Understanding Hydrologic Processes and Correlations using Modeling and Machine Learning with Remote Sensing and In-Situ Wireless Sensor Network Data

by

Germán Augusto Villalba Fernández de Castro

Bachelor of Science in Civil Engineering, Universidad de Los Andes, 2002

Master of Science in Computer Engineering, Universidad de Los Andes, 2004

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Germán Augusto Villalba Fernández de Castro

It was defended on

July 15, 2019

and approved by

Vikas Khanna, Ph.D., Associate Professor, Department of Civil and Environmental Engineering

Jeen-Shang Lin, Ph.D., Associate Professor, Department of Civil and Environmental Engineering

Zhi-Hong Mao, Ph.D., Professor, Department of Electrical and Computer Engineering

Dissertation Director: Xu Liang, Ph.D., Professor, Department of Civil and Environmental Engineering

Copyright © by Germán Augusto Villalba Fernández de Castro

2019

Understanding Hydrologic Processes and Correlations using Modeling and Machine Learning with Remote Sensing and In-Situ Wireless Sensor Network Data

Germán Augusto Villalba Fernández de Castro, PhD

University of Pittsburgh, 2019

This work addresses three challenging issues about the overall applicability of hydrologic modelling. The first challenge is improving the collection of sub-surface data. Our approach uses a long-term deployment of wireless sensor network with environmental sensors. This approach is cost-effective when compared with the use of data-loggers and more flexible as it allows real-time monitoring of environmental variables. The plot scale environmental data is collected from our own WSN, deployed in western Pennsylvania, currently composed by 104 nodes and over 240 sensors including commercially available soil moisture, water potential and temperature sensors along with lab-made xylem sap flow sensors.

The second challenge is improving the availability and accuracy of continuous streamflow time-series estimates. The hydrometric network is modelled as a sparse Gaussian graphical model where each site represents a node in a graph. The graph model will have an edge between two sites only when their streamflow time-series are conditionally dependent given the other sites. A novel algorithm is presented, estimating a sparse graph by imposing sparsity to the precision (covariance inverse) matrix via the Graphical Lasso algorithm. The resulting graph is used for inference and a second algorithm determines which gauges can be removed with the least loss of information. The estimated streamflow time-series have better accuracy that other methods based on geographic proximity (least distance) or marginal correlation.

The third challenge is estimating the soil-water characteristics from biased and noisy observations of soil moisture. A novel method is presented for the simultaneous estimation of soil moisture and soil-related parameters. The simulation of soil moisture is performed using the Noah and the VIC models. The simulated site is a well-documented testbed in the state of Oklahoma. The calibration of the soil-related parameters uses Machine Learning techniques such as clustering, regression and classification, and soil-water correlations, providing physical and statistical constrains in the parameter space. Thus, the search is made within a reduced parameter space which makes the parameter calibration approach more effective and realistic. The performance of

the calibration algorithm is assessed regarding the quality of the soil moisture estimations while keeping the parameters in a feasible range.

Table of Contents

Acknowledgements ..... **xiii**

1.0 Introduction **1**

2.0 A Networked Sensor System for the Analysis of Plot-Scale Hydrology **3**

2.1 Introduction **3**

2.2 Materials and Methods **5**

2.2.1 Equipment **5**

2.2.1.1 Initial Equipment (MICAz, IRIS, MDA300) **7**

2.2.1.2 Updated Equipment (TelosB Motes and Custom Sensor Board) **7**

2.2.1.3 Enclosures **9**

2.2.2 Mote Application **10**

2.2.3 Deployment **13**

2.2.4 Calibration of Soil Moisture Measurements **15**

2.2.5 Hydraulic Properties from Soil Moisture Measurements **16**

2.2.6 Transpiration Calculations from Sap Flow Measurements **17**

2.2.7 Geostatistical Analysis of Soil Moisture and Soil Water Potential: Spatiotemporal Trends **20**

2.3 Results and Discussion **21**

2.3.1 Data Quality Assessment of WSN Sensors **21**

2.3.2 Hydraulic Properties Estimation **22**

2.3.3 Sap Flow Data and Transpiration Estimation **25**

2.3.3.1 Sap Flow Time Series **25**

2.3.3.2 Transpiration Estimates from Sap Flow Measurements 27

2.3.4 Exploration of Soil Moisture and Soil Water Potential: Spatiotemporal Trends 30

2.3.5 WSN Challenges and Utility in Hydrology 38

2.3.5.1 Power Management 38

2.3.5.2 Node Maintenance 40

2.3.5.3 Network Routing and Scaling 41

2.3.5.4 Heterogeneous Mote Reprogramming 42

2.3.5.5 Network Costs 43

2.3.6 Lessons Learned with Sap Flow 45

2.4 Conclusions 45

3.0 Estimation of Daily Streamflow from Multiple Donor Catchments with the Graphical Lasso 48

3.1 Introduction 48

3.2 Common Approaches for Transferring Streamflow at Ungauged Basins 53

3.2.1 Drainage Area Ratio 54

3.2.2 Scaling by the Mean (SM) 54

3.2.3 Scaling by the Mean and Standard Deviation (SMS) 55

3.2.4 Linear Regression 55

3.3 New Approach of Selecting Multiple Donor Gauges Via Graphical Models 56

3.3.1 Multiple Linear Regression (MLR) 57

3.3.2 Concept of Gaussian Models and MLR 60

3.3.3 Relationship between MLR, the Covariance, and the Precision Matrices .. 61

3.3.4 The Graphical Lasso 65

3.3.5 Graphical Model Selection 66

3.3.5.1 Imposition of Sparsity to the Underlying Graphical Model 67

3.3.5.2 Preparation of Data Sets 68

3.3.5.3 Estimation of Training Covariance and Sparse Precision Matrices. 68

3.3.5.4 Estimation of Regression Coefficients and Streamflow Validation .. 69

3.3.5.5 Score Function and Validation Error 70

3.3.5.6 Selection of Graph Model Algorithm (SGM) 71

3.3.6 Stream Flow Inference 74

3.3.6.1 Inference of Daily Streamflow Time Series with Graph (SGM) 74

3.3.6.2 Inference of Daily Streamflow Time Series using Distance and Correlation Approaches 75

3.3.6.3 Estimation of the Test Error 76

3.3.6.4 Estimation of Inference Accuracy 77

3.3.7 Removal of Streamflow Gauges with the Least Loss of Information 77

3.4 Study Area and Data Sets 79

3.5 Results and Discussion 80

3.5.1 Inference on Streamflow 80

3.5.2 Removal of Gauges with the Least Loss of Information 87

3.6 Conclusions 97

4.0 Estimation of Soil Type and Related Soil Parameters for Land Surface Models based on Soil Moisture Observations 100

4.1 Introduction 100

4.2 New Methods 102

4.2.1 Estimation of Soil Parameter Data-Sets 104

4.2.1.1 Background on Soil Parameter Estimation 104

4.2.1.2 Soil Type Lookup Table Approach to Estimate Soil Parameters ... 105

4.2.1.3 Global Approach to Estimate Soil Parameters 105

4.2.1.4 Local Approach to Generate Soil Parameters 110

4.2.2 Soil Moisture Content Simulation through Land Surface Models 112

4.3 Estimating Soil Type and Soil Parameters based on SMC Observations 114

4.3.1 Forcing Data 115

4.3.2 Soil Parameter Estimations and Generation of Ground truth SMC 115

4.3.2.1 Soil Parameter Estimations 115

4.3.2.2 Generation of Ground truth SMC Time Series 116

4.3.3 Generation of Soil Moisture for the Noah Model 119

4.3.4 Optimized Estimation of Soil Type and its Associated Parameters 120

4.4 Results 120

4.4.1 Spatial Distribution of the SMC and Inference of Soil Types 122

4.4.2 Comparison of Estimated Soil Parameters 125

4.5 Summary and Conclusions 129

5.0 Conclusions 132

Bibliography 135

List of Tables

Table 2.1 Soil parameter calibration results for the Clapp and Hornberger, and Van Genuchten PTFs 23

Table 2.2 Comparison of silver maple (Acer saccharinum) sapwood area from [72] measurements and Equation (2.6) estimations in site 2 of the ASWP network 28

Table 2.3 AS/AG calculations based on the field survey within the three survey regions in site 2 of the ASWP 28

Table 2.4 RMSE of the interpolated surfaces 37

Table 2.5 Distribution of nodes by type of application 44

Table 3.1 List of 34 Streamflow Gauges Over the Ohio River Basin 80

Table 4.1 – Estimated mean soil parameters or properties for each given soil type class. 110

Table 4.2 – Summary of the relevant inputs for LSMs used for the simulations 113

List of Figures

Figure 2.1 - Schematics of the environmental sensors deployed at the ASWP network 6

Figure 2.2 - Custom sensor boards for TelosB motes 9

Figure 2.3 - Examples of node types and their enclosures in the ASWP network 10

Figure 2.4 - Mote application architecture 12

Figure 2.5 - Map of the six sites of the ASWP testbed (October 2016 configuration) 14

Figure 2.6 - Field survey area within site 2 for the determination of a representative AS/AG ratio 19

Figure 2.7 - Comparison of soil temperature, matric water potential, and volumetric soil moisture 22

Figure 2.8 - Estimation of the soil hydraulic parameters from data 24

Figure 2.9 - Sap flow results for node 2084 between 20 and 27 July 2016 26

Figure 2.10 - Transpiration (τ) calculations in the ASWP site based on the measurements 28

Figure 2.11 - Comparison of the mean and standard deviation of volumetric soil moisture (SM) 31

Figure 2.12 - Interpolated surfaces (Kriging method) showing the average seasonal variation in volumetric soil moisture (SM) and soil water potential (WP) 33

Figure 2.13 - Interpolated surfaces (Kriging method) showing a comparison between the average fall 35

Figure 2.14 - MobileDeluge, a hand-held mobile mote reprogramming tool 43

Figure 3.1 - Result of Running the SGM Algorithm with the Ohio River Basin Dataset 85

Figure 3.2 - Comparison of the generated graphs 86

Figure 3.3 - Comparison of the observed and inferred daily streamflow time series..... 89

Figure 3.4 - Comparison of the inference accuracy on the removable gauges with the RG algorithm..... 90

Figure 3.5 - Scatter plots between the observed and inferred daily streamflow time series 93

Figure 3.6 - Spatial distributions of the elevation, slope, soil type, and land cover over the study region with graph SGM(25)..... 96

Figure 4.1. Regression equations to estimate the soil type related parameters using the sand and clay 107

Figure 4.2. Regression for the Quartz Content, QTZ..... 109

Figure 4.3. Spatial distributions of the soil types and vegetation types for the study area with 32 by 32 118

Figure 4.4. Time series of root zone SMC, spatially aggregated for each member of the calibrationsimulations, ... 121

Figure 4.5. Spatial distributions of the root zone SMC averaged over the one year Noah model simulation for 122

Figure 4.6. Spatial Distribution of Soil Moisture and Inference of Soil Types over the Study Area 124

Figure 4.7. Soil related Parameters Calibration Comparison 128

Acknowledgements

First, I would like to thank my advisor and committee chair, Dr. Xu Liang, for her guidance through all these years of graduate school and for her financial support. This work was partially supported by the U.S. National Science Foundation under the grants CNS-1319331 and CNS-1320132 to the University of Pittsburgh and IUPUI, respectively. I am also grateful that professor Liang used funds from her William Kepler Whiteford Professorship to provide additional support for me.

Second, I also want to thank the members of my dissertation committee: Dr. Vikas Khanna, Dr. Jeen-Shang Lin and Dr. Zhi-Hong Mao for their time, feed-back and support to my dissertation.

Third, I want to acknowledge the contributions from my collaborators Tyler W. Davis, Shugong Wang, Fernando Plaza and Thomas A. Slater from the University of Pittsburgh; and Dr. Yao Liang, Miguel Navarro, Xiaoyang Zhong and Yimei Li from IUPUI, respectively. In addition, I want to thank my friends Felipe Hernández and Daniel Luna for their support. I am also grateful with my friend Daniel Salas who encouraged me to pursue a PhD in the University of Pittsburgh.

Last by not least, I want to thank my mother Margarita who has always been a supporting figure in my life and my father Faustino who supported my dream of becoming an engineer. I am grateful with my wife Gloria who always supported me despite the hardships of living abroad and my children Sebastian and Valentina for waiting for me to play with them.

1.0 Introduction

The accurate estimation of water and energy fluxes are very important for the optimal operation of water resources assets, water supply management, hydropower generation, forecasting of floods and droughts, estimation of agricultural yield, ecological flow assessment, navigation, the design of engineering structures such as highways and reservoirs, among many other applications. The methods for the estimation of water fluxes can be grouped into two broad categories: physically-based hydrologic models and statistical methods.

The physically-based hydrologic models simulate the water and energy fluxes on the land surface using as input an initial estate and meteorological forcing data measurements. The evolution of the simulated fluxes is driven by equations derived from physical principles. The Noah[1] and the VIC[2] are examples of modern land surface models that consider soil and vegetation related parameters. Those models allow the estimation of variables such as soil moisture content (SMC), soil temperature, snow coverage; in addition to surface run-off that can be used to estimate streamflow at the watershed outlet if a digital elevation model (DEM) is used along with the hydrologic model.

The statistical methods used to be the main approach for estimating water fluxes before cheap computing was available. Its use declined during the 1990s and 2000s. The last decade or so has seen a renaissance of statistical methods due to data-driven approaches such as machine learning gradually becoming main stream. This shift is mainly due to the increase in computing power and the availability of massive data sets.

The ever-increasing availability of data in the form of remote sensing benefit both, the application of hydrological models and the statistical methods. Some relevant examples of remote sensing for environmental monitoring are the space-borne satellites for: MODIS[3], GPM[4] and SMAP[5], for spectroradiometer, precipitation, and soil moisture, respectively.

Despite the improvement in data availability over time, both, the physically-based hydrologic models and the statistical methods often times do not have access to high quality, reliable, high resolution data for a given time and space domain for all the relevant variables. This is particularly true for sub-surface fluxes where the application of remote sensing is severely limited. This dissertation addresses the challenges of collecting data for sub-surface environmental variables such as soil moisture and soil temperature, improving parameterization of hydrological model simulations and improving the availability and accuracy of streamflow time series estimations.

The dissertation is composed by 5 chapters. Chapter 2 addresses the challenge of collecting sub-surface environmental data at the plot scale via the application of wireless sensor networks (WSNs). Chapter 3 addresses the challenge of inferring streamflow time series and the removal of gauges from a hydrometric network with the least loss of information using machine learning. Chapter 4 addresses the challenge of improving the estimation of soil moisture and soil related attributes for physically-based hydrologic models. Chapter 5 presents the summary and conclusions.

2.0 A Networked Sensor System for the Analysis of Plot-Scale Hydrology

Most of the content in this chapter was previously published in the open access journal, *Sensors*, under the title "A networked sensor system for the analysis of plot-scale hydrology" [6].

2.1 Introduction

The sustainable condition of our freshwater resources partially depends on our understanding of the natural system in which it is cycled [7]. It has long been known that physically-based distributed hydrologic models require an understanding of the spatiotemporal variability of environmental data, which is difficult without an abundance of ground-based measurements for calibration and validation [8]. Soil moisture and transpiration play a fundamental role in the soil-atmosphere interactions and eco-hydrological processes. Moreover, the impacts of these and other hydrological parameters on regional hydrologic and climatologic conditions need permanent in situ measurements. Exploring the variability of soil moisture and transpiration at the plot scale and qualifying such measurements statistically can help improve estimates (including flux and storage components) of water budgets at the regional/watershed scale conditions need permanent in situ measurements. Exploring the variability of soil moisture and transpiration at the plot scale and qualifying such measurements statistically can help improve estimates (including flux and storage components) of water budgets at the regional/watershed scale [9]–[11].

Ground-based measurements and monitoring of environmental variables have been impacted over the past decade by wireless sensor network (WSN) technology. Traditional data logging methods use cumbersome equipment that are expensive to operate, and inconvenient to maintain, leading to limited spatial coverage capabilities. Because of the high expense of sensors and data logging equipment, researchers are often forced to either forgo data loggers for high spatial density measurements with poor temporal resolutions (i.e., hand measurements) or obtain high temporal resolution at a limited number of strategically located data loggers.

Small, inexpensive, wireless monitoring devices are pervading beyond networking and communications research fields. These devices are providing scalable, high resolution data at a declining cost [12], [13] and have found applications in a variety of environmental monitoring fields, including: habitat monitoring [14]–[16], microclimate monitoring [17], [18], seismology [19], [20], understory sunlight studies [21], agriculture [22]–[30], ecology [31]–[34] and hydrology [35]–[39]. High-resolution sensor networks of plot-scale hydrology is a growing application for WSNs due, in part, to the increasing demand for calibrating and characterizing sub-grid variability of airborne and space-borne measurements [40], [41].

A long-term (over six years) WSN has been measuring edaphic (e.g., moisture, water potential and temperature) and arboreal (i.e., xylem sap flow) hydrological properties in a forested nature reserve at the Audubon Society of Western Pennsylvania (ASWP) [38]. The original motivation of the ASWP network was to determine the feasibility of using WSNs to continuously and reliably collect hydrological data under natural outdoor conditions. Following the successful deployment of the network, which has been running on TinyOS 2.1.2 [42] and CTP [43], comes a new stage of research, starting in 2015, aimed towards network expansion and improvement of data collection and processing.

The novelty of this WSN study includes: (1) a data acquisition board design and software driver for integrating digital environmental sensors into the wireless hardware platform; (2) an energy efficient and balanced routing protocol CTP + EER [44]; and (3) a heterogeneous over-the-air mote-reprogramming tool. Furthermore, our WSN enables the study of the following: (1) assessment of the quality of the data collected from soil moisture, soil water potential and soil temperature sensors attached to WSN nodes using a specially designed sensor board; (2) retrieval of high quality sap flow measurements from our lab-made [45] Granier-style [46], [47] sap flow sensors; (3) the application of the collected data to: (a) estimate soil hydraulic properties; (b) calculate transpiration based on sap flow; and (c) explore spatiotemporal patterns of soil moisture and soil water potential; and (4) evaluation of the utility of WSNs for environmental monitoring applications. To the best of our knowledge, this is the first comprehensive study to address these important questions from a single network perspective.

2.2 Materials and Methods

2.2.1 Equipment

Three types of external sensors were used throughout the study site. The first two are the MPS-1 and EC-5 sensors (Decagon Devices, Pullman, WA, USA) which provide measurements of matric water potential (WP) and volumetric soil moisture (SM) (Figure 2.1 a, b, respectively). The Decagon Devices MPS-2 digital sensor [48], which provides measurements of soil temperature in addition to WP, is also deployed within the network due to the discontinuation of the MPS-1 sensor. The third sensor is a pair of Granier-style thermal dissipation (constant heat)

sap flow sensor probes (Figure 2.1 c), which were made and calibrated following [45]. The sap flow probes are connected to the mote's sensor board via a control circuit (Figure 2.1 d) to amplify and condition the thermocouple response voltage as shown in [49]. The sap flow control circuit is operated by a 12 V lead-acid battery to accommodate the additional power requirements of the thermal dissipation method.

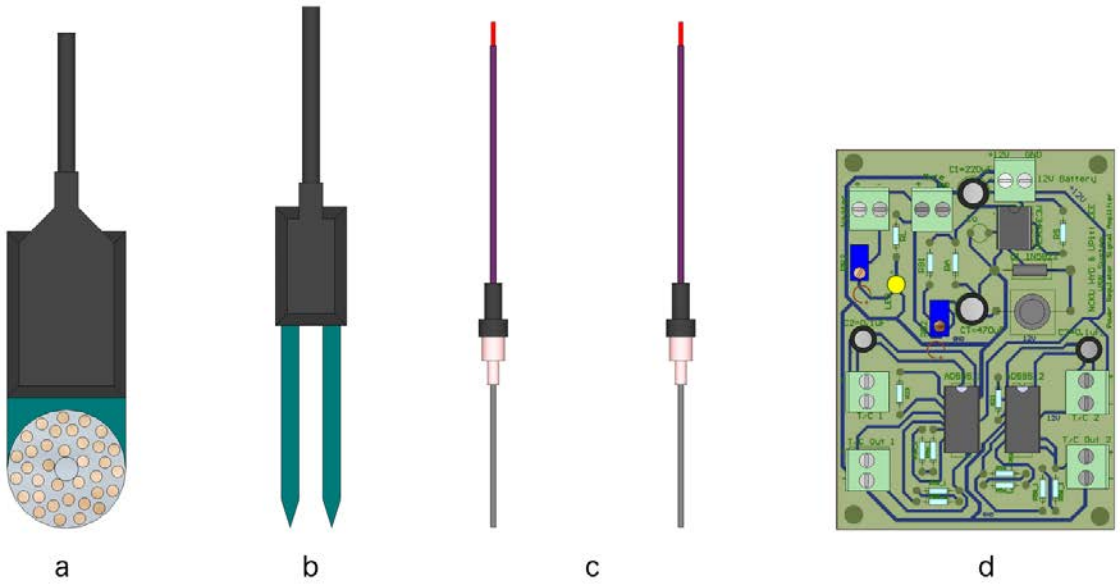

Figure 2.1 - Schematics of the environmental sensors deployed at the ASWP network (a) Decagon Devices MPS-1/MPS-2 soil water potential sensor; (b) Decagon Devices EC-5 soil moisture sensor; (c) thermometric sap flow sensor probes; and (d) sap flow sensor circuit

2.2.1.1 Initial Equipment (MICAz, IRIS, MDA300)

The network was built on an existing investment in WSN hardware, manufactured by Crossbow Technology (now MEMSIC, Inc.), which includes the MPR2400 (MICAz) and XM2110 (IRIS) processor and radio boards (i.e., wireless motes) and the MDA300 sensor board. The wireless motes are powered using rechargeable nickel-metal hydride (NiMH) batteries (size AA and D). The batteries, after recharging, are sorted based on their rested voltages to avoid deploying partially charged or uncharged batteries (see Section 2.3.5.1 for details). The data collection software, including data sampling and packet routing, is developed based on the state-of-the-art open-source WSN platform TinyOS [42].

2.2.1.2 Updated Equipment (TelosB Motes and Custom Sensor Board)

Starting in 2015, the wireless motes have gradually been updated to the CM5000-SMA (TelosB) by Advanticsys (Madrid, Spain) with our specially designed sensor board for data acquisition, forming a heterogeneous WSN consisting of MICAz, IRIS and TelosB motes. TelosB motes incorporate the 2.4 GHz CC2420 transceiver and the MSP430 microcontroller with 10 KB of RAM and provide a 16-pin expansion interface to connect external sensors. For the sensors described above, an excitation voltage is required before a reading. In the standard method, general-purpose input/output (GPIO) pins are used as excitation pins and ADC pins are used to gather sensor readings; however, when powering sensors directly using GPIO pins, the excitation voltage is unstable and can change under different workloads and battery levels. Since analog sensor readings are proportional to the excitation voltage, the readings might be inconsistent even in the same environment.

To address this problem, a novel custom sensor board was designed for TelosB motes (Figure 2.2) using voltage regulators to provide a stable excitation voltage. In the literature, TelosB acquisition boards are usually designed for specific sensors, such as motion sensors [50]–[52], physiological sensors [52], and SM sensors [53]. In contrast, our design is a generic sensor board for the TelosB mote. Our sensor board has two distinguishing features. First, it has ADC channels for analog sensor output and a UART channel for digital sensor output. Second, it provides two levels of stable excitation voltage that enables different combinations of external sensors to be attached based on the application configuration, which is essential in heterogeneous networks, such as the ASWP network.

The sensor board is attached to the TelosB mote's 16-pin expansion, providing screw wire connectors, ADC channels, UART0 serial port, and two excitation voltages. Analog sensors, such as EC-5 and MPS-1, can be attached to ADC channels and powered through a 2.5 V excitation voltage, obtained by using a TLV70025 voltage regulator (Texas Instrument, Dallas, TX, USA). Digital sensors can be connected to the UART0 serial port, such as the MPS-2, which generates a byte stream representing ASCII characters as its sensor readings. The MPS-2 requires an excitation voltage between 3.6 V and 15.0 V. A 5 V voltage booster (U1V11F5, Pololu, Las Vegas, NV, USA) is included in the custom sensor board to provide a 5.0 V excitation voltage from the 3.6 V nominal battery supply to power the MPS-2 sensors.

Figure 2.2 - Custom sensor boards for TelosB motes

2.2.1.3 Enclosures

The wireless motes and all necessary electronics are housed inside water-tight polycarbonate high-impact enclosures, connected to an external omni-directional high-gain (4.9 dBi) antenna, and are discretely hung from tree limbs, attached to PVC posts, or mounted to the sides of trees (Figure 2.3). Motes with sap flow sensors deployed during the 2015 and 2016 growing seasons were powered by the sap flow control circuit's power regulator; therefore, these motes did not require the AA or D rechargeable batteries.

Figure 2.3 - Examples of node types and their enclosures in the ASWP network
(a) relay nodes hanging from a tree branch; (b) sap flow node mounted to the side of a tree; and (c) soil sensor node mounted to a PVC pipe

2.2.2 Mote Application

The motes application was developed in TinyOS 2.1.2 [42] with an adoption of the newly developed routing protocol CTP+EER [44]. TinyOS is the most widely used WSN operating system and is found in 60% of WSN deployments [54], [55]. Owing to its popularity, TinyOS has a larger community, better documentation, and well-tested drivers and protocols compared to other WSN operating system alternatives [48].

CTP+EER is an efficient and balanced routing protocol that extends CTP [43]. CTP is the de facto standard collection routing protocol in TinyOS, in which each mote finds the best route to the sink (i.e., base station). Unfortunately, with this protocol, network data traffic tends to concentrate on a few specific motes that provide the best routes within the network. Thus, these motes experience heavy congestion and deplete their batteries faster than their neighbors. In contrast, CTP + EER, while maintaining the best primary route within the network, also allows motes to select suboptimal routes from a parent set; therefore, it can reduce the data traffic at the busiest motes and provide better overall energy efficiency and balance. CTP + EER has been evaluated though analytical modeling, simulations, and testbed experiments. Compared with CTP, CPT + EER achieves better packet reception ratio, load balance, and energy efficiency. Please see [44] (and the references herein) for more details.

The wireless motes form a multi-hop collection network operating in asynchronous low-power listening (LPL) [56]. The logical architecture of the application is presented in Figure 2.4. Each node uses CTP + EER to deliver two types of packets: data packets (DataSenderC) and summary packets (SumSenderC). Data packets are periodically sampled at a base interval of 30 min (DataTimerC). Randomness is added to each mote to avoid bursty network traffic (RandTimerC). Summary packets enable efficient network diagnosis (NetStatsC) and are generated every two hours (SumTimerC), which includes the network statistics such as retransmission, dropped packets, and information about the routing control traffic.

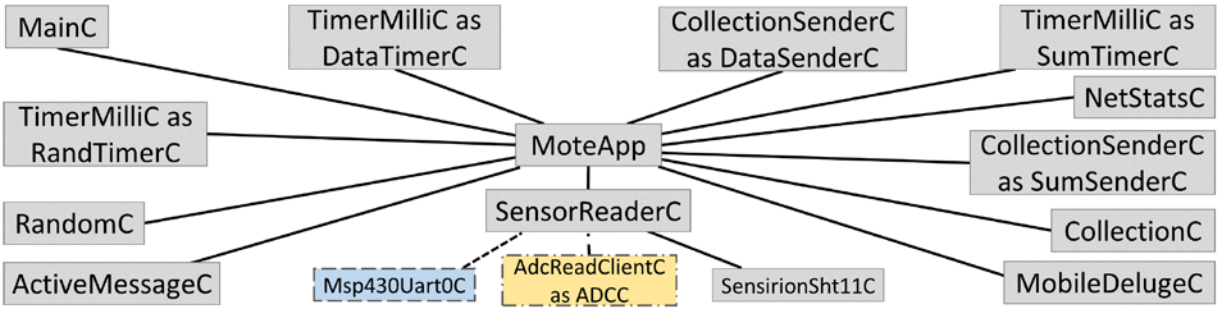

Figure 2.4 - Mote application architecture

The mote application was adjusted based on the sensor types attached to each individual mote. There are two types of nodes: relay nodes and sensor nodes. Relay nodes have no external sensors and are used at advantageous locations to improve communication throughout the network (e.g., hanging in trees as shown in Figure 2.3 a). Sensor nodes are nodes with external environmental sensors (e.g., SM, WP, soil temperature or sap flow). Sensor nodes also participate in network routing and packet forwarding in data collection. Relay nodes only have temperature and humidity sensors (Sht11C, Sensirion, Zürich, Switzerland) [57]. Sensor nodes have analog sensors attached that utilize the ADC channels. In addition, some sensor nodes with TelosB motes also have a digital sensor that communicates through the UART0 serial port (Msp430Uart0C).

To facilitate the reprogramming of motes that are deployed in difficult-to-access enclosures, an over-the-air mobile mote-reprogramming tool was utilized such that direct access to the mote hardware was not necessary. While many over-the-air programming approaches have been proposed for WSNs, none of them apply to heterogeneous and low power WSNs [58], [59]. The novel mobile mote reprogramming tool, MobileDeluge [58] (see Section 2.3.5.4), was developed to overcome this limitation.

2.2.3 Deployment

The ASWP network was initially formed in 2010, which culminated in 2014 as a 52-node deployment located over five sites as described in [38]. The network has since doubled in size (i.e., 104 nodes) due to a 36-node addition during the summer of 2015 and a 16-node addition during the summer of 2016. The study area now includes six sites, of which five are designated areas for environmental monitoring. Figure 2.5 shows the locations of the relay (yellow circles) and sensor (red and blue circles) nodes throughout the six sites of the deployment. The base station (white square) is located in an office window of the nature reserve building in site 1.

During the initial four years of the network deployment, all sensor nodes consisted of one MPS-1 and two EC-5 sensors. A subset of these sensor nodes was also outfitted with a sap flow sensor and a control circuit (operated during the growing seasons). Since the summer of 2014, sensor nodes have been divided into two classes: soil sensor and sap flow sensor nodes.

The soil sensor nodes include two EC-5 sensors and one MPS-2 sensor (replacing the MPS-1). One of the two SM sensors is co-located with the WP sensor at a depth of 30 cm, separated by enough distance such that the measurements of one would not interfere with the other. In the same hole, the second EC-5 sensor is placed at a depth of 10 cm. Best efforts were made to avoid rocks and tree roots during sensor installation. Holes are drilled into the bottom of the sensor node enclosures to allow the sensor wires to be connected to the sensor board (Figure 2.3 c).

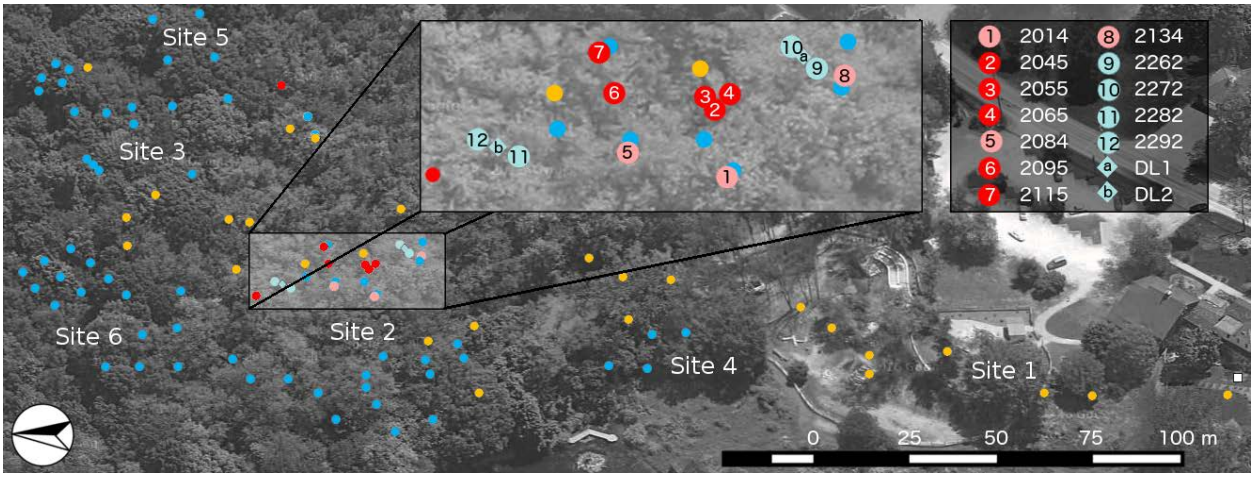

Figure 2.5 - Map of the six sites of the ASWP testbed (October 2016 configuration)
Relay nodes are represented as yellow circles, sap flow nodes are represented as red circles (the three pink circles in site 2 are used in this analysis), soil sensor nodes are represented as dark blue circles, and the base station is represented as a white square. The data loggers used for validation (i.e., DL1 and DL2) are shown as light blue diamonds and their corresponding nodes as light blue circles. The four-digit node numbers referenced in the analysis are indicated in the zoomed region of site 2

The sap flow sensor nodes are equipped with one set of sap flow probes connected to a sap flow control circuit. The sap flow node enclosure houses all the wireless and sensor electronics (Figure 2.3 b). Two holes are drilled through the back wall of the enclosure (i.e., side facing the tree), spaced approximately 10 cm apart for seating the sap flow probes into the tree. Before attaching the enclosures to the tree, a portion of the tree bark is stripped away to create a flat surface for the enclosure box and to increase the penetration depth of the probes into the tree's active xylem. Enclosures are attached to the trees using wood screws with the two enclosure holes aligned parallel to the tree's vertical growth axis. Once enclosures are attached to the tree, pilot holes for the sensor probes are drilled horizontally into the tree at the locations of the two enclosure holes and are given a prophylactic treatment of hydrogen peroxide. To aid in installation and removal while improving the thermal conductivity with the tree xylem, the probe needles are

coated with petroleum jelly. Once inserted into the tree, the probes are fixed inside the enclosure using a silicone adhesive, which also prevents water from entering the enclosure through the holes.

As a means of validating the WSN soil sensor measurements, two Decagon Devices EM50 data loggers were deployed during the summer of 2016 along the hill slope that stretches from the bottom of site 2 to the top of site 3 (i.e., light blue diamonds in Figure 2.5). Accompanying the data loggers are four additional nodes (i.e., light blue circles in Figure 2.5), two surrounding each data logger. The validation data loggers and nodes are located at approximately the midpoint of the hill (i.e., nodes 2282, 2292 and data logger DL2) and close to the lower part of the hill (i.e., nodes 2262, 2272 and data logger DL2) in site 2, respectively. Each validation node is connected to three soil sensors and each data logger is connected to five soil sensors in such a way that five out of the six node sensors are matched with a data logger sensor (i.e., the same location, sensor type and installation depth). The sensor type and installation depth are the same as the other soil sensor nodes in the network.

2.2.4 Calibration of Soil Moisture Measurements

The SM raw data is collected as a voltage (mV) from the EC-5 sensor attached to a sensor board via an ADC (analog-to-digital-converter). The raw data needs to be converted to SM using a conversion equation. Typically, the conversion equation is presented as a linear equation $\theta = c1 * ADC + c0$, where ADC is the raw voltage output (in mV) from the EC-5 sensor and c1 and c0 are the slope and intercept of the fitted linear regression model, respectively. The standard coefficients for non-Decagon data loggers at an excitation of 2.5 V for mineral soils, are $c1 = 0.00119$ and $c0 = -0.401$.

Estimates of SM based on the standard equation showed a bias towards drier conditions. To increase the accuracy of the estimation, the field data collected at the validation locations were calibrated by a linear regression using the ordinary least squares (OLS) method. The targets values are from the validation data logger and the input is the raw data (i.e., ADC in mV from EC-5 sensor) from the validation nodes. The EC-5 sensors were calibrated using an intercept of −0.360 and −0.367 for depths of 10 and 30 cm, respectively, and slopes of 0.0011 and 0.0012 for depths of 10 and 30 cm, respectively. The slope values were found to be similar to the standard value (i.e., 0.00119), while the intercept values are lower in magnitude.

2.2.5 Hydraulic Properties from Soil Moisture Measurements

One of the benefits of in situ plot-scale hydrology studies is the ability to estimate the hydraulic properties that govern the region. These hydraulic properties are important in characterizing a region with estimates to pedotransfer function (PTF) parameters that are utilized by hydrologic models to predict soil water retention properties based on available soil survey data [60], [61]. In this work, two PTFs are examined. The first PTF is the Clapp-Hornberger equation, which is given by the following power curve [62]:

$$s = \left(\frac{\psi}{\psi_s} \right)^{\frac{-1}{b}}, \tag{2.1}$$

where ψ is the hydraulic conductivity, ψ_s is the hydraulic conductivity at saturation, and $s = \theta / \theta_s$ is the soil wetness, a ratio of the SM, θ, to the saturated SM, θ_s (i.e., total porosity).

The second PTF is the van Genuchten equation, given by the following expression [63]:

$$\frac{\theta - \theta_r}{\theta_s - \theta_r} = \left[1 + (\alpha \,|\psi|)^n\right]^{-m}, \quad \textbf{(2.2)}$$

where θ_r is the residual SM, $m = 1 - 1/n$, and n and α are fitting parameters. Assuming, for simplicity, that θ_r is zero, the left-hand side of Equation (2.2) may be expressed in terms of s, such that:

$$s = \left[1 + (\alpha \,|\psi|)^n\right]^{-m}, \quad \textbf{(2.3)}$$

Field measurements of θ and ψ are used for fitting values of ψ_s and b in Equation (2.1) and α and n in Equation (2.3). The value for θ_s was determined experimentally based on the following methodology. A soil core sample was taken at a depth of 10 cm and 30 cm at the two node locations surrounding the validation data loggers (Figure 2.5). The soil samples were weighted under field conditions and then dried in an oven at 100 °C for 48 h before being weighted again. The bulk density was calculated as the dry weight divided by the volume of the soil sample core. The porosity was calculated as one minus the bulk density divided by the particle density, which was assumed as 2.65 g·cm^{-3}.

2.2.6 Transpiration Calculations from Sap Flow Measurements

In this study, the xylem sap flow (i.e., the velocity of the water being transported through the active sapwood of the tree) is calculated using an empirical equation based on daily temperature differences between a pair of heated and reference temperature probes [47], [64]:

$$Q_s = 0.000119 \times \left(\frac{\Delta T_0 - \Delta T}{\Delta T}\right)^{1.231}, \quad \textbf{(2.4)}$$

where ΔT is the temperature difference between the upper and lower probes (°C); ΔT_0 is the maximum daily value of ΔT (i.e., zero sap flow) (°C); and Q_s is the sap flux density ($m^3 \cdot m^{-2} \cdot s^{-1}$). Calculated quantities of Q_s are converted to transpiration based on the following equation [65]–[68]:

$$\tau = Q_s \left(\frac{A_s}{A_G} \right), \tag{2.5}$$

where τ is the rate of transpiration ($m \cdot s^{-1}$); A_s is the tree sapwood area (m^2); and A_G is the ground area (m^2). The ratio A_s / A_G depends on the study site and is indicative of the tree density and the predominant tree species. It has been shown that this ratio can be as low as 1 $m^2 \cdot ha^{-1}$ [69] and reach values as high as 25 $m^2 \cdot ha^{-1}$ [67] or 40 $m^2 \cdot ha^{-1}$ [70]. Estimates of A_s may be determined empirically, such as by the following allometric equation [71]:

$$A_s = B_0 \cdot d^{B_1}, \tag{2.6}$$

where d is the measured tree diameter at breast height (cm) and B_0 and B_1 are species-specific coefficients determined by regression techniques.

In a 2010 sap flow study at the ASWP site [72], 22 trees were surveyed to identify their species, take measurements of d, and estimate A_s from three core samples taken at breast height. Based on the results of that study, only the silver maple (*Acer saccharinum*) trees, predominantly in site 2, produced measurable sap flow quantities.

In 2016, another survey was conducted to estimate A_s / A_G for the silver maples in site 2. Figure 2.6 shows the surveyed area in site 2 for the A_s / A_G ratio estimation (approximately 1300 m^2) where three sap flow nodes (2014, 2084 and 2134) are located (pink circles in Figure 2.5). To

establish the boundary of the surveyed area, first, a preliminary perimeter was defined using trees located around the sap flow nodes. Then, this perimeter was displaced 5.33 m, which is the approximate mean distance between neighboring trees. The area of influence for each sap flow node was established using the Thiessen polygons criterion (i.e., the colored regions in Figure 2.6). These regions were used to calculate one A_s / A_G ratio for each node according to the number and diameter of trees within their area of influence. The diameters at breast height of 25 trees, including the three trees with sap flow sensors, were measured and the sapwood areas, calculated using Equation (2.6), are compared with the measurements made in 2010.

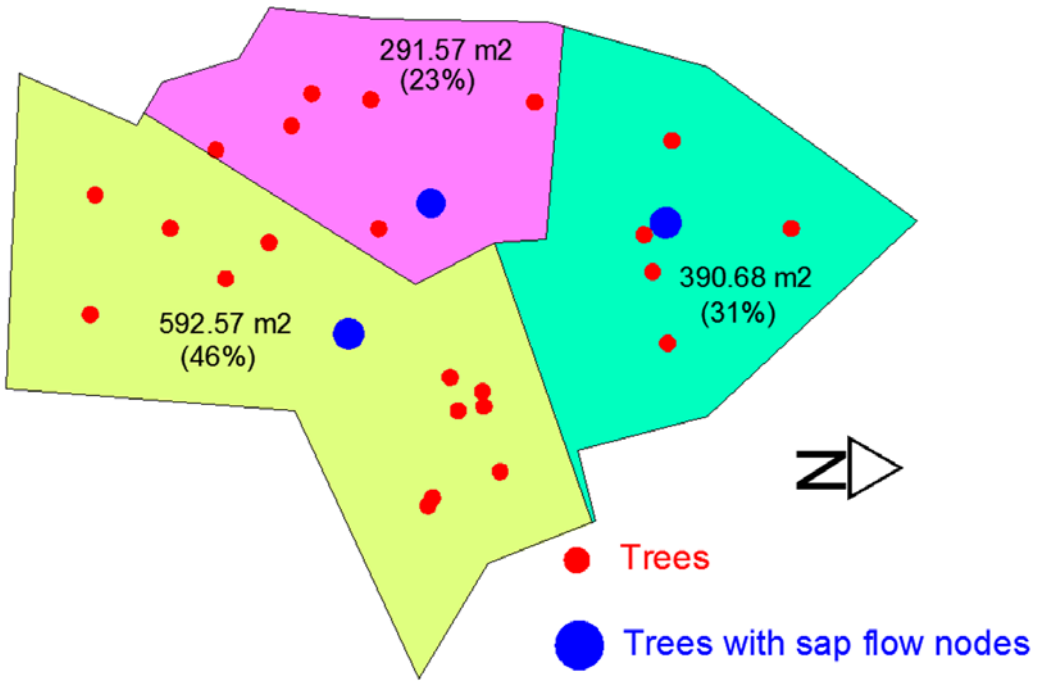

Figure 2.6 - Field survey area within site 2 for the determination of a representative AS/AG ratio

2.2.7 Geostatistical Analysis of Soil Moisture and Soil Water Potential: Spatiotemporal Trends

SM and WP are important variables in the water cycle within climate systems. Thus, the quantitative estimation of these parameters is fundamental for application fields such as weather forecast, hydrology and watershed management [73]. SM, for instance, usually shows strong spatial variability due to physical and geographic characteristics of the environment (e.g., topography, soil type, vegetation coverage) [74], [75]. Surface interpolation methods such as Kriging are widely used to assess the spatial characteristics of hydrologic variables [73], [76]–[81].

The Ordinary Kriging (OK) interpolation method is the most widely used geostatistical interpolation technique and is acknowledged as the standard approach for surface interpolation [41], [71], [76], [82]–[87]. OK assumes that the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. The spatial dependence is expressed by a semi-variogram. This method is appropriate when it is known that there is a spatially correlated distance or directional bias in the data, as is with SM and WP. The OK estimation equation is given by the following:

$$Z_{(s_0)} = \sum_{i=0}^{n} \lambda_i \cdot Z_{(s_i)}\,, \tag{2.7}$$

where $Z_{(s_i)}$ is the measured value at the *i*-th location, λ_i is an unknown weight for the measured value at the *i*-th location, $Z_{(s_i)}$ is the predicted value at the prediction location S_0, and *n* is the number of measured values. The weight, λ_i, depends on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location. In this study, an OK interpolation method with a spherical semi-

variogram model is used to estimate the spatiotemporal trends of SM and WP, since this model has been found to satisfactorily represent the spatial dependence in previous studies [73], [76], [82], [88]–[90]. The root mean square error (RMSE) is used to assess the performance of the selected interpolation method.

2.3 Results and Discussion

2.3.1 Data Quality Assessment of WSN Sensors

The data quality assessment of the WSN soil sensors was performed at the validation locations in site 2 where data logger measurements were accompanied by sensor node measurements at the same time and location. Only the validation results at the midpoint location of site 2 are presented here.

Figure 2.7 shows the comparison results of soil temperature, T (Figure 2.7 a), WP, (Figure 2.7 b), and SM, (Figure 2.7 c), for the time period between 29 July and 23 August 2016 for the data logger DL2 and the nodes 2282 and 2292 at two depths near the midpoint of the hill in site 2. The soil temperature measurements (based on the MPS-2) from the WSN peak slightly higher than the data logger measurements (light lines in Figure 2.7 a) and are indistinguishable at the validation location at the bottom of the hill (not shown). The WP measurements (also based on the MPS-2) are slightly lower at both depths from the WSN compared to the data logger (Figure 2.7 b). The SM measurements (based on the EC-5) are nearly indistinguishable between the WSN nodes and the data logger.

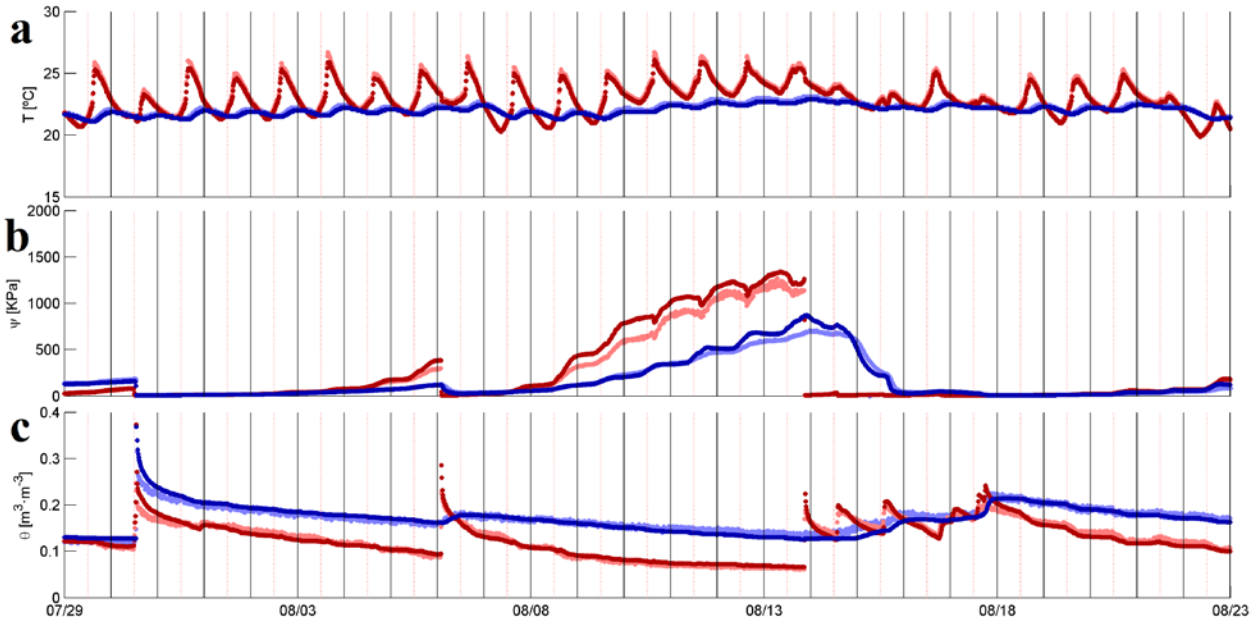

Figure 2.7 - Comparison of soil temperature, matric water potential, and volumetric soil moisture (a) soil temperature, T in Celsius degrees; (b) matric water potential (WP), ψ in kPa; and (c) volumetric soil moisture (SM), θ in m3·m−3 data collected by a data logger (DL2) and wireless nodes (2282 and 2292) from the ASWP network between 29 July and 23 August 2016. The variable at a depth of 10 cm is shown in dark red for the data logger and light red for the nodes. The variable at a depth of 30 cm is shown in dark blue for the data logger and in light blue for the nodes

2.3.2 Hydraulic Properties Estimation

Figure 2.8 shows the empirical relationship between the SM, θ, and the absolute value of the WP, in kPa, $|\psi|$ at a depth of 10 cm and 30 cm for the location close to the middle of the hill (i.e., nodes 2282, 2292 and data logger DL2) in site 2. The fitted equations lead to similar results for both equations. At a depth of 10 cm the fitted parameter for the Clapp and Hornberger equation are: ψ_s = 0.658 kPa and b = 4.49; For the van Genuchten equation: n = 1.215 and α = 1.808. The porosity value θ_s = 0.31 was calculated from a soil core taken in the same location. For the depth

of 30 cm, the parameters for the Clapp and Hornberger equation are ψ_s = 0.521 kPa and b = 5.87; For the van Genuchten equation: n = 1.154 and α = 3.51. The porosity value θ_s = 0.42 was calculated from a soil core taken in the same location. Table 2.1 summarizes these results, including the location close to the lower part of the hill (i.e., Node 2262, 2272 and Data logger DL1). The fitted equations were evaluated using the Nash-Sutcliffe efficiency (NSE) [91].

Table 2.1 Soil parameter calibration results for the Clapp and Hornberger, and Van Genuchten PTFs

Location	Depth (cm)	θ_s	b	ψ_s (kPa)	NSE (Clapp-Hornberger)	n	α	NSE (Van Genuchten)
DL1	10	0.32	9.63	0.045	0.914	1.093	62.34	0.929
DL1	30	0.44	10.69	0.00066	0.800	1.081	10814.0	0.821
DL2	10	0.31	4.49	0.658	0.921	1.215	1.808	0.922
DL2	30	0.42	5.87	0.521	0.801	1.154	3.51	0.812

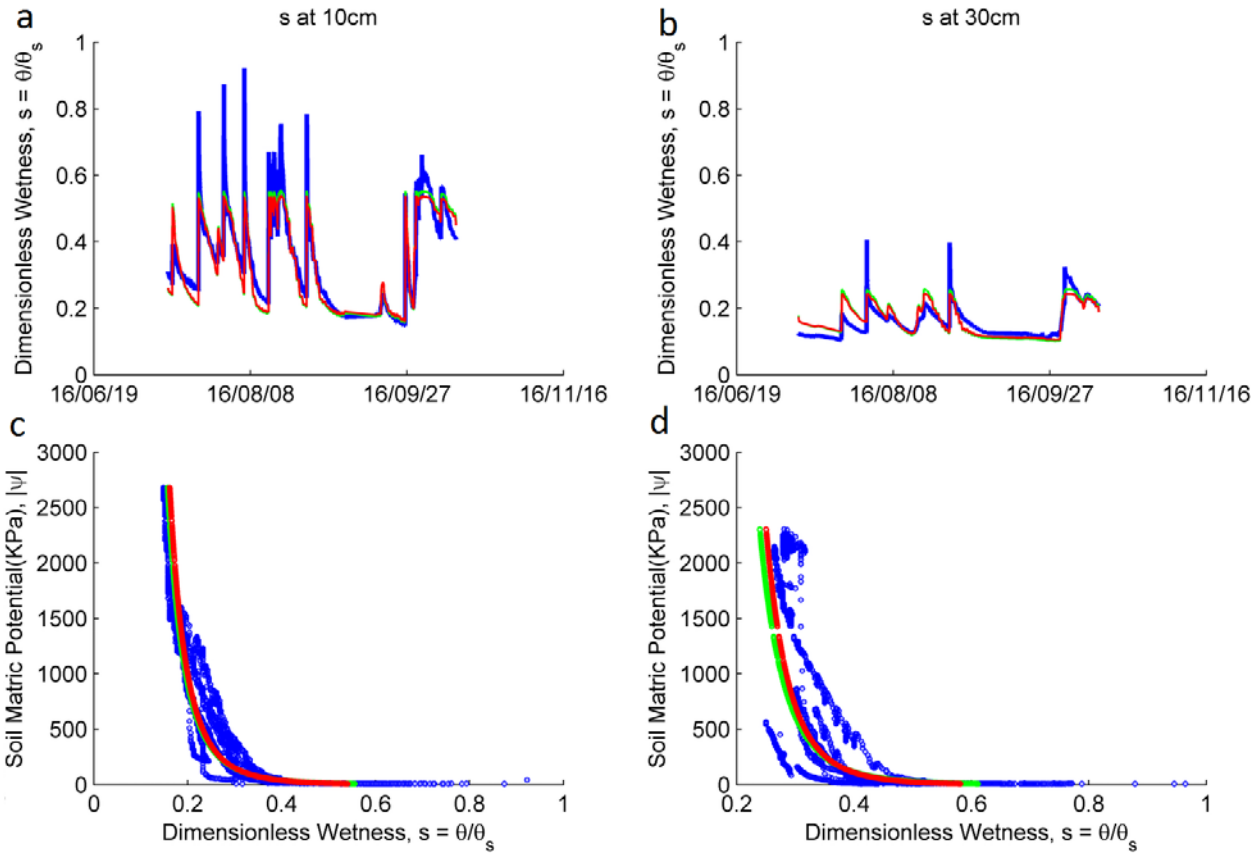

Figure 2.8 - Estimation of the soil hydraulic parameters from data at the location close to the middle of the hill in site 2 (i.e., Nodes 2282 and 2292, and data logger DL2). Wetness, *s*, in blue; Estimated wetness from measured matric water potential (WP) using the Clapp and Hornberger, and van Genuchten equations, in green and red, respectively. (a) Comparison of the wetness, *s*, time series, at a depth of 10 cm; (b) Same as part a, for a depth of 30 cm; (c) Relationship between the soil wetness, *s*, and the absolute value of the WP, in kPa, $|\psi|$ at a depth of 10 cm. The fitted Clapp and Hornberger equation is shown in green and the fitted van Genuchten equation in red; (d) Same as part c, for a depth of 30 cm

2.3.3 Sap Flow Data and Transpiration Estimation

2.3.3.1 Sap Flow Time Series

Figure 2.9 shows the results of the sap flow collected by one WSN node (i.e., node 2084) the week between 20 and 27 July 2016. Figure 2.9a shows the raw voltage measurements between 0 and 1000 mV from the sap flow probes collected by the same node; however, measurement noise produced raw voltage readings as high as 1500 mV (not shown).

To perform an accurate estimation of the voltages for each probe, a robust weighted local regression [92] is used. The robust weighted local regression smooths the raw data and it is not affected by a relatively small number of outliers. The smoothed results are shown as red and blue lines in Figure 2.9a for the heater probe (HP) and the temperature probe (TP), respectively. Figure 2.9b shows the temperature conversions from the smoothed raw measurements based on the individual calibrations for the HP and TP in red and blue, respectively. The difference in temperature between the HP and TP (i.e., $\Delta T = HP - TP$) in Celsius degrees is shown in Figure 2.9b, in magenta. Finally, Figure 2.9c shows the resulting sap velocity based on Equation (2.4).

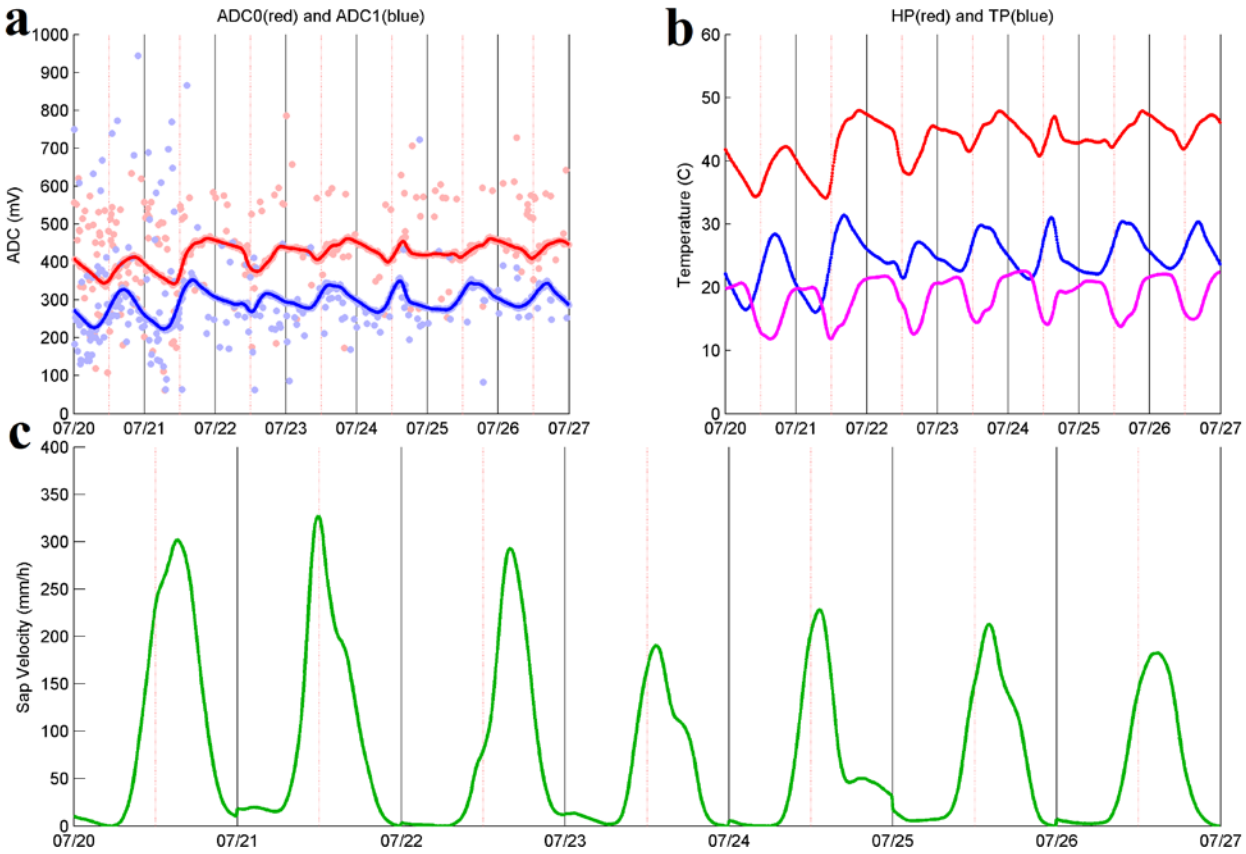

Figure 2.9 - Sap flow results for node 2084 between 20 and 27 July 2016

(a) Raw voltages (i.e., ADC in mV) from the HP (red scatter plot) and TP (blue scatter plot) between 0 and 1000, and smoothed plot for the HP and TP in red and blue, respectively; (b) Filtered and smoothed temperatures for the HP (red) and TP (blue) from the raw voltages ADC0 and ADC1 respectively, difference in temperature (HP-TP) in Celsius degrees (magenta); (c) Sap flow time series (mm/h)

2.3.3.2 Transpiration Estimates from Sap Flow Measurements

Table 2.2 shows a comparison between the A_s estimates of monitored silver maple trees in site 2 (Figure 2.5 for locations) and modeled sapwood area, $\hat{A}_s$, based on Equation (2.6), for which the same values of d were used from the field survey. The coefficients B_0 and B_1 were selected according to the tree species [67]. It is observed that the values of $\hat{A}_s$ are similar to the field estimates (RMSE = 78.6 cm^2); therefore, Equation (2.6) was used to calculate the sapwood area of the trees located inside the 1300 m^2 survey area.

Table 2.3 shows the results for the A_S / A_G ratios for the three sap flow nodes based on their regions of interest within the survey area (Figure 2.6).

Figure 2.10 shows the transpiration (τ) calculations using Equation (2.5) for the sap flow measurements from nodes 2014, 2084 and 2134 based on a mean-weighted A_S / A_G ratio of 12.06 $m^2 \cdot ha^{-1}$ from Table 2.3. The selected time period (i.e., from July to October) represents the time of year when, on average, most of the evapotranspiration occurs around the study site [93]. Figure 2.10a shows τ, in mm/h, every 10 min, which corresponds to the sap flow sensor's sampling interval. The peaks in Figure 2.10a represent a time close to noon on each day. Despite having few noticeable high peaks, τ is mostly within the range 0.2–0.4 mm/h.

Table 2.2 Comparison of silver maple (Acer saccharinum) sapwood area from [72] measurements and Equation (2.6) estimations in site 2 of the ASWP network

Node	d (cm) [a]	A_S (%) [b]	A_{total} (cm²) [c]	$\hat{A}_S$ (cm²) [d]	$\hat{A}_S$ (cm²) [e]
2045	34	77.4	839	650	700
2055	30.7	70	682	478	601
2065	31.5	76.7	719	552	633
2095	41.2	81.7	1250	1020	954
2115	24.3	83.1	415	345	394

[a] based on hand measurements made in 2010; [b] based on the average of three core samples taken in 2010; [c] assumes 0.64 cm bark thickness; [d] based on the estimated percentage of sapwood area times the total trunk cross-sectional area; [e] based on the regression equation of [61] where B_0 = 2.052 and B_1 = 1.654

Table 2.3 AS/AG calculations based on the field survey within the three survey regions in site 2 of the ASWP

Survey Regions	$\sum \hat{A}_S$ (m²) [a]	A_G (ha) [b]	A_S / A_G (m²/ha)
2134	0.375	0.0292	12.87
2084	0.38	0.0391	9.73
2014	0.784	0.0593	13.23

[a] sum of $\hat{A}_s$ values within each survey region, based on hand measurements taken at a height of 1.37 m in 2016 and Equation (2.6), where B_0 = 2.052 and B_1 = 1.654 [61]; [b] based on Thiessen polygon areas (Figure 2.6).

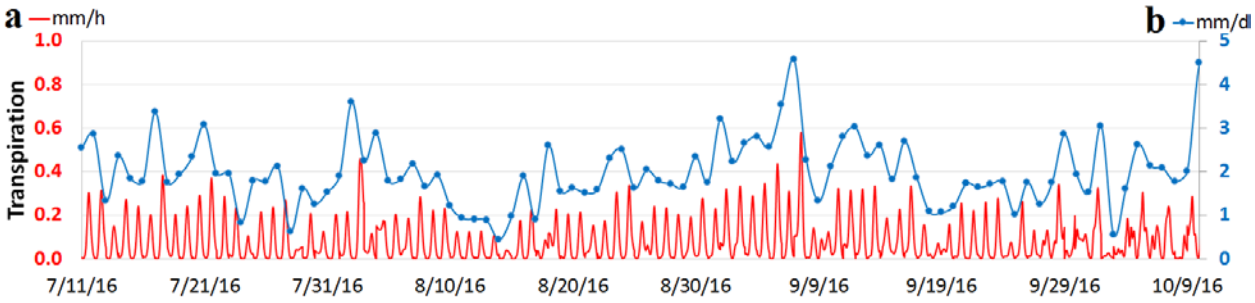

Figure 2.10 - Transpiration (τ) calculations in the ASWP site based on the measurements in nodes 2014, 2034 and 2134, from 11 July 2016 (7/11/16) to 11 October 2016 (10/11/16). (a) Transpiration rates in mm/h based on a 10-min interval; (b) Transpiration rates in mm/day based on a 24-h interval

Integrating the hourly τ rates in Figure 2.10a to monthly totals yields 40.8 mm from 11 to 31 July, 55.2 mm from 1 to 31 August, 65.4 mm from 1 to 30 September, and, 30.7 mm from 1 to 11 October. According to the NRCC, the monthly average potential evapotranspiration (PET) estimates for the greater Pittsburgh area are 110.7 mm for July, 96.3 mm for August, 66.3 mm for September and 39.12 for October. Considering the average values from NRCC as a reference, these results suggest that there could be an overestimation of the τ during September and October, since monthly τ calculated for September is very close to the estimated average PET (i.e., 65.4 mm compared to 66.3 mm) and the monthly τ calculated for 11 days of October (i.e., 30.7 mm) are projected to be higher than the estimated average PET (i.e., 39.12 mm). Besides two noticeable peaks in τ during these two months (i.e., 7 September and 10 October in Figure 2.10a), the remaining values are consistently higher than in the previous months (i.e., July and August). This suggests that for these two months, the τ rates were higher than average and there is not an overestimation of the transpiration. However, since the A_S / A_G ratio is a major factor controlling τ rates, an extension of the survey area can be considered for a better estimation.

Figure 2.10b shows the daily τ values, which is consistent with previous studies that have similar geographic and climatic characteristics to the ASWP site [66], [67], [94], [95]. The maximum, minimum and average values of daily τ during the specified period of time are 4.49 mm, 0.44 mm and 2.01 mm, respectively.

2.3.4 Exploration of Soil Moisture and Soil Water Potential: Spatiotemporal Trends

Determining and explaining the temporal and spatial hydrological patterns is one of the major challenges in the hydrological sciences, since the factors that control these patterns behave in a nonlinear way [35]. In this study, a spatial analysis was performed in order to show the variability of SM and WP.

Figure 2.11 shows the time series of mean SM and its standard deviation at two depths (10 and 30 cm) for sites 2 and 6. The mean and standard deviation were calculated using hourly time series for each node in sites 2 and 6. Site 6 is characterized by a steep hill slope, while the slope is moderate for site 2. Figure 2.11 a and b show that the mean SM at both depths in site 2 is generally higher than that at site 6. The SM is especially higher at 30 cm (Figure 2.11b). Figure 2.11c and d show the standard deviation at sites 2 and 6, at 10 and 30 cm, respectively. It is shown that site 2 has a higher standard deviation than site 6, especially at 10 cm.

Figure 2.11 illustrates that, due to the presence of significant heterogeneity within a small spatial scale (e.g., the two sites are only a few meters away from each other), individual measurements (e.g., SM in this case) from the nearby locations can be quite different. To capture the variability of SM within a small spatial scale would require many sensors within an area of study. The traditional approach of connecting one or a limited number of sensors to a single data logger is not practical as it would require a large number of data loggers that would make the installation and maintenance cost prohibitive. In contrast, the WSN approach, together with the network protocol used here, makes such applications feasible as the cost is relatively low and data processing is centralized and simplified.

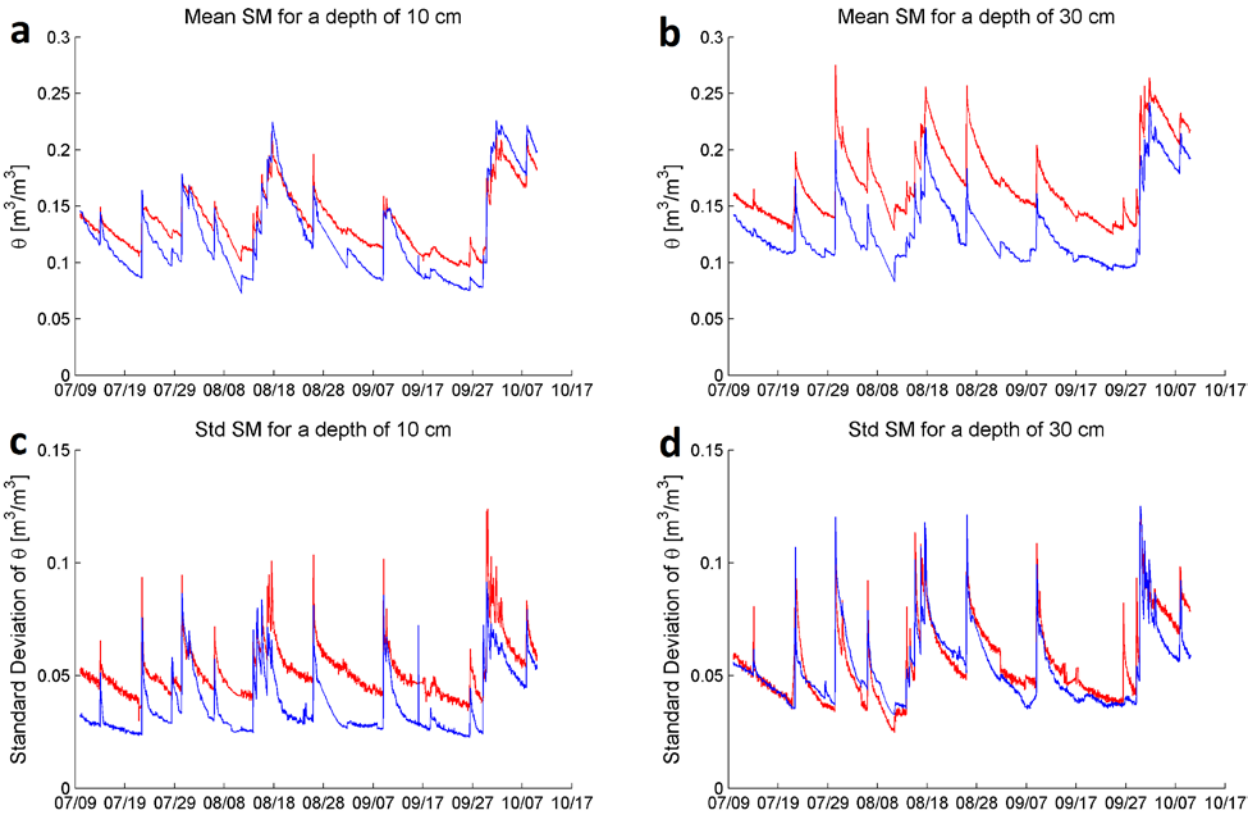

Figure 2.11 - Comparison of the mean and standard deviation of volumetric soil moisture (SM) θ in $m^3 \cdot m^{-3}$ at sites 2 and 6 in red and blue, respectively, in the ASWP WSN testbed between 10 July and 10 October 2016. (a) Mean SM at a depth of 10 cm; (b) Mean SM at a depth of 30 cm; (c) Standard deviation of the SM at a depth of 10 cm; (d) Standard deviation of the SM at a depth of 30 cm

SM and WP surfaces (1-m cell size) were generated to illustrate the average spatial and temporal variability of these two parameters. The surfaces were built using the OK method. Along with the interpolated surface, elevation contours were generated from a 2-m resolution LIDAR raster, in order to complement the surface analysis by providing elevation input. The interpolation boundary was defined based on the area extent (approximately 15,000 m^2) where the nodes are located. The highest and lowest elevations within the site are 365 and 346 m above mean sea level (m.a.m.s.l.), respectively.

Figure 2.12 shows the average-seasonal SM (at 10 and 30 cm) and WP (at 30 cm) surfaces from 2010 to 2016. Overall, it is noticed that the higher SM area is located in the lower part (at 346 m.a.m.s.l.), within a flatter region near a pond. Also, it is observed that SM, regardless of the elevation, is more homogeneous during winter than in the other seasons, which might be caused by snow accumulation and melting over the winter. This is more evident in the average winter SM at 10 cm (Figure 2.12a1), since the shallow soil is more influenced by the varying climatic conditions than the deeper soil [96].

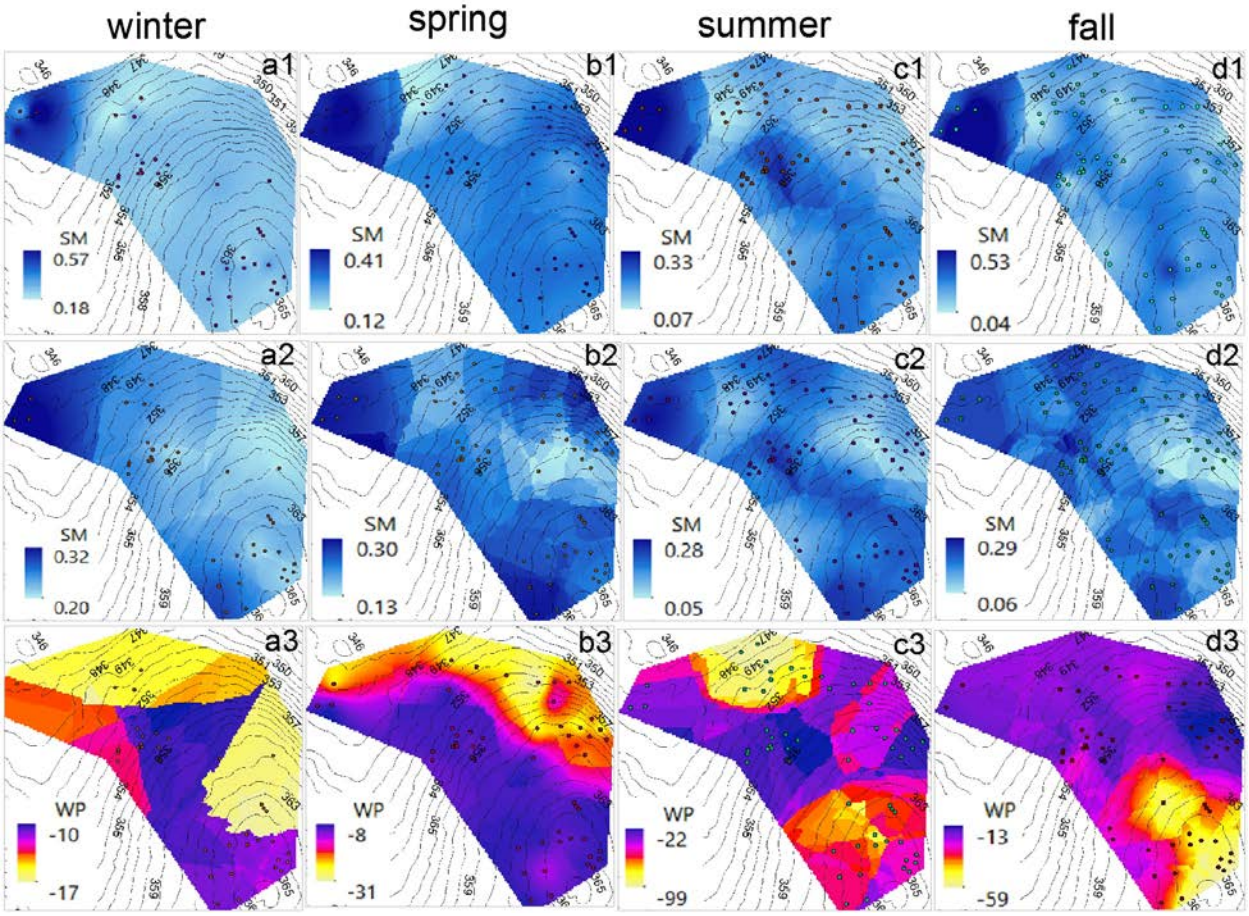

Figure 2.12 - Interpolated surfaces (Kriging method) showing the average seasonal variation in volumetric soil moisture (SM) and soil water potential (WP)

based on data retrieved from 2010 to 2016. (a) winter average (December–February): (a1) SM at 10 cm; (a2) SM at 30 cm; (a3) WP at 30 cm; (b) spring average (March–May): (b1) SM at 10 cm; (b2) SM at 30 cm; (b3) WP at 30 cm; (c) summer average (June–August): (c1) SM at 10 cm; (c2) SM at 30 cm; (c3) WP at 30 cm; (d) fall average (September–November): (d1) SM at 10 cm; (d2) SM at 30 cm; (d3) WP at 30 cm. SM is expressed in $m^3 \cdot m^{-3}$. WP is expressed in kPa. The elevation contours are expressed in m. The dots represent the nodes from which the data was retrieved

Another noticeable fact is that the average variation in SM is higher at 10 cm than at 30 cm, suggesting that the longer travel time to deeper soil reduces the spatial variability of SM. SM is lower during the summer than in the other seasons, which is consistent with the recorded daily rainfall data at the Pittsburgh Airport Meteorological Station [97] for the 2010–2016 period, where there is a dry period between the end of the spring and summer months (i.e., May–September). Finally, based on the SM surfaces from summer and fall (Figure 2.12 c1, c2, d1, d2), there seems to exist a water pathway (darker color in the surface) from the highest elevation to the lower part of the area, located at the right side of the SM surface. This pathway might be explained by the natural surface and subsurface water movement towards the creek located to the south of the study area which, in turn, drains into the pond (immediately downstream of the region with higher SM). Topography showed stronger influence on SM during the winter. Regarding WP, the interpolated surfaces show a variable behavior from one season to another, but WP is mostly higher in the regions with higher elevations, even though there are some lower regions with higher WP.

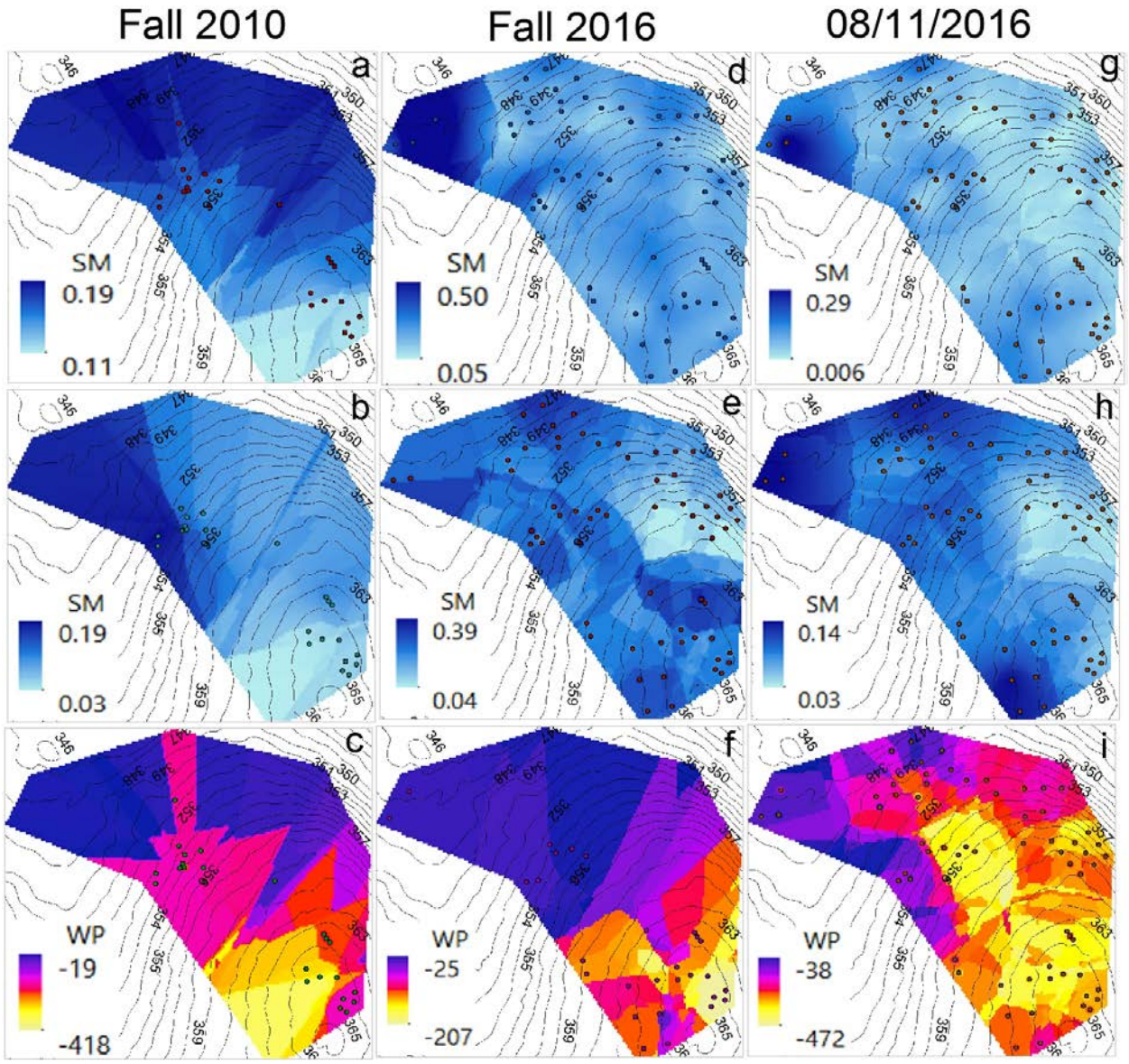

Figure 2.13 - Interpolated surfaces (Kriging method) showing a comparison between the average fall (September–November) SM and WP of 2010 and 2016, and average SM and WP on 11 August 2016 (08/11/2016) (a) SM at 10 cm (fall 2010); (b) SM at 30 cm (fall 2010); (c) WP at 30 cm (fall 2010); (d) SM at 10 cm (fall 2016); (e) SM at 30 cm (fall 2016); (f) WP at 30 cm (fall 2016); (g) SM at 10 cm (11 August 2016); (h) SM at 30 cm (11 August 2016); (i) WP at 30 cm (11 August 2016). SM is expressed in $m^3 \cdot m^{-3}$. WP is expressed in kPa. The elevation contours are expressed in m. The dots represent the nodes from which the data was retrieved

Figure 2.13 illustrates the improvement achieved by the network expansion. In highly complex and heterogeneous environments, the amount and quality of data is proportional to the amount of extractable knowledge [98]. SM and WP interpolated surfaces were created for the average fall conditions in 2010 and 2016. Additional surfaces were generated for the average of 11 August 2016, which is the day with the highest recorded alive nodes (102 nodes, including the relays). In general, these surfaces show similar patterns for each corresponding variable (i.e., SM at 10 cm and 30 cm and WP at 30 cm). However, the 2016 network size, with less scattering in the node locations, provides better estimations than the 2010 network size. It is observed that the patterns for fall 2016 and 11 August 2016 are more similar to each other than the patterns for fall 2010 (see Figure 2.13). This indicates that the higher node density in 2016 provides more detailed insights of the temporal and spatial variability of SM and WP. Overall, the analysis has shown the applicability of WSNs for short- and long-term hydrological patterns characterization at the catchment scale, for steep-forested environments.

In addition, Table 2.4 shows the RMSE obtained from the OK interpolation for the scenarios presented in Figure 2.12 and Figure 2.13. The RMSE has the same units as the analyzed variable (i.e., SM or WP). The number of nodes from which the data was extracted, along with the maximum and minimum SM and WP values are also included. In terms of the SM, it is observed that RMSE is lower at 30 cm than at 10 cm, which is consistent with what has been shown before (i.e., that the SM is much more variable in the near-surface soil than in the deeper soil). In the case of the different average seasonal conditions, even with a different density of nodes within the site, the RMSE did not experience a significant change, thus showing the robustness of the OK interpolation method.

Table 2.4 RMSE of the interpolated surfaces

Interpolated Surface		Winter 2010–2016	Spring 2010–2016	Summer 2010–2016	Fall 2010–2016	Fall 2010	Fall 2016	8/11/2016
SM ($m^3 \cdot m^{-3}$) at 10 cm	**# Nodes**	38	55	71	72	24	72	74
	Max SM	0.57	0.41	0.33	0.53	0.19	0.5	0.29
	Min SM	0.18	0.12	0.07	0.04	0.11	0.05	0.006
	RMSE	0.061	0.062	0.061	0.063	0.059	0.026	0.031
SM ($m^3 \cdot m^{-3}$) at 30 cm	**# Nodes**	38	57	72	72	23	72	74
	Max SM	0.32	0.3	0.28	0.29	0.19	0.39	0.14
	Min SM	0.2	0.13	0.05	0.06	0.03	0.04	0.03
	RMSE	0.03	0.048	0.052	0.051	0.058	0.057	0.024
WP (kPa) at 30 cm	**# Nodes**	37	51	59	57	24	69	74
	Max WP	−10	−8	−22	−13	−19	−25	−38
	Min WP	−17	−31	−99	−59	−418	−207	−472
	RMSE	2.12	4.94	17.46	6.01	41.91	17.88	26.53

The lowest RMSE in SM, for both depths, obtained for 11 August 2016, suggests that a shorter period of time and a higher density of nodes reduce the uncertainty of the SM estimation. The estimated RMSE in the WP surfaces showed more variability than in the case of SM, mostly due to larger differences in the WP ranges for the analyzed conditions. However, if considering the error as a percentage of the range, the 11 August 2016 and fall 2016 average scenarios have the lowest percentages, 0.061% and 0.098%, respectively. In summary, geostatistical tools such as the OK interpolation constitute an important complement to WSNs for environmental monitoring purposes, especially when it is intent to estimate the spatiotemporal behavior of hydrological parameters.

2.3.5 WSN Challenges and Utility in Hydrology

There are several challenges faced in outdoor environmental monitoring WSN deployments, including power management, node maintenance, network scaling, heterogeneous deployment, and overall network cost [99].

2.3.5.1 Power Management

It is of critical importance to maintain a constant power supply to the WSN nodes to ensure data collection and communication within the network. By far, the most common maintenance task is the replacement of batteries. Rechargeable nickel-metal hydride (NiMH) AA batteries were selected for powering the wireless motes as an environmentally friendly and cost-conscience means of maintaining the frequent battery changes of the network. Other alternatives such as the use of lithium-ion polymer battery (LiPo) were discarded due to budget constraints and the existing investment on a large number of AA (NiMH) batteries and chargers. Previous studies that analyzed the power efficiency of WSN motes using AA (NiMH) batteries showed that the expected autonomy of individual nodes is between 48 days [100] and 58 days [38]. More information about the energy profile for WSN nodes is available in [101]. The use of solar panels has been considered, but, with the dense forestation surrounding the majority of the network, it did not appear to have a sufficient return on investment; although, it might be suitable for other locations with more exposure to direct sunlight or during the winter months following tree leaf senescence.

Despite the benefits of rechargeable batteries, some drawbacks exist. Following a recharge, the NiMH batteries may have a significantly higher voltage (e.g., 1.4 V). This leads to circumstances where the combined voltage of three NiMH batteries (i.e., 4.2 V) is significantly greater than the recommended safe operating voltage for the wireless motes (i.e., 3.3 V). Also,

issues with irregular charging voltages were found in the NiMH batteries, which were sorted based on the recommended screening process described in [38]. In order to maximize the life span of each relay node for each battery cycle, it is recommended using D batteries that have a capacity of about 10,000 mAh or more.

Over the span of the project, two sorting strategies were used for the batteries: full and partial sorting. In the full sorting strategy, before a maintenance event, all recharged batteries are sorted by their standing voltage from low to high. Replacement batteries are then chosen as consecutive groups of three from the sorted group. In the partial sorting strategy, batteries are grouped based their standing voltage into bins (e.g., 1.25–1.30 V). Replacement batteries for a single mote are then taken from the same bin. In this method, only voltage bins with an adequate number of batteries are used, which often leads to unused batteries in bins with only one or two batteries. In this regard, the full sorting strategy is slightly better; however, it is more time consuming. As indicated in Figure 16 of [38], node battery life throughout the network improved following the adoption of a battery sorting strategy.

Another method for improving the battery life of wireless motes is to reduce its number of transmissions. This is due to the high-energy costs of transmitting wireless data [100], [102]. During the first years of this project, each node had a sampling rate of 15 min. This was a trade-off between the desired sub-hourly temporal resolution of the environmental data, the expected battery life for each power cycle of the motes' batteries, and the poor packet reception rate of the network, which was around 50% during the first years of deployment. With recent versions of the WSN protocol, the packet reception rate has significantly improved to over 90% [38]. In addition, the new WSN protocol allows for the customization of network parameters for individual nodes according to their intended use. In order to reduce power consumption, the sampling rate of relay

nodes (and some sensor nodes) was lengthened to 30 min. At the same time, to address measurement noise, the sampling rate for the sap flow nodes was shortened to 10 min. The increased sampling rate of the sap flow nodes was not a concern for battery life, as these nodes are powered by the 12 V lead-acid battery.

2.3.5.2 Node Maintenance

The maintenance of the data collection equipment depends on knowing the status of each individual node or data logger. However, the data loggers used in this study are not available on-line and therefore it is not possible to monitor the data collected with them in real time. In addition, in order to collect the data, the researcher needs to commute to the location where the data logger is located. There are some disadvantages with this approach. First, if a wire is loose, then data from one or several sensors attached to the data logger is lost. Second, if the batteries are depleted, then the data logger stops working. Third, there is no way to be aware of those issues until the data is downloaded and examined. Lastly, downloading the data from a data logger is time consuming and does not scale to the case of several locations because every location has to be downloaded independently. Also, the data for a single data logger generates a number of separate files (i.e., one at each location and time of downloading) that require further processing before analysis—as is the case for the Decagon Devices EM50 data logger used in this study. On the other hand, the data collected from our WSN is stored directly and automatically in a relational database that is available through a web-based integrated network and data management system for heterogeneous WSN site called INDAMS [103] for online monitoring.

One way to reduce the need for node maintenance is by using enclosures of high quality, even though they tend to be more expensive, their associated costs pays off in the long run as they are more resistant to environmental damage, less prone to water intrusion, easier to open and close, and, therefore, easier to maintain. In addition, high quality enclosures keep the sensing and communication equipment, and the batteries safer.

2.3.5.3 Network Routing and Scaling

In multi-hop large-scale WSN networking, the routing protocol plays an essential role for reliably collecting sensor data in real time. While WSN deployments appear promising due to the limitations of traditional data logging methods [104], the WSN scalability has proven to be a bottleneck in early studies. An increased network size introduces more data traffic, collisions and congestion in the network, resulting in network performance degradation. To mitigate this problem, starting from the summer of 2014, the ASWP network has adopted CTP + EER routing, which significantly reduces the workload of motes along efficient routes and thus extends the WSN lifetime.

During the network expansion, from 52 nodes in 2014, to 88 nodes in 2015, and finally to 104 nodes in 2016, the network performance has not been noticeably influenced while operating CTP + EER. The packet reception rate (PRR) of the network remains above 96%. The average packet path length is 3.95 hops during the 52-node network, 4.76 hops during the 88-node network, and 4.73 hops during the 104-node network. In the summer of 2015, both the width and length of the WSN deployment was expanded, which caused the increase of the average path length. In the summer of 2016, the major network change was its density, which caused a slight reduction in the average path length and PRR. This result demonstrates that with the proper routing protocol, the network is able to maintain high levels of performance over various deployment scales.

2.3.5.4 Heterogeneous Mote Reprogramming

The ASWP WSN deployment consists of three different mote platforms as well as multiple application versions (corresponding to various external sensors attached to individual motes). As an exploratory and evolving WSN deployment, the network application needs to be updated frequently to test new protocols and parameter configurations. Over-the-air reprogramming approaches become a natural choice since manually reprogramming the motes is cumbersome. The heterogeneous nature of the developed WSN with motes operating in LPL makes the existing reprogramming tools infeasible [58], [59].

Our developed MobileDeluge [58] is a novel hand-held mobile over-the-air mote reprogramming tool for outdoor WSN deployments (Figure 2.14). MobileDeluge builds a new control layer on top of Deluge [105]. It enables and disables Deluge services on demand, allowing for the selection of a subset of motes as targets when initiating a reprograming task. It then disables LPL in the targets for fast dissemination of the new application image, which usually consists of thousands of packets. The targets are also configured in a different radio channel to avoid interference with the rest of the network. MobileDeluge currently works with the motes within a one-hop range to avoid forwarding a bulk code image over intermediate nodes for mote energy conservation. Please see [58] (and the references herein) for more details.

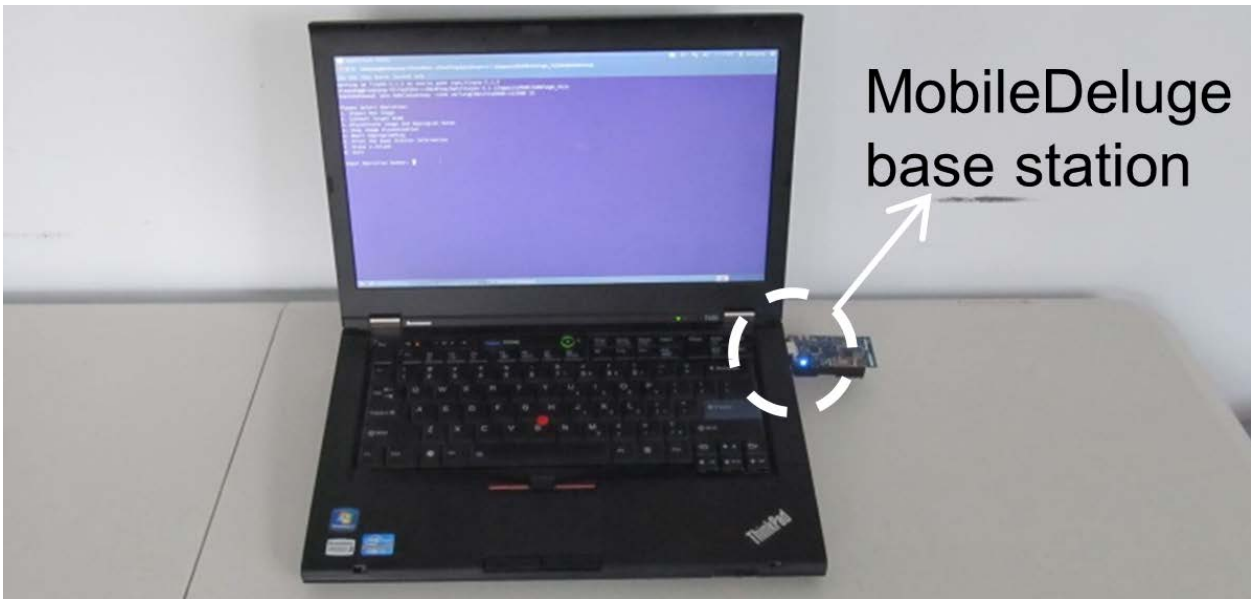

Figure 2.14 - MobileDeluge, a hand-held mobile mote reprogramming tool

MobileDeluge has significantly reduced the time and labor required to update the application in the outdoor WSN testbed. The manual reprogramming procedure would consist of getting the enclosure from the tree, opening the box, attaching the mote to the laptop and uploading the new application. For example, it usually takes a few days to reprogram the whole ASWP testbed (i.e., 104 motes). With MobileDeluge, in contrast, the reprogramming can be finished within one afternoon.

2.3.5.5 Network Costs

The 52-node MICAz and IRIS network at the end of 2014 had a cost of $31,500 for the wireless motes, gateway, sensors, and other peripherals [38]. The CM5000-SMA (TelosB) mote includes built-in humidity and temperature sensors and does not require the use of an acquisition board for relay nodes as opposed to the MICAz or IRIS motes that use the MDA300 acquisition

board ($179). Therefore, the adoption of the TelosB motes has significantly reduced the cost of relay nodes, from $330 (in the MICAz network) to $164, despite the TelosB ($110) being slightly more expensive than the MICAz or IRIS motes (both models about $99 each). These savings are also found for the soil sensor nodes (from $664 to $480), mainly due to the deployment of our inexpensive ($13) designed sensor boards (with 5 V voltage booster) instead of the MDA300, despite the increased cost for the MPS-2 (compared to the MPS-1) sensor. A new sap flow box design, which replaced the MDA300 by our sensor board ($9) without the 5 V voltage booster and does not require the use of AA or D batteries, has further contributed to the cost savings (from $464 to $257). The cost of the expanded 104-node (27 MICAz, 32 IRIS and 45 TelosB motes) network is approximately $50,000. Table 2.5 shows the distribution of sensors for each type of node.

Table 2.5 Distribution of nodes by type of application

App Type	Number
Relays	27
Soil Moisture Water Potential (EC-5 × 2, MPS-1 × 1)	31
Soil Moisture Water Potential (EC-5 × 2, MPS-2 × 1)	36
Sap Flow	10

For the sake of comparison, a Decagon Devices EM50 data logger costs about $476 and is roughly equivalent to our WSN soil sensor node ($177 without the sensors) in terms of its capability to host external sensors.

2.3.6 Lessons Learned with Sap Flow

Low-cost wireless sap flow monitoring is a challenge for environmental research. The delicate nature of the thermal dissipation sap flow sensor, not often surviving more than a single season, and the price of the commercial sap flow sensor, which is too high for large deployments with tight budgets, lead researchers to building their own sensors. While cost effective [45], there are challenges to manufacturing working sensors, which require a good deal of patience and careful attention to detail. Once manufactured, sensors should undergo calibration to account for slight variations in workmanship and care must be taken during transport and installation, during which time the heating filament can be easily damaged. The cost effectiveness of these self-made sensors outreaches the drawbacks of their tedious manufacturing and delicate installation.

There is also the issue regarding the integration of sap flow sensors into WSNs. Early WSN sap flow studies were based on experimental hardware burdened with power limitations and software development issues [34], [106]. These days, good wireless implementations are becoming more and more ubiquitous and more seamless in terms of user experience.

2.4 Conclusions

The environmental data collected with the WSN nodes were found to be similar to the data collected from the Decagon Devices Em50 data logger in terms of quality. However, the WSN nodes overcome some important limitations of traditional data loggers at a significantly lower cost. For instance, the data readings from the WSN nodes are automatically collected and stored in a relational database system, therefore all the environmental data are saved in a unified and

integrated repository, eliminating the need to manually download data at each location. In addition, the status of individual nodes is available in a web-based integrated network and data management system developed for heterogeneous WSN site called INDAMS.

This study has shown an effective application of WSNs to determine and explain spatiotemporal hydrological patterns. A specially designed sensor board provides stable excitation voltage for analog and digital sensors at only approximately 6% of the cost of the MDA300 acquisition board. MPS-2 sampling synchronization issues on sensor motes were solved with our driver software developed in TinyOS. Our exploratory study demonstrates how the innovative WSN routing protocol CTP + EER and the over-the-air reprogramming tool MobileDeluge can overcome the challenges of heterogeneous and large-scale multi-hop WSN for outdoor environmental morning. In particular, this study has presented the first of its kind comprehensive data analyses for the WSN monitored hydrological variables including soil temperature, WP, SM and sap flow. Two PTF parameters that are utilized by hydrologic models to predict soil water retention properties (i.e., the Clapp-Hornberger equation and the van Genuchten equation) were estimated with the retrieved SM and WP data with a high goodness-of-fit (i.e., NSE greater than 0.80). The improved installation design of the sap flow sensors allowed for the retrieval of high-quality data, which later were filtered using a robust weighted local regression to smooth the data without being affected by the outliers. At the same time, these sap flow data were used to estimate transpiration rates, which were highly consistent with previous studies in sites with similar geographic and climatic characteristics. The estimation was also consistent with the local measured data (meteorological stations). Moreover, a spatial analysis was performed to show the variability of SM and WP, which showed the applicability of WSNs for short- and long-term hydrological patterns characterization in a catchment scale in steep-forested environments.

It has also been shown that “out of site” procedures, such as sensor calibration methodologies and adequate data processing, provided a fundamental added value to the field work. Finally, despite the tremendous challenges posed by outdoor WSN deployments, including power management, node maintenance, routing scale, heterogeneous deployment, and overall network cost, the wireless sensor network approach (e.g., protocols, sensors, deployment tool, and acquisition) presented in this study has proved to be an effective (in terms of the data quantity and quality) and low-cost alternative for environmental monitoring. This helps pave the way to larger scale outdoor WSN developments in the future in order to ultimately study and answer the fundamental science questions for quantifying sub-grid heterogeneity and in understanding hydrologic parameters.

Future work should also consider continuing exploring materials and methods to lower the cost of the network without reducing the data quality and other complementary strategies such as the optimization of battery usage.

3.0 Estimation of Daily Streamflow from Multiple Donor Catchments with the Graphical Lasso

Most of the materials contained in this chapter are based on the manuscript submitted to the journal of *Water Resources Research*, under the title of "Estimation of daily streamflow from multiple donor catchments with the Graphical Lasso" by German A. Villalba, Xu Liang, and Yao Liang. The manuscript is currently under revision.

3.1 Introduction

Continuous daily streamflow time series are important for a wide variety of applications in hydrology and water resources. Such applications include water supply management, hydropower development, flood and drought control, forecasting of agricultural yield, ecological flow assessment, navigation, rainfall runoff model calibration, design of engineering structures such as highways and reservoirs, and many others [107]–[111]. However, continuous streamflow data are not available oftentimes due to either no existing streamflow gauges or data gaps in the recorded time series at gauged stations [112]. Also, data gaps of different time periods exist at different gauge locations within a large river basin [113]. Furthermore, there is an increasing decline in the hydrometric network density worldwide [114], [115]. For example, the U.S. Geological Survey (USGS) is discontinuing operations of some streamflow stations nationwide due to budget cuts [116] which has been a serious concern [117]–[120]. Therefore, it is critical to develop an effective

and general method to fill in data gaps, extend data records of those that have been or will be shut down, and even estimate data for ungauged locations.

The estimation of continuous daily streamflow time series techniques at ungauged or poorly gauged locations can be classified into two broad categories: (1) hydrologic model–dependent methods and (2) hydrologic model–independent methods [111]. The latter methods are also called statistical methods [121] or hydrostatistical methods [110]. Work related to the first category is abundant, but it is relatively limited for the second category (e.g. [110], [111], [122]), especially during the 1990s and 2000s. With an increase of various data types and computing power over the last couple of decades it is now possible to re-visit the challenging issues using data-driven approaches such as Machine Learning [123], [124], which also belongs to the second category of hydrostatistical methods. Farmer and Vogel [110] summarized the general procedure of the second category as a three-step process. Step 1, selection of one or multiple donor gauges based on some measure of hydrologic similarity. Step 2, estimation of the streamflow statistics, such as the mean and standard deviation, at the target location. Step 3, transference of the streamflow time series from the donor gauge(s) to the target site (e.g., partially gauged/incomplete or ungauged).

The accuracy of inferred daily streamflow estimations based on Step 3 is conditioned on the accuracy of a proper selection of the donor gauge(s) in Step 1. This selection is typically based on an assessment of the hydrologic similarity between the target and the donor gauge(s) and whether a single or multiple donor gauges are used. A number of approaches have been used so far with different levels of complexity, data requirements and accuracies (e.g. [107], [110], [114], [125]–[129]).

The approach of selecting the nearest gauge as the donor is a convenient and widely used method due to its simplicity and minimum data requirements (e.g., [110], [130]–[132]), For example, the work of Farmer and Vogel [110] adopted this simple distance-based method in the donor selection procedure (Step 1) in their study, where a number of methods using different streamflow statistics in Step 2 and Step 3 were investigated and compared. Archfield & Vogel [107] developed a procedure called Map correlation method that uses time series from several streamflow gauges in the study area to create a correlation map based on a kriging method and then uses that map to estimate the correlation between a given ungauged location and nearby gauges. They concluded that (1) the distance-based approach does not provide a consistent selection criterion; (2) the most correlated gauge is not always the closest one by distance; and (3) the accuracy based on the most correlated gauge outperforms the one based on the distance in most cases. The correlation-based approach is generally better than the distance-based approach because the streamflow data is more effectively used in the correlation-based approach, and the marginal independence between any pair of gauges can be easily determined [133]. Here, the pair-wise correlation between two gauges is used to evaluate their marginal independence. That is, the two gauges are assumed to be independent if their pair-wise correlation is below a given threshold. For example, Halverson & Fleming [129] set a correlation threshold of 0.7 in identifying whether two gauges in question are independent or not.

Although using a single donor gauge to estimate streamflow time series has been a dominant approach [134], Smakhtin et al. [125], [126] proposed to use more than one donor gauge from nearby gauges to improve the streamflow estimations for ungauged basins. Zhang and Chiew [127] and Arsenault and Brissette [135] also concluded that the estimation from multiple donor gauges is more accurate in general than that from a single donor gauge case. In these studies,

multiple donor gauges were investigated based on methods such as the degree of similarity of flow regimes between the donor and destination gauges, spatial proximity, physical similarity, simple arithmetic mean, inverse distance weighting, combinations of some of them, and an assignment of a fixed number of donor gauges. The challenges of these approaches include: (1) how to measure the similarity; (2) how to systematically determine which gauges should be the donor gauges; and (3) how many donor gauges each individual target gauge should have.

In addition to Archfield & Vogel [107], other previous work, (e.g. [134], [136], [137]) also showed that geostatistical methods such as kriging that uses multiple donor gauges, are an effective alternative. The kriging method is an spatial interpolation technique that estimates values at target locations as a linear weighted combination of the observations from different locations. The weights are assigned based on a variogram model which is usually fitted based on the variance between observations as a function of the distance between locations. The kriging method overcomes problems in terms of selecting the number of donor gauges and the individual donor gauges since all of them are used in a linear combination fashion. The kriging method is useful in transferring information from gauged to ungauged locations [138]. However, the accuracy of its estimation depends on the density and quality of the measurements of the gauged sites. Virdee & Kottegoda [139], noticed that a major problem with kriging is the lack of data with the needed density.

For a commonly encountered situation in which the density of streamflow network is sparse, kriging is not a good candidate. From the aforementioned various methods other than kriging, it appears that the correlation-based single donor method (i.e., pair-wise marginal independence approach) is less subjective and provides more consistent results while the multi-donor methods lead to better results with subjective selection process on the donor gauges.

Therefore, it is critical to develop a method that is less subjective in selecting a set of multi-donor gauges for each target location (i.e., Step 1). In this study, we present a novel method which draws on the strengths of existing methods but overcomes their weaknesses. More especially, we present an approach that can explicitly and effectively consider the correlation structure of the entire gauge network rather than the pair-wise correlation between two gauges at a time. The correlation-based single donor method only considers the pair-wise correlation between two gauges in which the marginal independence assumption is applied. It is basically a local approach which does not take advantage of the dependence structure of the daily streamflow distribution, based on conditional independence conditions, given by the underlying streamflow network. This is because the conditional independencies among gauges in the streamflow network are typically not apparent in the correlation matrix but in its inverse matrix, i.e., the precision matrix [133]. Thus, the existing correlation-based methods on multi-donor selection process are not effective and subjective. In this study, we use the precision matrix to extract dependence structure of the gauge network based on the concept of conditional independence conditions. We then use such identified dependence information to select donor gauges (Step 1). Since the donor gauges are selected based on the dependence structure of the entire gauge network, our method can be considered as a global approach as opposed to the existing local approach where only pair-wise correlations between gauges in the network are considered. Our method is generic and flexible, and it would be more effective since it can extract implicit information (i.e., conditional independence structure of the underlying streamflow network) using a sparse precision matrix instead of the correlation matrix that is commonly used. With this new method, we can infer daily streamflow for active gauges with data gaps and extend data for inactive gauges which are defined as those that are no longer collecting data but collected the data in the past (i.e., data extension). In addition, with this new

method of filling in data gaps and extending data records, we can estimate daily streamflow at ungauged sites or improve the estimation of daily streamflow at ungauged sites based on the kriging method. Furthermore, a new algorithm based on the conditional independence concept is presented to remove gauges from an existing streamflow network with the least loss of information.

The remainder of this chapter is organized as follows. Section 3.2 describes widely used approaches to transfer streamflow from a single donor to a target basin. Section 3.3 presents our new approach to systematically identify multiple donors based on the precision matrix and a new framework to infer daily streamflow time series based on the selected set of donor gauges. In addition, a new and general method to remove streamflow gauges from the existing hydrometric network with the least loss of information is presented in this section. Section 3.4 provides a brief description related to a study region over the Ohio River basin to evaluate the new approach presented in Section 3.3. Section 3.5 presents the results and discussions. Finally, Section 3.6 provides a summary of the main findings from this work.

3.2 Common Approaches for Transferring Streamflow at Ungauged Basins

This section briefly describes some common approaches used to transfer streamflow from a single donor to a target gauge (Step 2 and Step 3), assuming the single gauge is already identified by a method from Step 1. Let Q_j and Q_i represent the streamflow from a target and a donor gauge, respectively, and assume that the estimated streamflow time series at the target location ($\widehat{Q}_j$) is obtained by transferring the streamflow time series from a single donor gauged catchment by a scaling function such that $\widehat{Q}_j$ is an approximation of Q_j (e.g. [107], [110]).

3.2.1 Drainage Area Ratio

The drainage area ratio (DAR) is a simple scaling procedure that only requires the areas from the target and donor catchments along with the streamflow time series from the donor gauge:

$$\widehat{Q}_j = \frac{A_j}{A_i} Q_i \quad \textbf{(3.1)}$$

The DAR method represented by equation (3.1) assumes that the discharge per unit area is the same between the target Q_j and donor Q_i catchments at the same time step. A_j and A_i are the areas of the target and donor catchments, respectively. Unless the climate and hydrologic regimes at the target and donor sites are similar and the area is the only dominant factor affecting the streamflow, this assumption is unlikely to hold, because a number of factors can significantly change the scaling relationship in equation (3.1), such as, orographic effects where the site at a different elevation is likely to receive a different amount of rainfall and thus a different amount of runoff per unit area, a site on the windward side of the mountain versus the site in the rain shadow side where the rainfall characteristics are dramatically different, differences in slope, soil type, land cover and land use which can affect the conditions of runoff generation, leading to differences in the basin's response to rainfall, and differences in temperature that affect the evapotranspiration losses and runoff per unit area.

3.2.2 Scaling by the Mean (SM)

Scaling by the mean (SM), also called *Standardization by the mean streamflow* [110], is a method that requires the mean streamflow from the target and donor gauges in addition to the streamflow time series from the donor gauge:

$$\widehat{\mathrm{Q}}_{\mathrm{j}} = \frac{\mu_j}{\mu_i}\mathrm{Q}_{\mathrm{i}} \tag{3.2}$$

The SM method represented by equation (3.2) assumes that the discharge scaled by the mean streamflow is the same between the target $\mathrm{Q_j}$ and donor $\mathrm{Q_i}$ catchments at the same time step. μ_j and μ_i are the mean streamflow of the target and donor gauges, respectively.

3.2.3 Scaling by the Mean and Standard Deviation (SMS)

Scaling by the mean and standard deviation (SMS), also called *standardization with mean and standard deviation* [110], is a method that requires information of the mean and standard deviation from the streamflow for the target and donor gauges in addition to the streamflow time series from the donor catchment. This method was originally presented by Hirsch [140] and termed maintenance of variance extension (MOVE) and more recently reported by [107], [110] as:

$$\widehat{\mathrm{Q}}_{\mathrm{j}} = \frac{\sigma_j}{\sigma_i}(\mathrm{Q_i} - \mu_i) + \mu_j \tag{3.3}$$

The SMS method represented by equation (3.3) assumes that the discharge scaled by the mean and standard deviation from the streamflow is the same between the target $\mathrm{Q_j}$ and donor $\mathrm{Q_i}$ catchments at the same time step. μ_j, μ_i, σ_j and σ_i are the mean streamflow of the target and donor gauges and the standard deviation of the target and donor gauges, respectively.

3.2.4 Linear Regression

The linear regression method between streamflow time series of target and donor gauges is rarely used as a transfer method of streamflow due to the lack of streamflow data for the target catchment. However, this information is available for the case of inactive gauges. The regression

method (REG) is a simple least squares linear regression (e.g., [141]) where the regression coefficients for the slope γ_{ij} and the intercept γ_{0j} are estimated according to equation (3.4):

$$\widehat{\mathrm{Q}}_{\mathrm{j}} = \gamma_{0j} + \gamma_{ij} \cdot \mathrm{Q}_{\mathrm{i}} \quad \textbf{(3.4)}$$

3.3 New Approach of Selecting Multiple Donor Gauges Via Graphical Models

This section presents a novel approach that describes how our new algorithms: (a) select a set of multiple donor gauges for each target location over a study area (Step 1) by using a precision matrix obtained with a sparse graphical model, (b) estimate a matrix of regression coefficients that allow simultaneous inference of streamflow from the selected set of donor gauges to target gauges (Step 2) and, (c) perform the inference of the daily streamflow time series (Step 3). Finally, subsection 3.3.7 shows a novel method to remove gauges from the hydrometric network with the least loss of information. The streamflow inference is based on minimizing two main objectives: (1) the streamflow estimation error and (2) the model complexity. Here, model complexity refers to the number of donor gauges required to infer the streamflow at a given target location. Thus, the simplest model would be to select a single donor gauge while the most complex model would be to select all of the available gauges. We argue that there is a trade-off between the model complexity and the accuracy of the estimation, and that a complex model does not necessarily always result in more accurate estimated streamflow time series than a simpler model due to noises. This study aims at finding a suitable balance between the number of donor gauges and the accuracy to optimize these two objectives.

The remaining of this section is structured as follows. The building blocks for the development of our approach are described in sub-sections 3.3.1 to 3.3.4 . Our first algorithm:

"Selection of Graph Model" (called SGM algorithm hereafter) which selects a sparse model with low validation error (i.e., determination of model complexity), is provided in sub-section 3.3.5 . The model complexity determined in sub-section 3.3.5 is then used to train a multiple linear regression model. The inference of daily streamflow given the graph selected by the SGM algorithm is described in sub-section 3.3.6 . Finally, in sub-section 3.2.7, we present our second algorithm, "Removal of Streamflow Gauges" (called RG algorithm hereafter), for the removal of gauges with the least loss of information. The RG algorithm is developed based on the model complexity determined from sub-section 3.3.5 and the inferred results from sub-section 3.3.6 .

3.3.1 Multiple Linear Regression (MLR)

A simple multiple linear regression (MLR) approach is to extend the single linear regression of equation (3.4). That is, for a set of *p* available gauges with daily streamflow records over the study area, each location, *j*, is assumed as the target ungauged location, can be expressed by equation (3.5) as follows, while all the remaining gauges constitute the set of donor gauges:

$$\widehat{\mathrm{Q}}_{\mathrm{j}} = \eta_{0j} + \sum_{i=1, i\neq j}^{p} \eta_{ij} \cdot \mathrm{Q}_{\mathrm{i}} \tag{3.5}$$

where the estimated streamflow time series $\widehat{\mathrm{Q}}_{\mathrm{j}}$ at a target location, *j*, is computed by a linear combination of (*p* - 1) donor gauges, and η_{ij} and η_{0j} represent the multiple regression coefficients (slopes and intercept). Notice that because the donor and target gauges must be different, the target location, *j*, must be different from the donor location, *i*, and that all the available donor gauges are used.

Since the probability distribution of streamflow is often well approximated by a log-normal distribution (e.g. [142]), equation (3.5) can be modified and expressed by equation (3.6) in which Y_i follows a normal distribution and is related to Q_j by a logarithmic transformation, where β_{ij} and β_{0j} represents the regression coefficients for the slope and the intercept, respectively, in the logarithmic space. To avoid numerical issues with the logarithm of zero-valued streamflow, Farmer [134] assigned a small constant value (e.g. 0.00003 m^3/s), smaller that any non-zero value in the data set, to the zero-valued streamflow when applying a logarithmic transformation to the streamflow time series. Here, a different approach is followed. A value of one is added to the daily streamflow time series before the logarithmic transformation is performed, as expressed by equation (3.7), due to the fact that zero-valued streamflow is mapped to zero in the log-transformed variable and the transformation is reversible without loss of precision. Nevertheless, the results of applying either the Farmer's approach or the approach of equation (3.7) are almost identical as described in section 3.3.5.2.

$$\hat{Y}_j = \beta_{0j} + \sum_{i=1, i\neq j}^{p} \beta_{ij} \cdot Y_i \tag{3.6}$$

$$Y_i = log(Q_i + 1) \tag{3.7}$$

For convenience and simplicity, the standard score (Z-score) is used to define a new variable **Z**, as expressed by equation (3.8) where μ_{y_i} and σ_{y_i} are the mean and standard deviation of Y_i. Therefore, each vector Z_i has a mean of zero and a standard deviation of one. Equation (3.9) represents a Z-score regression, where the intercept is zero and the regression coefficient α_{ij} is the correlation between the *jth* target Z_j and the *ith* donor gauge Z_i (z-score of log-transformed) streamflow time series.

$$Z_i = \frac{Y_i - \mu_{y_i}}{\sigma_{y_i}} \tag{3.8}$$

$$\hat{Z}_j = \sum_{i=1, i \neq j}^{p} Z_i \cdot \alpha_{ij} \tag{3.9}$$

Note that equation (3.9) is valid for each of the *p* selected gauges. That is, the column vector $\hat{Z}_j$ is computed for $1 \leq j \leq p$. Equations (3.10) and (3.11) express equation (3.9) in a matrix form.

$$\hat{\mathbf{Z}} = \mathbf{Z} \cdot \mathbf{A} \tag{3.10}$$

$$\mathbf{A} = \begin{bmatrix} 0 & \alpha_{12} & \cdots & \alpha_{1p} \\ \alpha_{21} & 0 & \cdots & \alpha_{2p} \\ \cdots & \cdots & 0 & \cdots \\ \alpha_{p1} & \cdots & \cdots & 0 \end{bmatrix} \tag{3.11}$$

The linear system defined by equation (3.10) assumes that the *jth* gauge is the target and that the remaining (*p* - 1) gauges are the donor gauges for each of the *p* gauges. Equation (3.11) shows the elements of a *p* by *p* matrix **A** used in equation (3.10). The elements of matrix **A** are the regression coefficients α_{ij} in equation (3.9) and the *jth* column represents the vector of regression coefficients required to estimate the column vector $\hat{\mathbf{z}}_\mathbf{i}$. Note that the diagonal elements of **A** are zero. That is, $for \quad 1 \leq j \leq p: \quad \alpha_{jj} = 0$. Therefore, $\hat{\mathbf{Z}}$ is a matrix with the estimated streamflows computed from the observed streamflow **Z** that follows a standard normal distribution, and the squared matrix of regression coefficients **A**. $\hat{\mathbf{Z}}$ and **Z** are *n* by *p* matrices where *n* is the number of daily streamflow records. Note that if the streamflow data do not follow the log-normal distribution as assumed here, one can easily transform the data into the log-normal distribution, and thus equations (3.6)-(3.11) are applicable.

3.3.2 Concept of Gaussian Models and MLR

$\mathbf{Z}$ is a multivariate normal distribution over p random variables with covariance matrix $\mathbf{\Sigma}$ and zero mean vector such that $\mathbf{Z} = (\mathbf{z_1}, \ldots, \mathbf{z_p})$, where the random variable $\mathbf{z_j}$ represents the Z-score of the logarithm of the streamflow data at the *j*th gauge, therefore it follows a normal distribution with zero mean and unitary standard deviation for $1 \leq j \leq p$. $\mathbf{Z}$ defines an undirected graphical model known as a *Gaussian graphical model* where the underlying graph **G** is defined by a set of vertices **v** and edges **e** such that the graph **G** = (**v**,**e**) represents (conditional) independence assumptions among the random variables. This conditional independence means that if there is not an edge on the graph **G** between the i*th* and j*th* location, then these two gauges are independent from each other given the remaining gauges. The existence of such conditional independence implies that for a given target location some of the donor gauges are redundant or not correlated to the target location and therefore, they can be removed from the set of donor gauges for the target location under consideration. Equation (3.10) is a general relationship between each possible target and donor gauges. However, it does not explicitly show how to compute the matrix of regression coefficients **A**. One simple way would be to use an approach from the previous subsection. That is, computing the elements of the matrix **A**, column by column, by means of MLR as shown in equation (3.9) for each of the p target locations. However, this approach assumes that all of the (*p - 1)* donor gauges are included in the regression for each target location. That is, it impliest of having a graph **G** where each vertex is connected to all of the remaining vertices. In other words, it means a complete graph with $\frac{p^2-p}{2}$ edges. Therefore, this approach does not satisfy our second design objective which is to minimize the model complexity by reducing the number of donor gauges, equivalent to reducing the number of edges of the

underlying graph **G**. Furthermore, there are two problems with dense graphs. First, a more complex graph requires more training data to avoid overfitting. Second, only a single gauge or a small number of gauges could be removed from the hydrometric network because each gauge is estimated based on all the remaining gauges connected in the complex graph. Keeping a good balance between the complexity of the graph and the accuracy of an estimation is the goal of our new method. This is achieved by promoting sparsity while minimizing the estimation error.

3.3.3 Relationship between MLR, the Covariance, and the Precision Matrices

This subsection describes two methods to compute the regression coefficients of the matrix **A** represented by (3.11). The first method is based on the covariance matrix $\boldsymbol{\Sigma}$ and the second one is based on the inverse of the covariance matrix which is called the precision matrix $\boldsymbol{\Theta}$ as expressed by equation (3.12)

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} \tag{3.12}$$

Since the true covariance ($\boldsymbol{\Sigma}$) or precision ($\boldsymbol{\Theta}$) matrices are unknown, **A** can only be estimated from the noisy *p*-dimensional observed data from **Z** which often follow a normal distribution. One method is based on an estimated covariance matrix, represented by **W,** and the other method is based on an estimated the precision matrix, represented by $\widehat{\boldsymbol{\Theta}}$. Following Friedman et al. [143], the columns and rows of **W** can be permuted so that the target *j*th gauge is the last and then partition the matrices into four blocks composed by a square submatrix $\mathbf{W_{11}}$ with (*p - 1*) columns (and rows), a column vector $\mathbf{w_{12}}$ with (*p - 1*) elements, its transposed (row) vector $\boldsymbol{w_{12}^T}$ and a scalar

w_{22}. Similar partition scheme leads to the estimated precision matrix $\widehat{\boldsymbol{\Theta}}$ to define the four blocks $\widehat{\boldsymbol{\Theta}}_{\mathbf{11}}$, $\widehat{\boldsymbol{\theta}}_{\mathbf{12}}$, $\widehat{\boldsymbol{\theta}}_{\mathbf{12}}^{\boldsymbol{T}}$ and $\hat{\theta}_{22}$, respectively.

The relationship between the estimated covariance **W**, the estimated precision $\widehat{\boldsymbol{\Theta}}$ and the p by p identity matrix **I** is represented by $\mathbf{W} \cdot \widehat{\boldsymbol{\Theta}} = \mathbf{I}$. The block-wise expansion of this equation, adapted from Friedman et al. [143], leads to equation (3.13) as follows:

$$\begin{pmatrix} \mathbf{W}_{\mathbf{11}} & \mathbf{w}_{\mathbf{12}} \\ \boldsymbol{w}_{\mathbf{12}}^{\boldsymbol{T}} & w_{22} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\Theta}}_{\mathbf{11}} & \widehat{\boldsymbol{\theta}}_{\mathbf{12}} \\ \widehat{\boldsymbol{\theta}}_{\mathbf{12}}^{\boldsymbol{T}} & \hat{\theta}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^{T} & 1 \end{pmatrix} \tag{3.13}$$

Equation (3.14) shows the column decomposition of the matrix **A**:

$$\mathbf{A} = [\boldsymbol{\alpha}_{\mathbf{1}} \quad \cdots \quad \boldsymbol{\alpha}_{\mathbf{j}} \quad \cdots \quad \boldsymbol{\alpha}_{\mathbf{p}}] \tag{3.14}$$

There are several ways to compute the regression coefficients for each column of the matrix **A**. Equation (3.15) shows a method based on the estimation of the covariance matrix **W** and the partitioned matrices from equation (3.13). It computes $\boldsymbol{\alpha}_{\mathbf{j}}$ (equation (3.14)) for $1 \leq j \leq p$ using $\mathbf{W}_{\mathbf{11}}$ as the predictor matrix and $\mathbf{w}_{\mathbf{12}}$ as the response vector.

$$\boldsymbol{\alpha}_{\mathbf{j}} = \mathbf{W}_{\mathbf{11}}^{-\mathbf{1}} \cdot \mathbf{w}_{\mathbf{12}} \tag{3.15}$$

Alternatively, the regression coefficients can be computed from the estimation of the precision matrix $\widehat{\boldsymbol{\Theta}}$. Equation (3.16) shows the results with the latter approach.

$$\boldsymbol{\alpha}_{\mathbf{j}} = -\frac{1}{\hat{\theta}_{22}} \widehat{\boldsymbol{\theta}}_{12} \tag{3.16}$$

Equation (3.16) is derived by expanding the product from the first row and second column of Equation (3.13) such that $\mathbf{W}_{\mathbf{11}} \cdot \widehat{\boldsymbol{\theta}}_{\mathbf{12}} + \mathbf{w}_{\mathbf{12}} \cdot \hat{\theta}_{22} = \mathbf{0}$. After some algebra manipulation, it can be obtained that $\widehat{\boldsymbol{\theta}}_{\mathbf{12}} = -\left(\mathbf{W}_{\mathbf{11}}^{-\mathbf{1}} \cdot \mathbf{w}_{\mathbf{12}}\right) \cdot \hat{\theta}_{22}$. Substituting equation (3.15) into it leads to $\widehat{\boldsymbol{\theta}}_{\mathbf{12}} = -\boldsymbol{\alpha}_{\mathbf{j}} \cdot \hat{\theta}_{22}$. Finally, equation (3.16) is derived by clearing for $\boldsymbol{\alpha}_{\mathbf{j}}$. Equation (3.16) shows how the precision matrix $\widehat{\boldsymbol{\Theta}}$ and the matrix of regression coefficients, **A**, are related to each other. The

elements of the matrix **A** are computed as $\boldsymbol{\alpha_j}$ for $1 \leq j \leq p$. The vector of regression coefficients $\boldsymbol{\alpha_j}$ for the *j*th column of **A** is proportional to the vector $\widehat{\boldsymbol{\theta}}_{12}$ of $\widehat{\boldsymbol{\Theta}}$. The matrix $\widehat{\boldsymbol{\Theta}}$ is closely related to the representation of the underlying graphical model **G** as the zero elements in $\widehat{\boldsymbol{\Theta}}$ represent the missing edges in the graph **G**.

Thus, graph **G** can be represented by an adjacency matrix defined by equation (3.17) below where g_{ij} and $\hat{\theta}_{ij}$ represent the element of the *ith* row and *jth* column of **G** and $\widehat{\boldsymbol{\Theta}}$, respectively.

$$\mathbf{G} = \begin{cases} g_{ij} = 1 & if \quad |\hat{\theta}_{ij}| > 0 \\ g_{ij} = 0 & otherwise \end{cases} \tag{3.17}$$

Given that **Z** has a zero-mean vector, the calculation of the empirical covariance matrix **S** is simplified to equation (3.18) and the empirical precision matrix **T**, the inverse of matrix **S**, is defined by equation (3.19).

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Z^T} \cdot \mathbf{Z} \tag{3.18}$$

$$\mathbf{T} = \mathbf{S}^{-1} \tag{3.19}$$

Equation (3.20) shows how to calculate each column of the matrix **A** by replacing the estimated covariance matrix **W** in equation (3.15) with the empirical covariance matrix **S**, while equation (3.21) shows how to compute each column of matrix **A** by replacing $\widehat{\boldsymbol{\Theta}}$ with **T** in equation (3.16).

$$\boldsymbol{\alpha_j} = \mathbf{S_{11}}^{-1} \cdot \mathbf{s_{12}} \tag{3.20}$$

$$\boldsymbol{\alpha_j} = -\frac{1}{t_{22}}\mathbf{t_{12}} \tag{3.21}$$

Even if **T** is calculated by inverting **S**, Equation (3.21) is more efficient than equation (3.20) because it does not require the inversion of any additional matrix when **T** is known. In this work we describe a way to avoid computing the empirical precision matrix **T**, but to compute a sparse

precision matrix instead. Using equation (3.21) is perhaps the fastest way to estimate the coefficients of the matrix **A** and, therefore, the inferred streamflow from equation (3.10). This approach represents the case where each gauge is inferred based on the $(p - 1)$ remaining gauges through the MLR method. Thus, the apparent simplicity of the method is achieved at the expense of the model complexity.

If **G** is sparse, however, then the conditional independence assumptions imply that the precision matrix should also be sparse. In practice, both the covariance matrix $\boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Theta}$ are unknown and thus, they are approximated by the empirical covariance matrix **S** and empirical precision matrix **T** based on a finite number of noisy observations. But the empirical precision matrix **T** obtained is generally not sparse [144] due to the nature of the noisy data used to estimate **T**. Hence, the underlying graph **G** from the *Gaussian graphical model* is not sparse but a complete graph where each gauge depends (conditionally) on all of the remaining gauges in the hydrometric network. The MLR approach is thus often times associated with a complex model as MLR tries to use all of the predictor variables from a complete graph **G**. Since the objective is to select some, not all of the gauges in the network as the donor gauges to infer the streamflow values at the target location, it is appropriate to simply select the most relevant donor gauges to be included as the predictors. This is equivalent to making the graph **G** sparse. Therefore, our approach is to remove the least important edges from the graph **G** through a *Gaussian graphical model* by applying an algorithm known as the *Graphical Lasso*, through which we build a sparse graph while keeping a relatively low estimation error for the inferred streamflow values.

3.3.4 The Graphical Lasso

The *Graphical Lasso* (Glasso) is an algorithm defined initially by Friedman et al. [143] which imposes sparsity to the precision matrix by tuning a parameter λ. This algorithm has been actively used, analyzed and improved by several authors [145]–[147]. Our work used the glasso Matlab package (glasso) and also a more recent efficient implementation called GLASSOFAST [148].

The *Glasso* algorithm implements an efficient solution to the problem by maximizing the Gaussian log-likelihood according to the formulation given in equation (3.22), adapted from [143], where *det* and *tr* are the determinant and trace of a square matrix respectively, $||\widehat{\mathbf{\Theta}}||_1$ is the L_1 norm of estimated precision matrix $\widehat{\mathbf{\Theta}}$ and λ is the L_1 norm regularization parameter.

$$\widehat{\mathbf{\Theta}}_{\text{Glasso}} \equiv arg_{\widehat{\mathbf{\Theta}}} max\left[log\left(det\ \widehat{\mathbf{\Theta}}\right) - tr\left(\mathbf{S}\cdot\widehat{\mathbf{\Theta}}\right) - \lambda||\widehat{\mathbf{\Theta}}||_1\right] \tag{3.22}$$

The *Glasso* algorithm requires that the probability distribution of the input data be relatively well described by a multivariate Gaussian distribution as is the case for the multivariate random variable **Z**. The inputs required by the *Glasso* algorithm are the empirical covariance matrix **S** and the regularization parameter λ. The output from the *Glasso* algorithm is a sparse precision matrix estimate $\widehat{\mathbf{\Theta}}_{\text{Glasso}}$ optimized by equation (3.22). Equation (3.23) shows the inputs and output of the *Glasso* algorithm. The estimation of the regression coefficients of matrix **A** for the inference of streamflow time series via the *Glasso* is achieved by applying equation (3.16) in which $\widehat{\mathbf{\Theta}}$ is replaced by $\widehat{\mathbf{\Theta}}_{\text{Glasso}}$ as shown in equation (3.24) below.

$$\widehat{\mathbf{\Theta}}_{\text{Glasso}} = Glasso(\mathbf{S}, \lambda) \tag{3.23}$$

$$\boldsymbol{\alpha}_{\mathbf{j}} = -\frac{1}{\hat{\theta}_{Glasso_{22}}}\widehat{\boldsymbol{\theta}}_{Glasso_{12}} \tag{3.24}$$

If the regularization parameter λ is equal to zero, the estimated precision matrix $\widehat{\boldsymbol{\Theta}}_{\text{Glasso}}$ is equivalent to the empirical precision matrix **T** obtained by the (non-regularized) MLR approach with equation (3.19) and with the corresponding graph **G** is a complete graph. On the other hand, if the regularization parameter is very large, the underlying graph **G** would have zero edges. An algorithm (SGM) is presented in subsection 0 to select the λ parameter based on a multi-objective optimization procedure that minimizes the error metric and also the number of edges of the underlying sparse Gaussian Graphical Model.

3.3.5 Graphical Model Selection

Our approach in selecting a proper subset of donor gauges to be used for inferring each streamflow gauge (Step 1) is to apply the conditional independence assumptions encoded in the precision matrix. In other words, the idea of conditional independence is used to find a subset of donor gauges that are conditionally correlated to each target location. This proposed approach promotes sparsity on the precision matrix and, therefore, leads to an underlying graph **G** with fewer edges which is consistent with the parsimonious principle. That is, a simpler model that explains well the observations should be preferred over more complex models. Under such a context, the parsimonious principle implies a selection of an underlying graphical model that is as sparse as possible while keeping the estimation error relatively low. Subsections 3.3.5.1 to 3.3.5.5 outline how to compute a sparse underlying graphical model **G**, and how to compute the validation estimation error for given values of the parameters λ and τ. Finally, subsection 0 describes a novel algorithm called Selection of Graph Model (SGM), that uses several values of λ and τ, for the estimation of the optimal underlying graph model **G**.

3.3.5.1 Imposition of Sparsity to the Underlying Graphical Model

The sparsity is achieved by adjusting the regularization parameter λ for the *Glasso* algorithm in conjunction with a thresholding procedure that uses an additional parameter τ defined by equation (3.25) below, which is a modification of equation (3.17).

$$\mathbf{G} = \begin{cases} g_{ij} = 1 & if \quad |\hat{\theta}_{ij}| > \tau \\ g_{ij} = 0 & otherwise \end{cases} \tag{3.25}$$

The thresholding procedure is required in addition to the L_1 norm regularization because even though the L_1 norm of the precision matrix decreases monotonically as λ increases, the number of edges in the graph **G** does not necessarily decrease monotonically. Therefore, a multi-objective optimization is needed to minimize the mean error between the observed random variable **Z** and the inferred data matrix $\hat{\mathbf{Z}}$ from equation (3.10), and the number of edges of the underlying graph **G**. In addition to equation (3.25) for sparsity, there exist some situations, due to the problem setup, where a particular edge from the *ith* to the *jth* gauge needs to be removed from the underlying graphical model by setting the element g_{ij} to zero. One example of such situation is when both the *ith* and the *jth* gauges are known to be donor basins, therefore none of them need to be inferred and the corresponding edge in the graphical model should be removed. A similar case applies when both gauges are known to be the target gauges, the edge between them should not exist, as one gauge cannot be infered using the other as a donor. In such cases, the Glasso procedure with an optional parameter, graph **G**, in equation (3.26) allows removal of some edges. If that graph **G** is ommitted, as in Equation (3.23), it assumes that all edges are available. Therefore Equation (3.23) is equivalent to Equation (3.26), if Graph **G** is a full graph. Equation (3.26) is also useful because it allows one to compute the sparse precision matrix estimate with a prescribed sparsity pattern. In addition, if the regularization parameter λ is equal to zero, then this equation is

equivalent to a MLR where each target gauge is estimated by the donor gauges that share an edge with it in the graph **G**.

$$\widehat{\boldsymbol{\Theta}}_{train(\lambda,G)} = Glasso(\mathbf{S}_{\boldsymbol{train}}, \lambda, \mathbf{G}) \quad \textbf{(3.26)}$$

3.3.5.2 Preparation of Data Sets

The normalized standard Gaussian (Z-score of log-transformed) daily streamflow data set, **Z**, is sorted in ascending order by the timestamp of each daily record and then divided into three disjoint sets of (approximately) the same size. The subsets are used, respectively, for training $\mathbf{Z}_{train}$, validation $\mathbf{Z}_{val}$, and testing $\mathbf{Z}_{test}$, respectively. $\mathbf{Z}_{train}$ is used for training the inference model, by computing the regression coefficients for the matrix **A**. $\mathbf{Z}_{val}$ is used for choosing the λ and τ values that minimize the validation error and the number of edges of the underlying graph **G,** and $\mathbf{Z}_{test}$ is used for assessing the predictive capability of the streamflow inference algorithm through estimating the error based on the new data. The least recent two thirds of the daily streamflow records are randomly assigned to the training $\mathbf{Z}_{train}$ and validation $\mathbf{Z}_{val}$ data sets with a split ratio of 50%. The remaining one third of the data (most recent) is used as the test set $\mathbf{Z}_{test}$.

3.3.5.3 Estimation of Training Covariance and Sparse Precision Matrices

The initial training precision matrix, $\widehat{\boldsymbol{\Theta}}_{train(\lambda,G_{full})}$, for a given value of the regularization parameter λ, is computed by applying the *Glasso* algorithm of equation (3.23) using $\mathbf{S}_{train}$. The training covariance matrix, $\mathbf{S}_{train}$, was estimated by applying equation (3.18) along with the training dataset $\mathbf{Z}_{train}$. Alternatively, the initial precision matrix can be computed by using equation (3.26) with **G** equals to the full graph, $\boldsymbol{G_{full}}$. The initial sparsity of the training precision

matrix, $\widehat{\boldsymbol{\Theta}}_{train(\lambda,G_{full})}$, is determined by the regularization parameter λ. Additional sparsity is achieved by computing a sparse graph, $\boldsymbol{G_\tau}$, where a thresholding procedure for a given value of the truncation parameter $\boldsymbol{\tau}$, as defined in equation (3.25), is applied using the initial precision matrix. A new training precision matrix $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{train}(\lambda,\boldsymbol{G_\tau})}$, is then computed using equation (3.26) and the sparse graph $\boldsymbol{G_\tau}$. This sparse precision matrix has a value of zero on all elements where the graph $\boldsymbol{G_\tau}$ has missing edges.

3.3.5.4 Estimation of Regression Coefficients and Streamflow Validation

The training matrix of regression coefficients, $\mathbf{A}_{train}$, is computed by the matrix decomposition of the training sparse precision matrix, $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{train}(\boldsymbol{3.\lambda},\boldsymbol{G_\tau})}$, using equation (3.24), for $1 \leq j \leq p$, where *j* is the *jth* gauge.

The standardized validation (Z-score of log-transformed) streamflow time series, $\hat{\mathbf{Z}}_{val}$, are estimated by using $\mathbf{A}_{train}$ and the validation dataset, $\mathbf{Z}_{val}$, as expressed in equation (3.27) below:

$$\hat{\mathbf{Z}}_{val} = \mathbf{Z}_{val} \cdot \mathbf{A}_{train} \tag{3.27}$$

The estimated log-transformed validation streamflow data, $\widehat{\mathbf{Y}}_{val}$, is calculated using $\hat{\mathbf{Z}}_{val}$ in equation (3.8) and is shown in equation (3.28) below, for $1 \leq j \leq p$, where *j* is the *jth* gauge, $\mu_{y_{val_j}}$ and $\sigma_{y_{val_j}}$, represents, respectively, the mean and standard deviation of the vector $\widehat{\mathbf{Y}}_{val_j}$.

$$\widehat{\mathbf{Y}}_{val_j} = \hat{\mathbf{Z}}_{val_j} \cdot \sigma_{y_{val_j}} + \mu_{y_{val_j}} \tag{3.28}$$

The estimated validation streamflow data, $\widehat{\mathbf{Q}}_{val}$, is calculated by applying the exponential function to $\widehat{\mathbf{Y}}_{val_j}$, as shown in equation (3.29), for $1 \leq j \leq p$, where j is the jth gauge:

$$\widehat{\mathbf{Q}}_{val_j} = exp\left(\widehat{\mathbf{Y}}_{val_j}\right) - 1 \quad \textbf{(3.29)}$$

3.3.5.5 Score Function and Validation Error

Selection of the graphical model should maximize the quantity and the quality of the inferred daily streamflow time series. The goal is to estimate daily streamflow time series at the target gauges as accurate as possible so that these gauges can be potentially removed from the hydrometric network, with the least loss of information. The score function is designed to measure the accuracy of the inferred values at the target sites. Equation (3.30) defines a conditional goodness-of-fit metric that calculates the value of the coefficient of determination R^2 between the observed and estimated j*th* daily streamflow time series for the validation data set, where $R^2_{val_j}$ is the coefficient of determination, i.e., the square value of the correlation coefficient R^2, between the observed validation streamflow $\mathbf{Q}_{val_j}$ used for validation and the estimated streamflow $\widehat{\mathbf{Q}}_{val_j}$, for $1 \leq j \leq q$, where *j* is an index representing the *jth* gauge and *q* is the number of inferred gauges. By default, all of the gauges are considered as potential target sites, where *q* is equal to *p*. The score is non-zero only if $R^2_{val_j}$ is greater than an assigned threshold Γ, otherwise it is taken as zero. In this work the value of the threshold Γ was set to 0.7. Equation (3.31) defines the validation score and Equation (3.32) defines the validation error function used in our multi-objective optimization procedure.

$$score_{val_j} = \begin{cases} R^2_{val_j} = R^2\left(\mathbf{Q}_{val_j}, \widehat{\mathbf{Q}}_{val_j}\right) & if \quad R^2_{val_j} > \Gamma \\ 0 & otherwise \end{cases} \quad \textbf{(3.30)}$$

$$score_{val} = \sum_{j=1}^{q} score_{val_j}, \quad q \leq p \tag{3.31}$$

$$error_{val} = \frac{q - score_{val}}{q} \tag{3.32}$$

While the number of edges of the underlying graph indicates its sparseness, the validation error of the graphical model, is selected in such a way that it will maximize the validation score. The value of this validation error ranges over [0, 1], and is scale independent. It decreases as the validation score increases.

3.3.5.6 Selection of Graph Model Algorithm (SGM)

An algorithm called *Selection of Graph Model* (SGM) is developed to obtain an optimal underlying graph. A graph determined bt the SGM algorithm is represented by $\boldsymbol{G_{sgm}}$**.** The SGM algorithm implements a multi-objective optimization procedure where the optimization objectives include: (1) minimizing the error calculated by equation (3.32), and (2) minimizing the number of edges of the underlying graph. SGM generates a set of values for the regularization parameter λ between a minimum value of λ_{min} and and a maximum value of λ_{max}. For each regularization parameter value of λ, the truncation parameter, τ, in equation (3.25) is selected in such a way that the underlying graph has a given number of edges between a minimum, K_{min}, and a maximum, K_{max}, respectively. Given the multi-objective nature of the problem, a set of graphs corresponding to a set of non-dominated solutions on the Pareto front instead of a single solution is selected. Graph $\boldsymbol{G_{sgm}}$ thus represents one of the graphs from the set. A final graph(s) is (are) selected from the set of candidate solutions as the one (ones) that offers (offer) the desired trade-off between error and model complexity.

Algorithm 1 below briefly describes a code implementation the SGM algorithm. The parameter *res* is an integer number that represents the resolution of a sequence of sampling values to create a (1 x *res*) vector ***lamba_set*** with values between λ_{min} and λ_{max}. *DonorSet* and *TargetSet* are optional parameters that represent a set of identifiers of the gauges that are known to be donors or targets, respectively. The default values for *DonorSet* and *TargetSet, in Algorithm 1,* are empty sets. That is, any gauge can potentially be used as a *Donor* or *Target* gauge. If *DonorSet* or *TargetSet* are non-null sets, then the corresponding gauges are treated as donor gauges or target gauges, respectively. Therefore, computing the graph model $\mathbf{G}_{\tau}$ (i.e., graph G constrained by parameter τ) from equation (3.25), implies removing all the edges between the *ith* and *jth* gauge when both, *i* and *j,* belong to *DonorSet* or both belong to *TargetSet.* The *getSequence* function generates the vector ***lamba_set***. A simple way to implement this function is by using a linear sequence. This algorithm is summarized below as Algorithm 1.

Algorithm 1: *Selection of Graph Model (SGM)*

STEP 0. Define the SGM inputs (assignment of default values)

$\lambda_{min} = 0.01$; $\lambda_{max} = 0.10$; $K_{min} = 10$; $K_{max} = \frac{p^2-p}{2}$; *res = 30;* $\Gamma = 0.7$

DonorGroup:={}; TargetGroup:={}

Retrieve training ($\mathbf{Z}_{train}$) and validation ($\mathbf{Z}_{val}$) data sets;

STEP 1. Compute the empirical covariance matrix using equation **(3.18)** from the training set:

$n_{train} = length(\mathbf{S}_{train})$

$\mathbf{S}_{train} = \frac{1}{n_{train}-1} \mathbf{Z}_{train}{}^{\mathbf{T}} \cdot \mathbf{Z}_{train}$;

STEP 2. Generate Multi-objective optimization sampling points:

lambda_set *= getSequence(minVal=*λ_{min}*,* $maxVal = \lambda_{max}$*, res);*

for *r=1* ***to*** *res:*

λ_r*=* ***lambda_set****[r];*

Compute the initial precision matrix from $\mathbf{S}_{train}$ using equation **(3.23)**:

$\widehat{\mathbf{\Theta}}_{train_r} = Glasso(\mathbf{S}_{train}, \lambda_r)$;

for *k=* K_{min} ***to*** K_{max}:

choose $\tau_{r,k}$ to compute the underlying graph model with at most *k* edges using equation **(3.25)**:

$$\mathbf{G}_{r,k} = \begin{cases} g_{r,k_{ij}} = 1 & if \quad |\hat{\theta}_{train_{r_{ij}}}| > \tau_{r,k} \\ g_{r,k_{ij}} = 0 & otherwise \end{cases};$$

Compute the sparse training precision matrix, using equation **(3.26)**:

$\widehat{\mathbf{\Theta}}_{train_{r,k}} = Glasso\left(\mathbf{S}_{\boldsymbol{train}}, \lambda_r, \mathbf{G}_{r,k}\right)$;

Compute the training matrix of regression coefficients $\mathbf{A}_{train_{r,k}}$ from $\widehat{\mathbf{\Theta}}_{train_{r,k}}$, using equation **(3.24)**, for $1 \leq j \leq p$:

$\boldsymbol{\alpha}_{train_{r,k_{\mathrm{j}}}} = -\frac{1}{\hat{\theta}_{train_{r,k_{22}}}} \widehat{\boldsymbol{\theta}}_{train_{r,k_{12}}}$;

Compute the inferred Z-score log-transformed validation streamflow from equation **(3.27)**:

$\hat{\mathbf{Z}}_{val_{r,k}} = \mathbf{Z}_{val} \cdot \mathbf{A}_{train_{r,k}}$;

Compute the inferred log-transformed validation streamflow using equation **(3.28)** for $1 \leq j \leq p$:

$\hat{\mathbf{Y}}_{val_{r,k_j}} = \hat{\mathbf{Z}}_{val_{r,k_j}} \cdot \sigma_{y_{val_j}} + \mu_{y_{val_j}}$;

Compute the inferred validation streamflow using equation **(3.29)** for $1 \leq j \leq p$:

$\widehat{\mathbf{Q}}_{val_{r,k_j}} = exp\left(\hat{\mathbf{Y}}_{val_{r,k_j}}\right) - 1$;

Calculate the validation score using equation **(3.30)** and equation **(3.31)** for $1 \leq j \leq q$:

$$score_{val_{r,k_j}} = \begin{cases} \mathrm{R}^2_{val_j} = \mathrm{R}^2\left(\mathbf{Q}_{val_j}, \widehat{\mathbf{Q}}_{val_j}\right) & if \quad \mathrm{R}^2_{val_j} > \Gamma \\ 0 & otherwise \end{cases}$$

$score_{val_{r,k}} = \sum_{j=1}^{q} score_{val_{r,k_j}}, \quad q \leq p$;

Calculate the validation error using equation: **(3.32)**

$error_{val_{r,k}} = \frac{q - score_{val_{r,k}}}{q}$;

store the sampling results: *multi_objective_points* = [*k*, $error_{val_{r,k}}$], λ_r and $\mathbf{G}_{r,k}$.

STEP 3. Select the set of non-dominated solutions from *multi_objective_points*

STEP 4. From the set of non-dominated solutions, select a sparse graph (as the output), $\boldsymbol{G}_{sgm}$, with a suitable tradeoff between the number of edges and validation error and optionally the corresponding matrix of regression coefficients $\boldsymbol{A}_{sgm}$.

3.3.6 Stream Flow Inference

The inference task is greatly simplified once the underlying graph $\boldsymbol{G}_{sgm}$ is identified by the SGM algorithm. This graph $\boldsymbol{G}_{sgm}$ reveals conditional independence conditions between the streamflow gauges for the given hydrometric streamflow network. Therefore, the best set of donor gauges for each streamflow gauge is explicitly indicated by the graph $\boldsymbol{G}_{sgm}$. Such a set includes only the donor gauges for which the target station depends on conditionally.

3.3.6.1 Inference of Daily Streamflow Time Series with Graph (SGM)

Let matrix $\boldsymbol{A}_{sgm}$ represent the matrix A of equation (3.11) whose element α_{ij} (i.e., regression coefficient) is determined based on graph $\boldsymbol{G}_{sgm}$. The Z-score of the log-transformed streamflow time series for the test set $\hat{\mathbf{Z}}_{test}$, can then be estimated directly using the matrix $\boldsymbol{A}_{sgm}$ and the test dataset $\mathbf{Z}_{test}$. Thus, equation (3.10), can be expressed as follows:

$$\hat{\mathbf{Z}}_{test} = \mathbf{Z}_{test} \cdot \boldsymbol{A}_{sgm} \tag{3.33}$$

To obtain $\hat{\mathbf{Y}}_{test}$ from $\hat{\mathbf{Z}}_{test}$, the mean $\mu_{Y_{test_j}}$ and standard deviation $\sigma_{Y_{test_j}}$ for the test set are required, but they are unknown. One way to overcome this problem is to assume that the mean and standard deviation for the test set are the same as they are for the training set. Then, one can obtain $\hat{\mathbf{Y}}_{test}$ from which to obtain the original streamflow time series $\hat{\mathbf{Q}}_{test}$ by applying the exponential-transform function. Clearly, the assumption made here is not usually held.

An alternative approach is to perform an ordinary least squares multiple linear regression to estimate a new set of regression coefficients of β_{ij} (slope) and β_{0j} (intercept) over the log-transformed streamflow time series for the training data set over $1 \leq j \leq p$, using only the donors for the j*th* target site as expressed by equation (3.34), where $donors(j) = donors(\boldsymbol{G}_{sgm}, j)$.

$$\widehat{\mathbf{Y}}_{\text{test}_j} = \beta_{0j} + \sum_{i=1}^{size(donors(j))} \beta_{ij} \cdot \mathbf{Y}_{train_{donors(j)_i}} \quad \textbf{(3.34)}$$

The daily streamflow time series, $\widehat{\mathbf{Q}}_{\text{test}_j}$, is estimated by applying the exponential function to the log-transformed streamflow, $\widehat{\mathbf{Y}}_{\text{test}_j}$, for $1 \leq j \leq p$, as shown in equation (3.35).

$$\widehat{\mathbf{Q}}_{\text{test}_j} = exp\left(\widehat{\mathbf{Y}}_{\text{test}_j}\right) - 1 \quad \textbf{(3.35)}$$

The third alternative is to directly apply MLR to the non-transformed streamflow time series avoiding the logarithmic transformation. Among these three approaches, results from equation (3.34) should be either more accurate or more stable as indicated by Farmer [134] who found that the logarithmic transformation of the streamflow is generally the most stable predictand.

3.3.6.2 Inference of Daily Streamflow Time Series using Distance and Correlation Approaches

To evaluate the performance of our new method based on graph $\boldsymbol{G}_{sgm}$ in inferring daily streamflow time series, we compare our new method with two widely used methods, the distance-based method ("Dist") and correlation-based method ("Corr"). Two graphs, $\boldsymbol{G}_{dist}$ (distance-based) and $\boldsymbol{G}_{corr}$ (correlation-based), are constructed. The $\boldsymbol{G}_{dist}$ graph is built starting with an empty graph (i.e., none of the gauges in the study region are connected) and then adding edges (i.e., connecting gauges) between each target site and its nearest neighbor site. In this case, each target site has at least one donor site, expressed as $\boldsymbol{G}_{dist,1}$, and the constructed graph structure is determined by the number of edges added and their relative locations in the gauge network. For the case of having at least two donor sites, edges between each target site and its nearest and second nearest neighbor sites are added in the graph and is expressed as $\boldsymbol{G}_{dist,2}$. Graph $\boldsymbol{G}_{dist,3}$ represents the case where each target gauge has at least 3 donor sites. The graph of $\boldsymbol{G}_{corr}$ is built in a similar

way to $\boldsymbol{G}_{dist}$ except that the most correlated sites instead of the nearest sites are selected. For the case with one donor site, the built graph is represented by $\boldsymbol{G}_{corr,1}$. For the cases with two and three donor sites, the constructed graphs are represented by $\boldsymbol{G}_{corr,2}$ and $\boldsymbol{G}_{corr,3}$, respectively. The graphs of $\boldsymbol{G}_{dist,i}$ and $\boldsymbol{G}_{corr,i}$ (i = *1*, *2*, and *3*) are built in such a way to mimic the current practice in which a fixed and equal number of donors for each target site is used in both distance- and correlation-based approaches. Notice however, that the underlying graphs are undirected. Thus, the number of edges in the graphs $\boldsymbol{G}_{dist,i}$ and $\boldsymbol{G}_{corr,i}$ might be different for the same number of donors (i.e. 'i"). In comparison, an uneven number of donors for each target site is automatically determined and used in our new method. For both the distance- and correlation-based methods, the daily streamflow time series are inferred following the same procedure described in section 3.3.6.1 for our new method (i.e., SGM). The only difference is replacing the graph $\boldsymbol{G}_{sgm}$ by $\boldsymbol{G}_{dist,i}$ or $\boldsymbol{G}_{corr,i}$ (i = *1*, *2*, and *3*) in each case.

3.3.6.3 Estimation of the Test Error

The test error is computed in the same way as the validation error described in subsection 3.3.5.5, but using the test set as follows:

$$score_{test_j} = \begin{cases} \mathrm{R}^2_{test_j} = \mathrm{R}^2\left(\mathbf{Q}_{test_j}, \widehat{\mathbf{Q}}_{test_j}\right) & if \quad \mathrm{R}^2_{test_j} > \Gamma \\ 0 & otherwise \end{cases} \tag{3.36}$$

$$score_{test} = \sum_{j=1}^{q} score_{test_j}, \quad q \leq p \tag{3.37}$$

$$error_{test} = \frac{q - score_{test}}{q} \tag{3.38}$$

3.3.6.4 Estimation of Inference Accuracy

The accuracy of each of the inferred gauges associated with the graph from the SGM algorithm and with the graphs of $\boldsymbol{G}_{dist,i}$ and $\boldsymbol{G}_{corr,i}$ ($i = 1, 2,$ and 3) is evaluated, is evaluated using the Nash–Sutcliffe efficiency coefficient (NSE) [91] with the testing data set. The NSE of the testing set (NSE_{test_j}) is computed between the observed ($\mathbf{Q}_{\text{test}_j}$) and inferred ($\widehat{\mathbf{Q}}_{\text{test}_j}$) streamflow time series for $1 \leq j \leq p$ as shown in equation (3.39) below.

$$NSE_{test_j} = NSE\left(\mathbf{Q}_{\text{test}_j}, \widehat{\mathbf{Q}}_{\text{test}_j}\right) \tag{3.39}$$

3.3.7 Removal of Streamflow Gauges with the Least Loss of Information

The removal of streamflow gauges (RG), is a straightforward procedure once the model selection and inference stages are completed. The RG algorithm is designed to remove the gauges that can be inferred by other gauges with the highest efficiency, i.e. with the highest NSE (NSE_{test_j}) in the testing set, NSE_{test_j}. RG removes a non-marked gauge with the highest NSE for the testing data set. Thus, RG removed a gauge in the network with the highest NSE_{test_j} first, and then marks the removed gauge as a “target gauge” and each of its neighbors as a “donor gauge”. This process is repeated for the remaining available gauges in the network until all gauges are checked, with the exception of isolated gauges that should not be removed. Algorithm 2 below shows the details of the gauge removal process with the least loss of information.

Algorithm 2: Removal of Gauges (RG) Algorithm

STEP 0. Define RG inputs: [NSE_{test_1}, …, NSE_{test_q}], $\boldsymbol{G}_{sgm}$

STEP 1. Initialize the rank of removal: rank=0

STEP 2. Mark all the gauges with at least one edge as available for removal. Isolated nodes are marked as not available for removal.

STEP 3. Update the rank of removal (rank = rank + 1)

STEP 4. Define the *r*th gauge as the one with the highest Nash–Sutcliffe model efficiency coefficient from the currently available gauge set. Assign the *r*th gauge to the current rank of removal.

STEP 5. Mark the *r*th gauge and its neighbors on the underlying graph $\boldsymbol{G}_{sgm}$ as unavailable for removal.

STEP 6. Repeat from step 3 until there are no more available gauges for removal.

Equation (3.40) and Equation (3.41) defines a new score for the graph, based on the NSE, but it only includes removable gauges with a NSE value higher than the threshold **Γ**. For this study the value for **Γ** was set to 0.7. The constant *maxRemRank* represents the maximum number of gauges removable from the RG algorithm for a given graph. The $\boldsymbol{graph\_score_{test}}$ is useful to assess the quality and quantity of the inference of daily streamflow time series for the removable gauges from a given graph model. The higher the $\boldsymbol{graph\_score_{test}}$ is, the better.

$$graph\_score_{test_{remRank}} = \begin{cases} NSE_{test_{remRank}} & if \quad NSE_{test_{remRank}} > \Gamma \\ 0 & otherwise \end{cases} \tag{3.40}$$

$$graph\_score_{test} = \sum_{remRank=1}^{maxRemRank} graph\_score_{test_{remRank}} \tag{3.41}$$

3.4 Study Area and Data Sets

Our new method is applied to the Ohio River basin due to its size, relevance and good quality of long-term historical daily streamflow data. The Ohio River is, according to the discharge, the third largest river in the United States. It is the largest tributary of the Mississippi River and accounts for more than 40% of the discharge of the Mississippi River [149]. The Ohio River is located between the 77° and 89° west longitude and between the 34° and 41° north latitude.

Table 3.1 lists the National Weather Service Location Identifier (NWSLI) which is used in this work to index each gauge, along with the drainage area of the corresponding sub-basin, and the USGS station identifier of the 34 streamflow gauges. The naturalized daily streamflow data are taken from the United States Geological Survey (USGS)'s National Water Information System (NWIS: National Water Information System). This data set spans from January 1st, 1951 to December 31st, 1980 with a total of 10958 consecutive days (30 years) for all the 34 streamflow gauges. There are no missing streamflow records for any day or gauge over the selected study period.

Following the procedure described in subsection 3.3.5.2, the dataset was separated into 3 subsets. Data between 1951 and 1970 were used for the "training" and "validation". The training data set consists of 50% of the data randomly selected between 1951 and 1970. The remaining data over the period of 1951 and 1970 consists of the validation set. The data between 1961 and 1970 was used as the "test" set.

Table 3.1 List of 34 Streamflow Gauges Over the Ohio River Basin

#	NWSLI	USGS STAID	Drainage Area (Km^2)	#	NWSLI	USGS STAID	Drainage Area (Km^2)
1	ALDW2	3183500	3,533	**18**	GRYV2	3170000	777
2	ALPI3	3275000	1,352	**19**	KINT1	3434500	1,764
3	ATHO1	3159500	2,442	**20**	MROI3	3326500	1,766
4	BAKI3	3364000	4,421	**21**	NHSO1	3118500	453
5	BELW2	3051000	1,052	**22**	NWBI3	3360500	12,142
6	BOOK2	3281500	1,870	**23**	PRGO1	3219500	1,469
7	BSNK2	3301500	3,364	**24**	PSNW2	3069500	1,870
8	BUCW2	3182500	1,399	**25**	SERI3	3365500	6,063
9	CLAI2	3379500	2,929	**26**	SLMN6	3011020	4,165
10	CLBK2	3307000	487	**27**	SNCP1	3032500	1,368
11	CRWI3	3339500	1,318	**28**	STMI2	3345500	3,926
12	CYCK2	3283500	938	**29**	STRO1	4185000	1,062
13	CYNK2	3252500	1,608	**30**	UPPO1	4196500	772
14	DBVO1	3230500	1,383	**31**	VERO1	4199500	679
15	ELRP1	3010500	1,424	**32**	WTVO1	4193500	16,395
16	FDYO1	4189000	896	**33**	WUNO1	3237500	1,002
17	GAXV2	3164000	2,929	**34**	WYNI2	3380500	1,202

3.5 Results and Discussion

3.5.1 Inference on Streamflow

The inferred daily streamflow time series based on the new method (i.e., graph $\boldsymbol{G}_{sgm}$) and the distance- and correlation-based methods (i.e., graphs of $\boldsymbol{G}_{dist,i}$ and $\boldsymbol{G}_{corr,i}$ with $i = 1, 2$, and 3) are compared. For the latter two approaches, the three commonly used scenarios with 1, 2, and 3 donors per target gauge are considered. For our new method, the SGM algorithm was run with the

default parameters defined in *Algorithm 1*. That is, 30 different values of the regularization parameter λ were used for graphs with edges between 10 (very sparse) and 561 (complete graph). Thus, the number of sampling points is ((561 - (10+1)) *30) = 16560 (based on Step 2 of Algorithm 1).

The SGM algorithm selected 74 out of 16560 (0.45%) distinct graphs with different number of edges as the candidate solutions according to the multi-objective optimization procedure that minimizes both of the the validation error and the number of edges. Figure 3.1 (a) shows results with trade-offs between the number of edges and the validation error, $error_{val}$, defined by equation (3.32). The black dots represent the dominated solutions in the multiple-optimization space. The three red dots of the non-dominated solutions represent the graphs of SMG(25), SGM(47) and SGM(65) with 25, 47 and 65 edges, respectively. The remaining non-dominated solutions (i.e., solutions along the Pareto front) are represented by the green dots. Figure 3.1 (b) shows the comparison of the test error (Equation (3.38)) associated with the graphs $\boldsymbol{G}_{sgm}$, of SMG(25), SGM(47) and SGM(65), and graphs of $\boldsymbol{G}_{dist,i}$ (distance-based) and $\boldsymbol{G}_{corr,i}$ (correlation-based) with *i = 1*, *2*, and 3. More specifically, for the distance-based case, $\boldsymbol{G}_{dist,1}$ = Dist(24), $\boldsymbol{G}_{dist,2}$ = Dist(43), and $\boldsymbol{G}_{dist,3}$ = Dist(65) with 24, 43, and 65 edges in each corresponding graph. For the correlation-based case, $\boldsymbol{G}_{corr,1}$ = Corr(24), $\boldsymbol{G}_{corr,2}$ = Corr(47), and $\boldsymbol{G}_{corr,3}$ = Corr(68) with 24, 47, and 68 edges in each corresponding graph as well.

The shape of the pareto front (i.e. green and red dots) in Figure 3.1 (a) shows a large validation error when the graphs are very sparse. But the error decreases quickly as the number of edges increases until about 44 edges at which point, the error curve flattens with diminished change in the validation error up to about 93 edges where a significant decrease in the validation error occurs. Then, the error becomes flat again until about 158. The error then decreases gradually until

about 211 edges where there is other significant decrease in the validation error and then the error becomes close to a constant even with the increase in the number of edges. For this study region, it appears that a good trade-off between the sparsity and validation error is about having 44 or 45 edges, where the error is almost as low as the graph with 92 edges. Also, the error decreases dramatically at the beginning where an addition of a few more edges can significantly reduce the error. But for a graph with its number of edges starting around 45, an increase in the number of edges only reduces the error a little bit. When the number of edges increases to about 92 or more, the improvement in error reduction becomes almost unnoticeable. Figure 1(a) shows that the relationship between the error and the number of edges has a L-like-shape in which the error approaches almost a constant when the graph reaches an edge number around 93. The few "sudden" discontinuities in Figure 1(a) are due to the nature of the error function which includes conditional terms above/below a threshold that might affect the total validation error once the threshold has been reached. The full graph with 561 edges is not in the set of non-dominated solutions, which means that using all of the gauges available in the network to infer the streamflow for the target site gives worse results than many of the sparser graphs. This is likely related to the noisy correlation calculated due to the large noises involved in the data. In fact, Figure 1 (a) shows that using graphs with more than 222 edges is unlikely to reduce the validation error anymore. This result clearly shows that it is not the more complex the better.

The three graphs SMG(25), SGM(47) and SGM(65), represented by the three red points in Figure 3.1 (a), were selected from a set of non-dominated solutions that, in terms of the number of edges, approximately match the three graphs associated with 1-, 2-, and 3-nearest donors, Dist(24), Dist(43) and Dist(65) and the three graphs associated with 1-, 2-, and 3-most correlated donors, Corr(24), Corr(47) and Dist(68). These three graphs of SMG(25), SGM(47) and SGM(65) are

selected so that it makes a fair comparison among the three methods as they all have a similar graph complexity. Errors associated with these three different levels of sparsity are represented in Figure 3.1 (b), by the three red, green, and magenta bars for the graphs of $\boldsymbol{G}_{sgm}$, $\boldsymbol{G}_{dist}$, and $\boldsymbol{G}_{corr}$, respectively. Figure 3.1 (b) shows that the test error (equation (3.38)) is the lowest for the inferred daily streamflow time series using the $\boldsymbol{G}_{sgm}$ graphs from the SGM algorithm, and it is the worst for the inferred streamflow based on the distance-based approach. The inferred results for the correlation-based approach are between the two.

To test the statistical significance of these results shown in Figure 3.1 (b), the procedures described to infer the streamflow time series are repeated 30 times with random selection of the records for the training and validation sets (keeping the tests date set fixed). Running 6 single tailed t-tests using a significance level of 0.05, and a null hypothesis that the mean test error for the SGM graphs is equal to the Dist or Corr graphs (for the cases of 1-, 2-, and 3-donors, respectively), the null hypothesis was rejected on all cases (p-value < 0.0001), and the alternative hypothesis was accepted. That is, the mean test error with the test data set for SGM(25) is significantly lower than the mean error for Dist(24) and Corr(24); the mean error for SGM(47) is significantly lower than that for Dist(43) and Corr(47); and that the mean error for SGM(65) is significantly lower than that for Dist(65) and Corr(68). In other words, the results obtained using our new method of the SGM algorithm are significantly better than the results of using either the least distance-based or the maximum correlation-based approaches.

Figure 3.2 shows how each of the individual graphs look like using the SGM, least distance (Dist), and maximum correlation (Corr) approaches. For the latter two approaches, the three commonly used scenarios with 1, 2, and 3 donors per target site are illustrated. The graphs for a single donor are Dist(24) and Corr(24) with their equivalent counterpart of SGM(25) from the

SGM algorithm. For two donors they are Dist(43) and Corr(47), and their counterpart of SGM(47). Finally, for three donors they are Dist(65) and Corr(68), and their counterpart of SGM(65). The graphs in Figure 3.2 with green edges are for the distance-based approach (Dist), magenta edges for the correlation-based approach (Corr) and red edges for the SGM approach. It can be seen that the graphs associated with each of the three approaches are not the same although some features in their graphic structures are similar. From Figure 3.1 (b) and the hypothesis testing results, it is clear that the new SGM method is the best of the three. This is because our new method with the SGM algorithm accounts for the dependence structure in the entire streamflow network based on the concept of conditional independence conditions and employs the Glasso method to effectively extract such dependence structure through making the precision matrix sparse. Our results demonstrate that a good use of the conditional independence structure of the underlying streamflow network (i.e., use sparse precision matrix) is important and it outperforms the widely used correlation-based method (i.e., Corr) method which only directly uses the local correlation information. Comparing to the distance-based method, the correlation-based method is superior which is consistent with other results reported in the literature

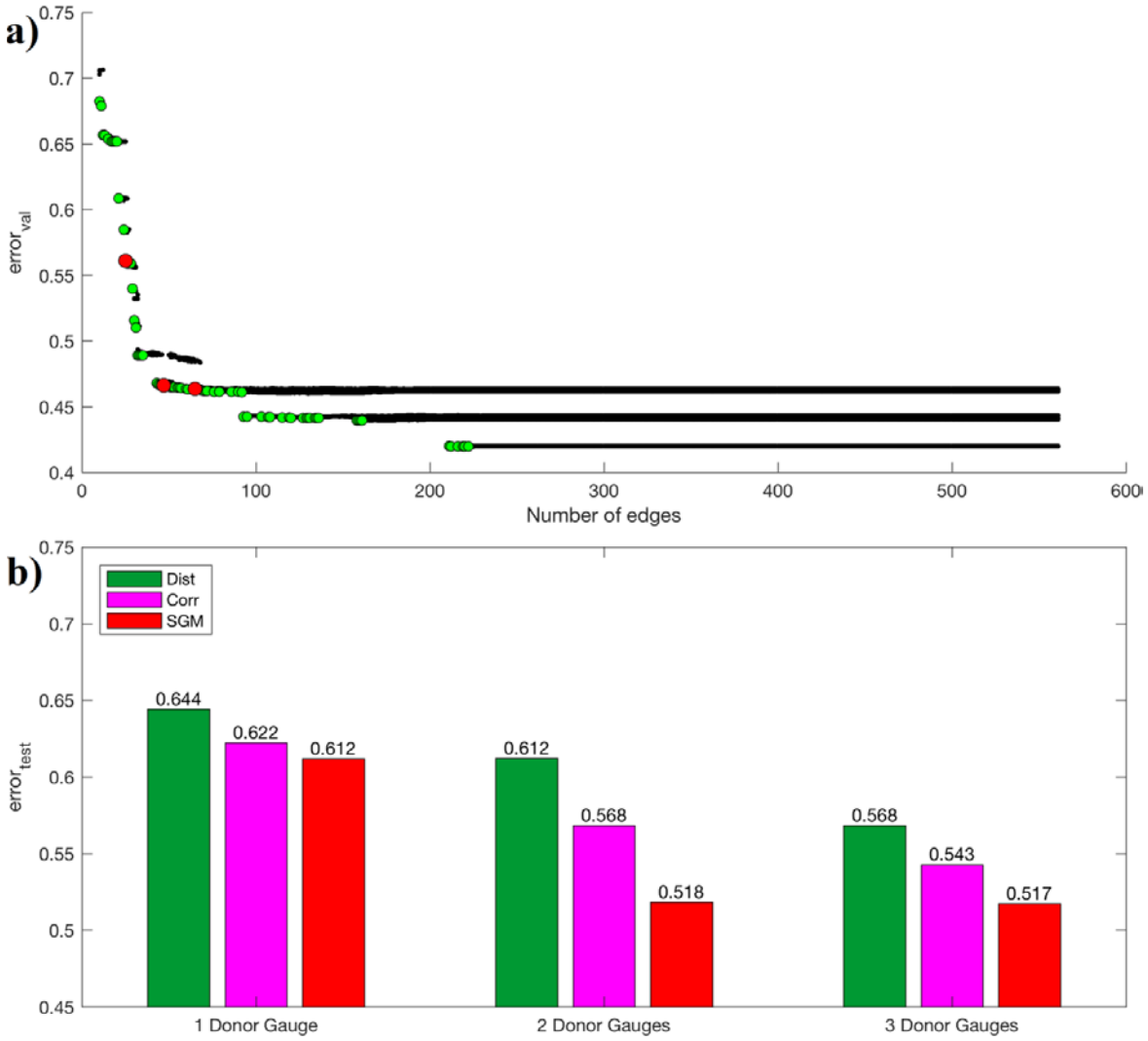

Figure 3.1 - Result of Running the SGM Algorithm with the Ohio River Basin Dataset

The training set is composed by a random selection of daily streamflow records between 1951 and 1970, while the validation set is composed by the remaining 50% for the same time span. The test data set is composed by the most recent 10 years of data (i.e. 1971-1980). (a) Validation error from the multi-objective optimization procedure of the SGM algorithm, between the observed and inferred daily streamflow time series vs the number of edges in the underlying graph (representing conditional independence assumptions between sites). The black dots represent sub-optimal (dominated) solutions. The Green dots represent the set of non-dominated (optimal) solutions. The red dots represent the graphs SGM(25), SGM(47) and SGM(65) with 25, 47 and 65 edges, respectively, chosen from the set of non-dominated solutions. (b) Comparison of the test error, between the SGM algorithm and the selection of donor gauges with the least distance (Dist) and maximum correlation (Corr) approaches, for 1, 2 and 3 donor sites. From left to right, comparison for one donor sites, the SGM(25) was selected to match the sparsity of Dist(24) and Corr(24). For two donor sites, graph SGM(47) was chosen to match Dist(43) and Corr(47). In the same way, SGM(65) is matched with Dist(65) and Corr(68)

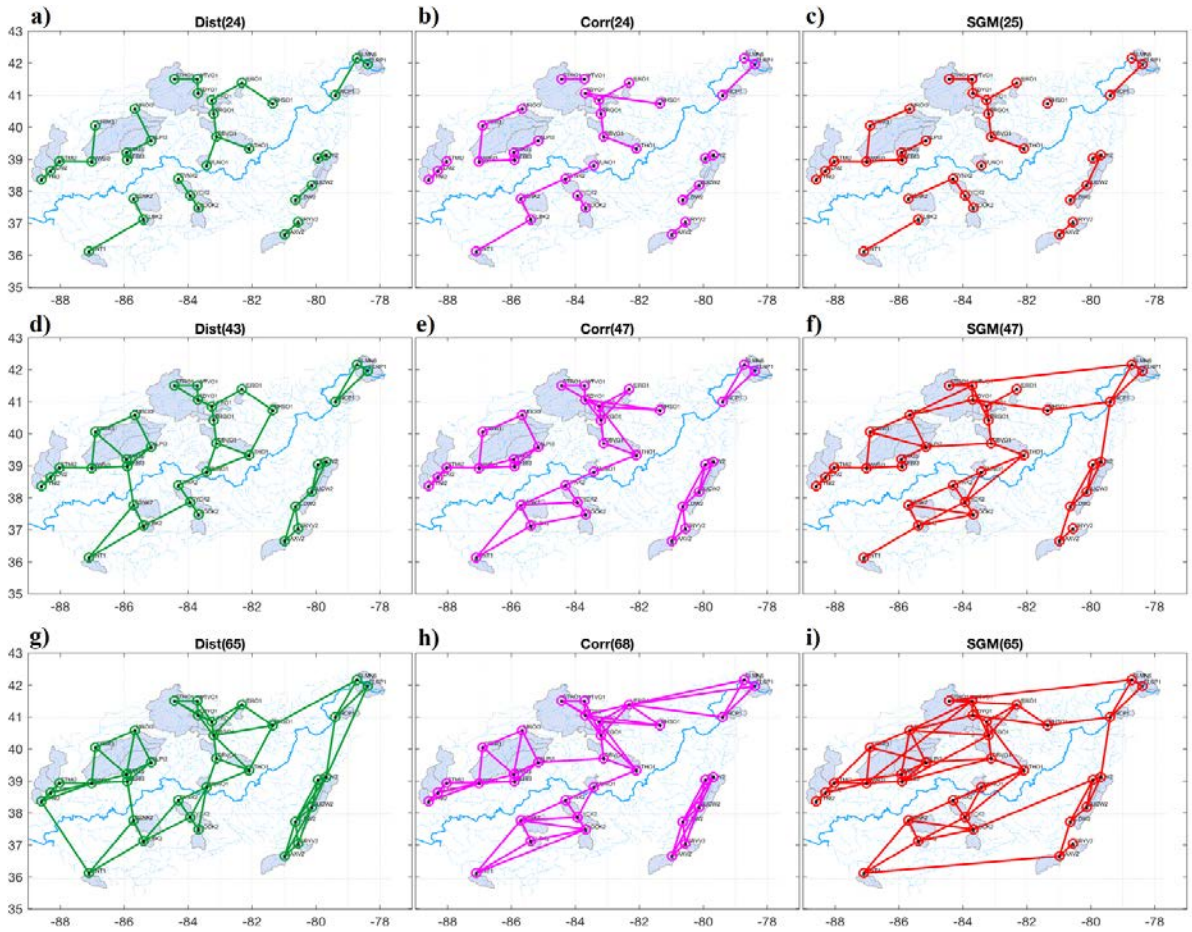

Figure 3.2 - Comparison of the generated graphs using the least distance (Dist) on left, the maximum correlation (Corr) on center column and the Selection of Graphical Model algorithm (SGM) on right, for 1 (on top), 2 (on middle row) and 3 (on bottom) donor sites, respectively. The graphs for "Dist" are built by adding 1, 2 or 3 edge(s) from each site to the site(s) with the least distance for 1, 2 and 3 donor sites, respectively. The graphs for "Corr" are built by adding 1, 2 or 3 edge(s) from each site to the site(s) with the highest correlation for 1, 2 and 3 donor sites, respectively. The number of edges for the "Dist" and "Corr" approaches is fixed depending on the number of donor sites used to build it, as opposed to the graphs from the SGM algorithm, where the number of edges can be selected from the set of non-dominated solutions. The SGM graphs are then selected to approximately match the sparsity for the "Dist" and "Corr" graphs for 1, 2 and 3 donor sites, respectively. (a) Least distance graph with a single donor site, Dist(24) with 24 edges. (b) Maximum correlation graph with a single donor site, Corr(24) with 24 edges. (c) Selection of graphical model algorithm graph, SGM(25), with 25 edges. (d) Least distance graph with two donor sites, Dist(43) with 43 edges. (e) Maximum correlation graph with two donor sites, Corr(47) with 47 edges. (f) Selection of graphical model algorithm graph, SGM(47) with 47 edges. (g) Least distance graph with three donor sites, Dist(65) with 65 edges. (h) Maximum correlation graph with three donor sites, Corr(68) with 68 edges. (i) Selection of graphical model algorithm graph, SGM(65) with 65 edges

3.5.2 Removal of Gauges with the Least Loss of Information

The objective here is to remove several gauges from the gauge network with the least loss of information. To do so, the graphs for a single donor: Dist(24), Corr(24), and SGM(25), for two donors: Dist(43), Corr(47), and SGM(47), and three donors: Dist(65), Corr(68) and SGM(65), are used for the sake of comparison among the least distance (Dist), Correlation (Corr), and SGM approaches. Each of the 9 graphs, shown in Figure 3.2, are selected as the input for the Removal of Gauges (RG) algorithm described in section 3.3.7 .The output of this algorithm shows that although there are 8 – 16 gauges that can potentially be removed, only about 7 – 8 gauges among them can be inferred with an NSE higher than 0.7.

Figure 3.3 shows the comparison of inferred daily streamflow inference results for the removable sites estimated by the RG algorithm, using the graphs shown in Figure 3.2 as inputs. Location of each gauge in Figure 3.3 is the same as that in Figure 3.2. Figure 3.3 highlights the removable gauges in a color-coded circle, indicating its relevant rank in terms of being adequately estimated by other gauges after being removed. The inference accuracy for each removed gauge is measured by an NSE value higher than or equal to 0.9 is depicted in blue; an NSE between 0.8 and 0.9 is depicted in green; between 0.7 and 0.8 in yellow; between 0.6 and 0.7 in orange; and below 0.6 in red.

Figure 3.4 (a) shows that the graph score for the single donor case is higher for the SGM graph than for the other two approaches. For the case with two donors, the SGM graph significantly outperforms the other two methods. The case for three donors is less clear, as there is a tie between the Corr and the SGM method, after reviewing the results, it was found that the chosen threshold ($\Gamma = 0.7$) is the culprit. For the three donors case, SGM(65) allows the removal of 11 gauges, but only 9 with an NSE greater than 0.6 and only 7 with a NSE greater than 0.7. On the other hand,

Corr(68) allows the removal of only 8 gauges, and 7 of them with an NSE greater than 0.7. Because the graph score only takes into account the gauges with an NSE higher than Γ, in this case the two additional gauges with NSE=0.65 were ignored, making the bar plot in Figure 3.4 (a) look like the same for the 3 donor gauge case, while in fact they are different as the SGM(65) graph allows for a removal of two additional gauges with just a slightly lower NSE value than the prescribed threshold. These two gauges are highlighted in orange in Figure 3.3 (i). Figure 3.4 (b) shows that the mean graph score is higher for the SGM method than those for the other two approaches, with the Corr method in second place and the Dist method in third. Figure 3.4 (c) shows a related but slightly different measure to that of Figure 3.4 (a), in assessing the quality and quantity of the inference results for the streamflow time series estimated by the three methods of SGM, Dist and Corr. In Figure 3.4 (c) the mean is taken from the 8 removable gauges with the highest NSE for each of the donor scenarios. The SGM has a higher mean among the top 8 removable gauges, for 1, 2 or 3 donors, than the other two approaches. Figure 3.4 (d) summarizes the results shown in Figure 3.4 (c). Clearly, the mean NSE for the top 8 removable gauges is higher for the SGM method than those for the other two methods.

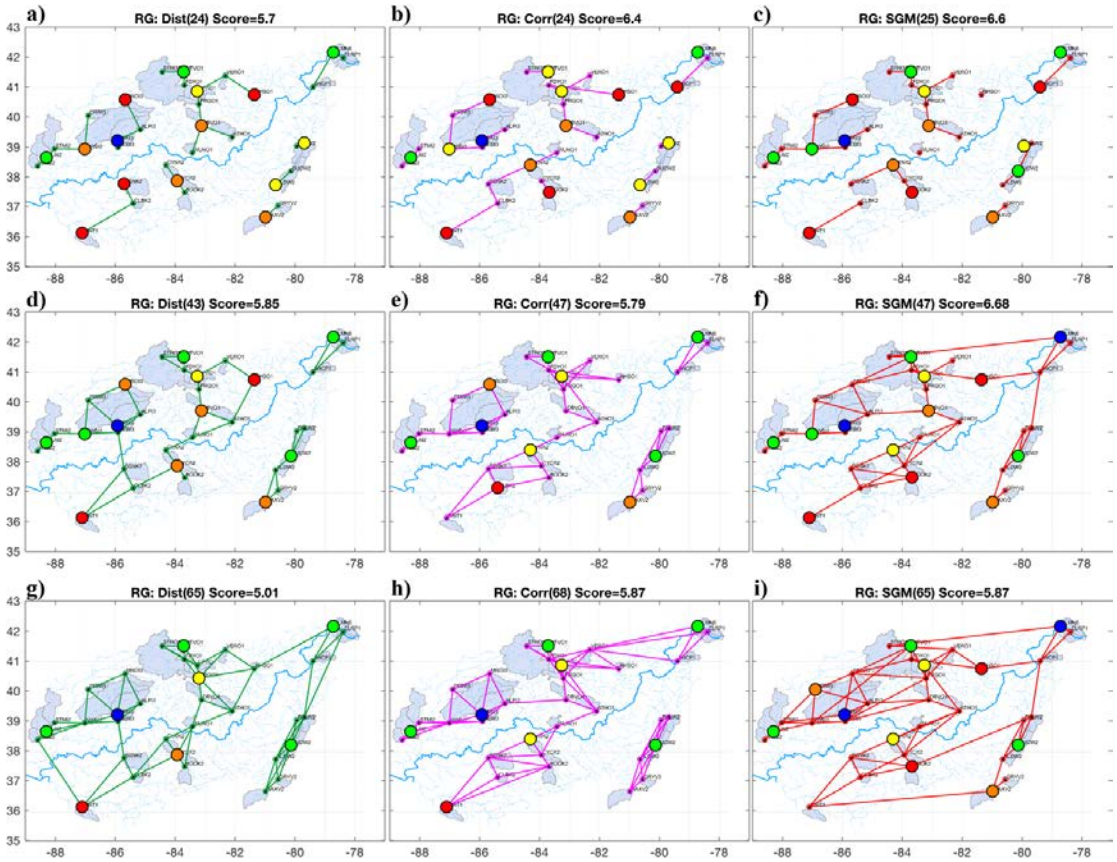

Figure 3.3 - Comparison of the observed and inferred daily streamflow time series in the test set (records between 1971 and 1980), for removable gauges estimated by the Removal of Gauges (RG) algorithm using the graphs shown on Figure 3.2 as input. Least distance (Dist) on the left, the maximum correlation (Corr) in the middle and the Selection of Graphical Model algorithm (SGM) on the right, for donor sites of 1 (on the top), 2 (in the middle row) and 3 (at the bottom), respectively. Note that for the SDM case, the number of donor sites are not fixed but automatically determined. The target sites chosen by the RG algorithm are highlighted in blue for Nash Sutcliffe efficiency (NSE) greater than or equal to 0.9, in green for NSE between 0.8 and 0.9, in yellow for NSE between 0.7 and 0.6, in orange for NSE between 0.6 and 0.7, and in red for NSE < 0.6. The graph score is the sum of the NSE for the subset of the inferred target sites with NSE > 0.7. The meaning of each plot is: (a) Least distance graph of Dist(24) for a single donor site with 24 edges and a score of 5.7. (b) Maximum correlation graph of Corr(24) for a single donor site with 24 edges and a score of 6.4. (c) SGM graph of SGM(25) with 25 edges whose sparsity is similar to the single donor case of graphs Dist(24) and Corr(24). It has a score of 6.6. (d) Least distance graph of Dist(43) for two donor sites with 43 edges and a score of 5.85. (e) Maximum correlation graph of Corr(47) for two donor sites with 47 edges and a score of 5.79. (f) SGM graph of SGM(47) with 47 edges and a score of 6.68. (g) Least distance graph of Dist(65) for three donor sites with 65 edges and a score of 5.01. (h) Maximum correlation graph of Corr(68) for three donor sites with 68 edges and a score of 5.87. (i) SGM graph of SGM(65) , with 65 edges and a score of 5.87.

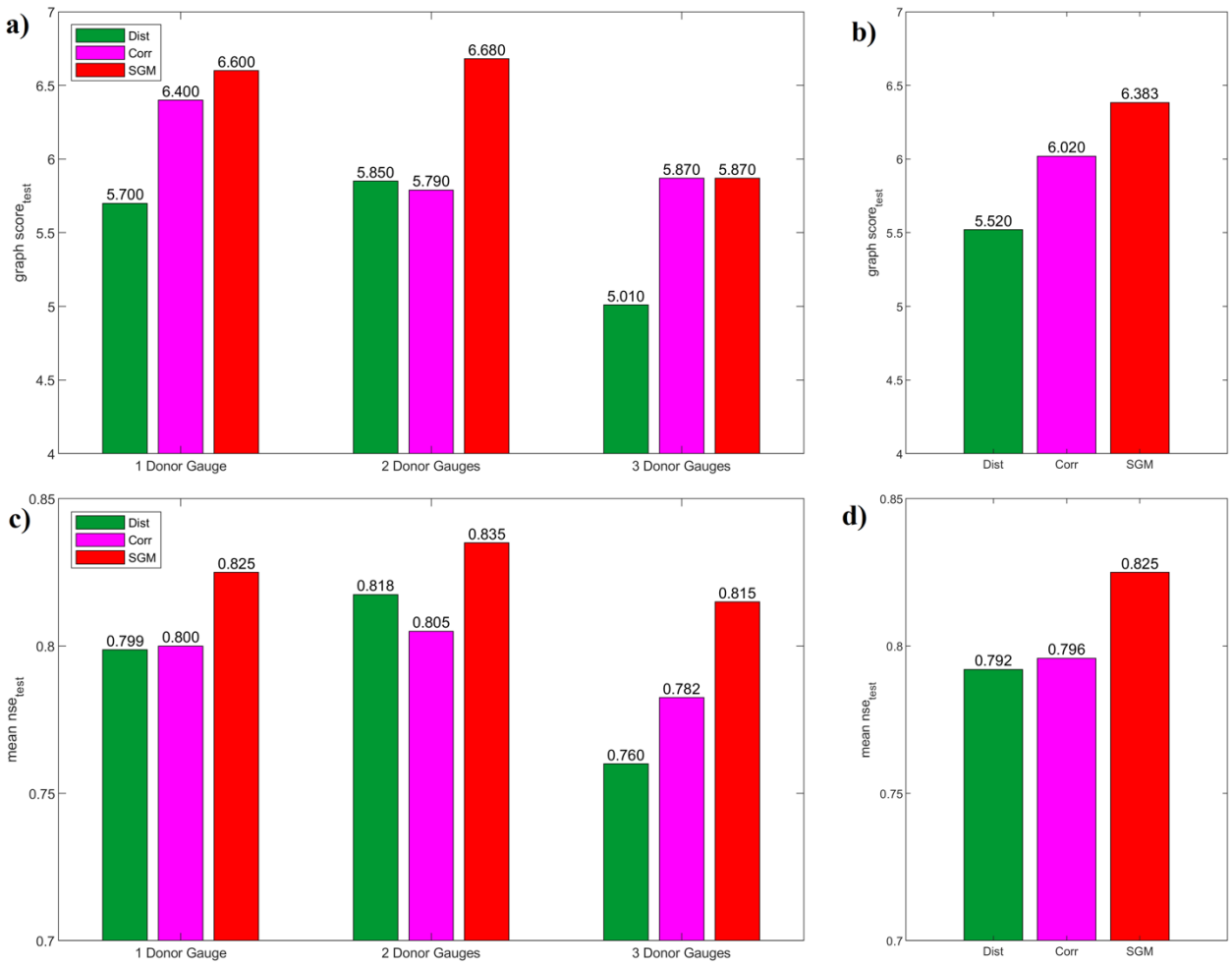

Figure 3.4 - Comparison of the inference accuracy on the removable gauges with the RG algorithm applied to graphs corresponding to the SGM, the least distance criterion (Dist) and the maximum correlation (Corr) methods. The meaning of each plot is: (a) Graph score (Equation (3.41)) for each of the individual graphs for the testing data set. (b) The mean graph score of the 1, 2 and 3 donors for each method based on values shown in (a). (c) Mean NSE of the 8 removable gauges with the highest NSE. (d) The mean NAE of the 1, 2 and 3 donors for each method based on values shown in (c).

In general, correlation-based approach (Corr) is more accurate than the distance-based approach (Dist). But the former requires more and better data to stablish the correlations. From the correlation perspective, the SGM method is more similar to the Corr approach than to the Dist approach which only depends on the geographical location of the sites. However, there is a fundamental difference between the SGM method and the Corr method. The widely used Corr method uses the marginal correlation to determine the edges between the sites. As a consequence, sites that have a decent correlation with other sites will end up with a relatively large number of edges associated with them. But some of these edges are redundant. On the other hand, the new SGM method takes advantage of the conditional correlation condition between sites as oppose to the marginal correlation used by the Corr approach. Therefore, the SGM method reduces the amount of redundant edges between sites and only connects a subset of these sites. In addition, the SGM method uses the precision matrix instead of the correlation matrix which makes it easier to extract the dependence structure among the sites within the entire network. These good characteristics associated with the SGM method, in turn, depict a simpler and more accurate dependence structure of the underlying gauge network for the study region. In practice, this simpler and better gauge network increases the number of sites that can be inferred without a significant loss in accuracy. That is the graph from the SGM method can distribute the "correlated flow" in a more efficient way so that when a site becomes a target, the neighbors for that site become donors. If the donors have high correlations, all of these sites are unavailable for removal because they are needed for the inference of the target site. Our results in Figure 3.4 have shown indeed that the accuracy of the inferred streamflow time series is improved and that the number of potentially removable sites is also increased compared to the Corr approach.

One clear example of the difference between the marginal (Corr) and the conditional (SGM) correlation methods is given by the relationship identified between the sites ALPI3, BAKI3, NWBI3 and SERI3 shown in Figure 3.3 (e) and (f). BAKI3 with a catchment area of 4421 Km^2 is a sub-basin of SERI3 with a catchment area of 6063 Km^2 along the main channel. Therefore, the catchment area of BAKI3 accounts for 73% of the catchment area of SERI3 and the correlation between them is the highest among the sites considered in the study area. The edge between them is present in all of the 9 graphs shown in Figure 3.3. The sites BAKI3 and SERI3 are also highly correlated to the sites ALPI3 and NWBI3. Figure 3.3 (e) shows the graph for Corr(47), with 5 edges: NWBI3-BAKI3, NWBI3-SERI3, BAKI3-SERI3, ALPI3-BAKI3 and ALPI3-SERI3. Figure 3.3 (f) shows the graph for SGM(47) with only 3 edges which are the same as shown in Corr(47), but having the following two edges, NWBI3-BAKI3 and ALPI3-SERI3, dropped. It is safe for SGM(47) to drop these two edges as NWBI3 is conditionally independent to BAKI3 given SERI3, and ALPI3 is conditionally independent of SERI3 given BAKI3. Figure 3.4 (a) and (c) show that the graph with the best trade-off, among the 9 graphs shown in Figure 3.3, between model complexity and accuracy is SGM(47).

Figure 3.5 shows the detailed comparison between the observed and inferred daily streamflow time series based on the testing set for the eight streamflow gauges with the highest NSE, when SGM(47), shown in Figure 3.3 (f), is chosen as the underlying graphical model.

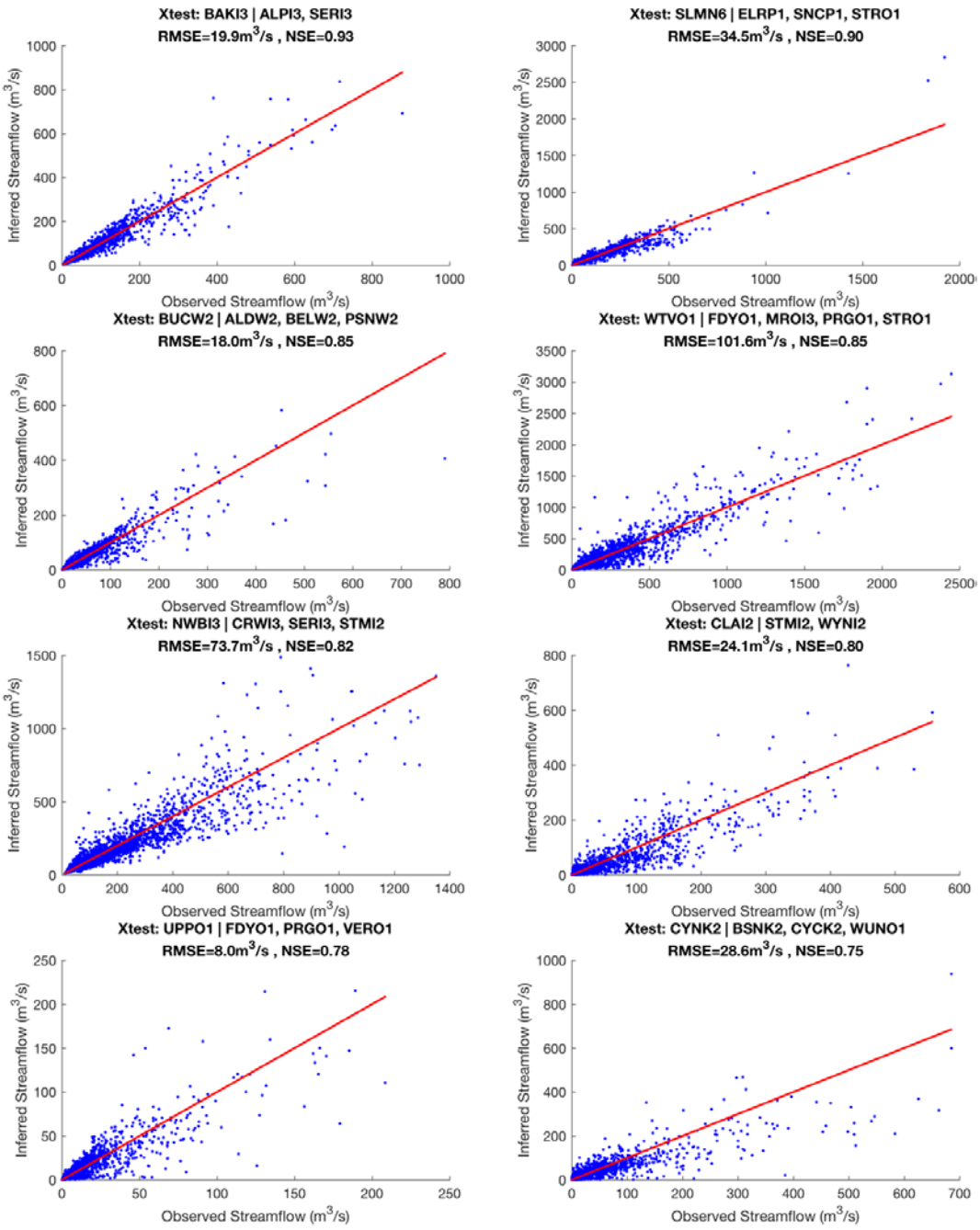

Figure 3.5 - Scatter plots between the observed and inferred daily streamflow time series over the test period of 1971-1980 (i.e., test data set). Each plot represents one of the eight gauges with the highest NSE values among the removable gauges shown in Figure 3.3. The RG algorithm is used in combination with the SGM(47) graph to identify the gauges to be removed. The MLR with Equations (3.34) and (3.35) is used to infer the daily streamflow shown in the plots. The root mean squared error (RMSE) and the NSE are shown for each gauge over the inferred period of 10 years. At the top of each plot, the name of the removed gauge is indicated on the left side of the divide line "|", and the names of gauges used to infer the streamflow of the removed gauge are indicated on the right side of the divide line "|".

For each of the 34 gauges, their corresponding watersheds were delineated using a Geographical Information System (GIS) to facilitate our understanding of the identified connections and isolated gauges based on the SGM method. Figure 3.6 (a) shows the elevation (NED: National Elevation Dataset), (b) slope (derived from elevation data), (c) soil type (Hybrid STATSGO/FAO Soil Texture) and (d) land cover (MRLC: Multi-Resolution Land Characteristics Consortium) along with the selected non-dominated graph SGM(25) obtained with a GIS tool and the corresponding cited data sets.

Using SGM(25), two sites are isolated. NHSO1 and WUNO1. Isolated sites should be maintained as much as possible to avoid loss of important regional information. Less sparse graphs, such as SGM(47) and SGM(65), can still have some marginal benefit from having some edges to those sites. NHSO1 has a significantly different land use comparing to other watersheds in the study region. For NHSO1, more than 50% of its drainage area is developed while others have less than 20% as developed. Thus, the hydrological response of this watershed to precipitation events is very different from other watersheds. In the case of WUNO1, its isolation in SGM(25) appears to be related to a combination of its geographic location, different land use from its neighboring watersheds, and its proximity to the main channel of the Ohio River. This last factor seems to be a natural separator of it. There are no edges crossing the Ohio River on the selected sparse graph, SMG(25) with 25 edges, shown in Figure 3.2 (c).

In general, the factors that impact the connections (i.e., conditional correlations) between gauges are complex and it is the integrated effect (e.g., the streamflow in this case) that determines the (conditional) correlations between the gauges. The first-order factors that contribute to the generation of streamflow in the study area seem to be the elevation, the slope and the catchment area. There is a relatively high correlation between the specific discharge (i.e. streamflow divided

by the catchment area) and the elevation (0.79), and between the specific discharge and the slope (0.76). The land cover also plays an important role, as the edges in SGM(25) are usually between sites with the same land cover class as shown in Figure 3.6 (d).

Results here have demonstrated again that it can be difficult to just use relatively simple and explicit functions to relate streamflow to different factors such as land cover, slope, soil type, drainage size in identifying their connections for complicated situations like this study case. Such a point has also been illustrated in the literature (e.g., [108]). On the other hand, these factors can sometimes help us understand why certain links exist while others do not. For example, the land cover types, elevation, and slopes appear to play more important roles than the soil type in this study region. It is worth pointing out that gauges are sometimes connected even if the correlations between them are not very high. They are connected simply because there are no other available gauges nearby with acceptably higher (conditional) correlations. In summary, the chosen graph SGM(25) does not have any edges crossing the Ohio River; there exist two gauges isolated from the rest, those gauges are geographically far from other gauges and one of them has a significantly different land use category distribution with more than 50% of its area being developed. Most of the area of the Ohio River basin belongs to the same soil type category and therefore, the soil type does not appear to contribute to the identification of the hydrologic similarity between sub-basins in this study case. On the other hand, most of the edges on the selected underlying graph SGM(25) are between watersheds with the same land use category. These results suggest that in the Ohio River basin, the land use is an important factor for the hydrologic similarity among the sub-basins.

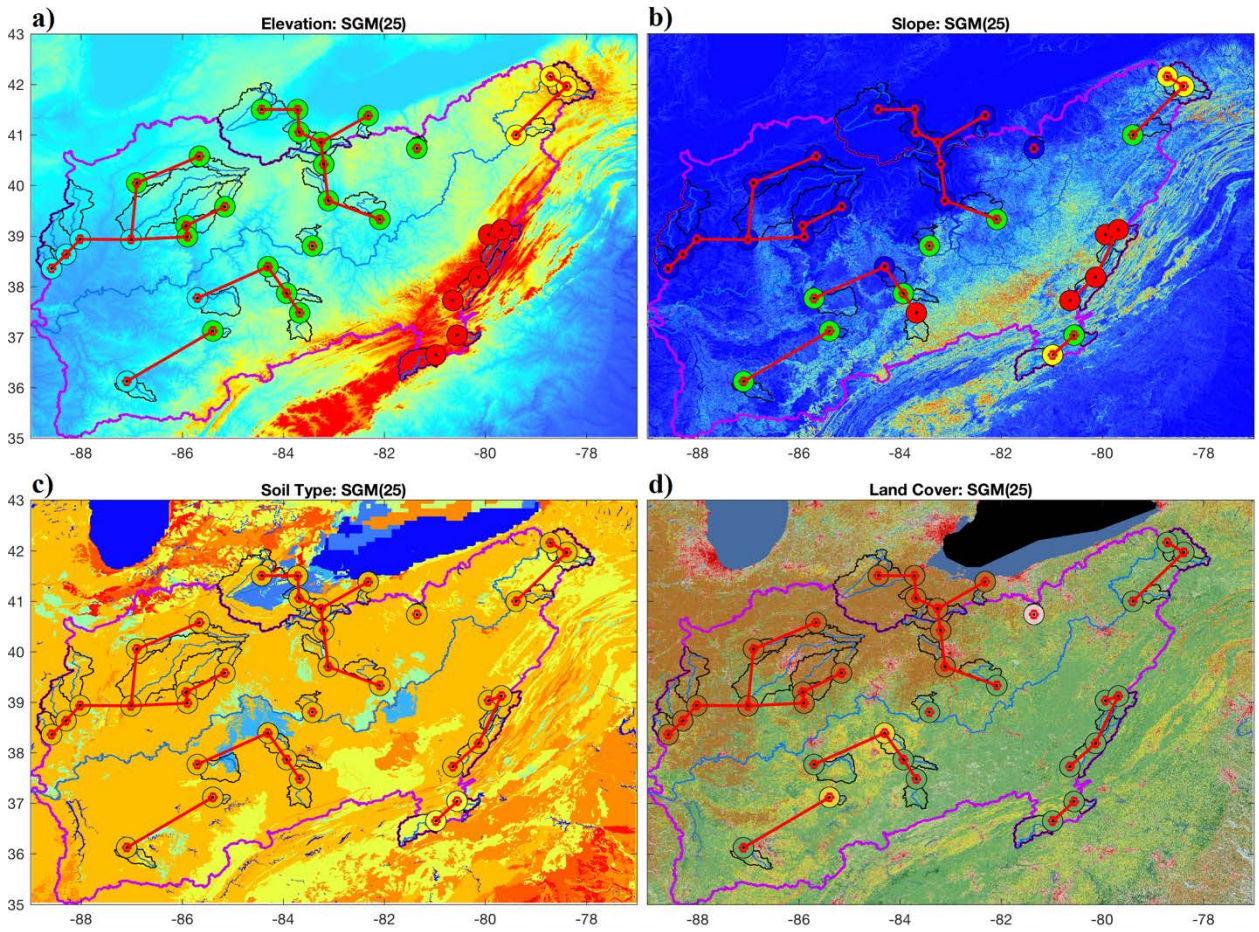

Figure 3.6 - Spatial distributions of the elevation, slope, soil type, and land cover over the study region with graph SGM(25)

(a) Elevation map showing a cluster of 4 categories indicated by a filled circle of cyan, green, yellow and red respectively. These different colors represent, respectively, "very low", "low", "high" and "very high" elevations based on the mean elevation of their corresponding watersheds. (b) Slope map showing a cluster of 4 categories indicated by a filled circle of blue, green, yellow and red respectively. The colors represent, respectively, "very low", "low", "high" and "very high" slope based on the mean slope of their corresponding watersheds. (c) Soil type map showing a cluster of 2 categories indicated by a filled circle of dark yellow and light yellow, respectively. These two colors represent, respectively, the "silt loam" and "loam" soil types. (d) Land cover map showing a cluster of 4 categories indicated by a filled circle of pink, green, yellow and brown, respectively. These different colors represent, respectively, the "developed, open space", "deciduous forest", "pasture/hay" and "cultivated crops" land cover types

3.6 Conclusions

In this study, we proposed a novel method, *Selection of Graphical Model* (SGM) algorithm, to select the set of multiple donor gauges for each inactive gauge (i.e., target gauge) to extend/infer its daily streamflow time series. This method generates a series of graphs that represent the set of potential donors for each site. The graphs generated by the SGM algorithm allow more accurate estimation of daily streamflow time series than other commonly used approaches based on the distance between sites (Dist) and the marginal correlation (Corr) between the streamflow time series. The main idea of our new method is to take advantage of the conditional independence structure encoded by an undirected graphical model known as Gaussian Graphical model, represented by the precision (i.e., covariance inverse) matrix. The SGM method selects multiple donor gauges by imposing sparsity to the precision matrix via the Graphical Lasso algorithm. The two parameters, the L1 norm regularization and a truncation threshold, in the SGM algorithm are determined by a multi-objective optimization procedure that minimizes both, the number of edges of the underlying graph and the error in the validation set to achieve a balance between the sparsity and connectivity/complexity for each graph. The resulting graphs from the non-dominated solution encode the set of donor gauges that are then used for the inference of the daily streamflow for each target gauge. We have illustrated in this study that for the gauge network composing of 34 daily streamflow gauges in the Ohio River basin, the graph with 47 edges selected based on our SGM algorithm has a good trade-off/balance between network sparsity and the estimation error. With our RG algorithm, a set of gauges can be removed from the hydrometric network with the least loss of information. In this study (e.g. Figure 3.3 and Figure 3.4), we have demonstrated that 8 out of 34 (25%) gauges can potentially be removed (NSE >= 0.75), and that from them, a group of 6 (18%) gauges can be inferred with relatively high accuracy (NSE >= 0.8). In addition, due to

the multi-objective nature of our proposed SGM algorithm, multiple graphs with different sparsity levels, can be identified for inferring daily streamflow that outperform the commonly used methods such as the least distance (Dist) approach and the maximum correlation (Corr) approach (see Figure 3.1 (b), Figure 3.3, and Figure 3.4). Depending on the number of gauges needed for removal, a balance between the inference accuracy and the gauge removal numbers can be achieved. In general, the sparser the graphs are, the more gauges one can remove. On the other hand, our study also demonstrates that the complete graph (i.e., with 561 edges) is not included in the set of non-dominated solutions, indicating that having more donor gauges does not necessarily achieve optimum results due to significantly more noises (inconsistency) introduced by the data and the inclusion of redundant edges. Therefore, not only can a suitable sparse graph achieve better inferring results through finding the most essential correlations, but also it is more practical because it requires a small but most relevant number of donor gauges in inferring the streamflow for the inactive gauges and a fewer observations to establish the relationship through the data training process. Furthermore, a graph with a fewer edges can reduce overfitting. Our method has two limitations. First, it requires a historical record of 2 to 5 years to characterize the relationships between the target and donor gauges. Second, the probability distribution of the daily streamflow should be approximated well by a log-normal distribution so that the log-transformed variable distributes normally. This second limitation, however, can be easily overcome through a common distribution transformation method if the log-normal assumption does not hold. In this study, the inference stage was performed with an ordinary least squares MLR approach due to its simplicity, although other approaches can also be used once a set of donor gauges is identified.

The most computationally expensive part of our new method is the SGM algorithm, as it relies on calling the Graphical Lasso method multiple times to find the optimal combination of the

regularization and truncation parameters. However, this complexity can be abstracted by calling efficient routines such as the glasso Matlab package (glasso) and also the GLASSOFAST [148] package which is a faster and more recent implementation of the Graphical Lasso algorithm. In addition, in this work, we performed a thoroughly search for the regularization parameter between 0 and 1, but it was determined that the best range was between 0.01 and 0.1, and that using just 10 values was almost as good as using 30 values in between. Also, we performed an almost exhaustive search for the truncation parameter to go from a very sparse graph with only 10 edges to a full graph with 561 edges (for a graph of 34 nodes), but we demonstrated that sparse graphs, under 65 edges, achieve better overall results, allowing the accurate inference of multiple sites with relatively high accuracy (NSE >= 0.8).

In this work, only contemporaneous daily streamflow records are considered. The methods explained here can be adapted to include lagged records for a finite set of days. However, for the sake of simplicity such approach was not followed. Related work [134], [136] found only marginal improvements when considering streamflow travel times into geostatistical analysis.

4.0 Estimation of Soil Type and Related Soil Parameters for Land Surface Models based on Soil Moisture Observations

The content in this chapter is based on a manuscript draft to be submitted to a hydrological journal under the title of "Estimation of Soil Type and Related Soil Parameters for Land Surface Models based on Soil Moisture Observations" by German A. Villalba, Xu Liang, and Yao Liang.

4.1 Introduction

Accurate estimation of soil moisture content (SMC) is very important in hydrological studies. For example, SMC, typically estimated by a land surface model, plays an important role in hydrologic forecasts. However, estimation of SMC is a challenging problem because there are a large number of factors affecting it, such as uncertainties associated with the model structure, forcing data, initial states and model parameters. Model parameters generally include those related to soil properties (called soil parameters hereafter), vegetation properties, and model structures. Therefore, there can be a large number of combinations of the values of these model parameters that can lead to the same or similar model simulated SMC given other conditions the same, making it very difficult to determine the appropriate model parameter values through model calibration technique. On top of it, there are also errors in observed data. To make things worse, there are many places where the observations are very scarce or even non-existent, which makes the calibration approach even harder to be applied. For situations where calibration is possible, model parameters can be estimated through the calibration process.

There are different techniques of automatic calibration used to calibrate the model parameters that have been developed by numerous researchers in the last decades [150]–[153]. Most of those methods rely on statistics of the time series of model simulated SMC and the observed SMC to perform Bayesian inference. Those methods have proved to be very useful in providing proper SMC estimates but the resulting set of model parameter values may not be necessarily physically meaningful, especially their combinations are not consistent with their physical meanings. This is the well-known equifinality problem. Indeed, obtaining consistent and physically meaningful parameters [154] is still a challenging task. To mitigate the problem, one widely acceptable approach is to reduce the number of model parameters to be calibrated. The soil parameters are typically considered as the ones that can be first removed from the calibration list (e.g., soil parameters included in the Noah model) since they are at least related to the different soil types. But the challenge is that information of the soil type map is typically obtained from compiled data sets such as the Hybrid STATSGO/FAO (30-second for CONUS /5-minute elsewhere) Soil Texture (top soil) dataset and it involves large uncertainties or errors due to limited measurements or ground surveys. Therefore, their corresponding soil parameters (e.g., Noah model) usually have to be calibrated due to the large uncertainties involved. The main purpose of this work is thus to explore a possibility of developing a physically based new method (inverse modeling) to more adequately estimate the spatial distribution of the soil types (i.e., the soil type map) based on observed soil moisture data and also to estimate a consistent set of soil parameters associated with these identified soil types.

In this chapter, we describe a novel approach to achieve this goal based on the assumption that the soil can be modeled according to the United States Department of Agriculture (USDA) soil texture triangle and therefore any soil can be described by a bi-dimensional model with the

sand and clay content as the two independent variables. The soil parameters associated with a specific soil type can then be obtained by performing a set of pedotransfer functions given information of the percentage sand and clay content for each soil type. The Noah land surface model [1], [155]–[158] (Noah LSM) was chosen to illustrate the new method and its relevant process in estimating the relevant soil parameters. In addition, the VIC model [2] was chosen to facilitate generating hypothetical observations to partially account for the effects of model structure uncertainties on the new method due to the significant model structure difference between VIC and Noah. A site in the state of Oklahoma was selected and all the forcing data were gathered. A series of experiments were performed to assess the efficiency of the new method. The test is based on a reference soil type map with which a time series of SMC is generated. This SMC is called “ground truth” and is used as an observed SMC time series. Our test is to see if the new method presented in this chapter is able to recover the reference soil type map using the generated SMC time series (i.e., ground truth). Once the soil type map is inferred, the relevant soil parameters can be then derived based on a set of pedotransfer functions presented in this chapter. Our results are encouraging as the new method is able to infer the soil types that are close to the original soil type map in most cases based on the observed SMC (i.e., generated ground truth).

4.2 New Methods

This section describes a new approach to identify soil types and then to estimate relevant soil parameters for each identified soil type to improve the simulation accuracy of soil moisture content using land surface models. The estimated soil parameters are intended to be physically-based, as opposed to randomly generated. The Noah LSM was selected as an example to illustrate

the idea, but the idea would work for other physically-based LSMs in a similar way as well. The new method can be summarized as a series of six steps: (1) Based on a few assumed potentially possible soil types/textures (called candidate soil types) for each modeling cell (as one does not know which soil type is more appropriate at this point), using pedotranfer functions to generate corresponding soil parameter data-sets. (2) Using the soil parameters from (1) to concurrently run a LSM over a study area to simulate the SMC time series. Thus, for each modeling cell, there are a few simulated SMC time series, each corresponding to a soil type candidate. (3) Calculating correlations between the SMC time-series obtained from (2) and the observed SMC time series for each modeling cell. Candidate soil type with the highest correlation is then selected as the proper soil type and its associated soil parameter values are selected as the proper parameter values for the individual modeling cell under consideration. Repeat this step for each model cell until the soil types of all of the modeling cells in the study area identified. (4) Computing the mean of the observed SMC time-series for each modeling cell. (5) Applying a clustering algorithm known as *K-means* to group the individual modeling cells based on the mean of observed SMC from (4) in such a way that the modeling cells with similar observed mean SMC are assigned to the same cluster. (6) Matching each of the resulting clusters into a dominant soil type with which a majority of the individual cells within the cluster are associated. For the individual cells whose original soil types identified in step (3) are not consistent with the dominant soil types, their original soil types are then adjusted to the dominant soil types and their corresponding soil parameter values are thus adjusted as well. These six steps are explained with details in sub-sections 4.2.1 , 4.2.2 and 4.3, respectively.

4.2.1 Estimation of Soil Parameter Data-Sets

This sub-section describes our approach to estimate consistent sets of soil parameters required for the simulation of soil moisture content in most of the LSMs using the Noah model as an example. Sub-section 4.2.1.1 provides context of some of the most relevant soil properties and the underlying motivation for our soil generation approach. Sub-section 4.2.1.2 describes a common approach to set soil parameters based on the soil type only. Sub-sections 4.2.1.3 and 4.2.1.4 show two approaches that are useful to generate soil parameters based on the percentage of sand and the percentage of clay, as opposed to just the soil type.

4.2.1.1 Background on Soil Parameter Estimation

A common representation of the soil-water relationship in most LSMs, are the Clapp and Hornberger equations [159] represented by Equation (4.1) and Equation (4.2).

$$K = K_s \cdot \left(\frac{\theta}{\theta_s}\right)^{2\beta+3} \tag{4.1}$$

$$\psi = \psi_s \cdot \left(\frac{\theta}{\theta_s}\right)^{-\beta} \tag{4.2}$$

where K is the hydraulic conductivity, K_s is the hydraulic conductivity at saturation, θ_s is the soil moisture content at saturation (total porosity), β is a curve fitting parameter and θ is the soil moisture content (volumetric water content). The matric potential ψ is a function of the saturated matric potential ψ_s and the soil moisture content θ.

4.2.1.2 Soil Type Lookup Table Approach to Estimate Soil Parameters

A common approach to specify/assign soil values to the soil parameters for LSMs is by retrieving a soil type map for the study area, and then use a lookup table to get a predefined set of soil values for each of the corresponding soil parameters related to a specific soil type. This approach is useful and simple. However, the parameter lookup table approach only provides sufficient accuracy to the case where the soil parameters happen to be at their "averaged values" for each soil type [154]. Thus, it is not an adequate method in determining the values of the soil parameters if their values are quite different from the averaged values [160]. Therefore, the associated soil parameter values so obtained might lead to a poor simulation of SMC, even if the initial soil type is correctly identified. Besides, the initial soil type can be wrong as well. Under such a situation, it becomes even more complicated as to what should be the adequate soil type and its related soil parameter values representing each modeling cell of the study area that can lead to a proper simulation of SMC.

4.2.1.3 Global Approach to Estimate Soil Parameters

This sub-section shows the development of a set of pedotransfer functions, hereinafter referred to as the "Global approach". The purpose is to describe a suitable method to estimate continuous values of soil parameters relevant to LSMs. This method is based on existing literature and it is the basis for a new approach explained in sub-section 4.2.1.4. Usually good estimates or direct measurements of the main soil properties are not available and therefore we need to rely on indirect methods to estimate them. Such functions that map the data we have into the data we need are sometimes called pedotransfer functions. These functions are usually based on empirical relationships that are only valid for the conditions where they were developed. This work was inspired by the relationships found by Cosby et al [161] where 1448 soil samples were taken

throughout several states of the United States of America. These samples came from the work of [162] and [163] and were analyzed in the laboratory to determine their soil types according to the grain sizes and particle distributions. The samples were then grouped according to their soil types. The mean and standard deviation for each soil type were computed. Finally, a set of linear regressions were performed to find relationships between the sand and clay content vs related soil parameter values. For more detailed information readers are referred to Cosby et al [161]. The same relationships were re-created in this work. However, we have changed the units to the International System of Units (SI) to match the units typically used by LSMs. Therefore, the specific values of the regression coefficients for unitless parameters are the same but are different otherwise to the values reported by Cosby et al. [161].

Cosby et al [161] found that the fitting parameter β from Equation (4.1) and Equation (4.2) is related to the percentage of clay (%clay) by a linear relationship. In a similar way, the saturated soil moisture content θ_s, the logarithm of the suction at saturation ψ_s, and the logarithm of the hydraulic conductivity at saturation K_s, are related to the percentage of sand (%sand) as shown on equations (4.3), (4.4), (4.5), and (4.6).

$$\beta = a_0 + a_1 \cdot \%clay, \quad a_0 = 2.9107 \quad a_1 = 0.1599 \tag{4.3}$$

$$\theta_s = b_0 + b_1 \cdot \%sand, \quad b_0 = 0.4889 \quad b_1 = -0.001259 \tag{4.4}$$

$$log_{10}(\psi_s) = c_0 + c_1 \cdot \%sand, \quad c_0 = -0.1179 \quad c_1 = -0.01317 \tag{4.5}$$

$$log_{10}(K_s) = d_0 + d_1 \cdot \%sand, \quad d_0 = -6.036 \quad d_1 = 0.1531 \tag{4.6}$$

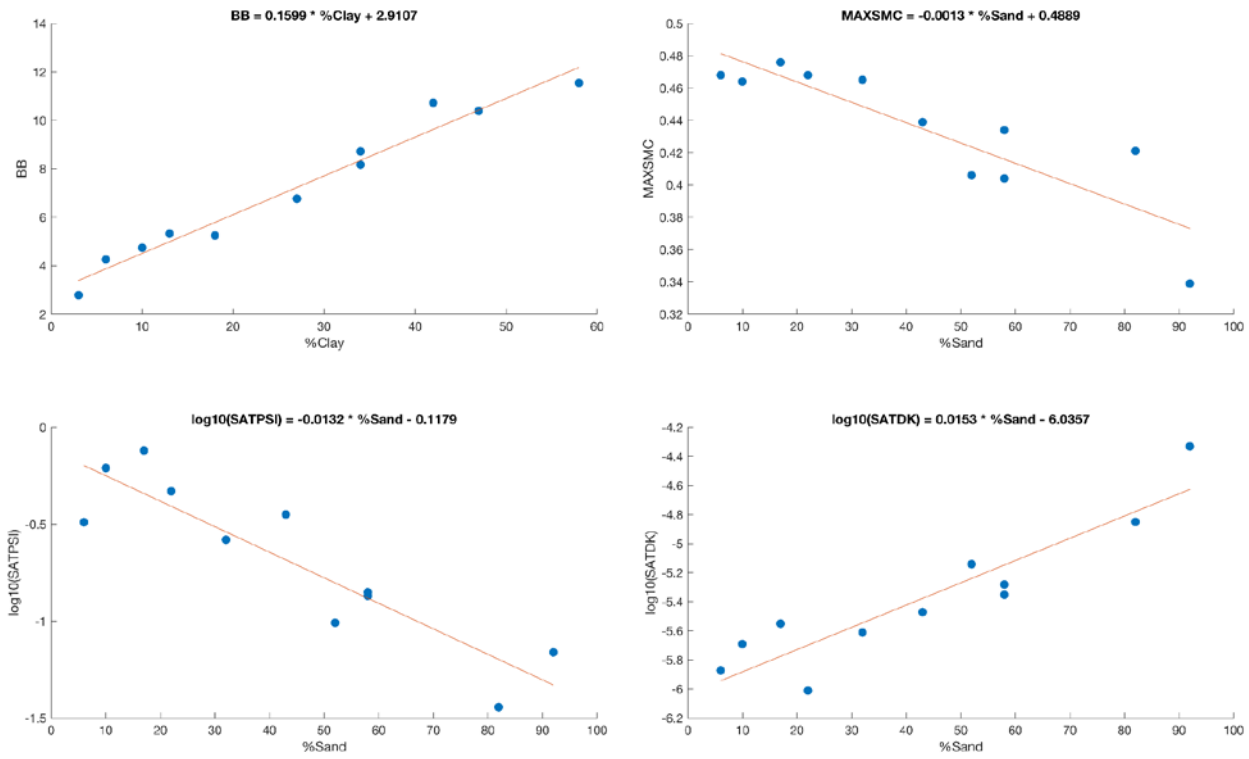

Figure 4.1. Regression equations to estimate the soil type related parameters using the sand and clay percentages as predictors: (a) Fitting parameter β as function of %clay; (b) Saturation soil moisture content (porosity) θ_s as a function of %sand.; (c) Logarithm of the suction at saturation ψ_s as a function of %sand; and (d) Logarithm of the saturated hydraulic conductivity K_s as a function of %sand.

Chen and Dudhia [157] used the four parameters β, θ_s, ψ_s and K_s from Cosby et al. [161] and defined equations to estimate two additional soil properties, the reference soil moisture content (field capacity) θ_{ref}, and the soil moisture content at wilting point, θ_w, for each soil type. The corresponding equations are shown in equation (4.7) and equation (4.8) below.

$$\theta_{ref} = \theta_s \left[\frac{1}{3} + \frac{2}{3} \left(\frac{5.79 * 10^{-9}}{K_s} \right)^{\frac{1}{2\beta+3}} \right] \quad \textbf{(4.7)}$$

$$\theta_w = \frac{\theta_s}{2} \left(\frac{200}{\psi_s} \right)^{-\frac{1}{\beta}} \quad \textbf{(4.8)}$$

The saturation soil diffusivity W_s, as defined in the Noah-LSM source code [164], is represented by equation (4.9):

$$W_s = \frac{\beta * \psi_s * K_s}{\theta_s} \quad \textbf{(4.9)}$$

In this work, we have assumed the dry soil moisture threshold θ_{dry}, to be equal to the wilting point soil moisture θ_w, as defined in equation (4.10).

$$\theta_{dry} = \theta_w \quad \textbf{(4.10)}$$

Finally, the soil quartz content, QTZ, as defined in Peters-Lidard et al [165], is computed by a linear regression expressed in equation (4.11) below using the percentage of sand (%sand) as a predictor:

$$QTZ = e_0 + e_1 \cdot \%sand, \quad e_0 = 0.03782 \quad e_1 = 0.009521 \quad \textbf{(4.11)}$$

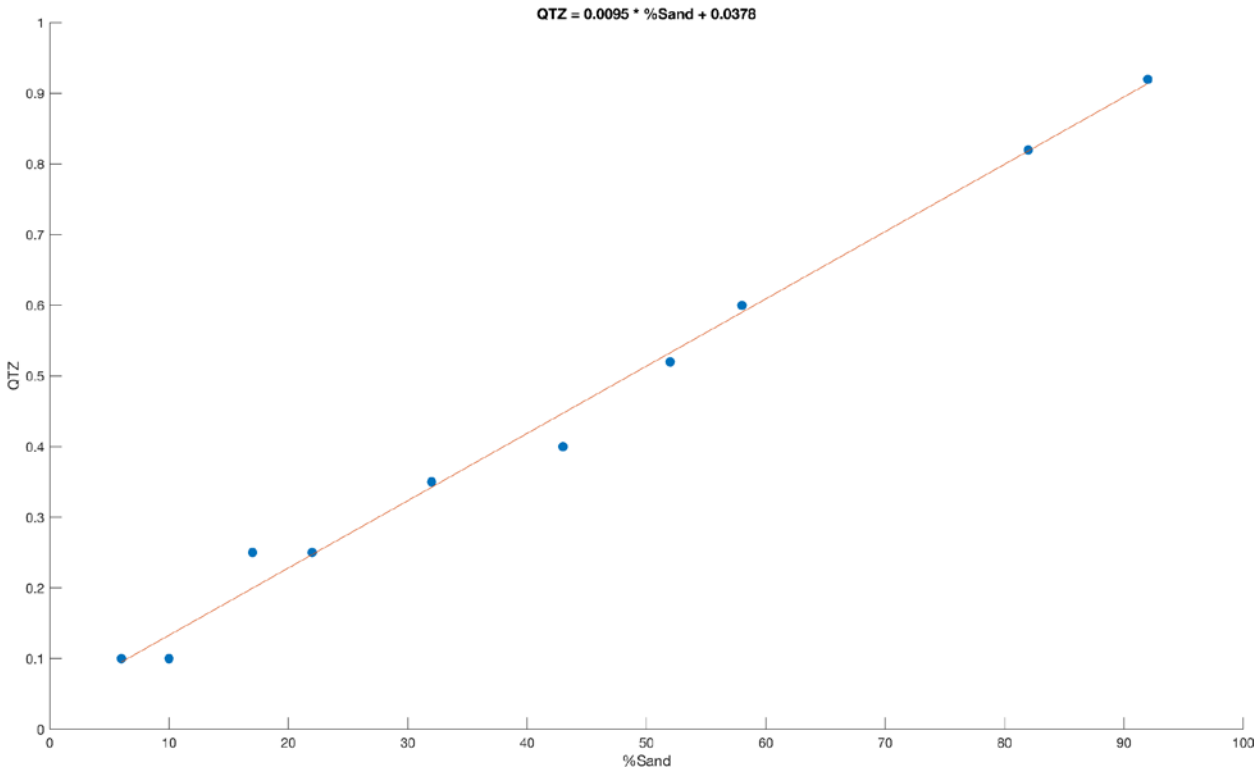

Figure 4.2. Regression for the Quartz Content, QTZ

Equations (4.3) to (4.11) describe some of the most important and commonly used soil parameters related to a soil type in LSMs as a function of the soil texture (i.e. %sand and/or %clay). Therefore, it is possible to assign values to the soil type related soil parameters for a given soil type based on information of the %sand and/or %clay that is close to or at the centroid of each given soil type as shown in Table 4.1 below.

Table 4.1 – Estimated mean soil parameters or properties for each given soil type class. The %sand and %clay given in the table represent the centroid location in the US Departement of Agriculture (USDA) soil map for each soil type and they are highlighted in blue. The soil properties of β, θ_s, ψ_s and K_s (highlighted in green) were adapted from Cosby et al. [161]**, except for the "Silt" soil type class which was not available in Cosby et al.** [161] **and is assumed here to tske the same values as the "Silty Loam" soil type class. The parameters θ_{ref} and θ_w (highlighted in yellow) were estimated using equations (4.7) and (4.8) from Chen & Dudhia** [157]**. The properties of W_s and θ_{dry} (highlighted in orange) were estimated using equations (4.9) and (4.10). The property QTZ, was taken from Peters-Lidard et al.** [165]**.**

index	name	%Sand	%Clay	[-] β	[m^3/m^3] θs	[m] Ψs	[m/s] Ks	[m^3/m^3] θref	[m^3/m^3] θw	[m^2/s] Ws	[m^3/m^3] θdry	[-] QTZ
1	Sand	92	3	2.79	0.339	0.069	4.66E-005	0.192	0.010	2.65E-005	0.010	0.92
2	Loamy sand	82	6	4.26	0.421	0.036	1.41E-005	0.283	0.028	5.14E-006	0.028	0.82
3	Sandy loam	58	10	4.74	0.434	0.141	5.23E-006	0.312	0.047	8.05E-006	0.047	0.60
4	Silty loam	17	13	5.33	0.476	0.759	2.81E-006	0.360	0.084	2.39E-005	0.084	0.25
5	Silt*	17	13	5.33	0.476	0.759	2.81E-006	0.360	0.084	2.39E-005	0.084	0.10
6	Loam	43	18	5.25	0.439	0.355	3.38E-006	0.329	0.066	1.43E-005	0.066	0.40
7	Sandy clay loam	58	27	6.77	0.404	0.135	4.45E-006	0.315	0.069	1.01E-005	0.069	0.60
8	Silty clay loam	10	34	8.72	0.464	0.617	2.04E-006	0.387	0.120	2.37E-005	0.120	0.10
9	Clay loam	32	34	8.17	0.465	0.263	2.45E-006	0.382	0.103	1.13E-005	0.103	0.35
10	Sandy clay	52	42	10.73	0.406	0.098	7.22E-006	0.338	0.100	1.87E-005	0.100	0.52
11	Silty clay	6	47	10.39	0.468	0.324	1.34E-006	0.404	0.126	9.64E-006	0.126	0.10
12	Light clay	22	58	11.55	0.468	0.468	9.74E-007	0.412	0.138	1.12E-005	0.138	0.25

4.2.1.4 Local Approach to Generate Soil Parameters

The previous sub-section described the "Global approach" used to estimate soil parameter values based on soil texture. However, that method has a limitation, it is based on a single predictor, either the %sand or the %clay for most of the soil parameters. One possible way to overcome this limitation is to use a multiple linear regression model. But Cosby et al. [161] reported that the second predictor is not very important in improving the accuracy of the estimated soil parameter values. This sub-section describes a new approach, hereinafter referred to as the "Local approach", to estimate the soil parameter values using information from the soil type. One

important consideration is that the centroid for each of the soil types was associated to the mean values from the multiple soil samples. Therefore, it contains valuable information we want to use in obtaining the soil property estimates. The “Local approach” is an extension of the “Global approach”, where instead of having a single equation to estimate a given soil property based on either the %sand or the %clay for most of them, a set of related equations is defined based on the soil type. The purpose is to better capture the variability of the soil properties as a function of both of %sand and %clay. Thus, instead of using a multiple linear regression approach, a single predictor regression model is used. However, instead of using the global intercept of the regression equations (4.3) to (4.11), a new local intercept is computed in such a way that the regression equation evaluated at the centroid of both, the percentage of sand (%sand) and the percentage of clay (%clay) passes through the centroid of the soil related parameter for each soil type class. That is, the new method uses the same formulations as those expressed in equations (4.3) to (4.11), but the local intercept value changes according to the soil type, i.e., each soil type has a different equation to estimate each soil property, where the only difference is the value of the regression intercept. The slope is kept the same as those used in equations (4.3) to (4.11). This approach is useful for the estimations of the soil parameters. The process to estimate soil-related parameters using this novel approach can be summarized by the following four steps:

(1) Estimate the soil type class given the percentage of sand (%sand) and the percentage of clay (%clay).

(2) Retrieve the soil type related parameter at the centroid of the soil type class from (1) (e.g. using Table 4.1).

(3) Compute the local intercept for each soil related parameter by subtracting the slope of the corresponding regression equation (from (4.3) to (4.11)) multiplied by either the %sand at the

centroid or the %clay at the centroid (%clay for β, %sand for the remaining soil related parameters), from the value of the soil related parameter at the centroid for the soil type class from (1).

(4) Estimate the soil related parameters using the formulations given by equations (4.3) to (4.11), but replacing the global intercept by the local intercept from (3).

4.2.2 Soil Moisture Content Simulation through Land Surface Models

The SMC was simulated using a tool, hereinafter referred to as "Noah-2D-SP", that was developed by our research group. Noah-2D-SP is a software that wraps the Noah LSM, which is a "1-D column model" [166], allowing the calculation of a grid of cells (2D) while specifying multiple custom soil parameter (SP) data-sets. The required inputs are: forcing data, model parameters and initial conditions. Some of the most relevant inputs for LSMs are summarized in Table 4.2. These inputs may contain large uncertainties.

Table 4.2 – Summary of the relevant inputs for LSMs used for the simulations

The initial conditions need to specify the "State" variables. The "Forcing" data is required for each day.

Type	LSM Variable	Description	Units
State	ALBEDO	Surface Albedo	(unitless)
State	CH	Exchange coefficient for heat and moisture	m*s^-1
State	CM	Exchange coefficient for momentum	m*s^-1
State	CMC	Canopy moisture content	m
State	SH2O	Volumetric fraction of liquid soil moisture content	m^3/m^3
State	SMC	Volumetric fraction of total soil moisture content	m^3/m^3
State	SNEQV	Liquid water-equivalent snow depth	m
State	SNOWH	Actual snow depth	m
State	STC	Soil temperature	K
State	T1	Effective skin temperature	K
Forcing	LWDN	Longwave downward radiation	W/m^2
Forcing	PRCP	Precipitation rate	Kg/(m^2*s)
Forcing	Q2	Mixing ratio at height ZLVL above ground	Kg/Kg
Forcing	SFCPRS	Pressure at height ZLVL above ground	Pa
Forcing	SFCSPD	Wind speed at height ZLVL above ground	m/s
Forcing	SFCTMP	Air temperature	K
Forcing	SOLDN	Solar downward radiation	W/m^2
Soil Parameters	β	Curve fitting parameter for the Clapp & Hornberger equation	(unitless)
Soil Parameters	θ_{dry}	Dry soil moisture threshold at which direct evaporation from top soil layer ends	m^3/m^3
Soil Parameters	θ_s	Soil moisture content at saturation (total porosity)	m^3/m^3
Soil Parameters	θ_{ref}	Reference soil moisture (field capacity), where transpiration begins to stress	m^3/m^3
Soil Parameters	ψ_s	Saturation suction	m
Soil Parameters	K_s	Hydraulic conductivity at saturation	m/s
Soil Parameters	W_s	Saturation soil diffusivity	m^2/s
Soil Parameters	θ_w	Wilting point soil moisture	m^3/m^3
Soil Parameters	QTZ	Soil quartz content	(unitless)

The forcing data include longwave downward radiation, precipitation rate, mixing ratio, pressure, wind speed, air temperature and solar downward radiation. In this study, the soil parameters are estimated using the "Local approach" from sub-section 4.2.1.4. and a soil type texture map for the study area. Finally, the estimation of the initial conditions is performed by assigning an initial value for each of the state variables. In particular, the initial conditions for the SMC related variables, follows the typical initialization procedure [166]. That is, each cell is initialized with a value of 30% of the porosity (θ_s from Table 4.1) for the cell's specified soil type. Then the Noah model is run repeatedly over the same year using the corresponding forcing data, until the state variables such as the SMC reach equilibrium. More details on how to set the initial conditions are provided in sub-section 4.3.3 . Sub-section 4.3.2 shows how to set the soil parameters.

4.3 Estimating Soil Type and Soil Parameters based on SMC Observations

This section illustrates our new approach to estimate optimized soil types and their related soil parameters based on given SMC observations. Because of limited availability of observed soil data and their corresponding SMC time series, we apply our new method to a set of hypothetically generated SMC time series based on given soil types and their associated soil parameters for a study area. In this way, we can test if our new method is able to identify the given soil types with their associated soil parameters based on the given SMC time series. The selected study area is bounded by Latitude 34.75° to 35° N and Longitude 97.9375° to 98.1875° W which is in the state of Oklahoma within the Little Washita watershed as described in [167], [168]. This area has been the subject of numerous studies due to its intensive field campaigns of collecting

various data including soil moisture. For this work a resolution of 1/128 degrees (approximately 780 m) is used which corresponds to 32 by 32 modeling cells.

4.3.1 Forcing Data

The forcing data from 1 January/1997 to 31 December/1997 was retrieved from the 0.125 Degree Hourly Primary Forcing Data for NLDAS-2 [169] which include hourly precipitation total [kg/m^2], longwave radiation [w/m^2], shortwave radiation [w/m^2], surface pressure [pa], 2-m above ground specific humidity [kg/kg], 2-m above ground temperature [k], 10-m above ground zonal wind speed [m/s], and 10-m above ground meridional wind speed [m/s]. All of the data were aggregated to a daily time-step and resampled to match the 32 by 32 modeling grids for the study area.

4.3.2 Soil Parameter Estimations and Generation of Ground truth SMC

4.3.2.1 Soil Parameter Estimations

The soil parameter values associated with the Noah model (i.e., %sand, %clay, BB, MAXSMC, SATPSI, SATDK, REFSMC, WLTSMC, SATDW, DRYSMC and QTZ) for each modeling cell in the study area are estimated using the Local approach described in sub-section 4.2.1.4 based on a prescribed soil type map. In this example, the soil type map obtained from the Hybrid STATSGO/FAO (30-second for CONUS /5-minute elsewhere) Soil Texture (top soil) dataset is assumed to be the true soil type map and is shown in Figure 4.3a for the study area.

4.3.2.2 Generation of Ground truth SMC Time Series

The observed SMC time series called ground truth is generated due to the limitation of available soil moisture time series. This ground truth SMC is generated by a combination of adding random noises and a utilization of two very different models, the VIC and Noah models, to make the generated ground truth SMC more complex than that by any single model alone. Since the model structures between VIC and Noah are very different, VIC is used to partially account for the uncertainties associated with model structures, because in reality, no model can simulate soil moisture time series identical to the observed soil moisture in the real world. There is always inconsistency between observed and model simulated SMC which is partially contributed by the model structures as no model at present can fully describe all the processes involved in the real world. The detailed steps of generating the ground truth SMC time series are summarized as follows:

(1) Add randomly generated gaussian noises with zero mean and a dynamically estimated standard deviation to the precipitation data time series to increase the uncertainties between the generated ground truth SMC and the SMC simulated by the Noah model. The standard deviation of the generated noise is computed one time-step at a time. For each time-step, the standard deviation of the added gaussian noise is estimated as 30% of the standard deviation of the spatial precipitation distribution.

(2) Estimate the soil parameters related to the given soil type map (i.e., Figure 4.3a). For the Noah model, the soil parameters (i.e., BB, MAXSMC, SATPSI, SATDK, REFSMC, WLTSMC, SATDW, DRYSMC and QTZ) were estimated based on the local approach described in subsection 4.2.1.4. Values of percentage of sand (%sand) and percentage of clay (%clay) used in the pedotransfer functions in subsection 4.2.1.4 are randomly generated

within a range suitable for the given soil type so that there are some variations in the soil parameter values for the same given soil type. In other words, the generated soil parameters are not constrained to a single value for the soil type class. Thus, two cells with the same soil type are likely to have similar but different soil parameter values. For the VIC model, the soil parameters were kept similar to the values as their counterparts in the Noah model when feasible, otherwise, VIC's default values were used.

(3) Estimate other model parameters. For parameters other than the soil parameters involved in both the Noah and VIC models, such as vegetation and model structure related parameters, their respective default values were used based on the given vegetation and other available information for the study area.

(4) Initialize each model's state variables in the same way as they are normally done. For example, for the Noah model, the initial conditions for soil moisture follow the initialization procedure of [166]. That is, each cell is initialized with a value of 30% of the MAXSMC from Table 4.1 according to the given soil type (i.e., Figure 4.3a) for that cell. The required initial conditions for the Noah LSM are given in Table 4.2 and their values were set as: ALBEDO = 0.699, CH = 0.024, CM = 0.04, CMC = 0.0, SNEQV = 0.0, SNOWH = 0.0, STC_0 = STC_1 = STC_2 = STC_3 = 273, T1 = 273. For the VIC model, their soil moisture values at different soil layers were set at equivalent values as the ones set for the Noah model. The other state variables in VIC were initialized in a similar way as those in the Noah model.

(5) Run the Noah and VIC models, respectively, over the one year forcing data from January 1st, 1997 to December 31st, 1997 repeatedly until both reached their equilibrium states. In this study, both models reached their equilibrium states after running for about 5 cycles of the one year forcing data from 1997 with their own initialization processes.

(6) Create the ground truth SMC time series. The one-year SMC time series at equilibrium from both the Noah and VIC model simulations based on the above steps are combined to form one SMC time series. Each of the models contribute 50% to the combined time series. To make the problem closer to the real world and the testing case scenario more challenging and complex for the proposed new method, additional randomly generated noises were added to the combined SMC time series. The added noise is gaussian with zero mean and with a standard deviation of 30% the standard deviation of the porosity reported by Cosby et al [161]. Thus, the standard deviation of the added noise is 30% of 0.061, that is 0.0183. This final one-year SMC time series with noises added is used as the assumed ground truth and served as the observed SMC time series corresponding to the given soil type map shown in Figure 4.3a. The ground truth SMC time series and the observed SMC time series are used interchangeably throughout this chapter.

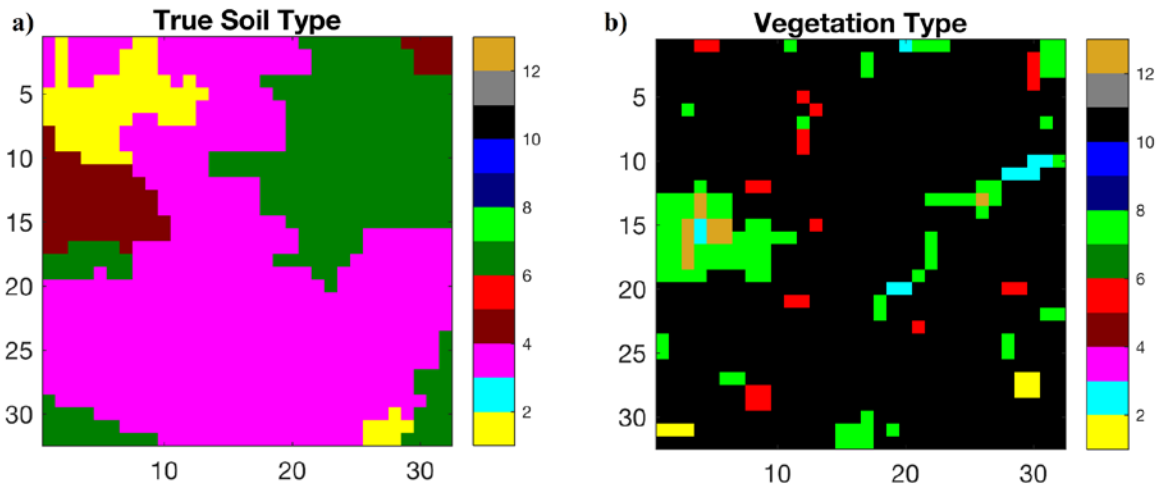

Figure 4.3. Spatial distributions of the soil types and vegetation types for the study area with 32 by 32 modeling cells : (a) Assumed true soil type distribution extracted from the Hybrid STATSGO/FAO (30-second for CONUS /5-minute elsewhere) Soil Texture (top soil); and (b) True spatial distribution of the vegetation types for the study area.

4.3.3 Generation of Soil Moisture for the Noah Model

The Noah model is used to illustrate the new method. The soil parameters (i.e., BB, MAXSMC, SATPSI, SATDK, REFSMC, WLTSMC, SATDW, DRYSMC and QTZ) of the Noah model were estimated in the same way as described in step (2) under the subsection 4.3.2.1. The only difference is that in this case, the values of percentage of sand (%sand) and percentage of clay (%clay) corresponding to the centroids of a given soil type defined in the USDA Soil Map are used in the pedotransfer functions described in sub-section 4.2.1.4 for simplicity. Additional soil texture values (i.e., %sand and %clay) can be used to improve accuracy in identifying the soil types at expense of more computational resources. Since the true soil type is assumed as unknown, we can simply assume that each of the 12 soil types is a possible candidate at each modeling cell. If one has more knowledge on the potentially possible soil type candidates, a fewer candidates of the soil types can be assumed for each modeling cell. In this study, we simply assume that all 12 soil types are possible candidates for each modeling cell. Thus, the soil parameter values corresponding to each of the 12 soil types have to be estimated based on the local approach described in sub-section 4.2.1.4 and the centroid location for obtaining the %sand and %clay information. After obtaining the soil parameter values for all of the 12 different soil types for each modeling cell, the Noah model was run 12 times per modeling cell for the entire study area using the forcing data. No random noises were added to the forcing precipitation time series in these Noah model simulations. These Noah simulation runs are called “calibration” simulations hereafter.

4.3.4 Optimized Estimation of Soil Type and its Associated Parameters

For clarity, we call the process of finding the appropriate spatial distribution of soil types and their related soil parameters given observed SMC time series a calibration process. The selection of a proper soil type and its related parameters is performed by computing the determination coefficient ($\boldsymbol{R^2}$) between the observed SMC time series (i.e., ground truth) and the Noah generated SMC time series (also called members hereafter) from subsection 4.3.3 for each of the 12 soil types for each modeling cell. Then the member with the highest $\boldsymbol{R^2}$ is selected as the estimated proper soil type for the cell and its related soil parameters are selected as the soil parameters for that given cell.

4.4 Results

Figure 4.4a shows that the root zone SMC from the Noah model simulations varies over a relatively large range among the 12 different soil types in the one year under equilibrium state. Figure 4.4b shows that most of the soil types have a relatively similar temporal patterns in standard deviation with an exception of the "Loamy Sand" soil type (light blue series). Figure 4.4a and Figure 4.4c show that the "Sand" (yellow) and "Loamy Sand" soil types are significantly dryer than other soil type classes. Figure 4.4d shows that, as expected, the "Clay" soil type is the wettest among the 12 soil types. Figure 4.4 also shows that the root zone SMC is likely to be clustered into approximately four distinguishable groups. Figure 4.5 shows that there are probably four groups according to the simulated root zone SMC averaged over the year for each cell as well. One group is sand (Figure 4.5a). One group is loamy sand (Figure 4.5b). The third group is the green-

colored ones which include sandy loam (Figure 4.5c), loam (Figure 4.5f), sandy clay loam (Figure 4.5g), and sandy clay (Figure 4.5j). The fourth group is the blue-colored ones which include silty loam (Figure 4.5d), silt (Figure 4.5e), silty clay loam (Figure 4.5h), clay loam (Figure 4.5i), silty clay (Figure 4.5k), and light clay (Figure 4.5l).

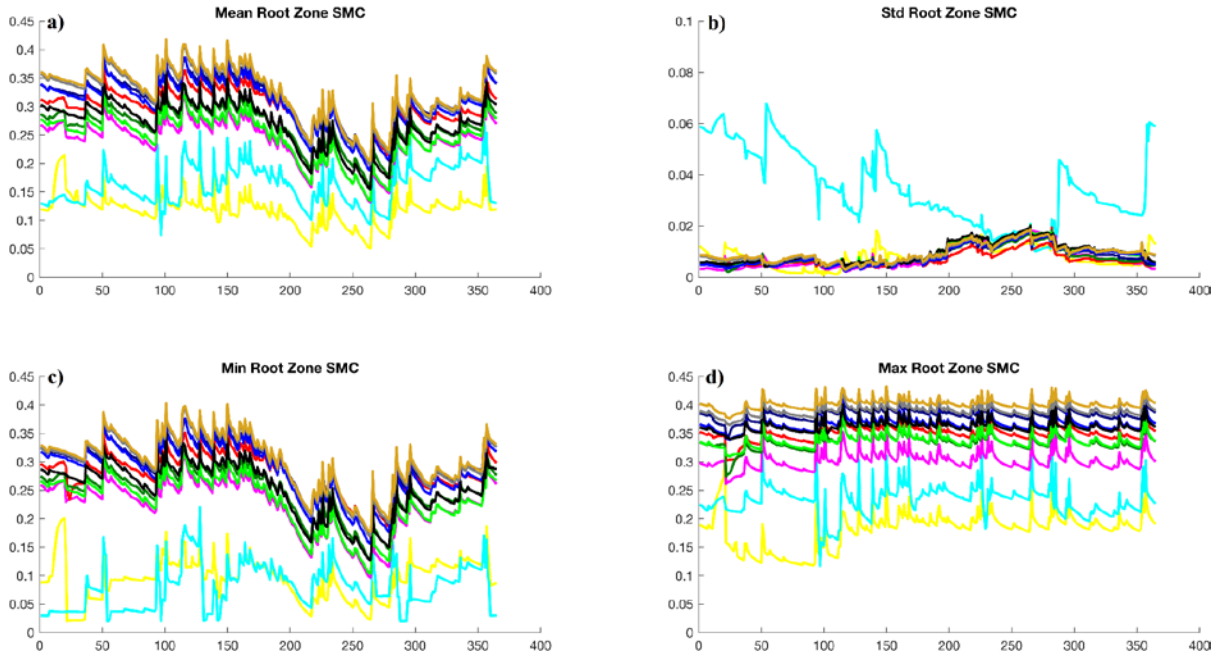

Figure 4.4. Time series of root zone SMC, spatially aggregated for each member of the calibrationsimulations, one time series for each soil type class: (a) Mean Root Zone SMC; (b) Standard Deviation of the Root Zone SMC; (c) Minimum value for the Root Zone SMC; and (d) Maximum value for the Root Zone SMC.

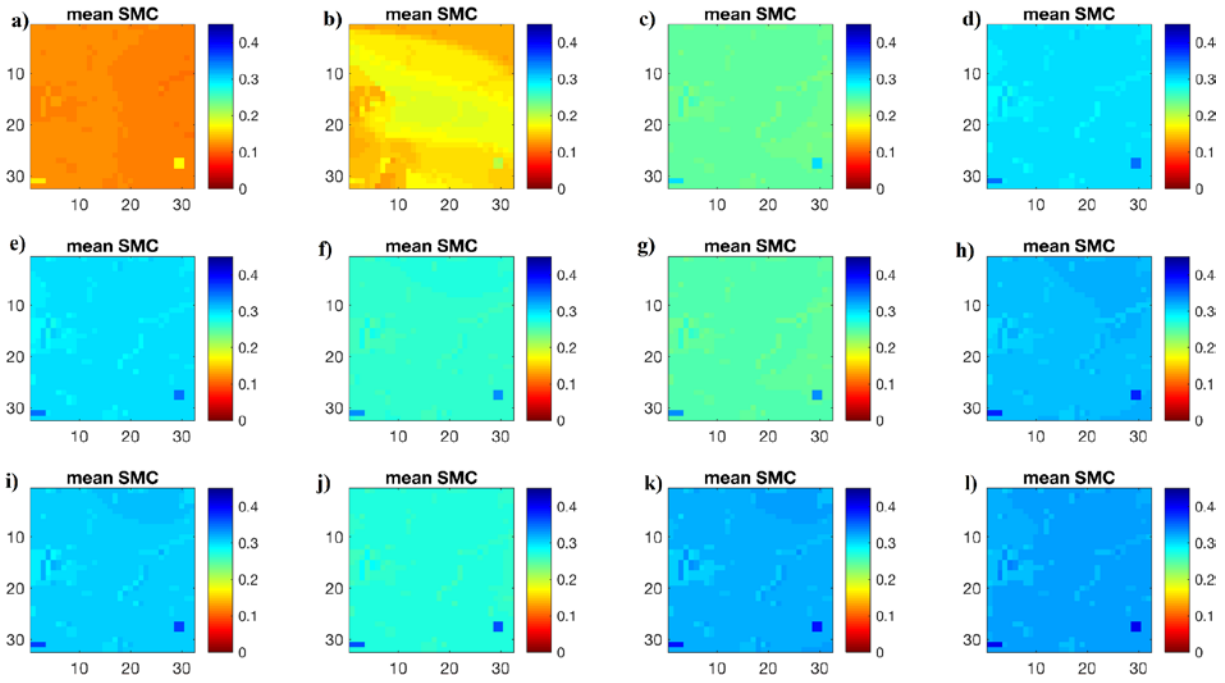

Figure 4.5. Spatial distributions of the root zone SMC averaged over the one year Noah model simulation for each of the 12 soil types: (a) 1. Sand; (b) 2. Loamy sand; (c) 3. Sandy loam; (d) 4. Silty loam; (e) 5. Silt; (f) 6. Loam; (g) 7. Sandy clay loam; (h) 8. Silty clay loam; (i) 9. Clay loam; (j) 10. Sandy clay; (k) 11. Silty clay (l) 12. Light clay.

4.4.1 Spatial Distribution of the SMC and Inference of Soil Types

Figure 4.6a shows the spatial distribution of the temporal mean of the observed SMC (i.e., ground truth). Figure 4.6b shows the mean of the K groups resulting from clustering with the K-means algorithm when K=3, using as input the spatial distribution of the temporal mean of the observed SMC shown in Figure 4.6a. Figure 4.6c shows the resulting clusters identified. The cluster with the lowest, intermediate and highest observed SMC mean is shown in red, green and blue, respectively. Figure 4.6d shows the resulting soil type distribution, estimated for each individual modeling cell, by selecting the SMC time series associated to a particular soil type class

with the highest correlation between the observed and simulated SMC time series. Figure 4.6e shows the resulting soil type distribution by clustering the observed mean SMC shown in Figure 4.6a. Figure 4.6e was derived by inputting the mean observed SMC (i.e., Figure 4.6a) to the K-means clustering algorithm with 3 groups (K=3), and using the soil types identified on each individual cell for each cluster to select the majority class. The majority class is then used to reassign the soil types for the cells that do not match the majority soil type class. Figure 4.6f shows the spatial distribution of the assumed true soil types for comparison with the groups from the clustering results of Figure 4.6e. The reason that Figure 4.6e has only three clusters is because K-means was run with only 3 groups, while Figure 4.6f has four soil types. In this case using 4 groups (K=4) for the clustering algorithm does not produce reliable patterns because two soil types, class 4 “Silty Loam” and class 6 “Loam”, have similar mean SMC and also because of the various random noises added to the precipitation and later to the combined model simulated soil moisture time series (i.e., from both Noah and VIC), and the combination of two different models. The aggregated effect is that only 3 clusters can reliably be identified from the mean SMC shown in Figure 4.6a and Figure 4.6b (i.e. red, light blue, dark blue).

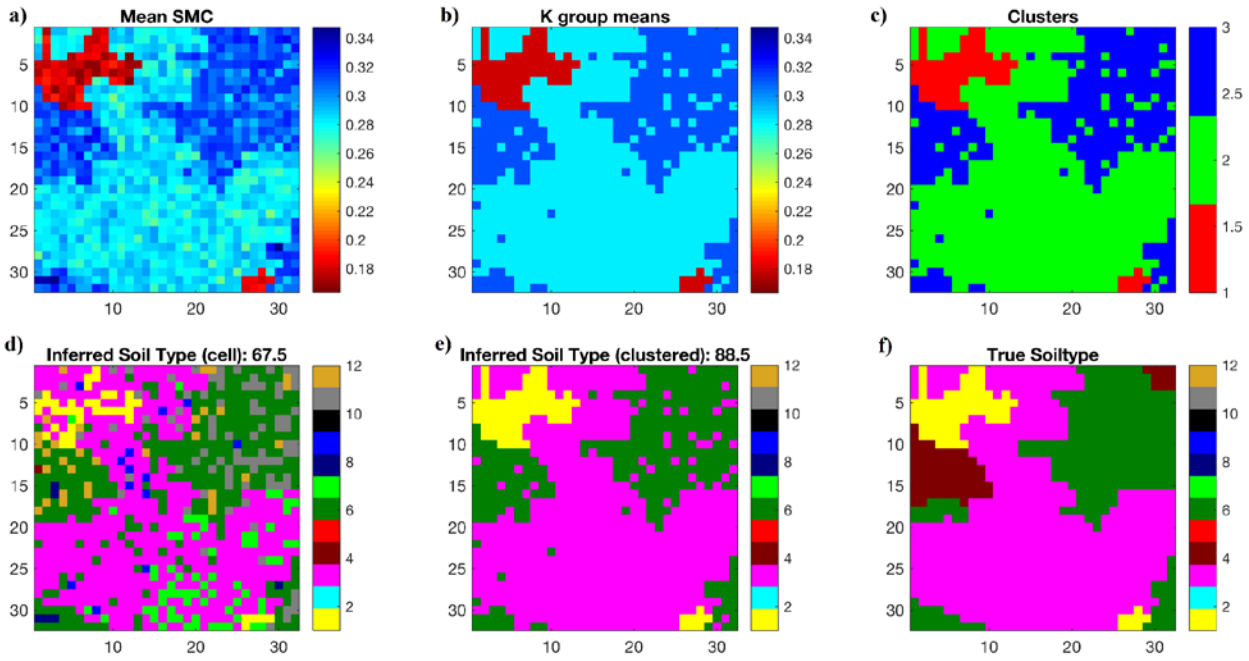

Figure 4.6. Spatial Distribution of Soil Moisture and Inference of Soil Types over the Study Area (a) temporal mean of the observed SMC (i.e., ground truth), (b) temporal mean of the K groups from the observed SMC, (c) resulting clusters based on the observed mean SMC shown in Figure 4.6a, (d) resulting soil types from comparing individual cells without conducting Kmeans clustering (%67.5 success rate), (e) final results after assigning clusters to soil types using 3 soil types (88.5% success rate) and (f) assumed spatial distribution of the true soil types.

4.4.2 Comparison of Estimated Soil Parameters

In this sub-section, we perform an assessment of the accuracy and efficiency of our new "local" approach to estimate soil related parameters from observed SMC time series. This assessment is based on a comparison with a commonly used approach for the calibration of hydrological models.

We selected a stochastic modelling approach for the sake of comparison with our new method for identifying soil types and soil parameters for LSMs. The calibration approach for this "common" method is summarized by the following steps:

(1) Select the number of members (concurrent simulations) and generations (number of cycles)

(2) Assign a valid range to each soil parameter.

(3) Scale the valid range of each soil parameter to a range between 0 and 1.

(4) Randomly initialize the normalized soil parameters (i.e. in the range 0 to 1).

(5) Scale the parameters back to its natural valid range.

(6) Run the LSMs to obtain the SMC time series.

(7) Use an error or objective function to assess if a new parameter set performs better than previous calibration simulations. In this study we used the Nash–Sutcliffe model efficiency coefficient between the observed and simulated SMC.

(8) Merge the historical soil parameters with the best performance (low error / high score, etc) according to the objective function from (7) with the most recently generated soil parameter sets.

(9) Discard the one half with the poorest performance and save/store the one half with the best performance.

(10) Generate new sets of soil parameters (in the range 0 to 1), one for each member.

(11) Repeat from step (5) until the number of generations is complete.

The commonly used method is based on an iterative approach. On the other hand, our new local approach is based on a “single shot” simulation, that is, only 1 generation is required and the number of member is given by the number of potential soil types considered, here we selected 12 soil type candidates, thus, we use 12 member for the simulations for our local approach.

Figure 4.7 shows the comparison of the results using the “common” approach vs our new “local” approach: (a), (b) and (c), shows the logarithm of the hydraulic conductivity at saturation (i.e. log10(SATDK)) for the commonly used approach (common), our new local approach (local) and the assumed ground truth (truth), respectively. The RMSE from our local approach (0.134) accounts for only 35% percent of the RMSE with the common approach (0.380). Also, the hydraulic conductivity distribution pattern is better described by our local approach. (d), (e) and (f), shows the logarithm of the saturation soil diffusivity (i.e. log10(SATDW)) for the commonly used approach (common), our new local approach (local) and the assumed ground truth (truth), respectively. The RMSE from our local approach (0.154) accounts for only 41% percent of the RMSE with the common approach (0.372). Also, the pattern of the distribution of the saturation soil diffusivity is better described by our local approach. (g), (h) and (i), shows the porosity (i.e. MAXSMC) for the commonly used approach (common), our new local approach (local) and the assumed ground truth (truth), respectively. The RMSE from our local approach (0.0264) accounts for only 63.7% percent of the RMSE with the common approach (0.0414). Also, the pattern of the distribution of the porosity is better described by our local approach. (j), (k) and (l), shows the curve fitting parameter for the Clapp & Hornberger equation (i.e. BB) for the commonly used

approach (common), our new local approach (local) and the assumed ground truth (truth), respectively. The RMSE from our local approach (1.17) accounts for only 35.5% percent of the RMSE with the common approach (3.29). Also, the pattern of the distribution of the curve fitting parameter is better described by our local approach.

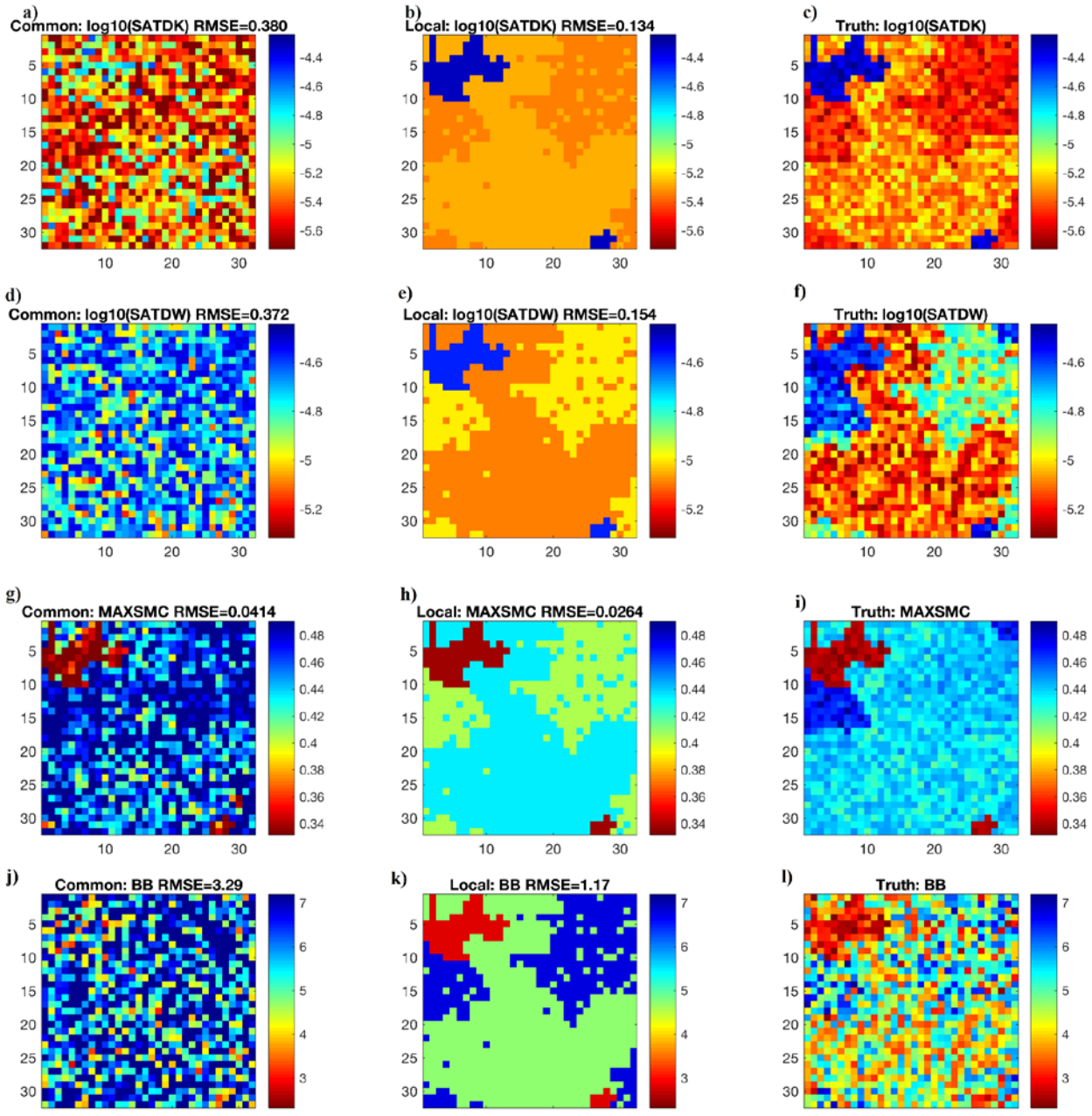

Figure 4.7. Soil related Parameters Calibration Comparison between a commonly used approach and our new clustering based local approach. Logarithmic distribution of the hydraulic conductivity at saturation for (a) the "common" approach (RMSE=0.338) (b) the "local" approach (RMSE=0.134), and (c) the ground truth, respectively. Logarithmic distribution of the saturation soil diffusivity for (d) the "common" approach (RMSE=0.372) (e) the "local" approach (RMSE=0.154), and (f) the ground truth, respectively. Distribution of the soil porosity for (g) the "common" approach (RMSE=0.0414) (h) the "local" approach (RMSE=0.0264), and (i) the ground truth, respectively. Distribution of the BB curve fitting parameter for the Clapp & Hornberger equation for (j) the "common" approach (RMSE=3.29) (k) the "local" approach (RMSE=1.17), and (l) the ground truth, respectively.

4.5 Summary and Conclusions

The experiments performed in this study suggest that it is possible to systematically retrieve meaningful and consistent soil types and soil type related parameters from the SMC observations (synthetically generated data in this case study). Our procedure can be extended to real observed soil moisture data although there are some challenges to gather and retrieve long term (e.g., one year or longer) root zone SMC time series over large spatial domains. In this work we presented a synthetically generated SMC time series (called ground truth) which consist of the model simulated SMC time series from two distinct land surface models with very different model structures (i.e. 50% from Noah simulation and 50% with VIC simulation). In addition, uncertainties (random noises) were added to the precipitation forcing data and to the combined simulated SMC time series as well. The soil parameters used for the ground truth simulations were obtained by random selection of the percentages of sand and clay within each of the corresponding given soil type class and the associated soil parameters were then computed using the pedotransfer functions described in subsection 4.2.1.4. The inferred soil type classes based on the new method had a success rate of 87.9% which is quite encouraging given the large uncertainties included in the ground truth SMC time-series to ensure a low correlation between the ground truth SMC time series and the Noah model SMC simulations for each of the 12 soil type classes. The assumed true soil type map has 4 distinct classes but only 3 types are retrieved with high confidence. The soil types between classes 4 (silty loam) and 6 (loam) are difficult to distinguish because their simulated SMC time series are similar to each other as indicated in Figure 4.5. Using the VIC and model Noah models instead of the Noah model alone to generate the ground truth (i.e., assumed observed SMC time series) is to introduce structure uncertainties as evidenced for the sand soil type, for example, in which the VIC model behaves very differently from that of the Noah model.

Adding uncertainties to the forcing data and then to the simulated SMC time series as well is an additional effort to make the final assumed observed SMC time series (used as ground truth) to be quite different from that of the Noah model simulations, making the problem of identifying the appropriate soil type classes for the study area using the assumed SMC observations more challenging. Our results show that the new method presented in this chapter is quite effective. Thus, the new method can be used to more adequately identify spatial distributions of the soil type classes for a study area which is very much needed for conducting adequate LSM simulations. For situations where the correlations are too low to distinguish one soil type from the other, it is suggested that some statistics of the SMC time series, such as the mean or median are used as the moisture index in the identification process. For example, "sand" is the driest soil type and it can be assigned to a cluster with the least mean SMC value. The procedure can be used for other soil classes if needed, although in some cases there might be a mismatch if the mean SMC is not distinct enough. For example, between soil class 4 (silty loam) and soil class 6 (loam).

In this study we presented a heuristic algorithm to match a cluster of a soil type, based on the correlation between the observed and the simulated SMC time-series. The correlation can be used to match a cluster to a soil type. In spite of the very different SMC time series created as the ground truth in this study, the new method presented in here was able to retrieve most of the soil types correctly (88.5 success rate). In addition, the general patterns of the soil-related parameters are better captured by our clustering-based local method and the RMSE are usually a fraction of the error from a commonly used approach such as the one described here for comparison purposes. Our calibration method allows automated estimation of soil types and soil-related parameters using only the observed SMC as input. This method is also more efficient as it requires only a limited

number of members, one for each soil type candidate and it only requires a single generation (i.e. single concurrent simulations).

5.0 Conclusions

This dissertation addressed three challenging issues for the overall reduction of uncertainty in hydrological modeling. First, the collection of sub-surface data at the plot-scale on real-time for environmental monitoring was performed by a long-term wireless sensor network in south-western Pennsylvania. This deployment is a cost-effective solution alternative to the more traditional use of data-loggers while allowing a large amount of environmental data collection. The lab-made sap flow sensors are significantly cheaper than industrial-grade ones, although the manufacturing process of the thermal probes and the circuitry needs to be done with care. The sap flow lab-made sensors need to be labeled and calibrated before the installation on the field. The data gathered by the WSN includes soil moisture content, water potential and soil temperature at multiple soil depths, in addition to xylem sap flow.

Second, the accuracy and availability of continuous streamflow time-series estimates was improved by a novel algorithm that uses daily streamflow data from a hydrometric network modeled as a graph where each site represents a node and the missing edges represent conditional independence assumptions between a given pair of sites given all the remaining sites. This approach takes advantage of the natural parsimony that arises from such a model, in such a way that a sparse graph is obtained by imposing sparsity to the Gaussian graphical model representation of the hydrometric network. The sparsity is achieved by selecting a representation of the Gaussian graphical model in terms of the precision (covariance inverse) matrix. This representation has the advantage that uses the conditional correlation, as oppose to the marginal correlation making it easier to interpret as the edges can be used to determine the optimal set of donor gauges for any target site. The optimal graph is found by a multi-objective optimization procedure using a

Machine Learning method, known as the Graphical Lasso to estimate a sparse precision matrix, along with a method to select an L-1 norm regularization parameter and a truncation operator. Once the optimal graph is known, a second novel greedy algorithm is introduced to estimate the gauges that can be removed from the hydrometric network with the least loss of information. This is particularly important given the budget cuts that challenge the appropriate time and space coverage of the measurements performed by agencies world-wide, such as the U.S. Geological Survey (USGS) in the United States of America. The inferred streamflow data was shown to be more accurate in terms of Nash-Suschife model efficiency and the coefficient of determination, R^2, between the observations and the inferred daily streamflow time-series, than other commonly used approached based on geographic proximity or marginal correlation. In general, the correlation approach is generally better in terms of accuracy, as compared with the geographical proximity criterion, but it requires additional data that might not be available. The conditional correlation is preferable over the marginal correlation, especially in the presence of multiple variables with high inter-correlation, as it reduces to the minimum the number of redundant predictors.

Third, a novel method for the simultaneous estimation of soil moisture content and soil-related parameters was presented based on dimensional reduction of the soil-parameter space by mapping it to a two-dimensional plane represented by the USGA soil-texture triangle. That is, any point on the triangle can be represented as a linear combination of sand and clay, those coordinates determine the soil texture class and also, by using the described pedotransfer functions, they can be mapped to a specific soil parameter value. This parameterization was actively used to generate synthetic simulations by running two distinct hydrologic land surface models, the Noah LSM and the VIC model. It was shown that a weighted average of the soil moisture output for those two models along with perturbations to the forcing data and observed soil moisture can create more

realistic soil moisture time series observations than any single model alone. This procedure was used to match the simulated SMC from the Noah LSM for each soil type. Then a clustering algorithm known as K-means was used to group the cells for the mean or median from the observed SMC. Finally, we proposed two heuristic algorithms to match the generated cluster to soil types, one based on correlation for individual cells and the other based on a moisture index. Depending on the correlation between the simulated and observed data, one heuristic outperforms the other. In general, it was observed that before calibration, the correlation between the simulated and the observed SMC is relatively low, therefore the moisture index heuristic is chosen unless the correlations are relatively high. Once the soil type is chosen, the last step is to generate additional samples within the selected soil types and choose the ones that improve the soil moisture time series for individual cells, but they are constrained to a particular soil type previously inferred.

Bibliography

[1] M. B. Ek *et al.*, "Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model," *J. Geophys. Res.*, vol. 108, no. D22, p. 8851, Nov. 2003.

[2] X. Liang, D. P. Lettenmaier, E. F. Wood, and S. J. Burges, "A simple hydrologically based model of land surface water and energy fluxes for general circulation models," *J. Geophys. Res.*, vol. 99, no. D7, p. 14415, Jul. 1994.

[3] W. L. Barnes, T. S. Pagano, and V. V. Salomonson, "Prelaunch characteristics of the Moderate Resolution Imaging Spectroradiometer (MODIS) on EOS-AM1," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 4, pp. 1088–1100, Jul. 1998.

[4] A. Y. Hou, G. Skofronick-Jackson, C. D. Kummerow, and J. M. Shepherd, "Global precipitation measurement," in *Precipitation: Advances in Measurement, Estimation and Prediction*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 131–169.

[5] D. Entekhabi *et al.*, "The Soil Moisture Active Passive (SMAP) Mission," *Proc. IEEE*, vol. 98, no. 5, pp. 704–716, May 2010.

[6] G. Villalba *et al.*, "A networked sensor system for the analysis of plot-scale hydrology," *Sensors*, vol. 17, no. 3, p. 636, 2017.

[7] M. Cumbers, *Achieving Sustainable Freshwater Systems: A Web of Connections*, vol. 11, no. 2. 2016.

[8] K. Beven, "Changing ideas in hydrology — The case of physically-based models," *J. Hydrol.*, vol. 105, no. 1, pp. 157–172, 1989.

[9] R. Oren, N. Phillips, G. Katul, B. E. Ewers, and D. E. Pataki, "Scaling xylem sap flux and soil water balance and calculating variance: a method for partitioning water flux in forests," *Ann. des Sci. For.*, vol. 55, no. 1–2, pp. 191–216, 1998.

[10] C. R. Ford, R. M. Hubbard, B. D. Kloeppel, and J. M. Vose, "A comparison of sap flux-based evapotranspiration estimates with catchment-scale water balance," *Agric. For. Meteorol.*, vol. 145, no. 3, pp. 176–185, 2007.

[11] J. S. Famiglietti, D. Ryu, A. A. Berg, M. Rodell, and T. J. Jackson, "Field observations of soil moisture variability across scales," *Water Resour. Res.*, vol. 44, no. 1, Jan. 2008.

[12] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Comput. Networks*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[13] P. Rawat, K. D. Singh, H. Chaouchi, and J. M. Bonnin, “Wireless sensor networks: a survey on recent developments and potential synergies,” *J. Supercomput.*, vol. 68, no. 1, pp. 1–48, Apr. 2014.

[14] P. Juang *et al.*, “Energy-efficient computing for wildlife tracking,” *ACM SIGARCH Comput. Archit. News*, vol. 30, no. 5, p. 96, Dec. 2002.

[15] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, “Wireless sensor networks for habitat monitoring,” in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications - WSNA ’02*, 2002, p. 88.

[16] R. Szewczyk, J. Polastre, and A. Mainwaring, “Lessons from a sensor network expedition,” *Eur. Work.*, 2004.

[17] G. Tolle *et al.*, “A macroscope in the redwoods,” in *Proceedings of the 3rd international conference on Embedded networked sensor systems - SenSys ’05*, 2005, p. 51.

[18] M. P. Hamilton *et al.*, “New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories,” *Environ. Eng. Sci.*, vol. 24, no. 2, pp. 192–204, Mar. 2007.

[19] G. Werner-allen, M. Welsh, and J. Lees, “Monitoring Volcanic Activity with a Wireless Sensor Network,” in *Proceeedings of the Second European Workshop on Wireless Sensor Networks*, 2004, pp. 1–30.

[20] G. Werner-Allen *et al.*, “Deploying a wireless sensor network on an active volcano,” *IEEE Internet Comput.*, vol. 10, no. 2, pp. 18–25, Mar. 2006.

[21] L. Selavo *et al.*, “LUSTER: Wireless sensor network for environmental research,” in *SenSys’07 - Proceedings of the 5th ACM Conference on Embedded Networked Sensor Systems*, 2007, pp. 103–116.

[22] J. Panchard, S. Rao, P. T. V., J.-P. Hubaux, and H. S. Jamadagni, “COMMONSense Net: A Wireless Sensor Network for Resource-Poor Agriculture in the Semiarid Areas of Developing Countries,” in *Information Technologies and International Development*, 2008, vol. 4, no. 1, pp. 51–67.

[23] T. Wark *et al.*, “Transforming Agriculture through Pervasive Wireless Sensor Networks,” *IEEE Pervasive Comput.*, vol. 6, no. 2, pp. 50–57, Apr. 2007.

[24] X. Li, Y. Deng, and L. Ding, “Study on precision agriculture monitoring framework based on WSN,” in *2nd International Conference on Anti-counterfeiting, Security and Identification, ASID 2008*, 2008, pp. 182–185.

[25] M. Martinelli, L. Ioriatti, F. Viani, M. Benedetti, and A. Massa, “A WSN-based solution for precision farm purposes,” in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2009, vol. 5, p. 469.

[26] J. A. López, A.-J. Garcia-Sanchez, F. Soto, A. Iborra, F. Garcia-Sanchez, and J. Garcia-Haro, “Design and validation of a wireless sensor network architecture for precision horticulture applications,” *Precis. Agric.*, vol. 12, no. 2, pp. 280–295, Apr. 2011.

[27] F. Viani, “Experimental validation of a wireless system for the irrigation management in smart farming applications,” *Microw. Opt. Technol. Lett.*, vol. 58, no. 9, pp. 2186–2189, Sep. 2016.

[28] F. J. Ferrández-Pastor, J. M. García-Chamizo, M. Nieto-Hidalgo, J. Mora-Pascual, and J. Mora-Martínez, “Developing ubiquitous sensor network platform using internet of things: Application in precision agriculture,” *Sensors (Switzerland)*, vol. 16, no. 7, p. 1141, 2016.

[29] F. Viani, M. Bertolli, and A. Polo, “Low-Cost Wireless System for Agrochemical Dosage Reduction in Precision Farming,” *IEEE Sens. J.*, vol. 17, no. 1, pp. 5–6, 2017.

[30] J. A. López, F. Soto, P. Sánchez, A. Iborra, J. Suardiaz, and J. A. Vera, “Development of a sensor node for precision horticulture,” *Sensors*, vol. 9, no. 5, pp. 3240–3255, 2009.

[31] K. Szlavecz *et al.*, “Poster: Life Under Your Feet: A Wireless Soil Ecology Sensor Network,” in *In Proc. 3rd Workshop on Embedded Networked Sensors*, 2008.

[32] A. Suri, S. S. Iyengar, and E. Cho, “Ecoinformatics using wireless sensor networks: An overview,” *Ecol. Inform.*, vol. 1, no. 3, pp. 287–293, 2006.

[33] P. W. Rundel, E. A. Graham, M. F. Allen, J. C. Fisher, and T. C. Harmon, “Environmental sensor networks in ecological research,” *New Phytol.*, vol. 182, no. 3, pp. 589–607, May 2009.

[34] S. S. Burgess, M. L. Kranz, N. E. Turner, R. Cardell-Oliver, and T. E. Dawson, “Harnessing wireless sensor technologies to advance forest ecology and agricultural research,” *Agric. For. Meteorol.*, vol. 150, no. 1, pp. 30–37, 2010.

[35] J. Trubilowicz, K. Cai, and M. Weiler, “Viability of motes for hydrological measurement,” *Water Resour. Res.*, vol. 46, no. 4, 2010.

[36] F. Ingelrest, G. Barrenetxea, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, “SensorScope: Application-specific sensor network for environmental monitoring,” *ACM Trans. Sens. Networks*, vol. 6, no. 2, p. 17, 2010.

[37] B. Kerkez, S. D. Glaser, R. C. Bales, and M. W. Meadows, “Design and performance of a wireless sensor network for catchment-scale snow and soil moisture measurements,” *Water Resour. Res.*, vol. 48, no. 9, Sep. 2012.

[38] M. Navarro *et al.*, “Towards Long-Term Multi-Hop WSN Deployments for Environmental Monitoring: An Experimental Network Evaluation,” *J. Sens. Actuator Networks*, vol. 3, no. 4, pp. 297–330, 2014.

[39] B. Majone *et al.*, “Wireless Sensor Network Deployment for Monitoring Soil Moisture Dynamics at the Field Scale,” *Procedia Environ. Sci.*, vol. 19, pp. 426–435, 2013.

[40] D. Ryu and J. S. Famiglietti, “Characterization of footprint-scale surface soil moisture variability using Gaussian and beta distribution functions during the Southern Great Plains 1997 (SGP97) hydrology experiment,” *Water Resour. Res.*, vol. 41, no. 12, pp. 1–13, Dec. 2005.

[41] H. R. Bogena, M. Herbst, J. A. Huisman, U. Rosenbaum, A. Weuthen, and H. Vereecken, “Potential of Wireless Sensor Networks for Measuring Soil Water Content Variability,” *Vadose Zo. J.*, vol. 9, no. 4, p. 1002, 2010.

[42] TinyOS, “Tiny OS.” [Online]. Available: http://tinyos.net/.

[43] O. Gnawali, R. Fonseca, K. Jamieson, M. Kazandjieva, P. Levis, and D. Moss, “CTP: An efficient, robust, and reliable collection tree protocol for wireless sensor networks,” *ACM Trans. Sens. Networks*, vol. 10, no. 1, p. 16, 2013.

[44] M. Navarro and Y. Liang, “Efficient and Balanced Routing in Energy-Constrained Wireless Sensor Networks for Data Collection,” *Proc. 2016 Int. Conf. Embed. Wirel. Syst. Networks*, no. August, pp. 101–113, 2016.

[45] T. W. Davis, C.-M. Kuo, X. Liang, and P.-S. Yu, “Sap Flow Sensors: Construction, Quality Control and Comparison,” *Sensors*, vol. 12, no. 12, pp. 954–971, Jan. 2012.

[46] A. Granier, “Une nouvelle methode pour la measure du flux de seve brute dans le tronc des arbres.,” *Ann. des Sci. For.*, vol. 42(2), no. 2, pp. 193–200, 1985.

[47] A. Granier, “Evaluation of transpiration in a Douglas-fir stand by means of sap flow measurements,” *Tree Physiol.*, vol. 3, no. 4, pp. 309–320, Dec. 1987.

[48] D. Devices, “MPS-2 User Manual.” [Online]. Available: http://manuals.decagon.com/Retired and Discontinued/Manuals/13755_MPS-2and6_Web.pdf.

[49] T. W. Davis, X. Liang, C. Kuo, and Y. Liang, “Analysis of power characteristics for sap flow, soil moisture, and soil water potential sensors in wireless sensor networking systems,” *IEEE Sens. J.*, 2012.

[50] J. Llosa, I. Vilajosana, X. Vilajosana, N. Navarro, E. Suriñach, and J. M. Marquès, “REMOTE, a wireless sensor network based system to monitor rowing performance,” *Sensors*, vol. 9, no. 9, pp. 7069–7082, 2009.

[51] E. Y. W. Seto *et al.*, “A wireless body sensor network for the prevention and management of asthma,” in *Proceedings - 2009 IEEE International Symposium on Industrial Embedded Systems, SIES 2009*, 2009, pp. 120–123.

[52] P. Kuryloski *et al.*, "DexterNet : An Open Platform for Heterogeneous Body Sensor Networks and Its Applications Allen Yang University of California at Berkeley DexterNet : An Open Platform for Heterogeneous Body Sensor Networks," in *Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, 2009, pp. 92–97.

[53] A. Cama, F. G. Montoya, J. Gómez, J. L. De La Cruz, and F. Manzano-Agugliaro, "Integration of communication technologies in sensor networks to monitor the Amazon environment," *J. Clean. Prod.*, vol. 59, pp. 32–42, 2013.

[54] G. Strazdins, A. Elsts, K. Nesenbergs, and L. Selavo, "Wireless Sensor Network Operating System Design Rules Based on Real-World Deployment Survey," *J. Sens. Actuator Networks*, vol. 2, no. 3, pp. 509–556, 2013.

[55] M. Amjad, M. Sharif, M. K. Afzal, and S. W. Kim, "TinyOS-New Trends, Comparative Views, and Supported Sensing Applications: A Review," *IEEE Sens. J.*, vol. 16, no. 9, pp. 2865–2889, 2016.

[56] D. Moss and P. Levis, "BoX-MACs: Exploiting physical and link layer boundaries in low-power networking," *Comput. Syst. Lab. Stanford*, 2008.

[57] Sensirion, "Sensirion SHT11." [Online]. Available: https://www.sensirion.com/fileadmin/user_upload/customers/sensirion/Dokumente/0_Datasheets/Humidity/Sensirion_Humidity_Sensors_SHT1x_Datasheet.pdf.

[58] X. Zhong, M. Navarro, G. Villalba, X. Liang, and Y. Liang, "MobileDeluge: Mobile code dissemination for wireless sensor networks," in *Proceedings - 11th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2014*, 2015, pp. 363–370.

[59] Z. Xu, T. Hu, and Q. Song, "Bulk data dissemination in low power sensor networks: Present and future directions," *Sensors (Switzerland)*, vol. 17, no. 1, 2017.

[60] W. M. Cornelis, J. Ronsyn, M. Van Meirvenne, and R. Hartmann, "Evaluation of Pedotransfer Functions for Predicting the Soil Moisture Retention Curve," *Soil Sci. Soc. Am. J.*, vol. 65, no. 3, p. 638, 2001.

[61] H. Lin, W. Zhang, and H. Yu, "Hydropedology: Linking Dynamic Soil Properties with Soil Survey Data," in *Application of Soil Physics in Environmental Analyses*, Cham: Springer International Publishing, 2014, pp. 23–50.

[62] R. B. Clapp and G. M. Hornberger, "Empirical equations for some soil hydraulic properties," *Water Resour. Res.*, vol. 14, no. 4, pp. 601–604, Aug. 1978.

[63] M. T. van Genuchten, "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1," *Soil Science Society of America Journal*, vol. 44, no. 5. p. 892, 1980.

[64] W. R. N. Edwards, P. Becker, and J. Eermak, "A unified nomenclature for sap flow measurements," *Tree Physiol.*, vol. 17, no. 1, pp. 65–67, Jan. 1997.

[65] A. Granier, R. Huc, and S. T. Barigah, “Transpiration of natural rain forest and its dependence on climatic factors,” *Agric. For. Meteorol.*, vol. 78, no. 1–2, pp. 19–29, Jan. 1996.

[66] K. B. Wilson *et al.*, “A comparison of methods for determining forest evapotranspiration and its components: sap- ow, soil water budget, eddy covariance and catchment water balance,” *Agric. For. Meteorol.*, vol. 106, no. 2001, pp. 153–168, 2008.

[67] S. D. Wullschleger, P. J. Hanson, and D. E. Todd, “Transpiration from a multi-species deciduous forest as estimated by xylem sap flow techniques,” *For. Ecol. Manage.*, vol. 143, no. 1–3, pp. 205–213, 2001.

[68] D. E. Pataki and R. Oren, “Species differences in stomatal control of water loss at the canopy scale in a mature bottomland deciduous forest,” *Adv. Water Resour.*, vol. 26, no. 12, pp. 1267–1278, 2003.

[69] H. Asbjornsen, M. D. Tomer, M. Gomez-Cardenas, L. A. Brudvig, C. M. Greenan, and K. Schilling, “Tree and stand transpiration in a Midwestern bur oak savanna after elm encroachment and restoration thinning,” *For. Ecol. Manage.*, vol. 247, no. 1–3, pp. 209–219, 2007.

[70] T. Kumagai *et al.*, “Effects of tree-to-tree and radial variations on sap flow estimates of transpiration in Japanese cedar,” *Agric. For. Meteorol.*, vol. 135, no. 1–4, pp. 110–116, 2005.

[71] R. A. Vertessy, R. G. Benyon, S. K. O. Sullivan, and P. R. Gribben, “Relationships between stem diameter, sapwood area, leaf area and transpiration in a young mountain ash forest,” *Tree Physiol.*, vol. 15, no. 9, pp. 559–567, 1995.

[72] T. W. Davis, “ENVIRONMENTAL MONITORING THROUGH WIRELESS SENSOR NETWORKS,” 2012.

[73] Jian Kang, Rui Jin, and Xin Li, “Regression Kriging-Based Upscaling of Soil Moisture Measurements From a Wireless Sensor Network and Multiresource Remote Sensing Information Over Heterogeneous Cropland,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 92–96, 2014.

[74] W. T. Crow *et al.*, “Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products,” *Rev. Geophys.*, vol. 50, no. 2, pp. 1–20, 2012.

[75] D. D. Bosch, V. Lakshmi, T. J. Jackson, M. Choi, and J. M. Jacobs, “Large scale measurements of soil moisture for validation of remotely sensed data: Georgia soil moisture experiment of 2003,” *J. Hydrol.*, vol. 323, no. 1–4, pp. 120–137, 2006.

[76] T. W. Ford and S. M. Quiring, “Comparison and application of multiple methods for temporal interpolation of daily soil moisture,” *Int. J. Climatol.*, vol. 34, no. 8, pp. 2604–2621, Jun. 2014.

[77] G. H. Zhang M, Li M, Wang W, Liu C, “Spatial and temporal variability of soil moisture based on multifractal analysis,” *Math. Comput. Model.*, vol. 58, no. 3–4, pp. 826–33, 2013.

[78] H. Zou, Y. Yue, Q. Li, and A. G. O. Yeh, “An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network,” *Int. J. Geogr. Inf. Sci.*, vol. 26, no. 4, pp. 667–689, Apr. 2012.

[79] J. Zhang, H. Chen, Y. Su, Y. Shi, W. Zhang, and X. Kong, “Spatial Variability of Surface Soil Moisture in a Depression Area of Karst Region,” *CLEAN - Soil, Air, Water*, vol. 39, no. 7, pp. 619–625, Jul. 2011.

[80] S. W. Lyon, R. S⊘rensen, J. Stendahl, and J. Seibert, “Using landscape characteristics to define an adjusted distance metric for improving kriging interpolations,” *Int. J. Geogr. Inf. Sci.*, vol. 24, no. 5, pp. 723–740, Apr. 2010.

[81] T. Lakhankar, A. S. Jones, C. L. Combs, M. Sengupta, T. H. vonder Haar, and R. Khanbilvardi, “Analysis of large scale spatial variability of soil moisture using a geostatistical method,” *Sensors*, vol. 10, no. 1, pp. 913–932, 2010.

[82] L. Mihaylova, A. Ali, B. Adebisi, and A. Ikpehai, “Location prediction optimisation in WSNs using Kriging interpolation,” *IET Wirel. Sens. Syst.*, vol. 6, no. 3, pp. 74–81, 2016.

[83] M. A. Oliver and R. Webster, “A tutorial guide to geostatistics: Computing and modelling variograms and kriging,” *Catena*, vol. 113, pp. 56–69, 2014.

[84] A. Konak, “Estimating path loss in wireless local area networks using ordinary kriging,” *Proc. - Winter Simul. Conf.*, no. Badman 2006, pp. 2888–2896, 2010.

[85] M. Umer, L. Kulik, and E. Tanin, “Spatial interpolation in wireless sensor networks: Localized algorithms for variogram modeling and Kriging,” *Geoinformatica*, vol. 14, no. 1, pp. 101–134, 2010.

[86] P. L. P. Corrêa, A. R. Hirakawa, C. E. Cugnasca, A. Camilli, and A. M. Saraiva, “From wireless sensors to field mapping: Anatomy of an application for precision agriculture,” *Comput. Electron. Agric.*, vol. 58, no. 1, pp. 25–36, 2007.

[87] A. Bárdossy and W. Lehmann, “Spatial distribution of soil moisture in a small catchment. Part 1: geostatistical analysis,” *J. Hydrol.*, vol. 206, no. 1–2, pp. 1–15, 1998.

[88] R. S. V. Teegavarapu and V. Chandramouli, “Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records,” *J. Hydrol.*, vol. 312, no. 1–4, pp. 191–206, 2005.

[89] S. M. Vicente-Serrano, M. A. Saz-Sánchez, and J. M. Cuadrat, “Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): Application to annual precipitation and temperature,” *Clim. Res.*, vol. 24, no. 2, pp. 161–180, 2003.

[90] P. Goovaerts, “Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall,” *J. Hydrol.*, vol. 228, no. 1–2, pp. 113–129, 2000.

[91] J. E. Nash and J. V. Sutcliffe, “River flow forecasting through conceptual models part I — A discussion of principles,” *J. Hydrol.*, vol. 10, no. 3, pp. 282–290, Apr. 1970.

[92] W. S. Cleveland, “Robust Locally Weighted Regression and Smoothing Scatterplots,” *J. Am. Stat. Assoc.*, vol. 74, no. 368, pp. 829–836, Dec. 1979.

[93] NRCC, “Northeast Regional Climate Center.” [Online]. Available: http://http//www.nrcc.cornell.edu.

[94] H. J. Tromp-van Meerveld and J. J. McDonnell, “On the interactions between the spatial patterns of topography, soil moisture, transpiration and species distribution at the hillslope scale,” *Adv. Water Resour.*, vol. 29, no. 2, pp. 293–310, 2006.

[95] J. C. Domec *et al.*, “A comparison of three methods to estimate evapotranspiration in two contrasting loblolly pine plantations: Age-related changes in water use and drought sensitivity of evapotranspiration components,” *For. Sci.*, vol. 58, no. 5, pp. 497–512, 2012.

[96] N. Chang and Y. Hong, *Multiscale hydrologic remote sensing: Perspectives and applications*. 2012.

[97] “Pittsburgh Airport Meteorological Station.” [Online]. Available: http://www.usclimatedata.com/climate/%0Apittsburgh/pennsylvania/united-states/uspa3601%0A.

[98] T. K. Thierfelder, R. B. Grayson, D. Von Rosen, and A. W. Western, “Inferring the location of catchment characteristic soil moisture monitoring sites. Covariance structures in the temporal domain,” *J. Hydrol.*, vol. 280, no. 1–4, pp. 13–32, 2003.

[99] L. M. Oliveira and J. J. Rodrigues, “Wireless Sensor Networks: a Survey on Environmental Monitoring,” *J. Commun.*, vol. 6, no. 2, Apr. 2011.

[100] T. W. Davis, X. Liang, M. Navarro, D. Bhatnagar, and Y. Liang, “An experimental study of WSN power efficiency: MICAz networks with XMesh,” *Int. J. Distrib. Sens. Networks*, vol. 2012, 2012.

[101] M. Navarro, Y. Li, and Y. Liang, “Energy profile for environmental monitoring wireless sensor networks,” in *2014 IEEE Colombian Conference on Communications and Computing, COLCOM 2014 - Conference Proceedings*, 2014, pp. 1–6.

[102] G. Ferrari, *Sensor Networks: where theory meets practice*. 2010.

[103] M. Navarro, D. Bhatnagar, and Y. Liang, “An Integrated Network and Data Management System for Heterogeneous WSNs,” in *2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems*, 2011, pp. 819–824.

[104] R. Ringgaard, M. Herbst, and T. Friborg, “Partitioning of forest evapotranspiration: The impact of edge effects and canopy structure,” *Agric. For. Meteorol.*, vol. 166–167, pp. 86–97, 2012.

[105] J. W. Hui and D. Culler, “The dynamic behavior of a data dissemination protocol for network programming at scale,” in *Proceedings of the 2nd international conference on Embedded networked sensor systems - SenSys ’04*, 2004, p. 81.

[106] S. Burgess, T. D.-P. and Soil, and U. 2008, “Burgess and Dawson 2008 Using branch and basal trunk sap flow measurements to estimate whole-plant water capacitance, a caution-Plant and Soil 305, 5.pdf,” *Springer*, vol. 305, no. 1–2, pp. 5–13, 2008.

[107] S. A. Archfield and R. M. Vogel, “Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments,” *Water Resour. Res.*, vol. 46, no. 10, pp. 1–15, 2010.

[108] L. M. Parada and X. Liang, “A novel approach to infer streamflow signals for ungauged basins,” *Adv. Water Resour.*, vol. 33, no. 4, pp. 372–386, Apr. 2010.

[109] C. Shu and T. B. M. J. Ouarda, “Improved methods for daily streamflow estimates at ungauged sites,” *Water Resour. Res.*, vol. 48, no. 2, p. n/a-n/a, Feb. 2012.

[110] W. Farmer and R. M. Vogel, “Performance-weighted methods for estimating monthly streamflow at ungauged sites,” *J. Hydrol.*, 2013.

[111] T. Razavi, P. Coulibaly, and M. Asce, “Streamflow Prediction in Ungauged Basins : Review of Regionalization Methods,” no. August, pp. 958–975, 2013.

[112] W.-C. Huang and F.-T. Yang, “Streamflow estimation using kriging,” *Water Resour. Res.*, vol. 34, no. 6, pp. 1599–1608, Jun. 1998.

[113] D. A. Hughes and V. Y. Smakhtin, “Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves,” *Hydrol. Sci. J.*, vol. 41, no. 6, pp. 851–871, 1996.

[114] A. Mishra and P. Coulibaly, “Developments in hydrometric network design: A review,” *Rev. Geophys.*, 2009.

[115] J. Samuel, P. Coulibaly, and R. A. Metcalfe, “Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods,” *J. Hydrol. Eng.*, vol. 16, no. 5, pp. 447–459, May 2011.

[116] USGS, “USGS 07327441 SCS Pond No. 26 near Cyril, OK.” [Online]. Available: https://waterdata.usgs.gov/nwis/uv?site_no=07327441. [Accessed: 15-May-2019].

[117] K. J. Lanfear and R. M. Hirsch, “USGS study reveals a decline in long-record streamgages,” *Eos (Washington. DC).*, vol. 80, no. 50, pp. 605–607, 1999.

[118] E. Stokstad, “HYDROLOGY: USGS Braces for Severe Budget Cuts,” *Science (80-.).*, vol. 292, no. 5519, pp. 1040a – 1040, May 2001.

[119] C. Vorosmarty *et al.*, “Global water data: A newly endangered species,” *Eos, Trans. Am. Geophys. Union*, vol. 82, no. 5, pp. 54–54, Jan. 2001.

[120] A. Witze, “US budget cuts hit Earth monitoring,” *Nature*, vol. 497, no. 7450, pp. 419–420, May 2013.

[121] A. Loukas and L. Vasiliades, “Streamflow simulation methods for ungauged and poorly gauged watersheds,” *Nat. Hazards Earth Syst. Sci.*, vol. 14, no. 7, pp. 1641–1661, Jul. 2014.

[122] Y. He, A. Bárdossy, and E. Zehe, “A review of regionalisation for continuous streamflow simulation,” *Hydrol. Earth Syst. Sci.*, vol. 15, no. 11, pp. 3539–3553, 2011.

[123] Y. B. Dibike and D. P. Solomatine, “River flow forecasting using artificial neural networks,” *Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos.*, vol. 26, no. 1, pp. 1–7, Jan. 2001.

[124] D. P. Solomatine and A. Ostfeld, “Data-driven modelling: some past experiences and new approaches,” *J. Hydroinformatics*, vol. 10, no. 1, pp. 3–22, Jan. 2008.

[125] V. Y. Smakhtin, “Generation of natural daily flow time-series in regulated rivers using a non-linear spatial interpolation technique,” *Regul. Rivers Res. Manag.*, vol. 15, no. 4, pp. 311–323, Jul. 1999.

[126] V. Y. Smakhtin, D. A. Hughes, and E. CREUSE-NAUDIN, “Regionalization of daily flow characteristics in part of the Eastern Cape, South Africa,” *Hydrol. Sci. J.*, vol. 42, no. 6, pp. 919–936, Dec. 1997.

[127] Y. Zhang and F. Chiew, “Relative merits of different methods for runoff predictions in ungauged catchments,” *Water Resour. Res.*, 2009.

[128] R. Arsenault and F. P. Brissette, “Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches,” *Water Resour. Res.*, vol. 50, no. 7, pp. 6135–6153, Jul. 2014.

[129] M. J. Halverson and S. W. Fleming, “Complex network theory, streamflow, and hydrometric monitoring system design,” *Hydrol. Earth Syst. Sci.*, vol. 19, no. 7, pp. 3301–3318, Jul. 2015.

[130] D. Emerson, A. Vecchia, and A. Dahi, “Evaluation of Drainage-Area Ratio Method Used to Estimate Streamflow for the Red River of the North Basin , North Dakota and Minnesota Scientific Investigations Report 2005 – 5017 Evaluation of Drainage-Area Ratio Method Used to Estimate Streamflow for th,” *Sci. Investig. Rep.*, p. 5017, 2005.

[131] W. Asquith, M. Roussel, and J. Vrabel, “Statewide analysis of the drainage-area ratio method for 34 streamflow percentile ranges in Texas,” 2006.

[132] Y. Mohamoud and M., “Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves,” *Hydrol. Sci. J.*, vol. 53, no. 4, pp. 706–724, Aug. 2008.

[133] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. 2009.

[134] W. H. Farmer, “Ordinary kriging as a tool to estimate historical daily streamflow records,” *Hydrol. Earth Syst. Sci.*, vol. 20, no. 7, pp. 2721–2735, 2016.

[135] R. Arsenault and F. Brissette, “Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches,” *Water Resour. Res.*, 2014.

[136] J. O. Skøien and G. Blöschl, “Spatiotemporal topological kriging of runoff time series,” *Water Resour. Res.*, vol. 43, no. 9, Sep. 2007.

[137] A. R. Solow and S. M. Gorelick, “Estimating monthly streamflow values by cokriging,” *Math. Geol.*, vol. 18, no. 8, pp. 785–809, Nov. 1986.

[138] J. Villeneuve, G. Morin, B. Bobee, D. Leblanc, and J. Delhomme, “Kriging in the design of streamflow sampling networks,” *Water Resour. Res.*, vol. 15, no. 6, pp. 1833–1840, Dec. 1979.

[139] T. S. Virdee and N. T. Kottegoda, “A brief review of kriging and its application to optimal interpolation and observation well selection,” *Hydrol. Sci. J.*, vol. 29, no. 4, pp. 367–387, Dec. 1984.

[140] R. M. Hirsch, “An evaluation of some record reconstruction techniques,” *Water Resour. Res.*, 1979.

[141] D. A. Sachindra, F. Huang, A. Barton, and B. J. C. Perera, “Least square support vector and multi-linear regression for statistically downscaling general circulation model outputs to catchment streamflows,” *Int. J. Climatol.*, vol. 33, no. 5, pp. 1087–1106, Apr. 2013.

[142] J. R. Stedinger, “Fitting log normal distributions to hydrologic data,” *Water Resour. Res.*, vol. 16, no. 3, pp. 481–490, Jun. 1980.

[143] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso.,” *Biostatistics*, vol. 9, no. 3, pp. 432–41, Jul. 2008.

[144] A. d’Aspremont, O. Banerjee, and L. El Ghaoui, “First-order methods for sparse covariance selection,” *SIAM J. Matrix Anal. …*, 2008.

[145] R. Mazumder and T. Hastie, “The Graphical Lasso : New Insights and Alternatives,” pp. 1–14, 2012.

[146] S. Sojoudi, “Equivalence of Graphical Lasso and Thresholding for Sparse Graphs,” pp. 1019–1025, 2014.

[147] D. M. Witten, J. Friedman, and N. Simon, “New Insights and Faster Computations for the Graphical Lasso,” *J. Comput. Graph. Stat.*, vol. 20, no. 4, pp. 892–900, 2011.

[148] M. a. Sustik and B. Calderhead, “GLASSOFAST: An efficient GLASSO implementation,” pp. 2–4, 2012.

[149] A. C. Benke and C. E. Cushing, *Rivers of North America*. Academic Press, 2011.

[150] H. Moradkhani, K. L. Hsu, H. Gupta, and S. Sorooshian, “Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter,” *Water Resour. Res.*, vol. 41, no. 5, pp. 1–17, 2005.

[151] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser, “Dual state–parameter estimation of hydrological models using ensemble Kalman filter,” *Adv. Water Resour.*, vol. 28, no. 2, pp. 135–147, Feb. 2005.

[152] C. Montzka, H. Moradkhani, L. Weihermüller, H.-J. H. Franssen, M. Canty, and H. Vereecken, “Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter,” *J. Hydrol.*, vol. 399, no. 3–4, pp. 410–421, Mar. 2011.

[153] S. V. Kumar, R. H. Reichle, K. W. Harrison, C. D. Peters-Lidard, S. Yatheendradas, and J. A. Santanello, “A comparison of methods for a priori bias correction in soil moisture data assimilation,” *Water Resour. Res.*, vol. 48, no. 3, Mar. 2012.

[154] J. a. Santanello *et al.*, “Using remotely-sensed estimates of soil moisture to infer soil texture and hydraulic properties across a semi-arid watershed,” *Remote Sens. Environ.*, vol. 110, pp. 79–97, 2007.

[155] F. Chen *et al.*, “Modeling of land surface evaporation by four schemes and comparison with FIFE observations,” *J. Geophys. Res. Atmos.*, vol. 101, no. D3, pp. 7251–7268, Mar. 1996.

[156] F. Chen, Z. Janjić, and K. Mitchell, “Impact of Atmospheric Surface-layer Parameterizations in the new Land-surface Scheme of the NCEP Mesoscale Eta Model,” *Boundary-Layer Meteorol.*, vol. 85, no. 3, pp. 391–421, Dec. 1997.

[157] F. Chen and J. Dudhia, “Coupling an Advanced Land Surface – Hydrology Model with the Penn State – NCAR MM5 Modeling System . Part I : Model Implementation and Sensitivity,” *Mon. Weather Rev.*, vol. 129, pp. 569–585, 2001.

[158] V. Koren, J. Schaake, K. Mitchell, Q.-Y. Duan, F. Chen, and J. M. Baker, “A parameterization of snowpack and frozen ground intended for NCEP weather and climate models,” *J. Geophys. Res. Atmos.*, vol. 104, no. D16, pp. 19569–19585, Aug. 1999.

[159] R. B. Clapp and G. M. Hornberger, “Empirical Equations for Some Soil Hydraulic Properties,” vol. 14, no. 4, 1978.

[160] E. D. Gutmann and E. E. Small, “The effect of soil hydraulic properties vs. soil texture in land surface models,” *Geophys. Res. Lett.*, vol. 32, no. 2, p. L02402, Jan. 2005.

[161] B. J. Cosby, G. M. Hornberger, R. B. Clapp, and T. R. Ginn, “Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils,” vol. 20, no. 6, pp. 682–690, 1984.

[162] H. Holtan, C. England, G. Lawless, and G. Schumaker, *Moisture-tension Data for Selected Soils on Experimental Watershed*. Agricultural Research Service, U.S. Dept. of Agriculture, 1968.

[163] W. J. Rawls, P. Yates, and L. Asmussen, “Calibration of selected infiltration equations for the Georgia Coastal Plain,” *Rep. ARS-S-113 July 1976. 110 p, 2 fig, 8 tab, 25 ref, 1 append.*, p. 110, 1976.

[164] Research Applications Laboratory, “unified-noah-lsm.” [Online]. Available: https://ral.ucar.edu/solutions/products/unified-noah-lsm. [Accessed: 23-Sep-2018].

[165] C. D. Peters-Lidard *et al.*, “The Effect of Soil Thermal Conductivity Parameterization on Surface Energy Fluxes and Temperatures,” *J. Atmos. Sci.*, vol. 55, no. 7, pp. 1209–1224, Apr. 1998.

[166] R. Shrestha and P. Houser, “A heterogeneous land surface model initialization study,” *J. Geophys. Res.*, vol. 115, no. D19, p. D19111, Oct. 2010.

[167] L. M. Parada and X. Liang, “Optimal multiscale Kalman filter for assimilation of near-surface soil moisture into land surface models,” *J. Geophys. Res. D Atmos.*, vol. 109, no. 24, pp. 1–21, 2004.

[168] L. M. Parada and X. Liang, “Impacts of spatial resolutions and data quality on soil moisture data assimilation,” *J. Geophys. Res. D Atmos.*, vol. 113, no. 10, p. D10101, May 2008.

[169] NASA, “0.125 Degree Hourly Primary Forcing Data for NLDAS-2.” [Online]. Available: https://hydro1.gesdisc.eosdis.nasa.gov/dods/NLDAS_FORA0125_H.002. [Accessed: 01-Oct-2018].