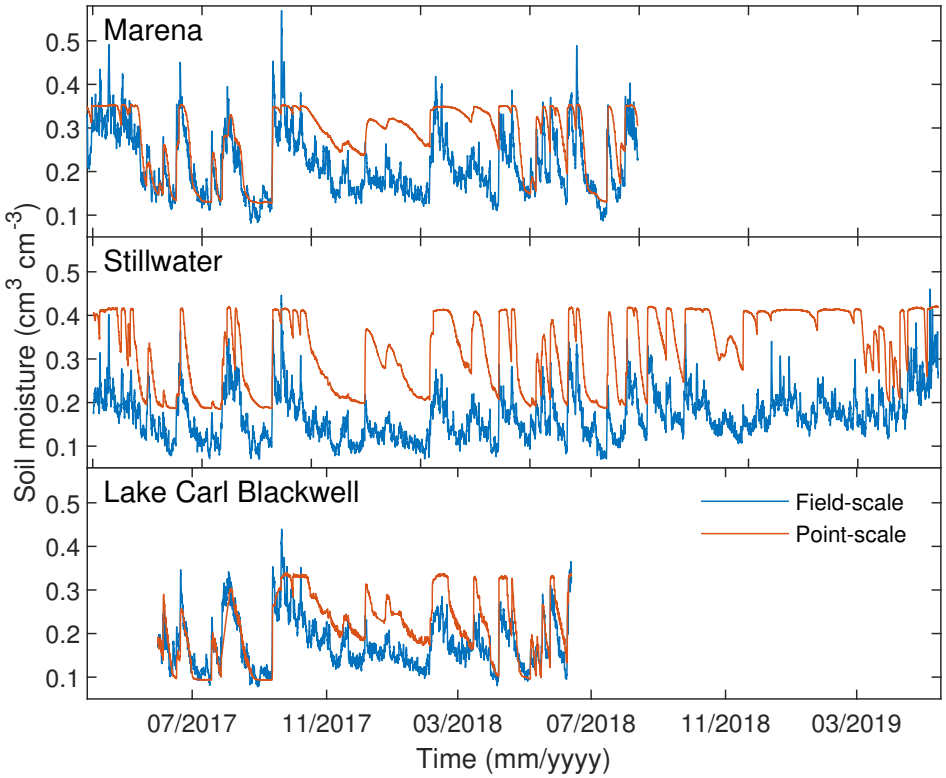


Figure 3.1: Time series for the three cosmic-ray neutron probes (CRNPs) at Stillwater, Lake Carl Blackwell, and Marena. Blue lines represent Mesonet 5-cm soil moisture, and red lines represent soil moisture measured by CRNPs.

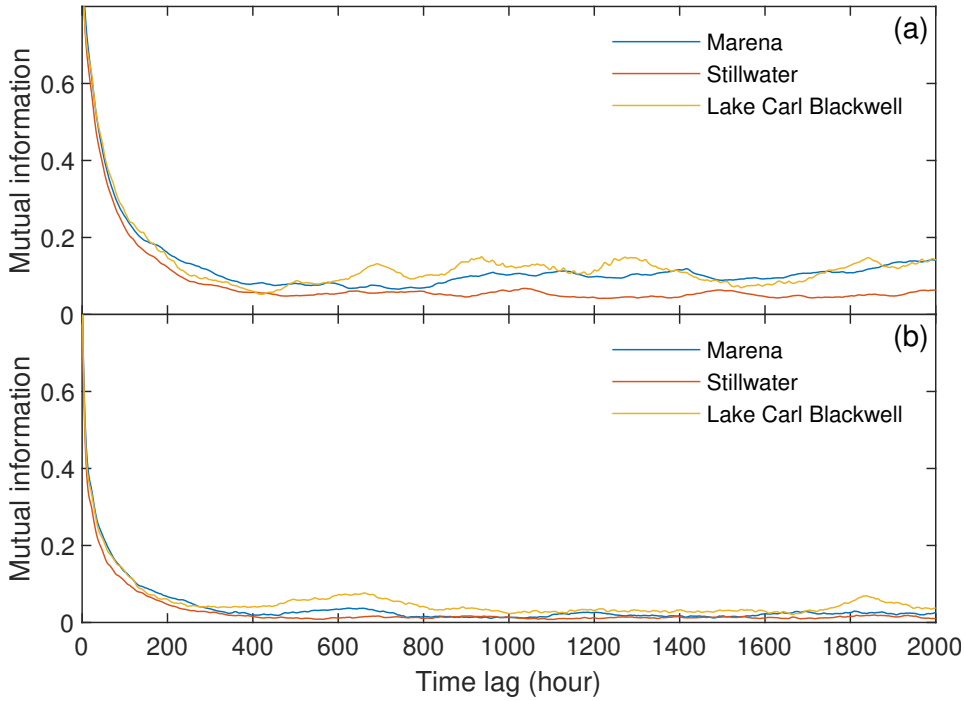


Figure 3.2: Average mutual information for the three sites at point-scale (a) and field-scale (b)

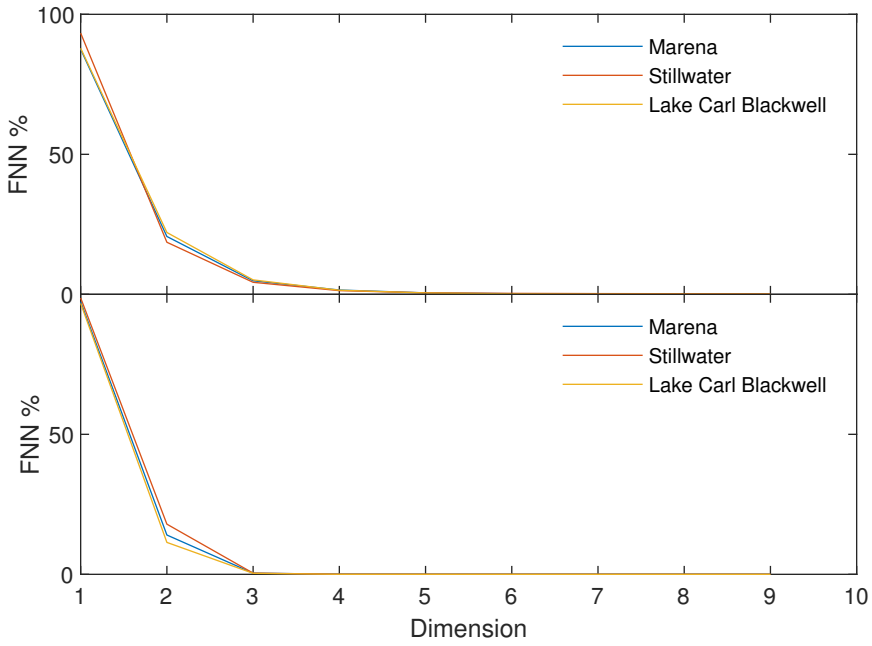


Figure 3.3: Fraction of false nearest neighbors for all sites and scales.

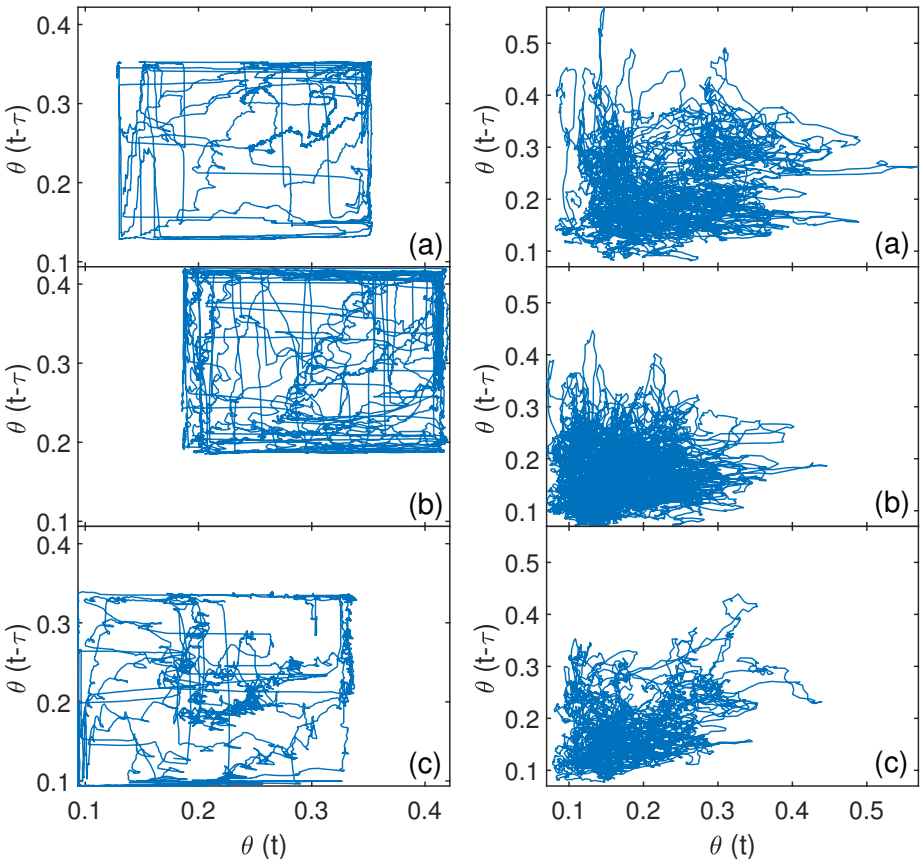


Figure 3.4: Two-dimensional phase portraits of soil moisture time series at point-scales (left column) and field-scale (right column) . Three sites are Marena (a), Stillwater (b), and Lake Carl Blackwell (c).

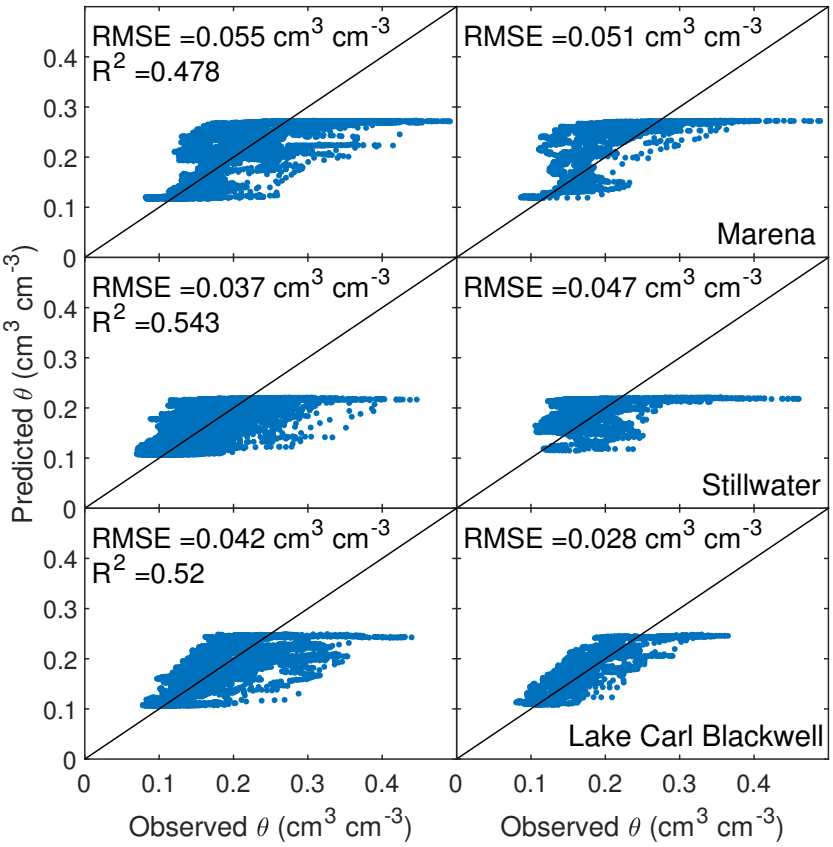


Figure 3.5: Calibration and validation results for the linear regression method. The left column is the calibration data set, and the right column is the validation data set.

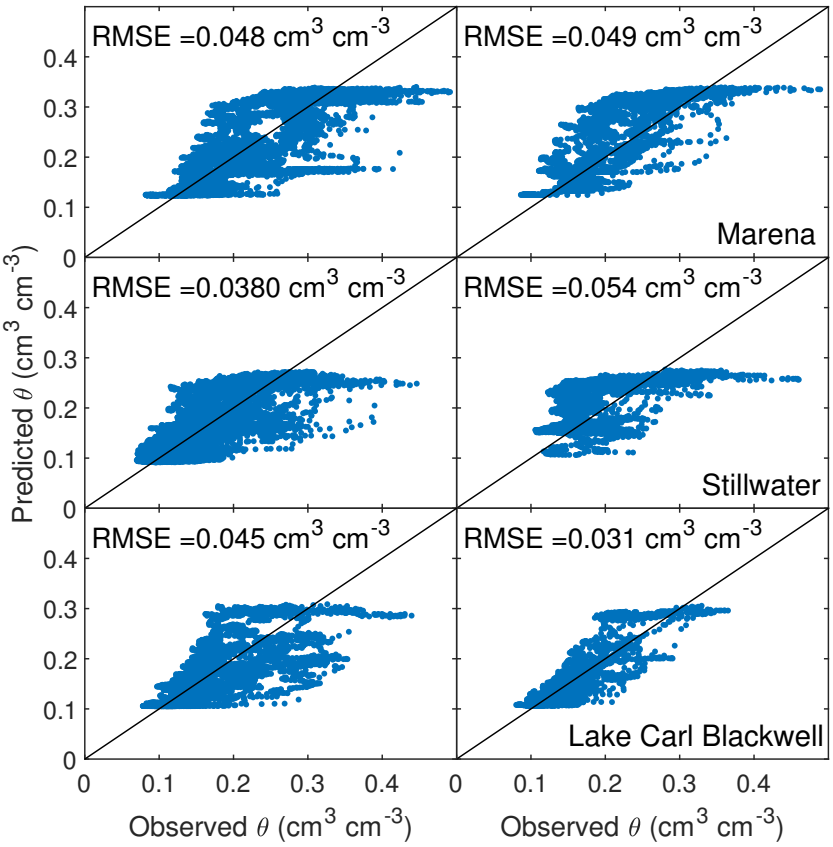


Figure 3.6: Calibration and validation results for the CDF matching method. The left column is the calibration data set, and the right column is the validation data set.

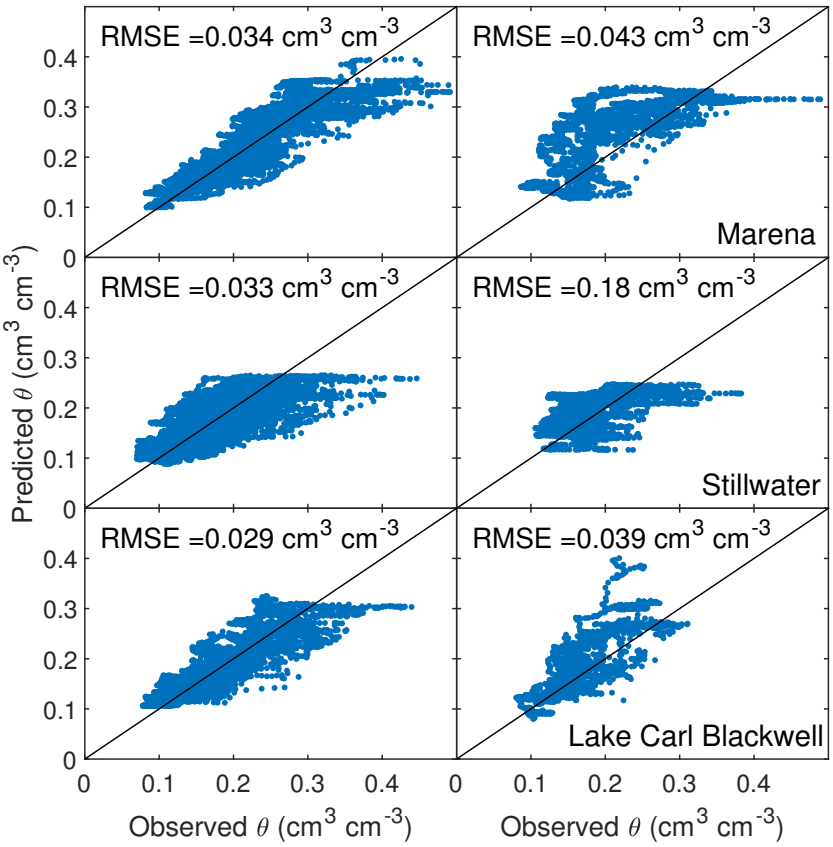


Figure 3.7: Calibration and validation results for the local polynomial map method. The left column is the calibration data set, and the right column is the validation data set.

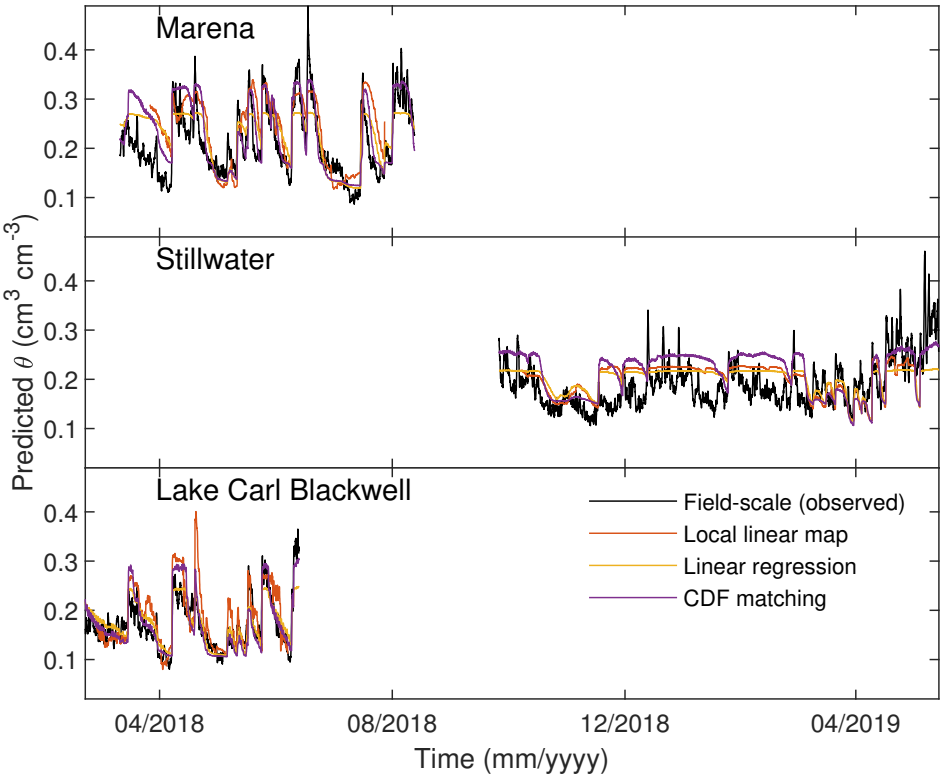


Figure 3.8: Time series predictions by the three methods. The methods are indicated by colors. The observed CRNP soil moisture is shown solid black.

Supplemental materials

Maximal Lyapunov exponent

Chaos is aperiodic long-term behavior in a deterministic system that exhibits sensitive dependence on initial conditions (Strogatz, 2001). A chaotic system is sensitive to its initial conditions, which means with small initial separation, the trajectories of the two points may diverge exponentially fast over some time (Kantz and Schreiber, 2004). The rate of this separation is defined as Lyapunov exponent λ (Eckmann and Ruelle, 1985). This definition can be expressed as follows

$$|\delta x(\Delta t)| \approx |\delta x(0)| e^{\lambda \Delta t} \tag{3.13}$$

In Eq. 3.13 , $|\delta x(0)|$ is the initial distance between two points in the phase space of variable x, and $|\delta x(\Delta t)|$ is the distance between the two points after time Δt. For every dimension of x, there can be defined a Lyapunov exponent, among which the largest one is named the maximal Lyapunov exponent λ_1. If λ_1 is positive, the system is chaotic. If λ_1 is zero or negative, the system is stable (limit cycle or fixed point). Thus, λ_1 quantifies the sensitivity to the initial conditions of a dynamical system, and characterizes the predictability of a chaotic system .

In order to determine if each soil moisture dynamical system is chaotic, we calculated the maximal Lyapunov exponents for each site and each scale using the algorithm introduced by Rosenstein et al. (1993).

The positive maximal Lyapunov exponents indicate that soil moisture as a dynamical system is chaotic, but the strengths are different (Table 3.1). The range of λ_1 for the point-scale soil moisture is 0.029-0.050, which is comparable to 0.011-0.066 that reported in the study of di2. All field-scale soil moisture have higher λ_1 than point-scale, which means that the field-scale soil moisture is more chaotic, i.e. more difficult to predict. This may also imply, in the perspective of nonlinear dynamics, the

difficulty in predicting soil moisture between the two scales, which is tested discussed in the following sections.

Generalized synchronization of chaotic systems

The synchronization of dynamical systems considers the relationship between the phase spaces of two coupled systems - a response system $\boldsymbol{r}(t)$ and its driving system $\boldsymbol{d}(t)$. Synchronization of the two systems simply means they are equal, i.e. no matter how different the initial conditions are, eventually $\boldsymbol{r}(t) = \boldsymbol{d}(t)$ as time progresses (Pecora and Carroll, 2015). A broader generalized synchronization is indicated by

$$\boldsymbol{r}(t) = \boldsymbol{\psi}(\boldsymbol{d}(t)) \tag{3.14}$$

where $\boldsymbol{\psi}$ is a transformation function(Rulkov et al., 1995). If $\boldsymbol{\psi}$ exists, the relationship of the two systems is called generalized synchronization.

A method called *mutual false nearest neighbors* (MFNN) were developed by Rulkov et al. (1995) to determine if two systems exhibit generalized synchronization. The basic idea is similar to the FNN method - two close neighbors in phase space of the driving system correspond to a pair of points in the phase space of the response system, which should also be close. Based on this idea, an MFNN parameter $P(n, d_r, d_d)$ was designed for testing synchronization of $d(t)$ and $r(t)$. This parameter is close to 1 when the system trajectories are synchronized, and much larger than 1 when the trajectories are not synchronized.

To keep it brief and clear, we use notations in Rulkov et al. (1995). The driving system $d(t)$ and response system $r(t)$ in this research are the point-scale soil moisture $\boldsymbol{\theta_{(p)}}$ and the field-scale soil moisture $\boldsymbol{\theta_{(f)}}$. The measurements of driving and response system are embedded with dimension d_d and d_r. respectively and the corresponding vectors are represented as $\boldsymbol{d}(\,\cdot\,)$ and $\boldsymbol{r}(\,\cdot\,)$. The driving system is also embedded in

the space of dimension d_r, which creates vectors $\boldsymbol{d}'(\cdot)$.

$$P(n, d_r, d_d) = \frac{|\boldsymbol{d}'(n) - \boldsymbol{d}'(n'_{NND})|^2}{|\boldsymbol{d}'(n) - \boldsymbol{d}'(n_{NND})|^2} \frac{|\boldsymbol{r}(n) - \boldsymbol{r}(n_{NND})|^2}{|\boldsymbol{r}(n) - \boldsymbol{r}(n_{NNR})|^2} \tag{3.15}$$

where $\boldsymbol{r}(n_{NNR})$ is the nearest neighbor of $\boldsymbol{r}(n)$ in the response embedding space, n_{NND} is the time index of the nearest neighbor of point $\boldsymbol{d}(n)$ in the embedded driving system with dimension d_d, and n'_{NND} is the time index of the nearest neighbor of point $\boldsymbol{d}'(n)$ in the embedded driving system with dimension d_r. The MFNN parameter is suitable for time series with fixed length and has stable behavior across different embedding dimensions. As demonstrated by Rulkov et al. (1995), in practice, the average value of $\bar{P}(d_r, d_d)$ can be used to distinguish systems' synchronized and unsynchronized behaviors.

Generalized synchronization test results

The values for the mean synchronization parameter $\bar{P}$ for all three sites generally exhibit a decreasing trend as the embedding dimension of the driving system (point-scale soil moisture) increases (Fig. S). As the embedding dimension increases, the value of $\bar{P}$ is expected to decrease to 1 if the two systems are synchronized (Rulkov et al., 1995). However, for these soil moisture time series $\bar{P}$ maintains values > 10 and does not tend to decrease to 1 at higher embedding dimensions. This suggests that the point- and field-scale data sets of soil moisture are not synchronized at these three sites. Although the point-scale and field-scale soil moisture are not synchronized in the generalized sense, local polynomial map predictions based on Eq. 3.8 can still be performed. The value of $\bar{P}(d_d)$ suggests to embed point-scale soil moisture in the dimension greater than 3 for better synchronization results. However, since $\bar{P}(d_d)$ doesn't decrease much as d_d increases, it would shorten the length of a data set and increase computation burden at higher embedding dimensions.

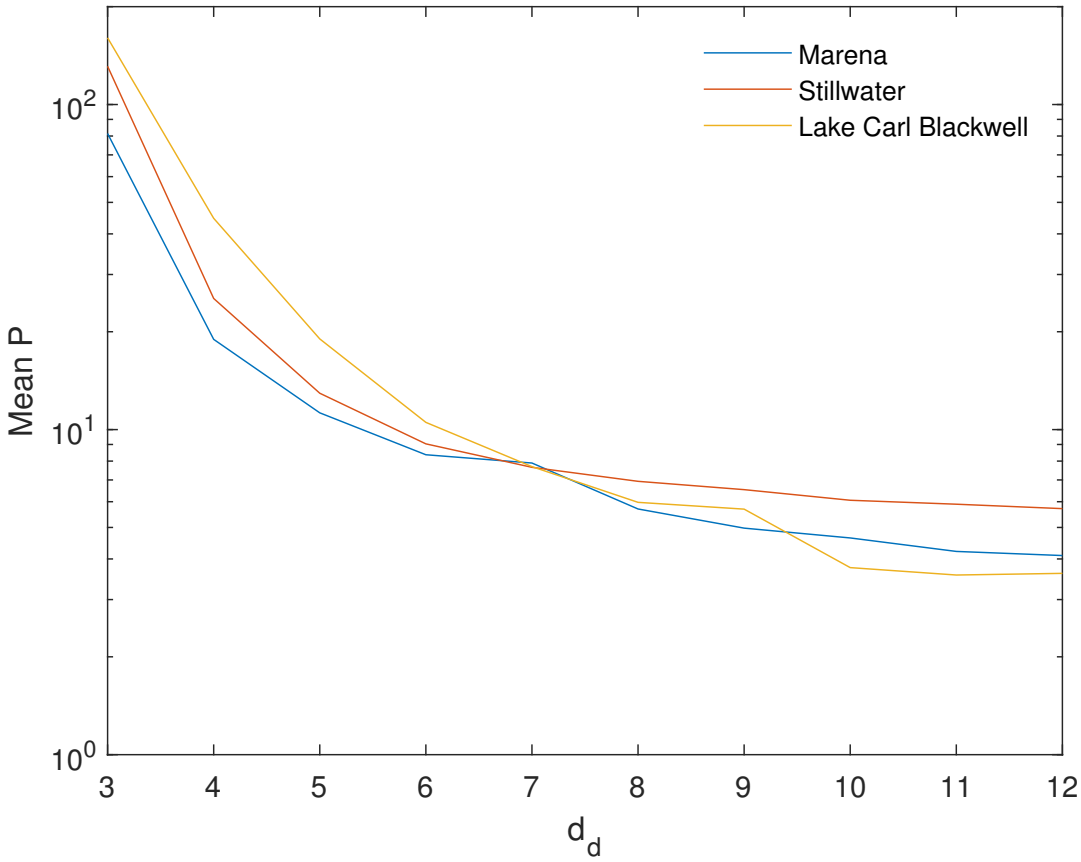


Figure S: Average MFNN parameter ($\bar{P}$) as a function of the embedding dimension of the driving system (d_d). Three polylines represent the three sites.

CHAPTER 4

Application of computational mechanics to the analysis of soil moisture data

4.1 Abstract

Soil moisture as a natural process computes - it stores and transmits information from its past to its future, and generates both randomness and structures at the same time. The aim of this research is to find structures hidden in the soil moisture process and to examine its structural complexity by applying the approach of computational mechanics. ϵ-machines were constructed for the zeroth, first, and second order derivatives of symbolized soil moisture time series to examine structures in multiple aspects. Based on the reconstructed ϵ-machines, soil moisture is complex, hidden, and unpredictable to some degree. Statistical complexities tend to increase with the orders of derivatives for soil moisture processes, which may be a result of noise in the data. Data resolution and the symbolization strategy may have strong effects on both finding patterns and constructing ϵ-machines. Further studies on highly diverse second order derivatives are needed, which will contribute to seeking the factors that controls the topology of ϵ-machines for soil moisture.

4.2 Introduction

In research related to mass and energy exchange at land surface, soil moisture has been recognized as an important variable, as various atmospheric and hydrological processes, such as infiltration, evapotranspiration, and drainage, leave their own signatures in soil moisture time series. With various interactions between these pro-

cesses, the resulting dynamics of soil moisture may exhibit complex chaotic behaviors, which complicate the processes of interpreting data, discovering patterns, and making predictions (Rodriguez-Iturbe et al., 1991). These interacting hydrologic processes and the resulting structures and patterns in soil moisture dynamics can be interpreted as the hydrologic system's "intrinsic computation", which is the way the system stores, structures, and transforms information temporally and spatially (Feldman et al., 2008).

The concept of intrinsic computation implies a potential of discovering new patterns in the hydrologic processes thus simplifying descriptions of them. To fulfill this, it is necessary to effectively identify and quantify the structures and complexity in soil moisture dynamics. In this study, we employ approaches from the field of computational mechanics, which are designed for studying how dynamical systems store and process information (Shalizi and Crutchfield, 2001; Crutchfield, 2012, 2017). Computational mechanics approaches are established upon theories of information and computation and are able to discover patterns and to quantify structural complexity of a process (Crutchfield, 2012). In this case, the goal is to reflect causal structures of soil moisture dynamics by explicitly representing states and transitions with their probabilities in the form of an ϵ-machine, which is a type of Hidden Markov Model (Crutchfield, 1994). Previously, computational mechanics have been successfully applied to geomagnetism (Clarke et al., 2003), wind speed (Palmer et al., 2000), and stock market data (Park et al., 2007), but not yet for soil moisture.

Soil moisture is a state variable of the hydrologic system (Koster et al., 2009; Houser et al., 1998) and here we consider the first and second derivatives of soil moisture with respect to time, along with soil moisture itself (zeroth order derivative). The first derivative ($\frac{\Delta\theta}{\Delta t}$) of soil moisture represents the rate of change in soil water content and is a key term in water balance models (Laio et al., 2001). The first derivative of soil moisture for a particular soil layer is directly proportional to the

net water flux into or out of that layer. In this research, we also explored the 2nd derivative of soil moisture ($\frac{\Delta^2\theta}{\Delta t^2}$), which is an indicator of the rate of change in the net water flux. The objectives of this research are (1) to construct ϵ-machines for soil moisture time series to reveal possible emergent causal structures in soil moisture dynamics, and (2) to compare causal structures of zeroth, first and second order of derivatives of soil moisture time series.

4.3 Materials and Methods

4.3.1 ϵ-machine

The goal of the subsequent analysis is to build a model that can both predict the future state of the system and describe the mechanisms of the underlying system which produces the observable process (Ellison et al., 2009). Computational mechanics provide a way of inferring models of a hidden processes with some observable behaviors, and ϵ-machines are one type of model employed in computational mechanics.

In order to fulfill the goal of effective prediction, we first look for ways of effectively representing the system. In this research, a *process* refers to a sequence of random variables X_i, so a process can be defined as a one-dimensional bi-infinite chain $\overleftrightarrow{X}$

$$\overleftrightarrow{X} \equiv \ldots X_{-2}X_{-1}X_0X_1X_2\ldots \tag{4.1}$$

where the subscript $i \in \mathbb{Z}$, positive integers represent the future, and negative integers represent the past. Realizations of X_i are written in lower case letters x_i. All x_i are drawn from a countable set $\mathcal{A}$, which is also called an *alphabet*. The process can be viewed as a *communication channel*, which transmits information from the past $\overleftarrow{X}$ to the future $\overrightarrow{X}$ by storing information in the present.

First we partition the space of all pasts (or histories) $\overleftarrow{X}$ into subsets which are mutually exclusive and jointly comprehensive to the whole set. A subset of history

space is called an effective state R_i. An effective state R (a particular partition of history) is also a random variable. Its distribution can be written as

$$\Pr(\mathbf{R} = R_i) = \sum_{\overleftarrow{x} \in R_i} \Pr(\overleftarrow{x}) \tag{4.2}$$

Any function defined on the whole set of histories partitions it, and the function maps specific histories into effective states. Here we introduce an equivalence relation ($\sim$) (Appendix C) which partitions the histories into certain types of effective states, named *causal states*. This equivalence relation does not differentiate histories which lead to the same predictions of the future. Here, a prediction refers to a distribution $\Pr(\overrightarrow{X}|\overleftarrow{x})$ of possible futures $\overrightarrow{X}$ given a particular past $\overleftarrow{x}$.

The formal definition of this causal equivalence relation is defined as follows. Two histories are equivalent if and only if they have the same conditional distribution of futures:

$$\overleftarrow{x}' \sim \overleftarrow{x}'' \Longleftrightarrow \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}'') \tag{4.3}$$

where $\overleftarrow{x}'$ and $\overleftarrow{x}''$ are particular pasts, $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}')$ and $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}'')$ are their conditional distributions of the future.

A function mapping a particular past $\overleftarrow{x}'$ into its equivalence class S is defined as ϵ-function,

$$\epsilon : \overleftarrow{\boldsymbol{X}} \to \mathcal{S}, \quad \epsilon(\overleftarrow{x}) = S = \{\overleftarrow{x}' : \overleftarrow{x}' \sim \overleftarrow{x}\}. \tag{4.4}$$

When a new observation x_0 is made, a history $\overleftarrow{x} = ...x_{-3}x_{-2}x_{-1}$ becomes a new history $\overleftarrow{x}' = ...x_{-2}x_{-1}x_0$. This new history belongs to a particular equivalence class, which is also the current causal state S of the process. By applying the ϵ-function to this new history, the process transits from a state $\mathcal{S} = \epsilon(\overleftarrow{x})$ to another state $\mathcal{S}' = \epsilon(\overleftarrow{x}')$ with a transition probability $\Pr(X = x, S'|S)$. Consequently, the dynamic of the process can be defined as state-to-state transitions. All the state-to-state transitions of an ϵ-machine can be expressed as transition matrices whose

elements are defined as the transition probabilities $T_{\mathcal{S}\mathcal{S}'}^{(x)} = \Pr(X = x, \mathcal{S}|\mathcal{S}')$, where (x) represents a symbol in the alphabet $\mathcal{A}$. The transition probabilities can be estimated through the process of ϵ-machine reconstruction, which is introduced in Section 4.3.5. Thus, the ϵ-machine of a process is defined as a set of causal states and transition matrices, i.e. $\mathcal{M} = \{\mathcal{S}, \{T^{(x)}, x \in \mathcal{A}\}\} = \{\mathcal{S}, \mathcal{T}\}$.

4.3.2 Properties of ϵ-machine

We selected ϵ-machines to model soil moisture dynamics because ϵ-machines have multiple important and beneficial properties. These properties are briefly described below. The proofs can be found in Shalizi and Crutchfield (2001).

The past and the future are independent given causal states. This can be written as $\Pr(\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}) = \Pr(\overleftarrow{X}|\mathcal{S})\Pr(\overrightarrow{X}|\mathcal{S})$.

ϵ-machines are first-order Markov chains, i.e. $\Pr(S_t|...S_{t-2}S_{t-1}) = \Pr(S_t|S_{t-1})$. This can be derived from the previous property.

ϵ-machines are unifilar (deterministic). For any state $S_i \in \mathcal{S}$ and any $x \in \mathcal{A}$ of an ϵ-machine, there is at most one successor state S_j. This implies that an observed sequence $...x_{-3}x_{-2}x_{-1}...$ has a 1-1 mapping relationship with a causal states sequence.

ϵ-machines are optimal predictors. Here, optimal means that information entropy of a sequence conditioning on causal states $\mathcal{S}$ is no larger than the information entropy conditioning on any rival effective states $\mathcal{R}$, i.e. $H[\overrightarrow{S}^L|\mathcal{R}] \geqslant H[\overrightarrow{S}^L|\mathcal{S}]$. Therefore, knowing the ϵ-machine and the current causal state is as good as knowing the entire past $\Pr(\overrightarrow{X}|\mathcal{S}) = \Pr(\overrightarrow{X}|\overleftarrow{X})$.

Causal states are minimal, which means that an ϵ-machine has less statistical complexity than any rival effective states with the same predictive power.

The ϵ-machine is unique for a certain process.

All the above properties imply that ϵ-machine is optimal in simplifying and modeling a process in the sense of information theory. By applying this approach to

soil moisture processes, some hidden patterns in their dynamics could be revealed effectively.

4.3.3 Information-theoretic measures

Computational mechanics were developed with the aid of the concepts, notions, and conclusions from information theory. In previous sections, the state and the dynamics of a process have been defined as random variables and communication channels. The properties of the process can thus be explored by defining and estimating some information-theoretic measures, which can also be called structural complexity measures.

In the framework of information theory, the uncertainty of a process is represented by entropy $H(X)$ (Appendix B), so the uncertainty of the future of a process can be written as $H(\overrightarrow{X})$. For a sequence of finite length, a quantity *block entropy* $H(X^L)$, or simply $H(L)$, is defined to represent the uncertainty of L consecutive symbols,

$$H(L) \equiv -\sum_{x^L \in \mathcal{A}^L} \Pr(x^L) \log_2 \Pr(x^L) \tag{4.5}$$

where L is positive and x^L represents all possible blocks of symbols with length L. Then the entropy rate is defined as the production of information of the process

$$h_\mu \equiv \lim_{L \to \infty} \frac{1}{L} H(L) \tag{4.6}$$

The entropy rate h_μ quantifies the irreducible randomness in a process (Crutchfield and Feldman, 2003).

Because of the unifilarity of ϵ-machines, the entropy rate can be directly calculated

from the ϵ-machine (Ellison et al., 2009)

$$h_\mu = H(X|\mathcal{S}) = -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \sum_{\{x\}}^{(x)} T^{(x)}_{\mathcal{S}\mathcal{S}'} \log_2 T^{(x)}_{\mathcal{S}\mathcal{S}'}. \tag{4.7}$$

The excess entropy is defined as the mutual information between the past and future,

$$\begin{aligned} \mathbf{E} &= I[\overleftarrow{X}; \overrightarrow{X}] \\ &= H(\overrightarrow{X}) - H(\overrightarrow{X}|\overleftarrow{X}) \end{aligned} \tag{4.8}$$

By definition, it can be interpreted as the reduction of uncertainty of the future by knowing the past. Since theoretically an ϵ-machine is optimally predictive, which means knowing the current state is as good as knowing the entire past $\overleftarrow{X}$, the excess entropy $\mathbf{E} = I[\mathcal{S}; \overrightarrow{X}]$. Excess entropy can be interpreted as the effective information capacity under the assumption of treating the process as a communication channel (Crutchfield et al., 2009).

Another formulation for excess entropy $\mathbf{E}$ is

$$\mathbf{E} = \lim_{L\to\infty} [H(L) - Lh_\mu] \tag{4.9}$$

which is equivalent to Eq. 4.8 (Crutchfield and Feldman, 2003). This formulation highlights another interpretation of $\mathbf{E}$ as a measure of memory, which eventually "explains" the apparent randomness in the process by considering longer blocks (Crutchfield and Feldman, 2003, 1997). The statistical complexity is defined as the entropy of the states and can also be directly calculated from the ϵ-machine

$$C_\mu \equiv H(\mathcal{S}) = \sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S}). \tag{4.10}$$

The statistical complexity C_μ characterizes the minimal amount of information that is required to transmit the excess entropy from the past to the future (Ellison et al.,

2009).

4.3.4 Estimation of information-theoretic measures

For an ϵ-machine, the transition matrix T is the summation of all transition matrices for all symbols, $T = \sum_x T^{(x)}$. π is the time-asymptotic probability of all states, which is the normalized principal eigenvector.

$$\pi = \pi T \tag{4.11}$$

where π is a row vector $(\Pr(\mathcal{S} = \mathcal{S}_1), \Pr(\mathcal{S} = \mathcal{S}_2), \Pr(\mathcal{S} = \mathcal{S}_3), ...)$ and its elements are normalized in probability, i.e. $\sum_\sigma \pi_\sigma = 1$ where σ denotes the state index. Thus, by knowing the transition matrix T, the state probabilities π can be estimated, and h_μ and C_μ can be calculated using Eq.4.7 and Eq.4.10.

Since $\mathbf{E}$ cannot be estimated through an explicit expression (Eq. 4.8 or Eq. 4.9), the concept *retrodiction* needs to be introduced to estimate $\mathbf{E}$. Retrodiction of a process is the opposite of prediction, which means using the future to predict the past. The formalism of the retrodiction is the same with prediction but scanning the measurements in reverse time direction. To differentiate the retrodiction from the prediction, we need to add $+$ and $-$ as superscripts to denote the direction of scanning a sequence. For example, C_μ^+ and C_μ^- represent the statistical complexity of the prediction and retrodiction of a process respectively, which are called predictive and retrodictive statistical complexities. By constructing ϵ-machines for both directions, $\mathbf{E}$ can be estimated. It has been proven that the excess entropy is the mutual information between the predictive and retrodictive causal states (Eq. 4.8), which also implies that it is the same for both directions (Ellison et al., 2009).

$$\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] \tag{4.12}$$

By the definitions of $C_\mu^+(C_\mu^-)$ 4.10 and conditional entropy (Appendix B), $\mathbf{E}$ can be derived as,

$$\mathbf{E} = C_\mu^+ - H[\mathcal{S}^+|\mathcal{S}^-] = C_\mu^- - H[\mathcal{S}^-|\mathcal{S}^+] \tag{4.13}$$

In Eq. 4.13 , $C_\mu^+(C_\mu^-)$ can be easily estimated using Eq. 4.10 after constructing ϵ-machines for each direction. Then the problem of estimating $\mathbf{E}$ is converted to estimating the conditional entropy $H[\mathcal{S}^+|\mathcal{S}^-]$ ($H[\mathcal{S}^-|\mathcal{S}^+]$). A method called *mixed-state presentation* relates the forward and reverse causal states which can yield an explicit form of the conditional entropy and $\mathbf{E}$. A detailed description of this method can be found in Ellison et al. (2009).

Eq. 4.13 can also be interpreted as a decomposition of the statistical complexity $C_\mu^+(C_\mu^-)$ for each direction. The amount of information that is directly presented in the observed sequence is excess entropy $\mathbf{E}$, and the conditional entropy $H[\mathcal{S}^+|\mathcal{S}^-]$ ($H[\mathcal{S}^-|\mathcal{S}^+]$) is called crypticity $\chi^+(\chi^-)$ since it is hidden.

In the process of calculating $\chi^+(\chi^-)$ and $\mathbf{E}$, another quantity which is defined as *causal irreversibility* Ξ can be easily estimated.

$$\Xi \equiv C_\mu^+ - C_\mu^- \tag{4.14}$$

Ξ is the difference of statitiscal complexity between the forward and reversed causal states. It is a measure of the asymmetry of a process. Only when $\Xi = 0$, the process is reversible.

4.3.5 Data symbolization and machine construction

Soil moisture monitoring networks, like the Oklahoma Mesonet, provide a great resource for studying soil moisture dynamics across a variety of soil conditions. The Oklahoma Mesonet consists of ∼120 stations across the state, covering a wide range of soil properties and weather conditions. Daily average surface soil moisture data

(5-cm depth) for all available sites were converted from the calibrated delta-T data which were retrieved from the Oklahoma Mesonet. The calibrated delta-T data are measured by heat dissipation sensors (CS-229, Campbell Scientific, Logan, UT) and the detailed conversion method can be found in studies conducted by Scott et al. (2013). A simple linear interpolation was applied to fill missing data of short periods, i.e. the length of the period shorter than 3 days. Fifty-four sites with consecutive time series longer than 1500 days were selected, and the longest uninterrupted time series for each selected site was further analyzed.

In order to infer ϵ-machines for each selected site, the soil moisture time series must be symbolized. In this research, the symbolization of the time series includes three phases. The first phase is to discetize the soil moisture time series into binary sequences based on soil moisture values (0th derivative), i.e. values greater than the median were assigned 1s, and values less than the median were assigned 0s. The second phase is to calculate the first forward difference $\frac{\Delta\theta}{\Delta t}$ of the time series which represents the first-order derivative of soil moisture dynamics. Since the natural meaning of first derivative is instantaneous rate of change, 0s and 1s were assigned to negative and positive first differences respectively to represent drying and wetting events. The third phase is to calculate the second-order forward difference $\frac{\Delta^2\theta}{\Delta t^2}$ of the time series which represents the second-order derivative of the soil moisture dynamics. Similar to the second phase of symbolization, 0s and 1s were assigned to negative and positive $\frac{\Delta^2\theta}{\Delta t^2}$ respectively to represent the changing rate of drying and wetting events.

To construct ϵ-machines for the symbolized soil moisture sequences (forward process), an algorithm called Causal State Splitting Reconstruction (CSSR) was used (Shalizi and Shalizi, 2004). A history length of 4 days was selected a the window size in scanning the soil moisture sequences. This selection is based on the error levels for different history lengths versus time series lengths, which were tested for the even process (Shalizi and Shalizi, 2004). The resulting ϵ-machines were classified into

several types based on their state-transition structures. ϵ-machines for the reversed processes were derived from the ϵ-machines for the forward processes following the approach of Ellison et al. (2009). This derivation was only completed for three types of ϵ-machines of the 0th and 1st derivatives of soil moisture.

The space of possible h_μ, C_μ, and $\mathbf{E}$ for these three types of ϵ-machines were explored by enumerating possible transition probabilities of each type of ϵ-machine. For each type of ϵ-machine, 10^5 sets of transition probabilities were randomly generated. The information processing invariants h_μ, C_μ, and $\mathbf{E}$ were thus calculated using the same method described above.

4.4 Results and discussion

4.4.1 Basic properties of the soil moisture dynamics

An example of the time series of 0th order (θ), 1st order ($\frac{\Delta\theta}{\Delta t}$), and 2nd order ($\frac{\Delta^2\theta}{\Delta t^2}$) derivatives of daily soil moisture are displayed in Fig. 4.1. The 1st order derivatives basically show the slopes of the original (0th order derivative) soil moisture sequence with positive values representing soil moisture increases and negative values representing decreases. The corresponding symbolized sequences of the three time series are plotted as dots and empty space in Fig. 4.1.

The positive spikes of the 1st order derivatives generally have larger absolute values than the negative spikes, which indicates that a wetting process is usually quicker than a drying process. The positive and negative spikes of the 2nd order derivatives are generally balanced around zero, which indicates that most wetting events are one day or less. The soil moisture conditions before and after a wetting event are relatively stable ($\frac{\Delta\theta}{\Delta t}$ is close to zero), which leads to a large positive $\frac{\Delta^2\theta}{\Delta t^2}$ followed by a large negative $\frac{\Delta^2\theta}{\Delta t^2}$.

4.4.2 The structures of ϵ-machines

Three types of ϵ-machine topologies were found for both the 0th and 1st order difference of soil moisture processes. These ϵ-machines are displayed in Fig. 4.2. Some key similarities between the ϵ-machine topologies are evident. Each has one state with 0 self-loop and one state with 1 self-loop. For the convenience of comparisons between ϵ-machines, these two states were thus assigned with transition probability p and $1-q$. The State 2 and 3 of the Type 2 ϵ-machine can be considered as the result of splitting of State 2 of Type 1. Similarly, State 2 and 4 of the Type 3 ϵ-machine can be considered as the result of splitting State 2 of Type 2. These similarities in ϵ-machine topologies can also be reflected in Fig. 4.5 where the three types of ϵ-machines largely overlap with different possible transition probabilities.

All the soil moisture processes represented by the three types of ϵ-machine topologies are reversible. All the three types of derived ϵ-machines for the reversed processes are the same as those for the forward processes. Therefore, the statistical complexity C_μ^+ and C_μ^- are obviously identical. Based on Eq. 4.14, the irreversibilities for the processes are zero, which means the processes are all reversible.

The three orders of derivatives of soil moisture have distinct behaviors in statistical complexity and randomness production. As the order increases, the entropy rate h_μ generally tends to increase, and the statistical complexity C_μ becomes increasingly variable for the 1st and 2nd order differences (Fig. 4.3). The entropy rate h_μ ranges from 0 to 1 (Eq. 4.7), which means the system ranges from perfectly ordered and predictable to totally random and unpredictable. In Fig. 4.3, the increasing h_μ with the order of derivatives indicates the unpredictability increases with the order of derivatives.

The differences in statistical complexity between soil moisture processes are mainly topological. Soil moisture processes show distinct patterns of statistical complexity C_μ with different orders of derivatives (Fig. 4.3), which is directly related to different

types of ϵ-machine topologies (Table 4.2). For a certain topology (type), ϵ-machines tend to have similar statistical complexities and various entropy rate h_μ. This implies that the main factors influencing the values of C_μ and h_μ are probably the topology and the transition probabilities respectively. However, we still don't know the factors that influence the ϵ-machines' topology for soil moisture processes. As the order of derivatives increase, soil moisture tend to have more types of ϵ-machines (Table 4.2). This behavior can be partly attributed to the increasing noise level in soil moisture derivatives, since it is known that increasing the order of differences calculated based on the time series can amplify the noise existing in the original time series.

Both 0th and 1st order derivatives of soil moisture show relatively low entropy rates (mostly < 0.5) with high self-loop probability p (Table 4.1). The space of possible entropy rates vs. transition probabilities was explored and plotted in Fig. 4.4. Real data points are limited to a small area where transition probability p is large and the entropy rate h_μ is relatively small. For the zeroth order derivative of soil moisture (θ), the high probability of the self-loop p represents the tendency of the soil staying in dry conditions. Similarly, for the first order derivative of soil moisture ($\frac{\Delta\theta}{\Delta t}$), it represents the tendency of the system to prolong a drying event. This kind of behavior in both sequences contributes to their low randomness (high predictability).

4.4.3 The structures of the process

The structures of soil moisture processes can be reflected in the transition probabilities and the consequent state probabilities, which are partly determined by the symbolization strategy. The state probabilities for Type 1 ϵ-machine are $\pi = \{\frac{q}{1-p+q}, \frac{1-p}{1-p+q}\}$, which was derived from its transition probabilities using Eq. 4.11. For the 0th order derivatives, the state probabilities are always equal to 0.5 for both states, which indicates $q = 1 - p$. For the 1st and 2nd order derivatives, this equality does not exist. This can be partly attributed to the symbolization strategy for the 0th order deriva-

tives, which generates equal amount of zeros and ones. This symbolization also leads to the statistical complexity of Type 1 ϵ-machine always equal to 1. The symbolized soil moisture time series thus become a random telegraph process, which is a stochastic process with two possible outcomes and only the previous symbol matters to the next word (Clarke et al., 2003). When $p = q$, the two states are going to collapse into one, and the statistical complexity will be 0. For the 0th order derivatives of soil moisture, p is usually much larger than q (Table 4.1), which makes the statistical complexities always far from 0. This gives the insight into the nature soil moisture process that soil moisture process is structured and complex.

For the Type 1 ϵ-machine, the formula for crypticity χ^+ turn out to be the same as that for entropy rate h_μ^+. Eq. 4.13 can thus be written as

$$\mathbf{E} = C_\mu^+ - h_\mu^+. \tag{4.15}$$

This equation explains the relationship between excess entropy $\mathbf{E}$ and entropy rate h_μ^+, which is also reflected in Fig. 4.5. Most of the 0th order and some of the 1st and 2nd order data perfectly follow the line with slope of -1. These points correspond to Type 1 ϵ-machines. The interesting part is that the data points of the other two types of ϵ-machines also follow closely around the line, which is distributed in a relatively small area compared to the entire space of possible values. The small dots showing the randomly sampled space of possible excess entropy $\mathbf{E}$ vs. entropy rate h_μ exhibits some spatial patterns. The excess entropy $\mathbf{E}$ vanishes as the entropy rate h_μ reaches a maximum ($h_\mu = 1$) and many dots are distributed at this corner.

The first two orders of derivatives of soil moisture processes are cryptic ($\chi \neq 0$). The crypticity χ for all 54 sites of soil moisture processes (θ and $\frac{\Delta\theta}{\Delta t}$) are positive, which indicates that there is always some information hidden in the soil moisture process. This hidden information constitutes the structures of the process but is never transmitted or remembered by the process.

Intuitively, the irreversibility of soil moisture time series should be reflected in the second order derivatives, since the convexity of drying and wetting events are distinct. However, irreversibility was still not commonly found in the second order derivatives (Table. 4.2). Type 1-3 ϵ-machines are always reversible, which indicates that more than half of the stations are reversible for the second order derivatives of soil moisture. One possible reason could be the temporal resolution (1 day) is too coarse that the change in convexity cannot be captured.

4.5 Conclusion

This research examined the temporal patterns existing in soil moisture time series with a time scale (history length) of 4 days. The selection of this history length is limited by the total length of the time series. Therefore, the resulting ϵ-machines and their structural complexities only reflect the structure of recent history in binary soil moisture processes. In this context, soil moisture can be described by 2- to 5-state ϵ-machines. The reconstructed ϵ-machines for soil moisture provide insights into the structural complexity, crypticity and randomness of soil moisture dynamics. Generally speaking, as a dynamical system, soil moisture is complex ($C_\mu > 0$), hidden ($\chi > 0$), and unpredictable ($h_\mu > 0$) to some degree. The common structure of a self-loop with a high transition probability contributes to the relatively low level of randomness of soil moisture processes. Increasing orders of derivatives for soil moisture processes shows interesting patterns on structural complexity and predictability - with increasing randomness of the process, the variability in statistical complexity tends to increase. Irreversibilities were barely reflected in soil moisture processes based on daily time resolution and symbolization strategy. Further studies on highly diverse second order derivatives are needed, which will contribute to seeking the factors that controls the topology of ϵ-machines for soil moisture.

References

Clarke, R. W., Freeman, M. P., and Watkins, N. W. (2003). Application of computational mechanics to the analysis of natural data: an example in geomagnetism. *Physical Review E*, 67(1):016203.

Crutchfield, J. P. (1994). The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54.

Crutchfield, J. P. (2012). Between order and chaos. *Nature Physics*, 8(1):17.

Crutchfield, J. P. (2017). The origins of computational mechanics: A brief intellectual history and several clarifications. *arXiv preprint arXiv:1710.06832*.

Crutchfield, J. P., Ellison, C. J., and Mahoney, J. R. (2009). Times barbed arrow: Irreversibility, crypticity, and stored information. *Physical review letters*, 103(9):094101.

Crutchfield, J. P. and Feldman, D. P. (1997). Statistical complexity of simple one-dimensional spin systems. *Physical Review E*, 55(2):R1239.

Crutchfield, J. P. and Feldman, D. P. (2003). Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54.

Ellison, C. J., Mahoney, J. R., and Crutchfield, J. P. (2009). Prediction, retrodiction, and the amount of information stored in the present. *Journal of Statistical Physics*, 136(6):1005.

Feldman, D. P., McTague, C. S., and Crutchfield, J. P. (2008). The organization of intrinsic computation: Complexity-entropy diagrams and the diversity of natural information processing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(4):043106.

Houser, P. R., Shuttleworth, W. J., Famiglietti, J. S., Gupta, H. V., Syed, K. H., and Goodrich, D. C. (1998). Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resources Research*, 34(12):3405–3420.

Koster, R. D., Guo, Z., Yang, R., Dirmeyer, P. A., Mitchell, K., and Puma, M. J. (2009). On the nature of soil moisture in land surface models. *Journal of Climate*, 22(16):4322–4335.

Laio, F., Porporato, A., Ridolfi, L., and Rodriguez-Iturbe, I. (2001). Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress: Ii. probabilistic soil moisture dynamics. *Advances in Water Resources*, 24(7):707–723.

Palmer, A. J., Fairall, C. W., and Brewer, W. (2000). Complexity in the atmosphere. *IEEE Transactions on Geoscience and Remote Sensing*, 38(4):2056–2063.

Park, J. B., Lee, J. W., Yang, J.-S., Jo, H.-H., and Moon, H.-T. (2007). Complexity analysis of the stock market. *Physica A: Statistical Mechanics and its Applications*, 379(1):179–187.

Rodriguez-Iturbe, I., Entekhabi, D., Lee, J.-S., and Bras, R. L. (1991). Nonlinear dynamics of soil moisture at climate scales: 2. chaotic analysis. *Water resources research*, 27(8):1907–1915.

Scott, B. L., Ochsner, T. E., Illston, B. G., Fiebrich, C. A., Basara, J. B., and Sutherland, A. J. (2013). New soil property database improves oklahoma mesonet soil moisture estimates. *Journal of Atmospheric and Oceanic Technology*, 30(11):2585–2595.

Shalizi, C. R. and Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics*, 104(3):817–879.

Shalizi, C. R. and Shalizi, K. L. (2004). Blind construction of optimal nonlinear recursive predictors for discrete sequences. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 504–511. AUAI Press.

Table 4.1: Ranges of transition probabilities for the three types of ϵ-machine topologies for 0th and 1st order soil moisture processes.

	p	q	r	s
Type 1	0.90-0.97	0.032-0.21	-	-
Type 2	0.95-0.96	0.027-0.038	0.33-0.38	-
Type 3	0.87-0.94	0.015-0.12	0.25-0.63	0.033-0.34

Table 4.2: Number of stations for each type of ϵ-machine. The number of states are indicated in the first row. State-transition diagrams for Type 1-3 (bold) are plotted in Fig. 4.2

	2 states	3 states		4 states					5 states	6 states	
Type	**1**	**2**	4	**3**	5	6	7	8	9	10	11
0th order	51	2	0	1	0	0	0	0	0	0	0
1st order	19	0	0	35	0	0	0	0	0	0	0
2nd order	23	6	5	5	1	7	1	1	2	2	1

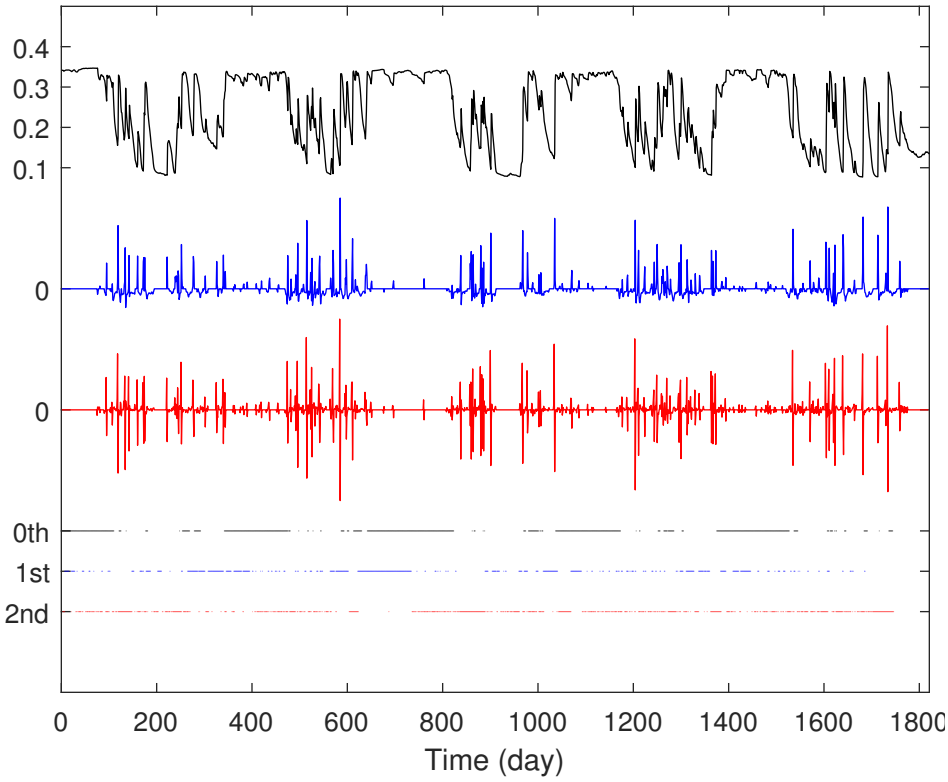
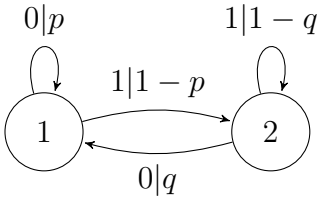
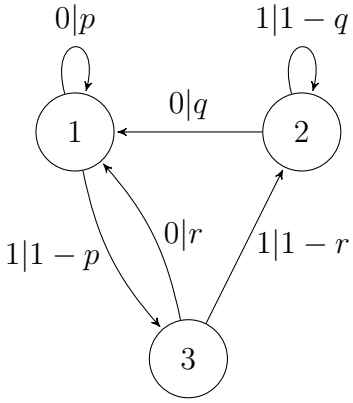


Figure 4.1: (Color) Time series and the corresponding symbolized sequences. The black line is the original soil moisture time series, the blue line is first order difference and the red line is the second order difference. The three corresponding symbolized sequences are displayed under the time series. The dots represent 1s and empty space represent 0s.

Type 1



Type 2



Type 3

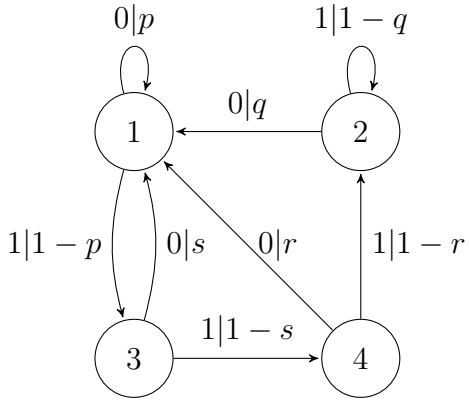


Figure 4.2: Three types of ϵ-machine representations for the soil moisture processes at 54 monitoring locations for the 0th order and 1st order differences. Circles represent states and arrows represent transitions. States are named by numbers (1-4). Transition probabilities are marked by letters (p, q, r, and s). Similar transitions across machines are marked with the same letters.

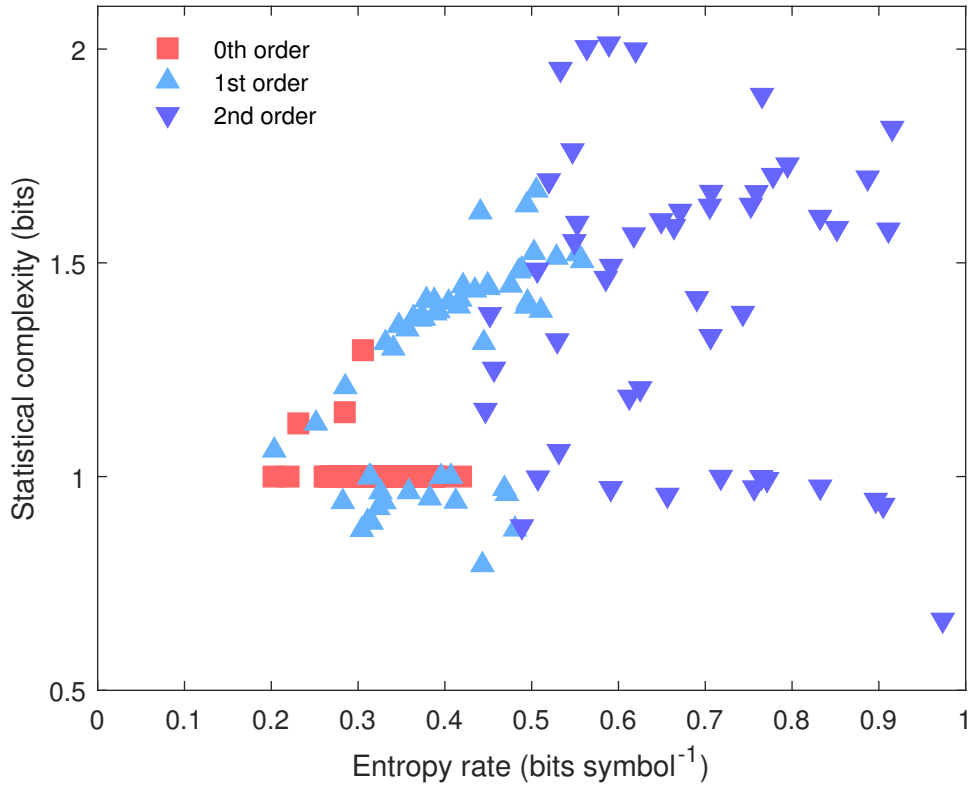


Figure 4.3: (Color) Statistical complexity vs entropy rate for all three orders of difference and all 54 sites. The sites with the same type of ϵ-machines are plotted with the same markers.

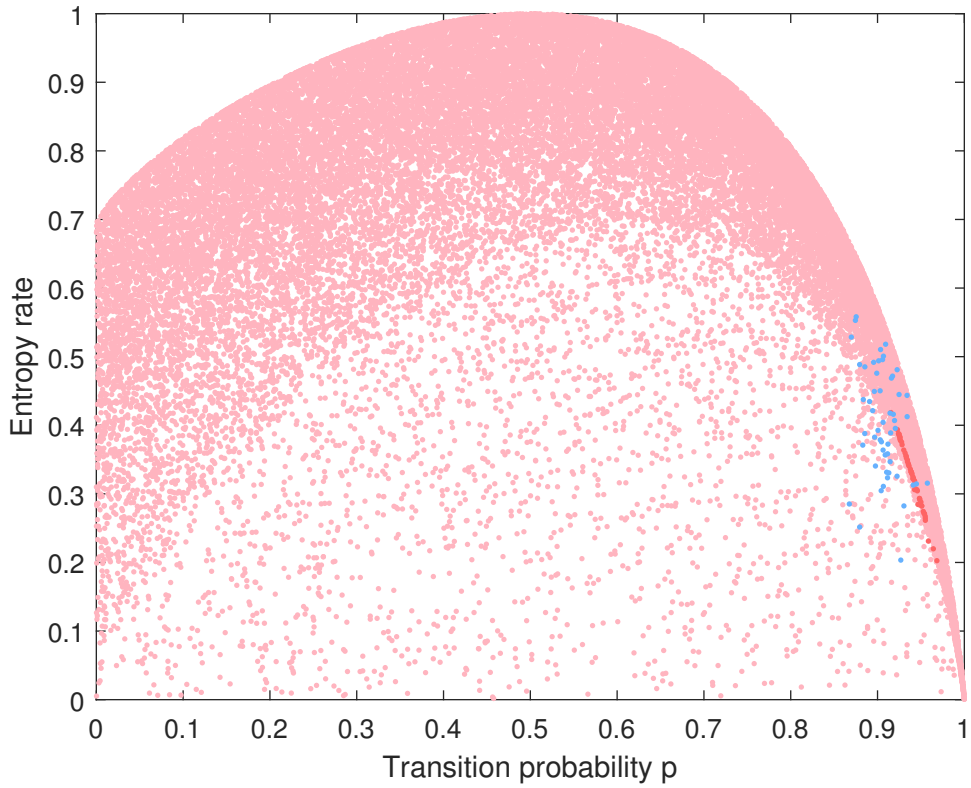


Figure 4.4: Entropy rate vs transition probability p. Pink dots are the results of simulations using the three types of generic ϵ-machines. Red dots represent the 0th order difference of soil moisture, and blue dots represent the 1st order.

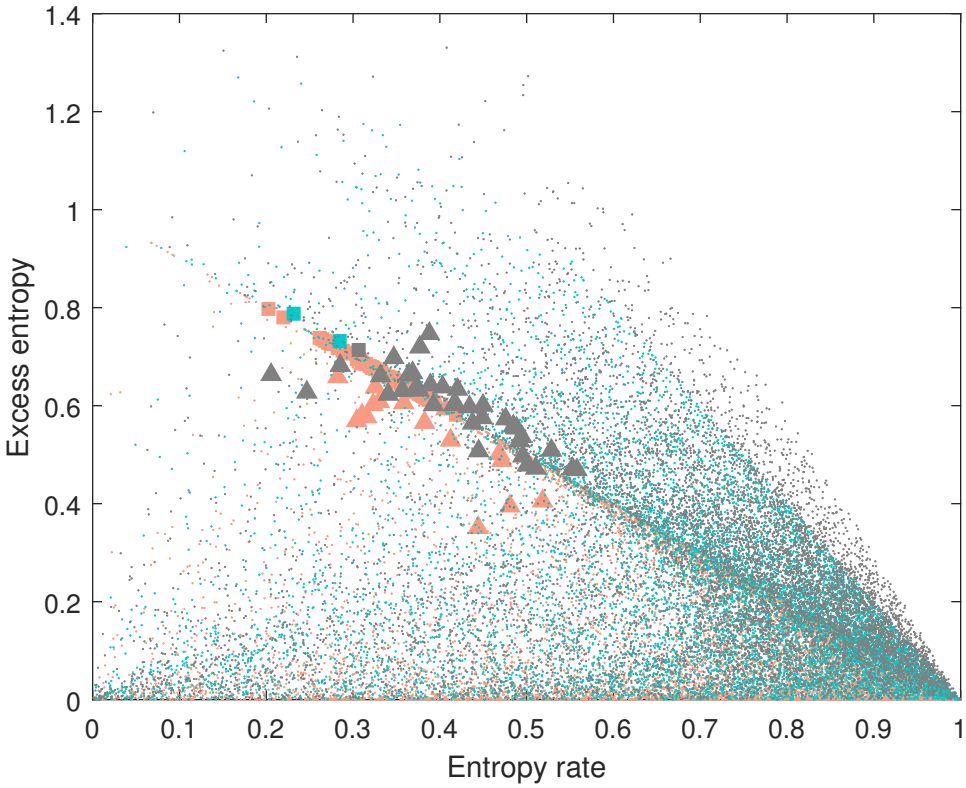


Figure 4.5: Entropy rate vs excess entropy for three types of ϵ-machines and for first two orders of differences for soil moisture. Small dots are simulated results using generic ϵ-machines. Triangles and squares represent 0th and 1st order difference of soil moisture measurements respectively.

CHAPTER 5

General Conclusion

This research investigated soil moisture temporal and spatial patterns from both observational and modeling perspectives. In Chapter 2, spatial patterns of soil moisture at the mesoscale were directly observed and analyzed. Chapter 3 examined the spatial scaling relationships of soil moisture with information extracted from its temporal patterns. In Chapter 4, computational mechanics models were constructed for multiple soil moisture sites distributed in a mesoscale environmental network.

Generally speaking, soil moisture studies still rely heavily on high-quality observations. Continued development of various kinds of improved measurements at different spatial and temporal scales is crucial in advancing knowledge on soil moisture. With the current limited amount of soil moisture data, methods like nonlinear phase space analysis may still be challenging to apply. New methods for pattern discoveries may be helpful in giving insights into the behaviors and even in understanding mechanisms of soil moisture dynamics.

Specifically, the key findings of this work are summarized as follows:

(1) Mesoscal soil moisture was more strongly correlated with sand content ($r = -0.536$ to -0.704) than with antecedent precipitation index (API) for most survey dates. Land surface characteristics exhibit coherent spatial patterns at scales up to 20 km, and those patterns often exert a stronger influence than do precipitation patterns on mesoscale spatial patterns of soil moisture.

(2) At the mesoscale, the correlation lengths of soil moisture, sand content, and API ranged from 12-32 km, 13-20 km, and 14-45 km, respectively.

(3) Upscaling of soil moisture using the method of phase space analysis is possible, but the local polynomial map method was unable to improve the prediction accuracy compared to linear regression and CDF matching.

(3) The unsynchronized behavior of the point-scale and field-scale soil moisture dynamics implies that some errors cannot be eliminated for upscaling due to not only randomness but also some deterministic reasons.

(4) Soil moisture can be described by 2- to 5-state ϵ-machines. As a dynamical system, soil moisture is complex ($C_\mu > 0$), hidden ($\chi > 0$), and unpredictable ($h_\mu > 0$) to some degree.

(5) Increasing orders of derivatives for soil moisture processes shows interesting patterns on structural complexity and predictability - with increasing randomness of the process, the variability in statistical complexity tends to increase.

NOMENCLATURE

$\overleftarrow{\boldsymbol{S}}$ ..... set of all pasts
$\mathbf{E}$ ..... excess entropy
$\mathcal{A}$ ..... alphabet
$\mathcal{S}$ ..... set of all states
$\mathcal{T}$ ..... set of all transitions
$\overleftarrow{S}$ ..... past of a process
$\overleftarrow{s}$ ..... a particular past, a realization of the random variable
$\overrightarrow{S}$ ..... future of a process
ρ ..... autocorrelation
θ_g ..... gravimetric water content
$\theta^{(f)}$ ..... field-scale soil moisture
$\theta^{(p)}$ ..... point-scale soil moisture
C_μ ..... statistical complexity
H ..... entropy
$H(\overrightarrow{S}^L)$ .. block entropy
h_μ ..... entropy rate
N ..... neutron counts
N_0 ..... neutron intensity over dry soil when all hydrogen sources within the footprint are taken into account
R ..... effective states
S ..... a state
T ..... a transition
w_{lat} ..... lattice water content

APPENDIX A

Stationary processes

A *stochastic process* is a series of random variables X_t, where t represents the time when the observation X_t is made. A stochastic process is *strict-sense stationary* if the joint distributions of the process for all k and τ are equal (Cressie, 1993),

$$f_X(x_{t_1}, x_{t_2}, ..., x_{t_k}) = f_X(x_{t_1+\tau}, x_{t_2+\tau}, ..., x_{t_k+\tau}). \tag{A.1}$$

A *wide-sense stationary process* is defined as a stochastic process if its mean (Eq. A.2) and covariance (Eq. A.3) are time-invariant,

$$E[X(t)] = \mu \tag{A.2}$$

$$E[X(t)X(t+\tau)] = C(\tau) \tag{A.3}$$

where the function $C(\,\cdot\,)$ is called a *covariogram*.

A process is *ergodic* if its time average $A[x(t)]$ is equal to its statistical mean,

$$E[X(t)] = A[x(t)] = \mu \tag{A.4}$$

where $A[x(t)]$ is defined as

$$A[x(t)] = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt \tag{A.5}$$

APPENDIX B

Information theory

The followings are basic quantities and formulas of information theory (Cover and Thomas, 2006). The entropy of a discrete random variable is defined as

$$H(X) = -\sum_{x\in\mathcal{A}} p(X=x)\log p(X=x) \tag{B.1}$$

where x are realizations of X. When the logarithm has its base of 2, the unit of entropy is *bit*.
The joint entropy $H(X,Y)$ of two random variables X and Y is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} p(X=x,Y=y)\log p(X=x,Y=y) \tag{B.2}$$

The conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{A}} p(X=x)H(Y|X=x) &\text{(B.3)}\\ &= -\sum_{x\in\mathcal{A}} p(X=x)\sum_{y\in\mathcal{B}} p(Y=y|X=x)\log p(Y=y|X=x) &\text{(B.4)}\end{aligned}$$

The mutual information is defined as

$$I(Y;X) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{B.5}$$

It can be derived that mutual information is the reduction of uncertainty in X given Y,

$$I(Y;X) = H(X) - H(X|Y) \tag{B.6}$$

For a stochastic process X_i, which is a sequence of random variables, the entropy rate is defined by

$$H(\mathcal{X}) = \lim_{L\to\infty}\frac{1}{L}H(X_1,X_2,\ldots,X_n) \tag{B.7}$$

when limit exists.

APPENDIX C

Equivalence relation

The definition and properties of equivalence relation are listed below (Martin, 2010). A relation $\sim$ on a set $\mathcal{A}$ is an equivalence relation if it is

1. Reflexive: $x \sim x, \forall x \in \mathcal{A}$
2. Symmetric: $x \sim y \Rightarrow y \sim x$
3. Transitive: $x \sim y \text{and} y \sim z \Rightarrow x \sim z$

The definition of equivalence relation induces the definition of equivalence class, written as $[x]$. For an equivalence relation R on a set $\mathcal{A}$, the subset $[x]$ contains all the elements equivalent to x, i.e.

$$[x] = \{y \in \mathcal{A} : y \sim x\} \tag{C.1}$$

The two definitions induce a theorem of partition: If $\sim$ is an equivalence relation on a set $\mathcal{A}$, the equivalence classes with respect to $\sim$ for a partition of $\mathcal{A}$, and two elements of $\mathcal{A}$ are equivalent if and only if they are elements of the same equivalence class.

A partition on a set $\mathcal{A}$ refers to

1. $[x_i] \neq \emptyset$
2. $\cup_i [x_i] = \mathcal{A}$
3. $[x_i] \cap [x_j] \neq \emptyset, i \neq j$

References

Cover, T. M. and Thomas, J. A. (2006). Elements of information theory 2nd edition.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics Applied probability and statistics. Wiley, New York, rev. edition.

Martin, J. C. (2010). *Introduction to Languages and the Theory of Computation*. McGraw-Hill Education, Inc., 4 edition.

VITA

Jingnuo Dong

Candidate for the Degree of

Doctor of Philosophy

Thesis: Spatial and temporal patterns of soil moisture: a study on soil moisture observation and modeling

Major Field: Soil Science

Biographical:

Personal Data: Born in Harbin, China in June 1988.

Education:

Completed the requirements for the degree of Doctor of Philosophy with a major in Soil Science at Oklahoma State University in April 2020.
Received a Masters' of Science in Plant and Soil Science at Oklahoma State University in May 2013. Received a Bachelors of Science in Resource and Environmental Science at China Agricultural University in July 2010.

Experience:

Research Associate - Oklahoma State University, 2014-2020
Volunteer - Santa Fe Institute, 2013-2014
Research Assistant - Oklahoma State University, 2010-2013