LARGE-SCALE ROOT ZONE SOIL MOISTURE ESTIMATION

LARGE-SCALE ROOT ZONE SOIL MOISTURE ESTIMATION USING DATA-DRIVEN METHODS

By XIAOJUN PAN, B. Eng

A Thesis Submitted to the School of Graduate Study

in Partial Fulfillment of the Requirement

for the Degree Master of Science

MASTER OF SCIENCE (2015) McMaster University

(Geography) Hamilton, Ontario

TITLE: Estimating Root Zone Soil Moisture for A Large Area Using Artificial Neural Networks

AUTHOR: Xiaojun Pan, B.Eng (Hohai University)

SUPERVISOR: Dr. Paulin Coulibaly

NUMBER OF GAGES: Xii, 117

Abstract

Soil moisture is an important variable in many environmental researches and application areas as it affects the interactions between atmosphere and land surface by controlling the energy and water exchange. The current measurement techniques are insufficient to acquire accurate large-scale root zone soil moisture (RZSM) data at the spatial resolution of interest. Though assorted models have been successfully applied in relatively small areas to estimate RZSM, the large-scale estimation is still facing challenges as it requires the flexibility and practicality of the models for the applications under various conditions. Though physically based soil moisture models are widely used, the errors in model physics affect the flexibility of these models meanwhile their large demand of data and computational resources reduces the practicality. On the contrary, the statistical and data-driven methods have high potential but their applications for large-scale RZSM estimation have not been fully explored.

To develop feasible models for large-scale RZSM estimation using the surface observations, artificial neural networks, specifically multilayer perceptrons (MLPs), were applied in this study to estimate RZSM at the depths of 20cm and 50cm, using the data of 557 stations in the United States. Two experiments including four models were developed and the input variables of the models were carefully selected. The sensitivity analysis found that surface soil moisture and the cumulative rainfall, snowfall, air temperature and surface soil temperature were important inputs. If given soil texture data as inputs, the models achieved better performance and were extremely sensitive to them. The results showed that the MLPs were effective and flexible for the estimation of soil moisture at 20cm under various climate types and were insensitive to the potential errors in soil moisture datasets. However, the results of the estimation at 50cm are not as good as that of the 20cm.

Acknowledgement

I would like to express my gratitude to many people for their support throughout the nearly two years of my graduate study.

First of all, I would like to thank my supervisor Dr. Paulin Coulibaly. I appreciate the chance you have offered me to pursue the research study here in McMaster University. Without the guidance and support from you, this thesis could not have been written. Thanks for the time, help and patience you have given to me.

A special thanks is given to Dr. Kurt C. Kornelsen. Your support for the collection of the data used in this study is gratefully appreciated. I also would like to thank you for your comments and suggestions which helped to improve the content of this thesis greatly. My gratitude is also given to Wouter Dorigo and Hualan Rui. Thank you for the patience and quick replies which helped me a lot to appropriately collect and organize the soil moisture data from the International Soil Moisture Network and the forcing data from the GLDAS products. I am also grateful to all the members in our Water Resources and Hydrologic Modeling Lab. I enjoy all the talks and discussions with you. The time we spent together always inspired me and helped me to learn more during these two years.

I would like to extend a special thanks to my family and friends. I thank my father Hanfu Pan for the great support and encouragement given to me for the pursuits of this degree. I am also grateful to my friends Susie Yu, Guessy Wang and Jianxiong Zhang. Thank you all very much for the help and the spiritual support that made me overcome lots of difficulties in my life and research study. Last but not least, I would like to express my gratitude to the professors and staffs who offered help to me. Thank you all for providing the enjoyable work and study environment in School of Geography and Earth Sciences.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABDOMEN	Approximate Buckingham-Darcy Equation for Moisture Estimation Model
ANFIS	Adaptive Neuro-Fuzzy Inference System
ARIMA	Autoregressive Integrated Moving Average Model
ARMA	Autoregressive Moving Average Model
BEACH	Bridging Event and Continuous Hydrological Modelling
CATHY	Catchment Hydrological Model
CLM	Community Land Model
CN	Crank-Nicolson Linear Finite Difference Scheme
CoLM	Common Land Model
CSSP	Conjunctive Surface–Subsurface Process Model
DA	Data Assimilation
DSUKF	Dual Standard-Unscented Kalman Filter
E	Evaporation
EKF	Extend Kalman Filter
EnKF	Ensemble Kalman Filter
EPR	Evolutionary Polynomial Regression
ET	Evapotranspiration
EX	Explicit Finite Difference Scheme
GP	Genetic Programing
GRNNs	Generalized Regression Neural Networks
HF	H8 Filter
HONNs	High-Order Neural Networks
ISBA	Interactions between Soil, Biosphere, and Atmosphere Surface Scheme
KNN	K Nearest Neighbours
LSM	Land Surface Model

LSP-DSSAT	Land Surface Process Model Coupled with Decision Support System for Agrotechnology Transfer Model
LSSVMs	Least Squares Support Vector Machines
MAE	Mean Absolute Error
MLPs	Multilayer Perceptrons
MLR	Multiple Linear Regression
NL	Backward Euler Finite Difference Scheme
NR	Net Radiation
PAMII	Prairie Agrometeorological Model Version II
PBMs	Physically based Models
PCA	Principal Component Analysis
PF	Particle Filter
PSO	Particle Swarm Optimization Algorithm
R	Rainfall
RE models	Models Using Richards Equation or Its Approximation
RH	Relative Humidity
RMSE	Root Mean Squared Error
RZSM	Root Zone Soil Moisture
RZWQM	Root Zone Water Quality Model
S	Snowfall
SDDMs	Statistical and Data-Driven Models
SCE	Shuffled Complex Evolution Method
SCE-UA	Shuffled Complex Evolution Method Developed at the University of Arizona
SDW	System Dynamics Watershed Model
SH	Specific humiditY
SIM	SAFRAN-ISBA-MODCOU Hydrometeorological Model
SKF	Standard Kalman Filter
SM05	Soil Moisture Measured at the Depth of 5cm
SM20	Soil Moisture Measured at the Depth of 20cm

SM50	Soil Moisture Measured at the Depth of 50cm
SMR	Soil Moisture Routing Model
SOMs	Self-Organizing Maps
SR	Short Wave Radiation
SVAT	Soil-Vegetation-Atmosphere Transfer Model
SVMs	Support Vector Machines
SWAP	Soil–Water–Atmosphere–Plant Model
SWAT	Soil and Water Assessment Tool
Ta	Surface Air Temperature
Ts	Surface Soil Temperature
UKF	Unscented Kalman Filter
VIC	Variable Infiltration Capacity Model
VSMB	Versatile Soil Moisture Budget Model
WetSpa	Water and Energy Transfer in Soil, Plant and Atmosphere Model
WNNs	Wavelet Neural Networks
WRF-Noah	Weather Research and Forecasting Model Coupled with Noah Land Surface Model
WS	Wind Speed

Chapter 1

Introduction

1.1 Background

The moisture in vadose zone governs the exchange of water and energy between atmosphere and land surface, making it significant in numerous environmental studies (Entekhabi et al., 1996; Fischer et al., 2007; Huang et al., 1996; Koster et al., 2004; Pan & Mahrt, 1987). Through its influence on infiltration process, soil moisture can affect the downward propagation of atmospheric forcing to land surface system (Entekhabi et al., 1996; Pan & Mahrt, 1987). It also transfers the feedback from land surface to atmosphere by controlling the partition between sensible and latent heat flux (Entekhabi et al., 1996; Fischer et al., 2007; Koster et al., 2004; Pan & Mahrt, 1987).

For the importance of the temporal variations and spatial distribution of soil moisture, many techniques and instrumentations have been developed to acquire soil moisture data (Ochsner et al., 2013; Robinson et al., 2008; Vereecken et al., 2008). In-situ soil sensing instruments such as neutron probes, electromagnetic sensors and heat pulse sensors are conventionally used to obtain soil moisture data (Robinson et al., 2008). Large in-situ observation networks have been built to measure soil moisture in many countries (Dorigo et al., 2011; Ochsner et al., 2013). However, the land surface conditions are highly spatial heterogeneous, while the fine spatial extent of sampling limits the use of these in-situ observations to depict the spatial distribution of soil moisture. In addition, these observation networks have relatively low density in some regions and many ungauged areas still exist. The current in-situ observation networks cannot fully meet the demand of soil moisture data while the implementation of these

networks is costly. Therefore many remote sensing techniques have also been proposed to collect spatial soil moisture information (Kornelsen & Coulibaly, 2013; Ochsner et al., 2013; Robinson et al., 2008; Vereecken et al., 2008). These advances also lead to the launching of large projects such as the Soil Moisture and Ocean Salinity (SMOS) mission (Kerr et al., 2001) and Soil Moisture Active Passive (SMAP) mission (Entekhabi et al., 2010), which aim to acquire global soil moisture measurements using remote sensing techniques. However, the soil moisture measurements acquired by remote sensing techniques are limited to the top a few centimeters. Therefore, remote sensing techniques themselves can provide little information about root zone soil moisture (RZSM), which is particularly important for plant growth, transpiration, runoff generation as well as groundwater recharge. Hence, developing reliable modeling approaches are necessary to acquire the soil moisture information of interest.

The soil moisture information for a large area can help to understand the water and energy dynamics between land surface and atmosphere at a large scale, leading to the improved meteorological and climatic prediction (Koster et al., 2004; Li et al., 2007; Ni-Meister et al., 2005) as well as drought monitoring (Fischer et al., 2007; Wu & Kinter, 2009) at continental scale and global scale. These improvements will assist in mitigating the effects of related natural hazards. Various models have been successfully adopted to estimate RZSM in relatively small areas (shown in **Appendix 1** and **Appendix 2**) while the large-scale estimation still faces many difficulties. The large-scale RZSM estimation requires the high flexibility of the applied models since land surface and climatic conditions are highly heterogeneous. Studies have detected that soil moisture models may be sensitive to local climate (Kumar et al., 2009; Xia et al., 2014). In

addition, the practicality of the models is also important for large-scale estimation while some models required much computational resources and a large number of data (Chirico et al., 2014; Subbaiah, 2013).

1.2 Objectives and Scope

Motivated by the challenges mentioned above, this study aims to develop feasible models for large-scale RZSM estimation using the surface observations. Specific objectives of this thesis include:

1. The identification of a potential method for large-scale RZSM estimation based on a literature review of the merits and demerits of current soil moisture models.
2. The collection of appropriate datasets for the development and evaluation of the method selected through objective 1.
3. The development of models for large-scale RZSM estimation identified through objective 1, based on a selection of appropriate inputs.
4. The sensitivity analysis of the models developed in objective 3 to the selected inputs, which will verify the appropriateness of the model design and explore the importance of different variables.
5. The assessment of the effects from the potential errors in the soil moisture observations on the developed models.
6. The evaluation of flexibility and performance of the developed models to estimate RZSM for a large area.

This study only uses in-situ soil moisture observations. The remote sensing soil moisture measurements were not considered. On one hand, the remote sensed data provide little information of RZSM, as mentioned in **section 1.1**. This drawback makes it

difficult to evaluate the performance of models for RZSM estimation using remote sensed data. On the other hand, many factors cause uncertainty in remote sensed soil moisture data (Kornelsen & Coulibaly, 2013; Wang & Qu, 2009) and there are difficulties in the retrieval of accurate soil moisture data at high spatial resolution (Vereecken et al., 2008). Though data assimilation has been widely utilized to improve model performance, it was found that the noises in active microwave remote sensing measurements make the assimilation system performed no better than the open loop simulation (Hoeben & Troch, 2000). On the contrary, the in-situ soil moisture observations are considered relatively reliable and widely utilized for the calibration and validation of remote sensing methods (Crow et al., 2012; Mohanty et al., 2013).

1.3 General Methodology and Thesis Structure

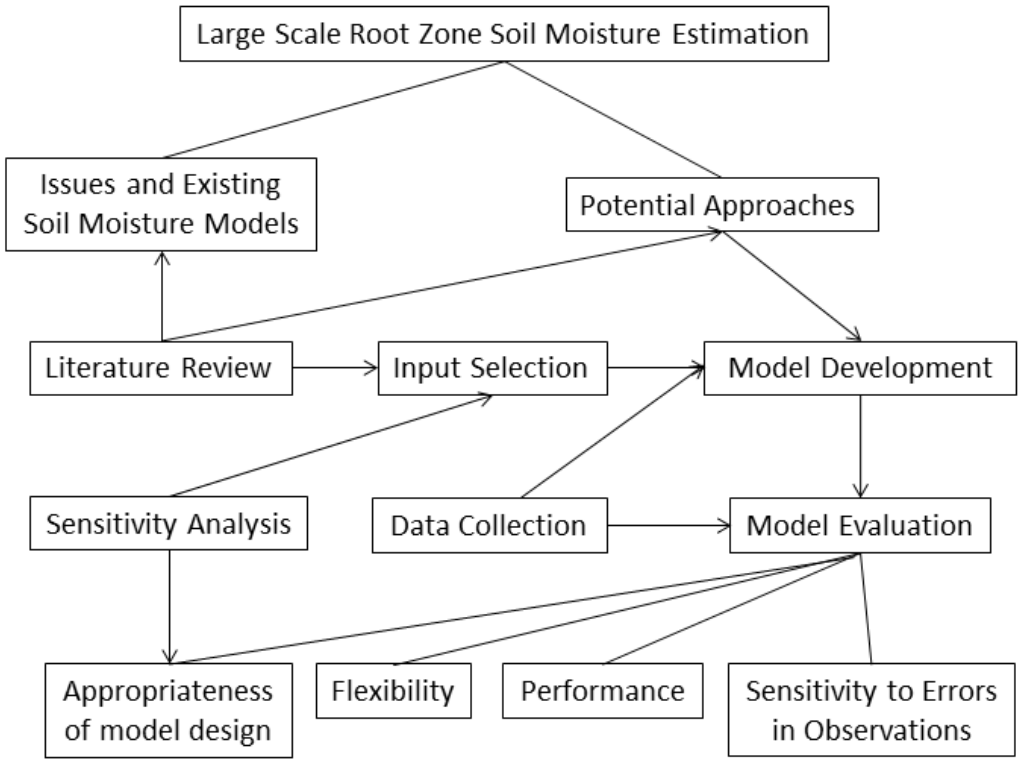

Figure 1 Flowchart of the general methodology

According to the objectives described in **section 1.2**, an overview of the general methodology for this study is depicted in **Figure 1**. A literature review, which will be presented in **Chapter 2**, was conduct to summarize the merits and demerits of the current methods as well as to discuss the issues in large-scale RZSM estimation. This review led to the identification of the potential methods and it also helped to select the input variables appropriately.

The conclusions of **Chapter 2** showed that artificial neural networks (ANNs), specifically multilayer perceptrons (MLPs), have the high potential for large-scale RZSM estimation and the following advantages:

1. The high ability for nonlinear input-output mapping. MLPs have been demonstrated that they are able to capture the nonlinearity in dynamic systems and therefore widely applied in hydrology (Abrahart et al., 2012; Abrahart & See, 2007; ASCE, 2000a, 2000b; Khan & Coulibaly, 2006). This characteristic makes them very suitable for RZSM estimation which involves high nonlinearity (Elshorbagy et al., 2010b).
2. The potential to be flexible tools. MLPs do not rely on physical assumptions nor require prior assumptions about data structures which often limit model reliability and flexibility. The MLPs trained with different soil moisture profiles can be used in a geographic region (Gill et al., 2006; Kornelsen & Coulibaly, 2014b), suggesting their applications for a larger area.
3. MLPs have a high computational efficiency. Once the MLPs are developed, they are computationally effective.

4. Low demand of soil properties data. MLPs do not require many soil properties data to specify soil hydraulic parameters (Kornelsen & Coulibaly, 2014b) which govern soil water dynamics.
5. The high adaptability to various types of data and input configuration (Ghedira et al., 2004).

Hence, MLPs were used in this study. The basic theory of ANNs and MLPs, the development of the models as well as the data collection and processing will be presented in **Chapter 3**. The development of the models included the processes of input variables selection and the specification of model configuration. As MLPs are data-driven methods, it is necessary to select the inputs appropriately. The findings from literature review (shown in **Chapter 2**) and the correlation analysis as well as sensitivity analysis were used to guide the selection of input variables. To achieve the objectives of this thesis, two experiments were designed and their details will also be presented in **Chapter 3**.

The evaluation of the models developed in **Chapter 3** will be shown in **Chapter 4**. The sensitivity analysis for the models in the first experiment will be presented to verify the appropriateness of the design of the models. To explore whether the models were sensitive to the potential errors in the soil moisture observations, a comparison between the two experiments will be given. The collected datasets were divided into two parts for training and independent validation, respectively (described in **Chapter 3**). The results of training and independent validation will also be summarized in **Chapter 4** to assess the flexibility and performance of the developed models.

Finally, the findings and conclusions as well as the proposed future work will be presented in **Chapter 5**.

Chapter 2

Review of Models for Root Zone Soil Moisture Estimation

2.1 Introduction

Hydrological models can be categorized with respect to the degree of representation of the involved physical processes (Jajarmizadeh et al., 2012). Theoretical models are developed based on equations derived from basic physics to simulate hydrological processes (Dingman, 2008; Jajarmizadeh et al., 2012). Conceptual models also apply physical laws but in a much simplified form. The statistical and data-driven models are not based on physical laws. They are empirical models developed by relating the input and output data, hence also known as input-output models. As the former two types of models attempt to represent the physical processes at different levels, they share some common features. Similarly, the latter two types of models relay on mathematical methods to relate inputs to outputs. Therefore, in this review the soil moisture models were classified into two categories, namely physically based models as well as statistical and data-driven models.

According to such classification, the objectives of this review includes: 1. reviewing the soil moisture models and identifying the key methods in the two categories of models; 2. summarizing the advances and common issues in the two categories of models for soil moisture estimation; 3. discussing the merits and demerits for the two categories of models to estimate large-scale RZSM. The conclusions of this review will be presented at the end of this chapter, leading to the determination of the method used in this study.

Note that for the reasons mentioned in **section 1.2**, this review focuses on the studies using in-situ soil moisture observations and/or synthetic data.

2.2 Physically Based Models

Here we define the physically based models (PBMs) for RZSM estimation as models that use equations based on physical laws to represent the physical processes governing soil moisture dynamics. A summary of the selected advances in the models of this category is presented in **Appendix 1**. Though **Appendix 1** may not be exhaustive, the author hopes this summary will lead readers to appropriate references. In this category, most of the models simulate soil moisture dynamics based on the principle of water balance (budget models), or the application of Richards equation and its approximation. In this section, the characteristics of these two types of models will be first summarized. A comparison will be conducted between these two types of models. Several issues that have received much attention in the researches using PBMs will be summarized. As data assimilation methods have been widely applied to enhance the accuracy of RZSM estimation by PBMs, the key data assimilation methods and some common issues in their applications will also be summarized and discussed.

2.2.1 Budget Models

The water budget models usually consider the soil layers as reservoirs and use water balance equations to simulate the processes of receiving and retaining water until the corresponding storage capacity is filled (Romano et al., 2011). The complexity of budget models varies, from models with a single soil layer only considering precipitation, evapotranspiration and runoff (Yamaguchi & Shinoda, 2002) to soil-vegetation-

atmosphere transfer (SVAT) models (Montaldo & Albertson, 2001; Montaldo et al., 2001; Sabater et al., 2007). The differences of these budget models for soil moisture estimation mostly depend on: 1) the hydrological components (e.g. precipitation, evapotranspiration, runoff, leakage and lateral flow) considered in the water balance equations and the methods applied to simulate these components; 2) the number of soil layers included in the models and the connections of the layers. With reference to a single-layer budget model, the water balance in root zone can be presented by the following equation (Guswa et al., 2002):

$$nZ_r \frac{dS}{dt} = I(S,t) - L(S) - T(S) - E(S) \quad (1)$$

where n is the porosity; Z_r is the depth of the root zone; S is the average saturation over the root zone; t is time; I is the infiltration rate into the root zone; L is the rate of leakage from the root zone; T and E are the transpiration and evaporation rates, respectively. In this single-layer budget model, the movement of the wetting fronts is ignored and the soil moisture is represented by a value of average saturation over the entire root zone (Guswa et al., 2002). The losses through leakage and evapotranspiration only depend on the average saturation of the root zone. The vertical spatial distribution of RZSM can be described by using multiple soil layers in budget models. The ISBA model and its variants are widely used SVAT models for RZSM estimation (Calvet & Noilhan, 2000; Montaldo & Albertson, 2001, 2003; Montaldo et al., 2007; Montaldo et al., 2001; Sabater et al., 2007; Sabater et al., 2008). A two-layer ISBA model consists of a near surface layer and a root zone soil layer. The water balance equation of the root zone soil layer in ISBA model is similar to (1), and the near

face soil moisture is related to the RZSM. More details of the ISBA model can be found in Calvet & Noilhan (2000) and Montaldo & Albertson (2001).

2.2.2 Richards Equation

Richards equation (Richards, 1931), which combines the mass conservation equation and Darcy's law, is widely applied in different hydrological models to describe the soil moisture dynamics (Baroni & Tarantola, 2014; Cornelissen et al., 2014; Crawford et al., 2000; De Lannoy et al., 2006; Greve et al., 2013; Jacques et al., 2002; Lü et al., 2010; Mertens et al., 2005; Mertens et al., 2006; Monsivais-Huertero et al., 2010; Paniconi et al., 2003; Rossler & Loffler, 2010; Schaedler, 2007; Starks et al., 2003; Yuan & Liang, 2011; Zhang et al., 2005). In Richards equation, the soil moisture movement is driven by total soil-water potential and root uptake. Richards equation can be written in three forms, namely h-based form, θ-based form and mixed form, as presented by equation (2), (3) and (4) respectively (Subbaiah, 2013):

$$C(h)\frac{\partial h}{\partial t} = \nabla[K(h)\nabla h] - \frac{\partial K}{\partial z} - S \tag{2}$$

$$\frac{\partial \theta}{\partial t} = \nabla[D(\theta)\nabla(\theta)] - \frac{\partial K}{\partial z} - S \tag{3}$$

$$\frac{\partial \theta}{\partial t} = \nabla[K(h)\nabla h] - \frac{\partial K}{\partial z} - S \tag{4}$$

where h is the pressure head; θ is the volumetric water content; t is time; ∇ is the gradient operator; $C(h) = \partial\theta / \partial h$ is the specific moisture capacity of the soil at h; K is unsaturated hydraulic conductivity; z is the elevation head; S is a sink term accounting

for the plant root water uptake; $D(\theta) = K(\theta)/C(\theta)$ is the unsaturated diffusivity. Compared with (4), equation (2) and (3) reduce the number of dependent variables to one, in the form of h or θ. However, (2) and (3) are not completely equivalent. As summarized in Subbaiah (2013), the h-based form suffer from mass non-conservation for the simulation of infiltration in dry and/or highly nonlinear soils but is useful when variably saturated flows and the flows in layered or spatially heterogeneous soils are involved. The θ-based form faces difficulty in mass conservation at the boundaries and the simulation of flows in saturated and layered soils, but is suitable to simulate infiltration processes in dry heterogeneous soil profile and useful for horizontal flow problems (Subbaiah, 2013). To solve Richards equation, it is necessary to characterize the water retention curve and hydraulic conductivity function, and to define the initial and boundaries conditions. Richards equation is highly nonlinear and difficult to solve, owing to the nonlinearity in water retention curve and hydraulic conductivity function. Methods used to solve Richards equation can be classified into three categories (Subbaiah, 2013) including analytical methods, semi-analytical methods and numerical methods. Though numerical methods require intensive computation for the fine temporal and spatial discretization, they do not rely on the linearization of Richards equation and are more flexible to realistically describe natural flow systems under various conditions (Subbaiah, 2013). For these advantages, numerical methods such as finite element method (Hurkmans et al., 2006; Lü et al., 2011b; Paniconi et al., 2003) and finite difference method (Lü et al., 2010; Medina et al., 2014a, 2014b; Mertens et al., 2005; Romano et al., 2011) are more commonly applied. However, numerical methods also suffer from instability problems in certain conditions. Comprehensive reviews on

the solutions of Richards equation as well as the advantages and limitations of these methods are available, and readers are referred to these papers (e.g.Subbaiah (2013)) for more details. For brevity, models based on Richards equation or its approximation are referred to as RE models in the later content.

2.2.3 Comparison between Budget Models and Models Using Richards Equation

Both budget models and RE models are valid approaches to estimate soil moisture dynamics and have been successfully applied at various scales as presented in **Appendix 1**. Compared with RE models, the description of soil water fluxes in budget models is simpler. In budget models, the soil layers are considered as reservoirs receiving and retaining incident water until the corresponding storage capacity is fully filled. This simplifies the processes of infiltration and redistribution. In some budget models, the effects of rain intensity are neglected and the runoff is generated until the vadose zone is saturated (Romano et al., 2011; Sheikh et al., 2009) or the water in soils exceeds the field capacity (Yamaguchi & Shinoda, 2002). The other budget models may apply simple empirical methods to indirectly reflect the effects of rain intensity and infiltration rate on runoff generation (Chen et al., 2011; Nishat et al., 2007; Panigrahi & Panda, 2003; Tavakoli & De Smedt, 2013). In the budget models with consideration of lateral flow, it is often assumed that the subsurface flow is controlled by surface topography (Chen et al., 2011; Sheikh et al., 2009), which may not be valid. In addition to the simplified description of hydrological processes, the vertical distribution of soil moisture is not realistically described. This can be caused by the limited number of soil layers designed in the budget models, the simplified connections between soil layers for the models with multiple soil layers and the assumption that the root zone reservoirs

reach equilibrium instantaneously. For example, in the ISBA model the near surface soil moisture is related to the RZSM through the equilibrium surface volumetric moisture content. This variable depends on RZSM and the soil hydraulic properties and it describes the hypothetical state when the gravity and the capillary forces are balanced (Montaldo & Albertson, 2001). As this approach is based on the vertical contrast in matric potential, it is flawed for layered soils where the relation of matric potential and soil moisture varies with soil types. Montaldo & Albertson (2001) modified the ISBA model by rescaling the RZSM to an “equivalent” RZSM to reduce the effects of this issue. In comparison with budget models, RE models more realistically simulate the soil water dynamics driven by various force such as gravity, pressure and suction.

However, there are also some limitations in the applications of RE models. The numerical schemes of RE models require extensive soil property data which are usually difficult to acquire especially for large-scale estimation. The numerical solutions of Richards equation require intensive computation for the fine temporal and spatial discretization. On the contrary, the budget models are more efficient for large-scale estimation of land surface processes and hence are widely applied when coupled with General Circulation Models (GCMs) (Romano et al., 2011). In addition, RE models are relatively less easy to use, requiring expertise to achieve numerical convergence and stability. The accurate representation of soil water dynamics by RE models heavily relies on the appropriate estimation of effective soil hydraulic parameters at the selected scale (Vereecken et al., 2008). Considering the high spatial variability of soil properties for large-scale estimation, this results in the need to develop scaling methods and

measurement technologies to obtain macroscale system parameters (Romano, 2014; Vereecken et al., 2008).

As the appropriateness of these two types of models is affected by various factors, comparative studies have been carried out to assess the feasibility of budget models and RE models for soil moisture estimation under certain circumstances (Guswa et al., 2002; Romano et al., 2011). Guswa et al. (2002) conducted a study to compare a single-layer budget model (shown in equation (1)) with a RE model which more accurately represent the processes of infiltration, drainage, evaporation and plant uptake, for the predictions of soil and vegetation behaviours in an African savanna. The study showed that the differences in the results of the two models lay in the vertical distribution of soil moisture. When the climate got wetter and the plants were able to compensate for the spatial variability in the soil moisture profile, the match of the two models was improved. If the compensation was small the match was poor. Romano et al. (2011) compared the same budget model (Guswa et al., 2002) with a Richards equation based Soil Water Atmosphere Plant (SWAP) model to test the prediction capacity of the budget model under the conditions relevant to Mediterranean land areas where seasonal changes in rainfall are out-of-phase with the variations in plant transpiration. Their study showed that worse predictions from the budget model occurred when dealing with coarser soils and the field capacity for this soil type was parameterized according to a certain point of the soil water retention curve. These comparative studies indicated that the simple budget models may achieve results similar to those obtained by RE models under specific circumstances but RE models were more flexible to represent soil moisture dynamics under various conditions.

However, the budget model used in these studies is very simple. In the last decades, some advanced SVAT model (e.g. ISBA model) and land surface model (LSM) (e.g. Mosaic model (Berg et al., 2005) based on the water balance equation have also been shown valid for large-scale estimation. The prediction capacity of budget models and RE models for large-scale soil moisture estimation needs further study.

2.2.4 The Common Issues in the Applications of Physically Based Models

Despite the distinctions in different types of PBMs, they all rely on the appropriate model physics for the description of the dominant processes as well as the correct specification of model parameters and initial conditions to achieve accurate soil moisture. Hence, there are some common issues which cause uncertainty in soil moisture estimation by PBMs.

2.2.4.1 Model Physics

In the last decade, the importance of appropriate model physics for soil moisture estimation has been emphasized by several studies (Kumar et al., 2009; Monsivais-Huertero et al., 2010; Walker et al., 2001a). The errors of the models physics can be caused by the inappropriateness in the description of dominant processes, the selection of model dimension and the specification of boundary conditions. The effects because of incorrect description of infiltration and redistribution processes have been shown in the **section 2.2.3**. In addition, the selection of soil hydraulic functions can affect the estimation of soil fluxes (Braun & Schadler, 2005; Schaedler, 2007). The study carried out by Walker et al. (2001a) showed the necessity to include a root water uptake term. Though their model was a good approximation to Richards equation, the model

overestimated the soil moisture at the deeper soil layers for extreme drying events when the root water uptake term was neglected. Xia et al. (2014) evaluated the performance of four land surface models for soil moisture simulation in the North American Land Data Assimilation System phase 2 (NLDAS-2). They reported the low simulation skills of Noah model in the northeast of America, which may be related to a strong constraint on the surface turbulent exchange coefficient. This constraint was applied to improve the estimation of snow water equivalent. However, it also underestimated the evaporation in the northeast of America during spring and early summer, leading to higher total runoff, wetter soils and smaller seasonal change in soil moisture of the top two soil layers. Hence, the performance of PBMs heavily relies on the appropriate description of dominant processes.

To simplify the estimation and reduce the computational cost, most of studies applied one-dimensional (1D) models assuming that the soil water flows occur only vertically. However, this assumption may not be valid in wet soils where lateral flows can occur. The inappropriate selection of model dimension may affect the estimation of soil hydraulics parameters as well as the model performance for soil moisture simulation. Several studies were conducted to apply three-dimensional (3D) models for soil moisture estimation (Cornelissen et al., 2014; Hurkmans et al., 2006; Walker et al., 2002). Walker et al. (2002) applied a distributed 3D soil moisture model coupled with a modified Kalman filter to retrieve soil moisture profile in a 6ha catchment using near surface soil moisture measurements. They found that the differences in the estimates generated by the 1D model and the 3D model were small. This may be because vertical redistribution was more important than lateral redistribution in the Nerrigundah

catchment (Walker et al., 2002). However, De Lannoy et al. (2006) reported that the 1D model failed to capture the large difference in RZSM before and after dry-out when lateral flow occurred. As there is still a lack of 3D soil moisture data to calibrate and validate soil water models (Cornelissen et al., 2014), the effects of model dimensionality are not fully explored.

The inappropriate specification of boundary conditions can cause effects on model performance (Cornelissen et al., 2014; Lü et al., 2011b; Schaedler, 2007). At the bottom of soil moisture models, the gravitational drainage condition is commonly used. For RE models, this is also known as “free drainage” bottom boundary condition (Lü et al., 2010; Medina et al., 2014a; Romano et al., 2011; Sheikh & van Loon, 2007; Wegehenkel, 2005), which depends on the hydraulic conductivity. The study of Lü et al. (2011b) showed that the free drainage boundary condition may cause the underestimation of soil moisture. Using observed heads to specify bottom boundary condition generated better simulation results (Lü et al., 2011b). The free drainage boundary condition assumes that the water table is deep enough and the groundwater does not affect the soil moisture variations. This assumption may not be valid for the regions with shallow water table. However, the data reflecting the realistic bottom boundary conditions are usually limited. Some studies suggested the integration of soil water model and groundwater model (Maxwell & Miller, 2005; Yuan & Liang, 2011; Zeng & Decker, 2009). Maxwell & Miller (2005) coupled the Common Land Model (CoLM) with the ParFlow groundwater model to simulate hydrological processes at Usadievskiy Watershed, Valdai, Russia and achieved more realistic RZSM dynamics, compared with those of the uncoupled model. To reduce the effects of the inappropriateness in free drainage

condition for the Coummunity Land Model version 3.0 (CLM3.0), Zeng & Decker (2009) developed a new bottom boundary condition using the equilibrium soil moisture distribution, which leads to the coupling between surface water and groundwater. Yuan & Liang (2011) evaluated the application of the conjunctive Surface-Subsurface Process (CSSP) model over the contiguous United States. In comparison with CoLM and CLM3.5, CSSP model had critical advantages that it incorporated a scalable representation of subgrid topographic control on the dynamics of soil moisture, an explicit treatment of flow interaction between surface and subsurface and the comprehensive land surface boundary conditions based on the beast available observations (Yuan & Liang, 2011). The results showed that CSSP model was superior to CLM3.5 in representing the seasonal and interannual variations of RZSM.

2.2.4.2 The Determination of Soil Hydraulic Parameters

To determine soil hydraulic parameters, one can: 1) refer to the values used in literatures; 2) conduct field measurement or laboratory experiment; 3) implement calibration manually or using optimization algorithms; 4) apply pedotransfer functions. The effective parameters of the selected models and the measured parameters may be different. This can be caused by several factors (Mertens et al., 2005) including scaling issue, parameter measurement techniques, the achievement of global optimized parameters, the uncertainty inherent in the modeling exercises. In addition, the data of soil properties are usually difficult to acquired, especially for large-scale soil moisture estimation, due to the high spatial heterogeneity of land surface and the high cost of field measurement. Therefore, calibration is usually applied to achieve the required model parameters. As manual calibration is more suitable for the cases with a small

number of parameters, the automatic optimization methods are commonly used for parameter calibration. As summarized in Vereecken et al. (2008), various global optimization methods have been applied to estimate the unsaturated soil hydraulic properties, including the annealing-simplex method, genetic algorithms, multilevel grid sampling strategies, ant colony optimization, shuffled complex methods, multi-objective search methods, and simultaneous multi-method genetically adaptive optimization. Unlike the local optimization methods (e.g. Levenberg-Marquardt method), which may be stuck in a local optimum, the global optimization methods can achieve global optimum from all the possible solutions. Despite the advances in the applications of optimization methods, the difficulties in determining soil hydraulic parameters still exist, due to the large number of uncertain parameters (Mertens et al., 2005) and the correlation between the parameters (Medina et al., 2014a). The sensitivity analysis is commonly used to identify the influential parameters in soil water models. Many pedotransfer functions (PTFs) have been developed to relate the soil hydraulic functions to soil properties which can be more easily measured (Vereecken et al., 2010). Most PTFs are based on the van Genuchten–Mualem (VGM) soil hydraulic functions, such as those in ROSETTA (Schaap et al., 2001). For further details about the PTFs using VGM model, the reader is referred to the review carried out by Vereecken et al. (2010).

2.2.4.3 The Initialization of Soil Moisture Profile

For the lack of prior knowledge of the soil profile state and soil properties, the model performance for soil moisture estimation can be affected by the poor initialization (Lü et al., 2011b; Schaedler, 2007; Walker et al., 2001a; Zhang et al., 2005). Schaedler (2007) applied the VEG3D soil–vegetation model for soil moisture simulation over a period of

more than 9 months for a site in Germany. They investigated how the model adjust to the initial soil water content taken from the European Centre for Medium-Range Weather Forecasts 40-Yr Reanalysis (ERA-40) dataset and found that it took about five months for the model to “forget” the poor initialization. Such long adjustment time may be related to the large biases of initial soil moisture at the deeper soil layers (Schaedler, 2007). In the last decade, studies have shown that assimilating soil moisture observations can reduce the effects caused by the incorrect initialization (Lü et al., 2010; Medina et al., 2014b; Montaldo et al., 2001; Walker et al., 2001a; Walker et al., 2002; Zhang et al., 2005). The advantages of applying data assimilation methods will be further discussed in the next section.

2.2.5 The Integration of Physically Based Models and Data Assimilation Methods

Data assimilation (DA) methods are used for merging the complementary information from observations with estimates from models to obtain optimal values of the geophysical variables of interest (Reichle, 2008). The DA methods that have been applied for RZSM estimation include variational assimilation methods (Calvet & Noilhan, 2000; Sabater et al., 2007; Sabater et al., 2008), Kalman filtering techniques (Chen et al., 2011; Chirico et al., 2014; De Lannoy et al., 2007a; De Lannoy et al., 2007c; Han et al., 2012; Kumar et al., 2009; Lü et al., 2011a; Medina et al., 2014a, 2014b; Monsivais-Huertero et al., 2010; Montaldo et al., 2007; Nagarajan et al., 2011; Nearing et al., 2013; Walker et al., 2001a, 2001b; Walker et al., 2002; Zhang et al., 2005), Newtown nudging methods (Hurkmans et al., 2006; Paniconi et al., 2003), direct insertion (Han et al., 2012; Lü et al., 2011b; Zhang et al., 2005), H_∞ filter (Lü et al., 2010), particle filter (Nagarajan et al., 2011) and hierarchical Bayesian network (Qin et al., 2013). As the Kalman

filtering techniques are most commonly used, a brief introduction of it will be first given in this section. Then the author will review the applications of DA methods for RZSM estimation in the last decade. A discussion will also be presented with respect to the issues in the applications of DA methods.

2.2.5.1 Kalman Filtering Techniques

The Kalman Filter (KF), also known as sequential DA, is a recursive filter applied for the estimation of the state in a dynamic model, using a series of measurements with noises involved. Though it was originally proposed by Kalman (1960) for linear systems, several variants have been developed for the applications in non-linear systems. The ensemble Kalman Filter is most widely used to deal with the high nonlinearity in soil water models. The Kalman Filter assumes that the noises of the model and the noises of the measurements follow the Gaussian distribution. According to Bayesian theory, model state $\boldsymbol{x}_n$ evolves over time. The KF can generate a posterior estimate of the first two moments of the state distribution: the mean $\hat{\boldsymbol{x}}_n = E[\boldsymbol{x}_n]$ and the covariance of the state distribution (Chirico et al., 2014). When the measurements $\boldsymbol{y}_n$ are available, KF can optimize a prior model estimate $\hat{\boldsymbol{x}}_n^-$ into the posterior estimate $\hat{\boldsymbol{x}}_n$, also known as update, by the following equations (Chirico et al., 2014):

$$\hat{\boldsymbol{x}}_n = \hat{\boldsymbol{x}}_n^- + \boldsymbol{K}_n(\boldsymbol{y}_n - H_n(\hat{\boldsymbol{x}}_n^-)) \tag{5}$$

where $\boldsymbol{K}_n$ is the Kalman gain at time step n, defining the weights of the prior model estimates and measurements. H_n is the measurement model which is used to relate the current model state to the measurements. The variants of KF use different methods

to obtain the covariances, which contain the error information of the dynamic system and measurements and are used for the calculation of Kalman gain. To handle the nonlinearity in dynamic systems, ensemble Kalman Filter generates an ensemble prediction by applying perturbations following Gaussian distribution for the noises of the model and measurements (Chirico et al., 2014). In this way, it solves the equation for the time evolution of the model-state probability density through the Markov Chain Monte Carlo method (Evensen, 2009). Given an ensemble with L members, the updated state is calculated by the mean of the updated ensemble predictions (Chirico et al., 2014):

$$\boldsymbol{x}_{n,i} = \boldsymbol{x}^{-}_{n,i} + \boldsymbol{K}_n(\boldsymbol{y}_{n,i} - \boldsymbol{H}_n(\boldsymbol{x}^{-}_{n,i})) \qquad i = 1...L \qquad (6)$$

The KFs that have been used to retrieve RZSM include the standard Kalman Filter (SKF) (Chirico et al., 2014; Walker et al., 2001a), the extended Kalman Filter (EKF) (Chirico et al., 2014; Lü et al., 2011a; Sabater et al., 2007), the ensemble Kalman Filter (EnKF) (Chen et al., 2011; Chirico et al., 2014; De Lannoy et al., 2007a; De Lannoy et al., 2007c; Han et al., 2012; Kumar et al., 2009; Medina et al., 2014a, 2014b; Monsivais-Huertero et al., 2010; Montaldo et al., 2007; Nagarajan et al., 2011; Nearing et al., 2013; Sabater et al., 2007; Zhang et al., 2005) and the unscented Kalman Filter (UKF) (Chirico et al., 2014; Medina et al., 2014a, 2014b). As SKF is only suitable for linear systems, it is seldom applied to retrieve RZSM because of the high nonlinearity of subsurface systems. To meet the linearity assumption of SKF, Walker et al. (2001a) proposed a vertical distribution factor in the simplified soil moisture model to approximate the typical moisture retention relationships with no nonlinear term involved.

The results showed that the retrieval by applying SKF was just as good as the model description of soil moisture dynamics and its calibration (Walker et al., 2001a). EKF adopts the tangent linear operator (Jacobian) to linearize the dynamic system and observation model (Evensen, 2003). EnKF was developed to overcome the two major drawbacks in EKF: 1) EKF suffers from instability because of the linearization; 2) the large computational cost limits the applications of EKF to low dimensional dynamical models (Evensen, 2009). Comparative studies indicated that EKF may not provide accurate solutions for all conditions (e.g. coarser textured soils) (Chirico et al., 2014) and EnKF was more efficient than EKF (Sabater et al., 2007). Instead of using ensemble predictions, UKF applies a set of sigma points whose mean and covariance are equal to those of the model state. The calculation of prior model estimates and the required convariances are based on the sigma points transformed through the nonlinear function. In this way, UKF reduces the difficulty in transforming the probability density function through the nonlinear function. Unlike the uncertainty found in the determination of the ensemble size in EnKF, the number of the sigma points are clearly defined by the system dimension, but the identification of these sigma points requires much computation (Chirico et al., 2014). In the recent study conducted by Chirico et al. (2014), UKF was shown to be as feasible as EnKF for the application on the models of small dimensionality (1D Richards equation).

2.2.5.2 The Applications and Limitations of Data Assimilation Methods

Many studies have proven that the applications of DA methods can improve soil moisture estimation, reducing the effects from the inappropriate initialization of soil moisture profile. Walker et al. (2001a) tested a SKF on a simplified soil moisture model

to deal with the poor initialization of soil moisture profile and found that SKF quickly brought the estimation on track for soil layers of the top 235mm. They applied similar methods for three-dimensional soil moisture profile retrieval and found that the results were independent of initial conditions (Walker et al., 2002). Lü et al. (2010) also found the estimation was insensitive to the initial soil moisture when using a H_∞ filter on an 1D Richards equation and the effects of poor initialization were eliminated after 5-7 days. However, several studies indicated that the improvements from the DA methods can be impacted by the uncertainty in the model parameters. Calvet & Noilhan (2000) reported that the parameter reflecting vegetation coverage had an impact on assimilation efficiency because the soil moisture estimation by ISBA model was sensitive to this parameter. Han et al. (2012) recommended well calibration of model parameters to better characterize the soil profile properties before implementing DA. To reduce the effects of the uncertain wilting point in ISBA-A-g_s model during assimilation, Sabater et al. (2008) reset the wilting point when the RZSM was below it.

With respect to the impacts of uncertain model parameters on DA, studies have been carried out to dynamically adjust parameters during simulation. Montaldo & Albertson (2003) updated the saturated hydraulic conductivity, which was a sensitive parameter, according to the observations of persistent bias in the simulated RZSM. As the update interval of the parameter was different from that of the model state, this method was referred to as multi-scale assimilation. The study found that multi-scale assimilation provided great improvements, superior to the methods with state update only. The improvements were also reported, when the particle swarm optimization (PSO) algorithm was used for the calibration of soil hydraulic parameters in the DA systems

using direct insertion assimilation scheme (Lü et al., 2011b) and EKF (Lü et al., 2011a). In addition to different dynamic calibration methods, DA methods have also been utilized to update soil hydraulic parameters (Medina et al., 2014a; Monsivais-Huertero et al., 2010; Nagarajan et al., 2011). However, these studies demonstrated that there were difficulties in retrieving a few soil hydraulic parameters simultaneously. Both of the studies carried out by Monsivais-Huertero et al. (2010) and Nagarajan et al. (2011) estimated soil moisture using the LSP-DSSAT model and considered the uncertainty caused by the four parameters including porosity, saturated hydraulic conductivity, air entry pressure and pore-size index. These two studies indicated that neither the particle filter (PF) nor the EnKF could achieve the parameter convergence to the true values when implementing simultaneous state-parameter estimation, except for the porosity which is highly correlated to RZSM. Medina et al. (2014a) tested dual standard-unscented Kalman filter (DSUKF) and dual ensemble Kalman filter (DEnKF) on an 1D Richards equation for simultaneous state-parameter estimation, with respect to saturated hydraulic conductivity, and the two empirical scale shape parameters α and n. They found the difficulty in retrieving saturated hydraulic conductivity because of its strong correlation with the other parameters.

Regardless the advances in DA methods, the biases from the physics of the PBMs are still hard to correct by DA. Walker et al. (2001a) found the model errors caused by the lack of root water uptake term and concluded that appropriate model physics were more important than a suitable update frequency. Kumar et al. (2009) demonstrated that the vertical coupling strength of the soil layers in the models had impacts on the DA scheme to propagate the surface information into root zone. This finding was confirmed in the

study of Chen et al. (2011) who reported that the insufficient vertical coupling in the SWAT model resulted in limited updating of the deep soil moisture, independent to the model parameterization. Lü et al. (2011a) found the assimilation system was only able to retrieve the RZSM of the top 50cm when the assumption of homogeneity for the soil between the depth 5cm and 100cm was in conflict with the actual situation. Considering the effects from model errors, some researchers proposed to include a bias correction process in the assimilation system (De Lannoy et al., 2007a; De Lannoy et al., 2007c). For an EnKF with bias correction, the bias-corrected model state was generated by adding a posteriori bias estimate to the posteriori state estimate (De Lannoy et al., 2007a). De Lannoy et al. (2007c) further investigated the applications of different algorithms in EnKF to estimate the model state and forecast biases. Their result showed that with the bi-weekly assimilation of soil moisture observations for the entire profile, the proposed method reduced the root mean squared error (RMSE) by about 60%, compared with EnKF without bias correction. However, the selection of the algorithms depended on the nature of the model bias and the need of users, and there were also some limitations in the proposed method (De Lannoy et al., 2007c). The applications of DA methods to correct errors caused by inappropriate model physics need further investigation.

In addition to these influencing factors, the appropriate selection of DA methods plays an important role to achieve good model performance. Comparative studies have been carried out to investigate the feasibility of different DA methods. Though direct insertion, which directly replaces the model state with available observations, is simple and easy to used, a few studies found that the improvements from it were very limited (Han et al.,

2012; Walker et al., 2001b; Zhang et al., 2005). The direct insertion relies on the infiltration and exfiltration processes represented in the soil moisture model to propagate the information of surface observations into the deeper soil layers, which limited its effects on more around the observation depth (Walker et al., 2001b). On the contrary, EnKF has shown its superiority not only over the other Kalman filters mentioned in **section 2.2.5.1** but also some other types of assimilation methods such as direct insertion (Han et al., 2012; Zhang et al., 2005) and PF (Nagarajan et al., 2011). Though the DSUKF was more effective than DEnKF for the case with ensemble size of the same order, DEnKF achieved higher accuracy when given larger ensemble size (Medina et al., 2014a). However, in the study of Sabater et al. (2007) EnKF was outperformed by the one-dimensional variational DA scheme. This may be explained by the limited nonlinearity involved in the system and the uncertain inflation factor in EnKF which was used to avoid the collapse of the ensemble (Sabater et al., 2007). Nearing et al. (2013) demonstrated that the EnKF did not use all the information involved in observations and suggested to develop directed assimilation strategies based on Bayes' Law. Though Qin et al. (2013) successfully implemented DA using a hierarchical Bayesian network algorithm, similar studies are still rare.

It can be detected in the content above that the improvements for the estimation are related to the proper implementation of assimilation system such as appropriate initialization (e.g. covariances) and selection of parameters in the assimilation methods. As EnKF applies the Monte Carlo method to estimate the error covraiances, the ensemble size of EnKF can directly affect the model performance. Though larger ensemble size tends to achieve higher accuracy, however, the differences would be

small when the ensemble size is increased to a certain number. Furthermore, the large ensemble size leads to extensive computational cost. Therefore, this issue has been widely investigated in the studies using EnKF by comparing the results generated by the filters with different ensemble size (Medina et al., 2014a; Monsivais-Huertero et al., 2010; Zhang et al., 2005). The optimal ensemble size tends to be related to the complexity of the studies. For most of the studies using synthetic data, the optimal ensemble size is small, varying from 12 to 50 (Chen et al., 2011; Chirico et al., 2014; Kumar et al., 2009; Medina et al., 2014a; Zhang et al., 2005). For most of the studies using observations, the optimal ensemble size is much larger and varies in a wider range, between 100 and 500 (Chen et al., 2011; Han et al., 2012; Monsivais-Huertero et al., 2010; Montaldo et al., 2007; Nagarajan et al., 2011). Other assimilation methods are also affected by the parameters in the filters, such as the parameter G in Newtonian nudging which defines the relative strength of the nudging term (Hurkmans et al., 2006; Paniconi et al., 2003), the weighting coefficient and the error attenuation parameter in H_∞ filter (Lü et al., 2010), the number of particles in PF (Nagarajan et al., 2011) and the assimilation window in the variational approach (Sabater et al., 2007). It is still not clear how to optimize these parameters but is usually heuristically chosen. In addition, Medina et al. (2014a) pointed out that the initialization of the covariances for the model parameters and states demands more caution when implementing dual state-parameter estimation through DA. However, usually there is short of prior knowledge to initialize these convariances. These factors would cause uncertainty in DA systems.

The update frequency also places effects on the performance of the DA systems and many efforts have been given to explore factors impacting the determination of

appropriate update frequency. The update frequency differs between studies, such as daily updating (Chen et al., 2011; Han et al., 2012), an update interval of 3 days (Monsivais-Huertero et al., 2010; Nagarajan et al., 2011) and 6 days (Hurkmans et al., 2006). Walker et al. (2002) demonstrated that the appropriate update frequency was related to the errors in model physics and forcing data. This was confirmed by the findings in Han et al. (2012). They reported that the update interval larger than 4 days provided limited improvements for the simulation at the upper soil layers (5 and 20cm depths) but the update interval of 1-2 days produced worse simulation for deeper layers (40 and 60cm depths). They explained that this may be due to the model errors (Han et al., 2012). When the bias correction was included in the assimilation system, the required update frequency was smaller (e.g. bi-weekly updates) (De Lannoy et al., 2007a; De Lannoy et al., 2007c). The required update frequency also depends on the need of users and the adopted assimilation methods. Zhang et al. (2005) showed that it took the EnKF about 16 hours and 15 days to retrieve the full soil moisture profile with hourly and daily updates, respectively. They pointed out that frequent updates are required if one wants to achieve full soil moisture profile retrieval in a short time. They also found that the direct insertion was able to retrieve the soil moisture profile within 12 days with hourly updates but failed to realize full retrieval with daily updates (Zhang et al., 2005). This finding indicates the appropriate update frequency is relevant to the performance of the selected DA methods. De Lannoy et al. (2007a) suggested to apply less intensive assimilation when the assimilation shows adverse impacts on soil layers outside the selected single assimilation layer. This study suggested the impacts of assimilation depth on the determination of update frequency. To summarize, these

studies indicated that the appropriate update frequency depends on several factors such as model errors, the need of users, the adopted assimilation methods and the assimilation depth.

As the soil moisture data acquired from the remote sensing techniques are limited to the top a few centimeter, the effects of the soil depth at which the assimilation is carried out have also been investigated (De Lannoy et al., 2007a; Medina et al., 2014b; Monsivais-Huertero et al., 2010; Zhang et al., 2005). De Lannoy et al. (2007a) found that the assimilation in the surface layer was less influential than the assimilation in the other layers . The assimilation of observations for the entire soil profile may tend to achieve better results (Monsivais-Huertero et al., 2010). However, the effects from the assimilation depth on the DA system are small (Monsivais-Huertero et al., 2010; Zhang et al., 2005), unless the surface-subsurface decoupling occurs (De Lannoy et al., 2007a; Medina et al., 2014b).

Though the applications of DA methods can help to reduce the effects from the uncertainty in initial model state and model parameters, improving the performance of PBMs for soil moisture estimation, the influencing factors mentioned above complicate its implementation for actual practice.

2.3 Statistical and Data-Driven Models (SDDMs)

In comparison to PBMs, statistical and data-driven models do not explicitly describe the physical processes which govern soil moisture dynamics, but rely on mathematical methods to relate the given inputs to the target values (e.g. observations). These models are also known as input-output models. In the last decade, the studies using

SDDMs are much fewer than those applying PBMs, a summary of the selected advances of this category is presented in **Appendix 2**. In this category, the ANNs, especially MLPs, and the support vector machines (SVMs) are most widely applied. In this section, a brief introduction will be given to these two methods. To gain insight into the different performance of these two methods, the findings in the related comparative studies will be summarized. Then some common issues and advances that were shown in the studies using SDDMs for soil moisture estimation will also be discussed.

2.3.1 Multilayer Perceptrons

In the last two decades, various types of ANNs, especially MLPs, have been widely applied in hydrology (Abrahart et al., 2012; ASCE, 2000a, 2000b). MLPs had the ability to capture the nonlinearity in systems (Abrahart & See, 2007). Given a three-layer perceptron with only one output neuron, its configuration includes one input layer, one hidden layer, one output layer, neurons in each layer, weights given to the connections between the neurons and a bias given to each of the former two layers. The input variables involved in the model are represented by the neurons in the first layer. The weighted inputs and a bias are added up and this sum is passed to the corresponding neuron in the hidden layer. The output of each hidden neuron is generated when the corresponding sum passes through the activation function built in the hidden neuron. Using the sum of the weighted outputs of the hidden neurons and the bias of the hidden layer, an estimate of the whole network is then generated after that sum passes through the activation function within the output neuron. This process can be represented by the following equation (Hagan et al., 2002):

$$y = f(\boldsymbol{W_2}\boldsymbol{g}(\boldsymbol{W_1}\boldsymbol{X} + b_1) + b_2) \quad (7)$$

where *y* is the output of the network; *f* and *g* are the activation functions for the output and hidden neurons, respectively; $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are the weights for the input variables and the outputs of the hidden neurons, respectively; $\boldsymbol{X}$ is the inputs; b_1 and b_2 are the biases of the input layer and hidden layer, respectively. To capture the nonlinearity of a system, MLPs relies on the nonlinear activation function used in the neurons, such as the sigmoid function (Elshorbagy & Parasuraman, 2008; Yang et al., 2009), and the tangent sigmoid function (Elshorbagy et al., 2010a; Gill et al., 2006; Kornelsen & Coulibaly, 2014a; Kornelsen & Coulibaly, 2014b; Wu et al., 2008). Considering each input/target (observation) pair as one example for the MLPs to learn, an error exists between the model output and the corresponding target for each example. The ability for MLPs to represent a system can be enhanced, after being presented a particular number of examples. This process, also known as training process, is accomplished by adjusting the weights of the connections between the neurons to minimize the errors between the outputs and the targets. The training process is implemented through the selected training algorithm such as Levenberg-Marquardt optimization and Bayesian regularization. More details about the basic theory of ANNs and MLPs will be presented in **Chapter 3**.

2.3.2 Support Vector Machines

The SVMs were statistical learning tools developed by Vapnik and the colleagues in the early 1990s originally for the applications in classification and then extended for regression (Vapnik, 1995, 1998). The development of a SVM for regression is to

estimate a functional dependency, $f(\boldsymbol{x})$, between inputs $\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x}_L\}$ $\boldsymbol{x} \in \boldsymbol{R}^K$, and outputs $\{y_1, y_2, ..., y_L\}$ $y \in \boldsymbol{R}$ taken from a set of independent and identically distributed observations, using the regularized function (Gill et al., 2006):

the minimization $$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{L}\left(\xi_i + \xi_i^*\right)$$

is subject to $$\left\{\begin{array}{l} y_i - \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - b \le \varepsilon + \xi_i \\ \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{array}\right\} \quad (8)$$

$$f(\boldsymbol{x}) = \sum_{j=1}^{K} w_j x_j + b \quad (9)$$

where K is the dimension of $\boldsymbol{x}$ and $\boldsymbol{w}$; L is the sample size; b is the bias; ε is used to define the bounds within which the errors are ignored, hence also known as ε-insensitive bounds; ξ_i and ξ_i^* are slack variables measuring the cost of the errors on the sample points outside the ε- insensitive bounds, C is a parameter determining the trade-off between the complexity of function $f(\boldsymbol{x})$ and the tolerance of errors. Usually, (8) is solved in dual form (Gill et al., 2006):

the maximization

$$W(\alpha^*, \alpha) = -\varepsilon\sum_{i=1}^{L}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{L}(y_i(\alpha_i - \alpha_i^*)) - \frac{1}{2}\sum_{i,j=1}^{L}(\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

is subject to $$\sum_{i=1}^{L}(\alpha_i^* - \alpha_i) = 0, \qquad 0 \le \alpha_i,\ \alpha_i^* \le C \tag{10}$$

and the approximation function is:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)k(\boldsymbol{x}, \boldsymbol{x}_i) + b \tag{11}$$

where α^*, α are Lagrange multipliers. For optimality, the product of the dual variables and constraints should vanish, according to Kuhn-Tucker conditions. α_i^* and α_i vanish if the corresponding points lying inside the ε- insensitive bounds. The points with non-vanishing coefficients are called support vectors. $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the kernel function which implicitly performs a non-linear mapping of the data into a feature space and it follows Mercer's theorem (Vapnik, 1995, 1998). To apply a SVM, it requires to specify a kernel function and its parameter, the insensitive parameter ε and the parameter C.

2.3.3 Comparing Performance of Different SDDMs

Several studies applied both MLPs and SVMs to compare their performance for soil moisture estimation (Deng et al., 2011; Elshorbagy et al., 2010a, 2010b; Gill et al., 2006; Wu et al., 2008; Zhao et al., 2014). A few studies found that SVMs and the variants (e.g. Least Square Support Vector Machines (LSSVMs) and LSSVMs coupled with Particle Swarm Optimization Algorithm (PSO-LSSVMs)) outperformed MLPs (Deng et al., 2011; Gill et al., 2006; Wu et al., 2008; Zhao et al., 2014). Deng et al. (2011) found SVMs had higher generalization ability than MLPs, as the MLPs outperformed SVMs during training while worse results were found during testing. Wu et al. (2008) also pointed out that SVMs can generate stable outputs while the outputs of ANNs are unstable due to

the various initialization. Despite these advantages, it has been noticed that SVMs are sensitive to the selection of kernel function, and the radial base function was proven to be suitable for soil moisture estimation (Elshorbagy et al., 2010b; Wu et al., 2008). In addition, Elshorbagy et al. (2010b) found that MLPs were superior to SVMs for highly nonlinear cases because it was shown that MLPs outperformed SVMs for the prediction of soil moisture at the upper peat layer and lower till layer. Though SVMs generated lower uncertainty due to the narrow range of residuals and overall small root mean squared error but they failed to capture the trend of the soil moisture data resulting in lower correlation coefficient between the estimates and observations (Elshorbagy et al., 2010b). These studies lead to different conclusions about the abilities of MLPs and SVMs to estimate RZSM. More investigation is needed to explore the advantages and limitations of these two methods for RZSM estimation.

For the other comparative studies, Zou et al. (2010) found that MLPs were superior to autoregressive integrated moving average (ARIMA) model for the prediction of averaged soil moisture profile and the soil moisture at 60cm. Elshorbagy & Parasuraman (2008) applied high-order neural networks (HONNs) for its ability to capture high-order correlations and compared the results with those of MLPs. The results revealed that the HONNs achieved relatively higher correlation coefficient than MLPs for some of the estimation. The genetic programming (GP) and evolutionary polynomial regression (EPR) were adopted by Elshorbagy & El-Baroudy (2009) and they found that neither of them showed superiority over the other. The study of Kornelsen & Coulibaly (2014a) demonstrated that MLPs and interpolation methods were more effective for infilling different types of gaps in the soil moisture datasets than EPR

and the other statistical methods. LSSVMs were found to have more stabilities and advantages in the simulation of soil moisture, outperforming not only MLPs but also the adaptive neuro-fuzzy inference system which combines the fuzzy logic theory and neural networks technique (Deng et al., 2011). Zhao et al. (2014) carried out a study for RZSM prediction using four different SDDMs including, PSO-LSSVMs, MLPs, the generalized regression neural networks and wavelet neural networks, with the best simulation achieved by PSO-LSSVMs. To gain insight into the predictive capabilities of data-driven methods in hydrology, a comprehensive study was conducted to compare the performance of different SDDMs, namely MLPs, GP, EPR, SVMs, M5 model trees, K-nearest neighbours (KNN), multiple linear regression and naïve models, for the simulation of evapotranspiration, soil moisture dynamics and rainfall-runoff events (Elshorbagy et al., 2010a, 2010b). The results revealed that MLPs were the most successful models for soil moisture simulation due to the ability to capture the high nonlinearity in the system, resulting in relatively lower errors and higher correlation coefficient. Following MLPs, GP was shown to have ability to adapt the model complexity to the given data, making it not only suitable for modeling soil moisture dynamics but also the other hydrological processes. KNN and M5 model trees also had the potential for soil moisture estimation while EPR may achieve performance close to GP for more linear cases and SVMs only slightly outperformed multiple linear regression method which has the poorest performance.

2.3.4 The Common Issues and Advances in the Applications of SDDMs

When applying SDDMs, there are some common issues similar to those in the implementation of ANNs as summarized in Abrahart et al. (2012), including the

unknown physical rationality and no fixed rules to guide the model development such as architectural design and data handling. As SDDMs simulate the hydrological processes by relating the inputs to the corresponding outputs, selecting input variables strongly related to the output variables is vital to achieve accurate estimates. For the studies summarized in **Appendix 2**, the variables commonly selected as inputs include air temperature, precipitation, relative humidity, wind speed, soil temperature, net or solar radiation, evaporation or evapotranspiration and soil moisture related to the outputs (e.g. soil moisture of the previous time steps or the upper soil layers). The input variables were usually selected based on the prior knowledge of the systems and the target outputs of interest. Some studies identified optimal combination of input variables through specific methods such as the time series analysis (Gill et al., 2006), the correlation analysis (Lamorski et al., 2013), the trial and error method (Elshorbagy & El-Baroudy, 2009; Elshorbagy & Parasuraman, 2008), and the principal component analysis (Zhao et al., 2014). It should be noted that the study of Elshorbagy & Parasuraman (2008) revealed that the soil thermal properties had a strong relation to the corresponding moisture status, suggesting the inclusion of soil temperature as input. Their results also showed that the use of cumulative inputs was able to achieve better simulation than the use of time-lag inputs. This finding was also confirmed in the study of Elshorbagy & El-Baroudy (2009) in which EPR and GP were adopted. Kornelsen & Coulibaly (2014b) used a sensitivity analysis on MLPs which were developed to estimate RZSM with meteorological data, surface soil moisture and soil texture information. They found that the important variables included surface soil moisture, soil clay content, potential evapotranspiration and solar radiation (Kornelsen & Coulibaly,

2014b). There are no well-established rules to guide the design of SDDMs for soil moisture estimation and usually the trial and error method is applied. Studies have suggested applying ensemble prediction using multiple model configurations, modeling techniques and modeling environments to achieve reliable outputs (Elshorbagy & El-Baroudy, 2009; Kornelsen & Coulibaly, 2014b). Few studies aimed to explore the physical rationality of SDDMs.

Despite these unresolved issues, advances have been achieved in the last decades mostly by using hybrid methods based on SDDMs. Gill et al. (2007) combined the SVMs with EnKF and found EnKF greatly reduced the errors and increased the correlation coefficient. Yu et al. (2012) also adopted SVMs coupled with DA methods and they combined EnKF and PF (EnPF) to take the advantages of these two filters (Yu et al., 2012). The results showed that the DA methods improved the prediction and EnPF outperformed EnKF and PF, but the performance of EnPF and PF may be degraded by the large resampling size. Comparing with the PBMs, the applications of DA on SDDMs for soil moisture estimation are still rare. Liu et al. (2008) combined the self-organizing maps (SOMs) with SVMs. The SOMs were used for classifying the data according to the dynamic variation features and the SVMs were applied to generate the prediction of each cluster. The estimates had good agreement with the observations and the hybrid method was superior to either SOMs or SVMs (Liu et al., 2008). To avoid the MLPs converging to local optimum, Huang et al. (2011) used the genetic algorithm to train the networks and achieve favourable precision.

2.4 Large-Scale Root Zone Soil Moisture Estimation

For large-scale RZSM estimation, the inevitable issue is that the land surface and climatic conditions are highly heterogeneous. This issue requires the flexibility of the selected models to capture the RZSM dynamics under various conditions. In the category of PBMs, various SVAT models and LSMs have been developed for large-scale simulation. Many modules simulating different processes are built in these models to describe the situations under various climatic types and land surface conditions, while watershed models may neglect some less important processes according to the characteristics of the study areas. Therefore, usually the PBMs used for large-scale estimation have high complexity. Compared to SDDMs, these models provide a clear relation between the inputs and outputs. This characteristic enables their application not only to understand the dynamic interactions of land surface and atmosphere, but also to achieve other variables from the corresponding modules, in addition to the RZSM. These advantages make PBMs more popular than SDDMs. The North American Land Data Assimilation System (NLDAS) (Xia et al., 2014), the Global Land Data Assimilation System (GLDAS) (Rodell et al., 2004) and ERA-Interim/Land (Balsamo et al., 2015) have used different LSMs to produce optimal values of several land surface states and fluxes. For SDDMs, there is a need to build a few models to achieve different variables of interest. However, when DA is implemented with PBMs, this advantage may be weakened. The water balance of the system may be broken up by DA (Han et al., 2012) and sometimes the large increments, which is used to optimise the soil moisture estimates during the implementation of assimilation, can result in unrealistic fluxes (De Lannoy et al., 2007c). In addition, the high complexity of LSMs makes the model

physics become indistinct and some anomalies in the outputs are usually hard to interpret.

The various modules, which used to simulate different processes governing energy and water dynamics, cannot ensure the flexibility of PBMs. Studies found that the LSMs in NLDAS may suffer from large errors in some regions (Xia et al., 2014) and the improvements from DA in GLDAS are sensitive to the regional climate and soil types due to the errors of model physics (Kumar et al., 2009). In terms of SDDMs, it has been demonstrated that they can be superior to some PBMs, such as system dynamics watershed model (Elshorbagy & Parasuraman, 2008) and HYRDUS-1D (Lamorski et al., 2013), for the applications in small areas. However, as empirical models SDDMs heavily rely on the given training data, which may lead to poor simulation for the cases outside the range of the training data. Training the SDDMs with soil moisture profiles of different areas may increase their flexibility. Gill et al. (2006) trained the SVMs and MLPs using the data from 10 stations lying in Little Washita River Experimental Watershed with an area of 610 km^2, and tested the models with data for an independent station in the watershed. Their prediction of soil moisture had good agreement with the observations. Kornelsen & Coulibaly (2014b) trained the MLPs with the data of different soil moisture profiles generated by HYDRUS-1D model using the forcing data from the lower Great Lakes region. They found that the MLPs were able to well represent the soil moisture dynamics of the independent testing sites from the same region, when the HYDRUS-1D estimates were close to the observations. These two studies indicate the flexibility of SDDMs for the estimation in a geographic region when the models are trained with different soil moisture profiles. SDDMs do not rely on many physical assumptions nor

require prior solution structures. This helps to avoid the impacts from the potential errors of model physics. Hence, SDDMs may have high potential to be flexible tools. However, the applications of SDDMs for large-scale RZSM estimation are still limited. Further investigation is needed to explore the feasibility of SDDMs to estimate RZSM for larger areas.

In the last decade, many large observation networks have been built to acquire in-situ soil moisture observations (Dorigo et al., 2011; Ochsner et al., 2013). The large amount of soil moisture data increases the possibility of applying SDDMs for large-scale estimation. As the soil moisture data collected by remote sensing techniques are limited to the top a few centimeters, the RZSM observations can only be acquire by in-situ measurement. Therefore, the estimates of RZSM by SDDMs, which are calibrated (trained) by in-situ observations, may be limited to point scale. These estimates may not reflect the spatial distribution of RZSM due to the high spatial heterogeneity of land surface. However, the point-scale RZSM information is still very useful. The in-situ RZSM observations have been widely used to calibrate and evaluate soil moisture models and remote sensing techniques. Though many large soil moisture observation networks have been built, the density of the networks may not be high and some regions still lack of in-situ observations. To fill in these gaps, it is necessary to use modeling methods to estimate soil moisture dynamics in the low-density areas and ungauged areas. In addition, approaches have been developed to scale up point measurements (Cosh, 2002; Crow et al., 2012; De Lannoy et al., 2007b; Teuling et al., 2006).

In addition to the flexibility, the models for large-scale RZSM estimation are desired to be efficient and easy to use. From this aspect, SDDMs have great advantages of high practicality and computational efficiency. First, SDDMs require less computation. As mentioned before, models using Richards equation usually apply numerical solutions which are computationally intensive, involving fine temporal and spatial discretization. Moreover, the large number of modules included in PBMs, the calibration of extensive model parameters and the implementation of DA further increase the computational cost. On the contrary, SDDMs are computationally efficient once the models are developed. Second, SDDMs have less demand of soil properties data. As summarized in **Appendix 2**, most of the studies using SDDMs did not require extensive soil properties data. Kornelsen & Coulibaly (2014b) found the soil texture data provided useful information to characterize RZSM dynamics while the addition of soil water retention parameters did not improve the performance of MLPs. Third, SDDMs have high adaptability for various types of data and input configuration (Ghedira et al., 2004). When ancillary data are available, it is usually difficult for PBMs to utilize these data while it can be easily implemented in SDDMs.

Despite the differences between these two types of models, they are all affected by the coupling strength between surface and deeper layers of the actual soil moisture profile. Walker et al. (2002) indicated that the RZSM were not able to be retrieved when the vertical decoupling occurred. As all modelling techniques rely on the connection between the surface and subsurface to retrieve RZSM using surface information, the surface-subsurface decoupling will become an inevitable challenge. However, surface-

subsurface coupling strength in nature for large areas is still unknown (Kumar et al., 2009).

2.5 Summary and Conclusions

According to the degree of representation of the involved physical processes, the soil moisture models can be classified into two categories, namely physically based models (PBMs) as well as statistical and data-driven models (SDDMs). In the category of PBMs, most of the soil moisture models are based on the principle of water balance (budget models) or the application of Richards equation (RE models) and its approximation. Budget models usually have simpler description for the relevant processes and the structure of the subsurface soil layers but they are more computationally efficient. Comparative studies indicated that the simple budget models may achieve results similar to those of RE models under specific circumstances but RE models were more flexible to represent soil moisture dynamics under various conditions (Guswa et al., 2002; Romano et al., 2011). Advanced budget models may be comparable to RE models but it needs further investigation. Regardless of these differences, there are common issues affecting the simulation by PBMs including the errors in model physics, the uncertainty of soil hydraulic parameters and the poor initialization of soil moisture profile.

The applications of data assimilation (DA) methods are effective for reducing the effects from poor initialization of soil moisture profile, bringing improvements to the performance of PBMs. Therefore, the integration of DA and hydrological models has been considered as the most promising approach (Kumar et al., 2009). Among various

DA methods, Kalman filtering techniques, especially the ensemble Kalman filter (EnKF), were most commonly used. The improvements from DA depend on the selected methods and the appropriate implementation such as: selecting parameters in the DA methods, initializing the required covariances and specifying suitable update frequency. Regardless of the advances, there are still difficulties for DA to reduce the impacts from errors in the parameters and model physics. Hence, considering the issues detected in PBMs, the errors in model physics and parameters require attention, for the applications of PBMs and DA to retrieve RZSM.

The studies adopting SDDMs are much fewer than those using PBMs. In this category, multilayer perceptrons (MLPs) and support vector machines (SVMs) are widely applied. Several studies showed that SVMs outperformed MLPs (Deng et al., 2011; Gill et al., 2006; Wu et al., 2008; Zhao et al., 2014). However, opinion is divided and it was found that MLPs were superior to SVMs for its ability to capture the high nonlinearity in the subsurface system (Elshorbagy et al., 2010b). The capacity of these two models needs further investigation. Advances can be detected in the applications of SDDMs and they are usually achieved by using hybrid methods. However, there are some unresolved issues in SDDMs including the unknown physical rationality and no fixed rules to guide the model development such as architectural design and data handling.

For large-scale RZSM retrieval, PBMs are more popular. They provide a clear relation between the inputs and outputs, which can help to understand the dynamic interactions of land surface and atmosphere. In addition, they can also be used to achieve other variables in specific processes but this advantage may be weakened when DA is implemented. It is hard for PBMs to avoid the errors in model physics which reduce their

flexibility for large-scale RZSM retrieval. Furthermore, the complex model physics make the anomalies in the outputs difficult to interpret. On the contrary, SDDMs do not rely on many physical assumptions or require prior solution structures. The findings in Gill et al. (2006) and Kornelsen & Coulibaly (2014a) suggests the flexibility of SDDMs for the applications in large areas. Moreover, SDDMs have high practicality due to the lower cost of computation, the much less demand of soil properties data and the high adaptability for various types of data and input configuration (Ghedira et al., 2004). Hence, SDDMs have high potential for large-scale estimation but have not been fully investigated.

Therefore, in this study, the focus of the author is to explore the feasibility of SDDMs to estimate RZSM for a large area. Specifically, MLPs are selected among various SDDMs for its proven high ability of nonlinear input-output mapping and wide applications in hydrology. Though SVMs may outperform MLPs in some cases, the model performance is affected by the selected kernel function and the model parameters which lack fixed rules to guide the selection. On the contrary, the design of MLPs has been more widely studied. The basic theory of ANNs and MLPs as well as the development of the models will be presented in **Chapter 3**.

Chapter 3

Methodology

This chapter will present a brief overview of the Artificial Neural Networks technique and Multilayer Perceptrons. The study area and the data used in this study will also be described. Then a full introduction will be given to the development of the models including the selection of input variables, the design of the models and the two experiments. The evaluation criteria for the models will also be presented.

3.1 Artificial Neural Networks (ANNs) and Multilayer Perceptrons (MLPs)

3.1.1 Introduction of ANNs

Artificial neural networks are machine learning tools inspired by biological neural systems. To some degree, they imitate the structure and learning process of the human brain which is highly complex and nonlinear but efficient to process different information (Haykin, 1999). The brain is made up of a large number of structural constituents—neurons. On one hand, an artificial neural network resembles the brain in the aspect that it consists of simple processing units which are conceptual models of neurons. On the other hand, it acquires knowledge under specific environment through a learning process accomplished by adjusting interneuron connection strengths, also called synaptic weights (Haykin, 1999). For these two characteristics, ANNs are able to learn the patterns between the input data and the corresponding outputs without making any assumptions for the input data (Haykin, 1999) and able to store the knowledge for future use. Therefore, ANNs, as model-free estimators, are powerful tools to model an unknown relation between inputs and outputs.

Generally, ANNs can be used for curve fitting, classification and clustering. There are three fundamentally architectures: single-layer feedforward networks, multilayer feedforward networks and recurrent networks; five basic learning rules: error-correction learning, memory-based learning, Hebbian learning, competitive learning and Boltzmann learning as well as two fundamental learning paradigms: supervised learning and unsupervised learning for ANNs (Haykin, 1999). In this study, the objective is to relate the surface observations to RZSM with overall small errors between the estimates and observed RZSM. It is a problem of curve fitting with error-correction learning under supervision. As the desired models will be used for large-scale RZSM estimation, the architecture of single layer is too simple to deal with the high nonlinearity in the subsurface systems while the recurrent networks are computationally intensive for large-scale estimation. Therefore the feedforward MLPs which use back-propagation algorithm is a suitable type of ANNs to achieve the goal of this study. For simplicity, the feedforward MLPs will be denoted as MLPs in the later text.

For the better understanding of MLPs, the introduction of the neurons in ANNs and the configuration of MLPs as well as the learning process of MLPs will be presented in **section 3.1.2** and **3.1.3** respectively.

3.1.2 Basic ANN Models and the Architecture of MLPs

The human brain is composed of a large number of neurons. The conceptual models of neurons are also the fundamental processing units in ANNs. **Figure 2** and **Figure 3** show the structure of a neuron and its conceptual model, respectively. As it is shown in **Figure 2**, generally a neuron consists of a cell body, braches of dendrites and an axon

with synapses at its end. The cell body receives the input signals from dendrites and processes these inputs. The output signal of a neuron is propagated through the axon to synapses. Synapses transfer the output to other neurons by releasing electrical or chemical signals which can be received by the dendrites of other neurons.

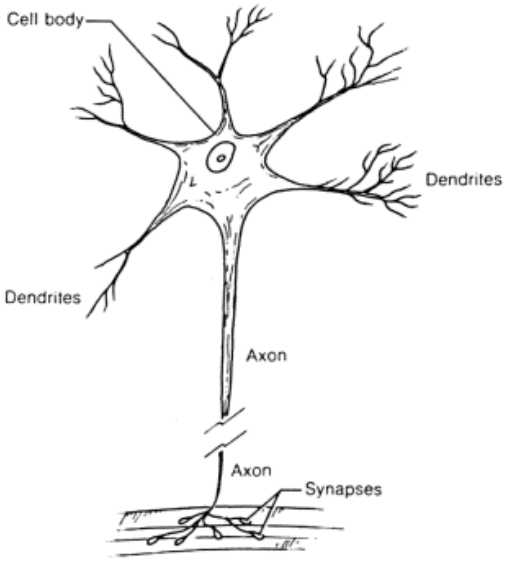

Figure 2 The structure of a neuron (Turchin, 1977)

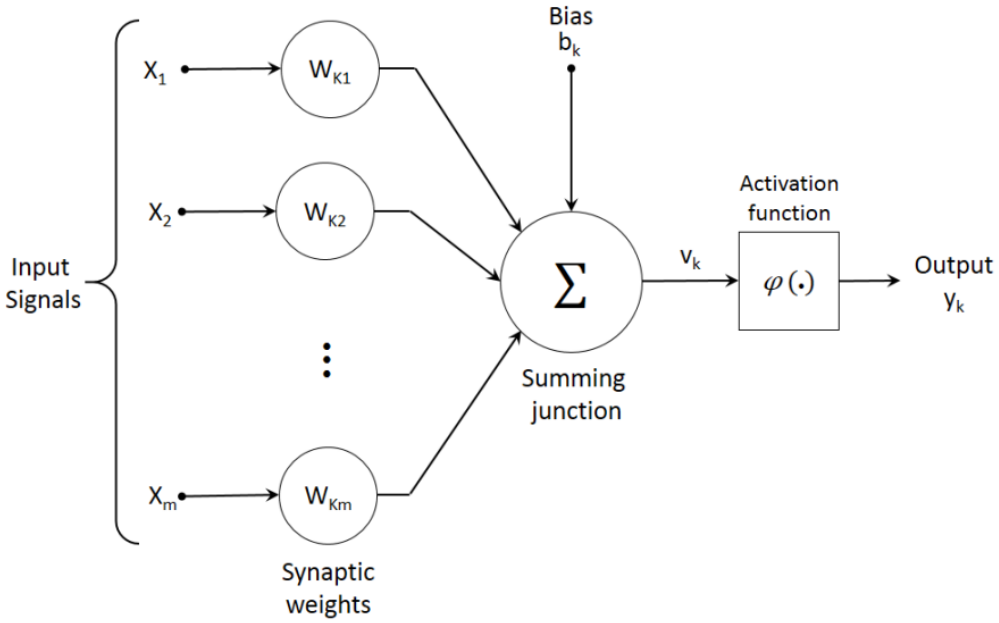

Figure 3 The conceptual model for a neuron k (reproduced from (Haykin, 1999))

The neuron model presented in **Figure 3** has three elementary components (Haykin, 1999):

1. A group of synapses characterized by the given weights. This component is used to transfer the signals of different strengths to the cell body (represented by the other two components). An input signal with a specific strength is represented by an input x_j multiplied by the corresponding synaptic weight w_{kj}.
2. An adder, also known as a linear combiner. It is used to combine all the inputs by adding up all the receiving weighted signals.
3. An activation function. It processes the result, which is generated by the linear combiner, and limits the output of a neuron model into a specific range.

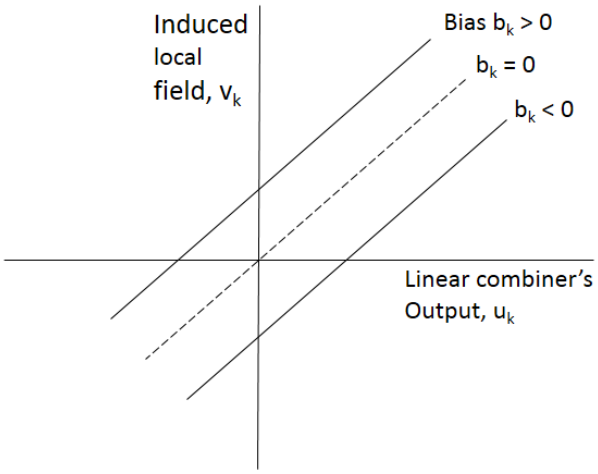

Figure 4 Affine transformation produced by the presence of a bias; note that $v_k = b_k$ when the sum of the weighted input signals is zero (reproduced from (Haykin, 1999))

In addition to these three components, there is also a bias which can be used to increase or decrease the net input for the activation function (Haykin, 1999). This bias can also be considered as a synaptic weight whose corresponding input is always one.

The impacts of this bias are shown in **Figure 4**. Therefore, the net input for the activation function in a neuron k can be written as (Haykin, 1999):

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \tag{12}$$

and the output of the neuron can be generated by:

$$y_k = \varphi(v_k) \tag{13}$$

where m is the number of the input signals, b_k is the bias, $\varphi(\cdot)$ is the activation function. There are assorted available activation functions. Many of them limits the output of a neuron in the interval [0,1], such as threshold function, piecewise-linear function and sigmoid function (Haykin, 1999), or in the interval [-1,1] like hyperbolic tangent function (Haykin, 1999). The hyperbolic tangent function is shown in **Figure 5**.

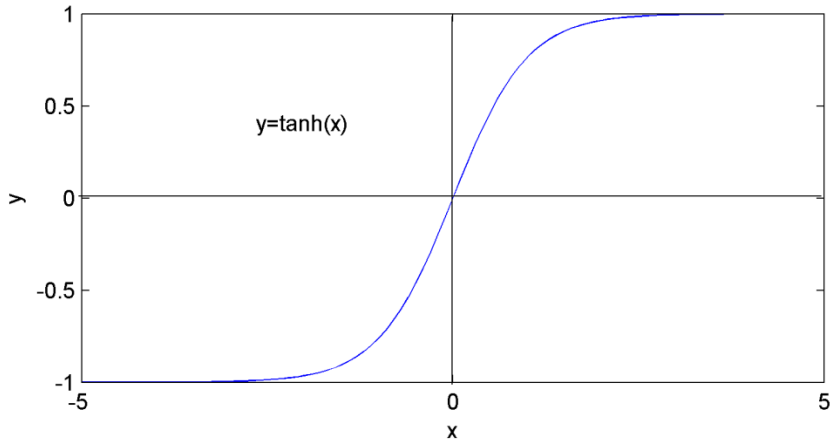

Figure 5 Hyperbolic tangent function

For MLPs, they consist of at least three layers with neurons in each layer, including one input layer, one output layer as well as one or more hidden layers between the input layer and output layer. Compared with single-layer networks, which only have one input layer and one output layer, the hidden neurons and hidden layers enable MLPs to

capture higher-order statistics (Haykin, 1999). The term "single-layer" only refers to the output layer as the input layer does not conduct any computation (Haykin, 1999). The topography of a MLP with one hidden layer is presented in **Figure 6**. Similar with the configuration of a neuron model, each neuron in the hidden layer and the output layer possesses an adder and an activation function. Each connection between the neurons in a MLP is also given a synaptic weight. For convenience, the equation (7) is reproduced here to present the calculation of an output of a MLP (Hagan et al., 2002):

$$y = f(\boldsymbol{W}_2 \boldsymbol{g}(\boldsymbol{W}_1 \boldsymbol{X} + b_1) + b_2) \tag{14}$$

where y is the output of the network; f and g are the activation functions for the output and hidden neurons, respectively; $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are the weights for the input variables and the outputs of the hidden neurons, respectively; $\boldsymbol{X}$ is the inputs; b_1 and b_2 are the biases of the input layer and hidden layer, respectively.

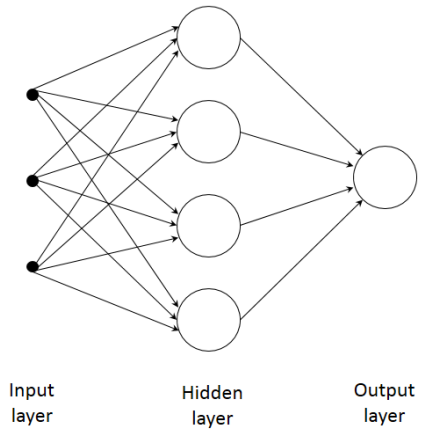

Figure 6 The topography of a MLP with one hidden layer

3.1.3 Training MLPs

MLPs apply error-correction learning with supervision to accomplish input-output mapping. At the beginning of a learning (training) process, the outputs of a MLP may

have large errors compared to the corresponding targets. After learning a particular number of examples, a MLP can enhance its performance by adjusting the synaptic weights to minimize the errors between the outputs and targets. In addition to the good performance for the presented examples, a MLP need to generalize well. A MLP can be considered to have good generalization when it also has good performance for the data never presented before (Haykin, 1999). Without good generalization, a MLP is not sufficient for practical use. The content in this section is to introduce the techniques to complete the training process and achieve good generalization.

3.1.3.1 Back Propagation Algorithm

Back propagation algorithm is a typical error-correction learning algorithm. It consists of two passes of computation including a forward pass and a backward pass (Haykin, 1999). The completion of the cycles of these two passes for all the examples in the training dataset is considered as an epoch. In the forward pass, the weights of a MLP remain the same and an epoch of training examples are presented to the MLP generating a series of outputs, as described in **section 3.1.2**. The differences between the estimates and the targets can be evaluated by a cost function. In the backward pass, the weights are modified to minimize the cost function. It starts from the output layer and the error signal is transferred backward, layer by layer, by recursively calculating the local gradient which is used to calculate the corrections of the weights (Haykin, 1999). Such process can be considered as a numerical optimization problem (Haykin, 1999). Many optimization algorithms such as Levenberg-Marquardt algorithm, Bayesian regularization algorithm, gradient descent algorithm and conjugate gradient algorithm can be used to solve this problem. The training process is conduct on an epoch-by-

epoch basis (Haykin, 1999). These two passes are iterated for a number of epochs until the stopping criteria are met.

3.1.3.2 Cross-Validation

At the beginning of the training process, a MLP can largely reduce the cost function for a number of epochs. However, the decrease of the errors becomes slow in the latter part of the training. The MLP may be trapped in a local minimum or start overfitting the training data. Therefore, appropriate stopping criteria are important to make the training stop at a suitable point when the weights of the MLP stabilize and the MLP has good generalization. A cross-validation can be used as an early stopping method (Haykin, 1999). For a given dataset, the examples are divided by random order into three subsets including a training subset, a testing subset and a cross-validation subset. The training subset is used to train the MLP with back propagation algorithm in a normal way as described in **section 3.1.3.1**. However, the training is stopped periodically for the testing with the validation subset. If the MLP fails to decrease the cost function in the test for a few times, it can be considered that the MLP becomes stable or starts overfitting and it is the right time to stop the training. As the test subset is not used in the training process, it can be used to evaluate the performance and the generalization of the MLP.

3.1.3.3 Data Pre-Processing and Post-Processing

As the activation functions limit the outputs to a specific range, an activation function may be saturated when given large inputs. For example hyperbolic tangent function (shown in **Figure 5**), when the inputs are larger than 3, the outputs of the activation

function are all close to 1. To avoid the activation function getting saturated, the corresponding weight must be kept small. This can lead to a very small gradient resulting in a slow learning process. In addition, the input variables may have different ranges. A MLP may treat these variables differently. Therefore, it is necessary to transform the inputs into the same range of the activation function before the training process (pre-processing). Such transformation is also known as normalization. To achieve meaningful results for the user, the outputs of the MLP need to be converted back to their actual range by de-normalization after the training process (post-processing).

According to the content described in this **section 3.1.3**, the training process with back propagation of a MLP can be summarized as follows:

1. Data pre-processing: All the data are normalized to an appropriate range according to the selected activation function.
2. Data division: The data are divided into three subsets by random order for training, cross-validation and testing.
3. Initialization: The weights in the MLP are given initial values using a specific initialization method, for example Nguyen-Widrow method (Nguyen et al., 1990).
4. Forward pass: All the examples in the training subset are presented to the MLP. All the corresponding outputs are generated and the cost function is calculated.
5. Backward pass: The weights of the MLP are adjusted according to the cost function calculated in step 4 using the selected numerical optimization algorithm.
6. Iteration: Iterate the forward and backward passes in step 4 and step 5 by presenting new epochs for a few cycles and then turn to the next step.

7. Cross-validation: Test the MLP on the validation subset. If the MLP fails to decrease the cost function in the test for a few times, stop training otherwise turn back to step 6.
8. Data post-processing: Convert the outputs of the MLP into their actual range by de-normalization.

3.2 Study Area and Data

3.2.1 Study Area

With the objective of examining the feasibility of MLPs to retrieve RZSM for a large area, the study area was chosen to be the United States where various climatic patterns exist along with soil moisture data. According to the updated global map of the Köppen-Geiger climate classification (Peel et al., 2007), there are 22 climate types in the United States (shown in **Figure 7**) including five main types: Dfc (cold climate without dry season having cold summer) (18%), Cfa (temperate climate without dry season having hot summer) (18%), Dfb (cold climate without dry season having warm summer) (17%), Bsk (Arid and cold steppe climate) (16%), Dfa (cold climate without dry season having hot summer) (12%). Cold climate without dry season seems to be the dominant climatic pattern. The details and criteria for each climate type are shown in **Table 1**. Another reason for choosing the United States is that a few large observation networks are available, providing a large number of soil moisture data measured at multiple depths (Ochsner et al., 2013).

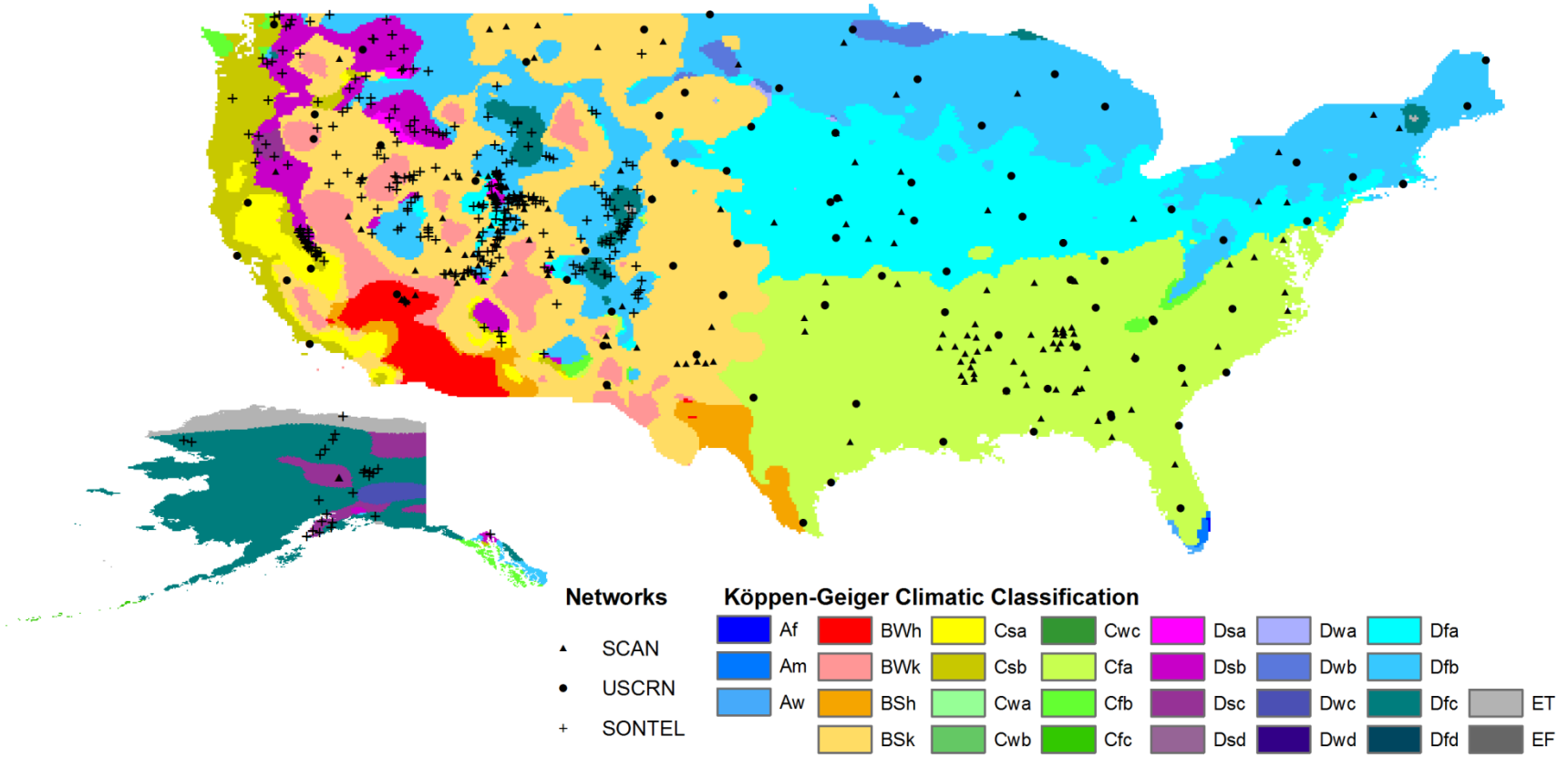

Figure 7 Köppen-Geiger climate classification and the selected stations in the United States

(This map was created using the updated world map of the Köppen-Geiger climate classification (Peel et al., 2007) and the administrative map of United State downloaded from http://gadm.org/country)

Table 1 Description of Köppen climate symbols and defining criteria (reproduced from (Peel et al., 2007))

1st	2nd	3rd	Description	Criteria*
A			Tropical	$T_{cold} \geq 18$
	f		- Rainforest	$P_{dry} \geq 60$
	m		- Monsoon	Not (Af) & $P_{dry} \geq 100 - MAP/25$
	w		- Savannah	Not (Af) & $P_{dry} < 100 - MAP/25$
B			Arid	$MAP < 10 \times P_{threshold}$
	W		- Desert	$MAP < 5 \times P_{threshold}$
	S		- Steppe	$MAP \geq 5 \times P_{threshold}$
		h	- Hot	$MAT \geq 18$
		k	- Cold	$MAT < 18$
C			Temperate	$T_{hot} > 10$ & $0 < T_{cold} < 18$
	s		- Dry Summer	$P_{sdry} < 40$ & $P_{sdry} < P_{wwet}/3$
	w		- Dry Winter	$P_{wdry} < P_{swet}/10$
	f		-Without dry season	Not (Cs) or (Cw)
		a	- Hot Summer	$T_{hot} \geq 22$
		b	- Warm Summer	Not (a) & $T_{mon10} \geq 4$
		c	- Cold Summer	Not (a or b) & $1 \leq T_{mon10} < 4$
D			Cold	$T_{hot} > 10$ & $T_{cold} \leq 0$
	s		- Dry Summer	$P_{sdry} < 40$ & $P_{sdry} < P_{wwet}/3$
	w		- Dry Winter	$P_{wdry} < P_{swet}/10$
	f		-Without dry season	Not (Ds) or (Dw)
		a	- Hot Summer	$T_{hot} \geq 22$
		b	- Warm Summer	Not (a) & $T_{mon10} \geq 4$
		c	- Cold Summer	Not (a, b or d)
		d	- Very Cold Winter	Not (a or b) & $T_{cold} < -38$
E			Polar	$T_{hot} < 10$
	T		- Tundra	$T_{hot} > 0$
	F		- Frost	$T_{hot} \leq 0$

*MAP = mean annual precipitation, MAT = mean annual temperature, T_{hot} = temperature of the hottest month, T_{cold} = temperature of the coldest month, T_{mon10} = number of months where the temperature is above 10, P_{dry} = precipitation of the driest month, P_{sdry} = precipitation of the driest month in summer, P_{wdry} = precipitation of the driest month in winter, P_{swet} = precipitation of the wettest month in summer, P_{wwet} = precipitation of the wettest month in winter, $P_{threshold}$ = varies according to the following rules (if 70% of MAP occurs in winter then $P_{threshold}$ = 2 x MAT, if 70% of MAP occurs in summer then $P_{threshold}$ = 2 x MAT + 28, otherwise $P_{threshold}$ = 2 x MAT + 14). Summer (winter) is defined as the warmer (cooler) six month period of ONDJFM and AMJJAS.

3.2.2 Data Sources

The soil moisture measurements used here were obtained from the International Soil Moisture Network (ISMN) (Dorigo et al., 2011). Among all the soil moisture networks in the United States, only the Soil Climate Analysis Network (SCAN) (Schaefer et al., 2007), the Snowpack Telemetry (SNOTEL) (USDA & NRCS, 2010, 2012) and the U.S. Climate Reference Network (USCRN) (Bell et al., 2013; Diamond et al., 2013) are under consideration because the stations of these networks are operating in a similar configuration while other networks provide observations at different depths. SCAN was established with a pilot project starting in 1991 focusing on the agricultural areas to measure soil moisture and temperature data as well as atmospheric data (Schaefer et al., 2007). Compared with SCAN and USCRN which cover most parts of the United States, SNOTEL stations are distributed in the western states and Alaska, collecting hydrologic and climatic data in cold mountainous region since the late 1970s (USDA & NRCS, 2010, 2012). USCRN aims to acquire data with high quality for the study of climate change (Bell et al., 2013; Diamond et al., 2013). Unlike SCAN and SNOTEL, the stations of USCRN are distributed evenly over the United States and have much shorter record of soil moisture data since its soil-probe deployment was accomplished in August 2011. The data acquired from USCRN through ISMN are not earlier than 2012. These three networks all apply electromagnetic techniques to measure soil moisture on an hourly basis. Data for total 557 stations (shown in **Figure 7**) from SCAN (144), SNOTEL (329) and USCRN (84) were selected as all these stations have soil moisture data at the depths of 5, 20 and 50cm with appropriate length of record and data quality. Among all the selected stations, 448 of them (80%) were used to build the MLPs while the other

(20%) (to be shown in **Figure 12**) were used for independent validation, in order to evaluate the flexibility of the MLPs. The validation stations were selected based to be evenly distributed both spatially and by variability in geographic characteristics (e.g. height). For consistency with the forcing data, only soil moisture measured after 2002 were collected for use.

In conjunction with the soil moisture observations, meteorological data as well as soil texture information were acquired to help the proposed models to infer soil moisture dynamics in root zone. The forcing data between 2002 and 2013 were extracted from the Global Land Data Assimilation System Version 1 (GLDAS-1) products based on the Noah land surface model with a temporal and spatial resolution of 3-hours and 0.25° respectively. GLDAS can generate optimal values of land surface states by integrating space- and ground- based observations through land surface models and data assimilation techniques (Rodell et al., 2004). Wang et al. (2011) have conducted a study in a mid-latitude area indicating that the high resolution GLDAS/Noah forcing data are reliable. In the study of Dorigo et al. (2013), GLDAS data were used to investigate the quality of ISMN data showing that the precipitation data provide by GLDAS/Noah products can be more reliable than the in-situ measurements in some cases. GLDAS-1/Noah data were selected because when compared with the data of other versions, GLDAS-1/Noah could provide a long term record with consistency (since mid-2001). The soil texture information with spatial resolution of 1° was obtained from the Webb et al. (2000) dataset in which specific fractions of slit, clay and sand were given to the defined 106 types of soil. In this dataset, soil texture data at various horizons are included but only the data of the top horizon were extracted to utilize herein.

3.2.3 Data Processing

Except for the soil moisture and soil texture data, all the data of the forcing variables were extracted from the GLDAS dataset. The gridded GLDAS data have the temporal and spatial resolution of 3-hours and 0.25°, respectively. The data for a specific grid can be extracted according to the longitude and latitude at the center of the grids. The data of the four grids around the selected stations were extracted according to the latitude and longitude, and the forcing data of the selected stations were calculated using the inverse distance weighting (IDW) approach. The author aimed to develop daily models to retrieve RZSM regardless the daily fluctuations in the related hydrological processes. All the selected state variables (forcing variables and soil moisture data) were transformed into daily values by calculating their sum or average as appropriate. Concerning the missing values in soil moisture datasets, the gaps which were smaller than 24 hours were infilled by linear interpolation before the temporal aggregation. This is reasonable because of the persistence of soil moisture and the linear interpolation was found to be effective for infilling small gaps in soil moisture datasets (Kornelsen & Coulibaly, 2014a). The data that are obviously incorrect (e.g. negative or larger than 1 m^3m^{-3}) were removed and treated as missing values. The daily records with missing values in the soil moisture datasets at either 5cm, 20cm or 50cm were not taken into consideration. In other words, only the date when daily soil moisture at depths of 5cm, 20cm and 50cm all exist were selected for further use. As the data used to build the models were selected into the three subsets (training, testing and cross-validation) by random order, the data did not need to be temporally continuous. The removal of the missing data would not affect the training processes.

3.3 Model Development and Testing

3.3.1 Selecting Input Variables

3.3.1.1 Basic Selection

Since MLPs are data-driven methods, the representativeness and size of the training dataset can have a large impact on the model performance. In this study, along with the surface soil moisture (soil moisture measured at the depth of 5cm) (SM05), the following forcing variables were considered as inputs: evapotranspiration (ET), specific humidity (SH), rainfall (R), snowfall (S), surface air temperature (Ta), shortwave radiation (SR) and wind speed (WS). As mentioned in **section 2.3.4**, these input variables have been widely applied in the studies using SDDMs. They have a close relation to the water and energy fluxes exchanged between atmosphere and land surface. In addition, the surface soil temperature (Ts) was also utilized, as it was found to be an effective input variable by Elshorbagy & Parasuraman (2008).

3.3.1.2 The Application of Correlation Analysis and Sensitivity Analysis

Root zone soil moisture, acting as a low-pass filter, has positive feedback mechanisms with atmosphere (Entekhabi et al., 1996). Instead of using the instantaneous values, the cumulative precipitation and temperature are more effective inputs to estimate RZSM (Elshorbagy & El-Baroudy, 2009; Elshorbagy & Parasuraman, 2008). Hence, accumulated values from the previous time steps were considered to be used as input variables in this study. As previous 13 days (Elshorbagy & Parasuraman, 2008) and 30 days (Jiang & Cotton, 2004) cumulative precipitation were found to be effective inputs, cumulative inputs from the previous 15 days and 30 days were considered in this study.

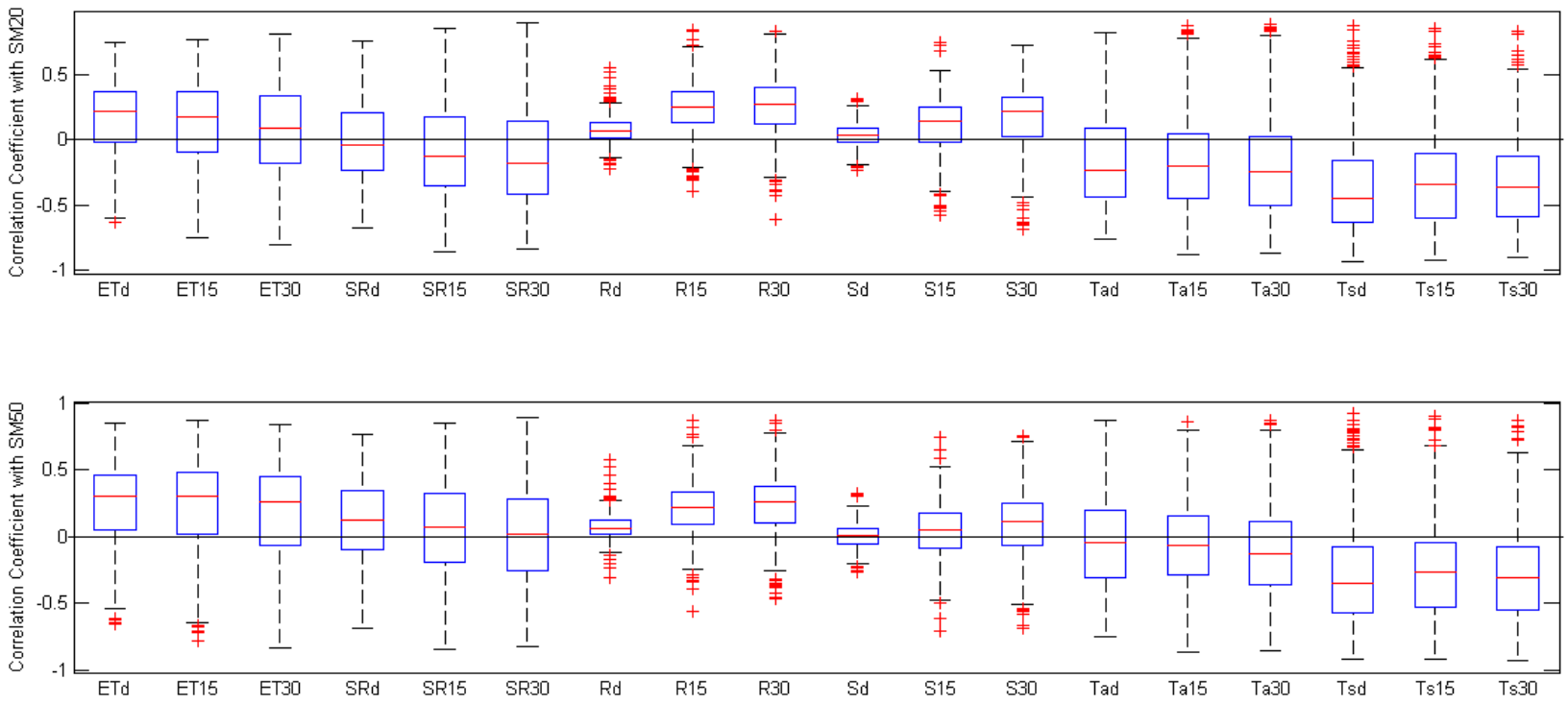

Figure 8 The correlation coefficients between input variables and root zone soil moisture

Note that: “d” means daily; “15” and “30” mean the accumulation values from the previous 15days and 30days respectively.

For simplicity, first a correlation analysis which was also used by Lamorski et al. (2013), was applied here to identify effective accumulative inputs. We calculated the correlation coefficient between the forcing variables and RZSM for the 557 stations. The results were depicted in **Figure 8** in the form of box plots. It can be detected that generally the cumulative precipitation (rainfall and snowfall) and air temperature have higher correlation with soil moisture at 20cm (SM20) and 50cm (SM50) than that without accumulation. In addition, the previous 30 days accumulated rainfall (R30), snowfall (S30) and air temperature (Ta30) have higher correlation than that of the previous 15days (R15, S15 and Ta15, respectively). In contrast, the daily ET tends to have higher correlation. For SR, the results for SM20 and SM50 are inconsistent. SR mostly reflects the energy dynamics. Such information can also be partially captured by using Ta. As the cumulative Ta was selected to be used because of its high correlation with RZSM, using the daily SR to reflect the short-term energy fluctuation is more preferable. In addition, usually estimating soil moisture at deeper layers has more challenges. Using daily SR instead of accumulative values may assist in the estimation for the deeper layer as it tends to have higher correlation with SM50. However, in terms of Ts the correlation coefficients of daily and accumulative values with RZSM are similar. The daily values and the accumulation of previous 30days (Ts30) have slightly higher correlation than that of the accumulation of previous 15 days (Ts15). For consistency with Ta, Ts30 is preferable to be used. In summary, the correlation analysis indicated that generally the cumulative precipitation and temperature have higher correlation coefficient with RZSM and the previous 30days accumulated values are more preferable than that of the previous 15days.

Since correlation analysis is more suitable in linear cases and it only reflects the influence of a variable on RZSM individually, the author also used a few combinations of inputs to build models and conducted a sensitivity analysis to confirm the findings described above. The methods of model development were the same with those to be presented in **section 3.3.2**. The sensitivity analysis was based on the method applied in Wang et al. (2000). In this method, the importance of a variable to a MLP was shown by its effects on model performance when all observations of the variable are replaced by its mean value (clamping). Once the models were developed, each input variable was individually clamped to its mean and the MLPs regenerated RZSM with the clamped inputs. The sensitivity index was calculated using equation (15) (Wang et al., 2000):

$$\xi(x_i) = 1 - \frac{p(\boldsymbol{X} \mid x_i = \overline{x_i})}{p(\boldsymbol{X})} \tag{15}$$

where $p(\boldsymbol{X} \mid x_i = \overline{x_i})$ is the model performance (RMSE used here) using the input dataset in which the values of variable x_i are replaced by its mean, $p(\boldsymbol{X})$ is the model performance without clamping. If a MLP was well trained, the information provided by each example and variable should be useful for the model to infer the RZSM dynamics. For a well-trained MLP, clamping the variables with their mean decreases the information for the MLP to estimate RZSM dynamics, resulting in worse performance compared to that without clamping. As we use RMSE here, $p(\boldsymbol{X} \mid x_i = \overline{x_i})$ should be larger than $p(\boldsymbol{X})$, generating a negative sensitivity index. According to (15), clamping a significant variable can lead to a great difference between $p(\boldsymbol{X} \mid x_i = \overline{x_i})$ and $p(\boldsymbol{X})$, resulting in a large magnitude of the sensitivity index. Therefore, the sign of a sensitivity

index can show whether a MLP is well-trained and the magnitude can identify the importance of the variables.

The sensitivity indices for the variables in the models with different combinations of inputs are presented in **Figure 9**. Comparing **a)** and **b)** in **Figure 9**, it can be detected that the cumulative inputs have larger impacts. This is consistent with the findings detected in Elshorbagy & El-Baroudy (2009) and the correlation analysis mentioned before. Besides, R30 and S30 are more influential than R15 and S15 respectively, as shown in **b)** and **c)**. Comparing **d)** with **e)**, Ta30 and Ts30 are more influential than Ta15 and Ts15 respectively. For the upper row of **Figure 9**, the input combination in **c)** seems to be the better choice as each variable carried some useful information for the model. For the lower row of **Figure 9**, the input combination in **e)** can be easily distinguished from the other two. It is unclear why SH has some impacts on the models when daily Ta and Ts instead of the accumulative values were applied and these impacts were larger than those from the evapotranspiration and precipitation. In the study of Kornelsen & Coulibaly (2014b), relative humidity was used as an input variable in the MLPs, but did not show high influences for RZSM estimation. Regardless of SH, **c)** and **e)** are very similar, but the sensitivity index of Ts30 is higher in **e)**. Though, the results for the training dataset showed that the performance of the MLPs using these combinations of inputs is similar, overall it is believed that the input combination of **e)** is better than the others. First, each variable in it provided some useful information to the model while in the other combinations a few variables seem to be useless. Second, comparing **c)** with **e)**, Ts30 have larger impacts on the model. Third, the results depicted in **e)** are in line with those found in the correlation analysis.

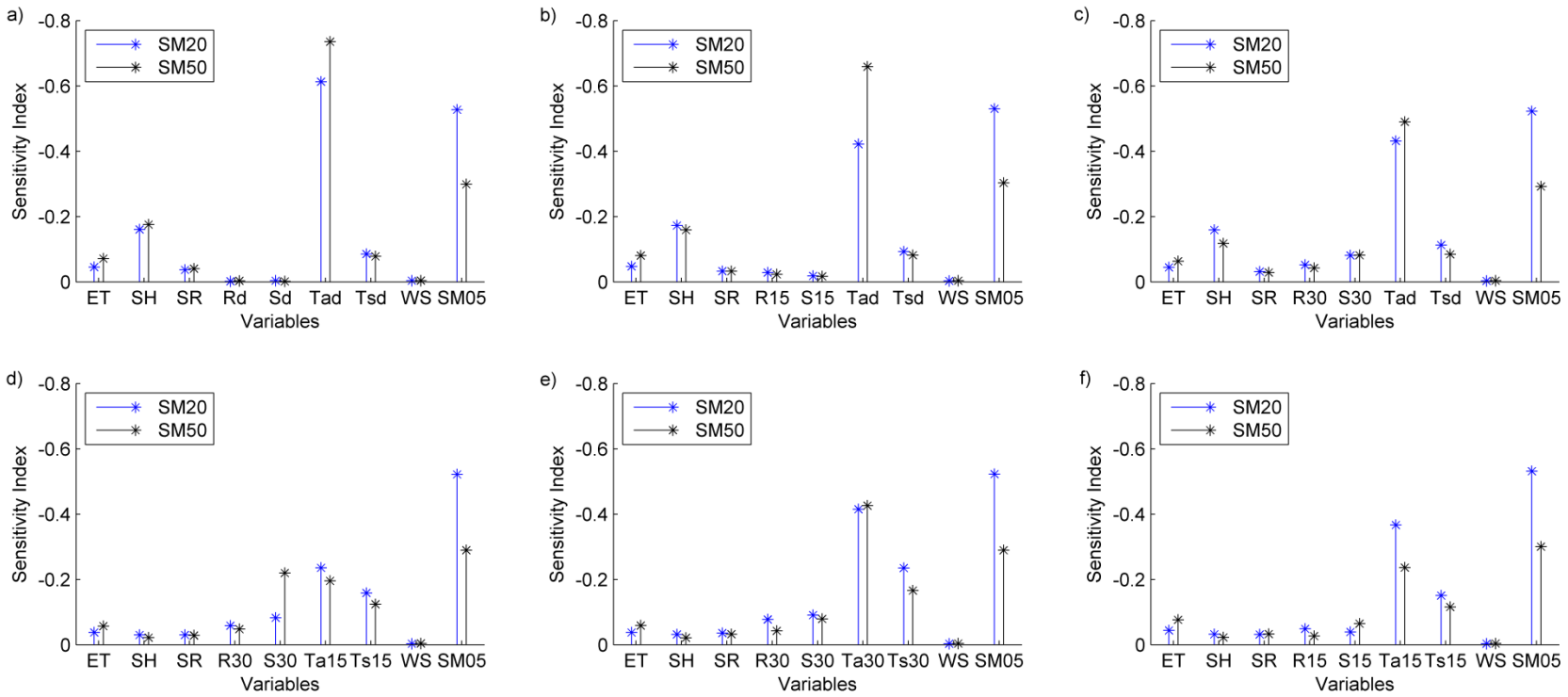

Figure 9 Sensitivity indices of the model input variables with different combinations ((a) to (f)) of inputs

Therefore, the accumulated previous 30 days rainfall, snowfall, surface air temperature, and surface soil temperature were chosen as input variables to provide the MLPs with memory of the past values while the other variables were not cumulated to reflect incremental daily effect.

3.3.2 The Design of MLPs

When designing the MLPs, this study mostly concerned about the selection of activation function, the model architecture and the training algorithm, as these settings should be varied according to specific problems. This study used the Neural Network Toolbox™ in Matlab® R2013a to build MLPs. The other detailed settings of the models used the default values provided in the toolbox. For the activation function, the ability of MLPs to capture the nonlinear characteristics mostly relies on the nonlinear activation function used in the hidden neurons. The tangent sigmoid function was chosen to be the activation function of the hidden neurons because it has the characteristic of anti-symmetry which may, in general, accelerate the learning process (Haykin, 1999) and was proven to be a better choice in hydrologic modeling (Yonaba et al., 2010). The linear function was selected for the output layer as the utilization of a nonlinear function was not found to improve model performance (Yonaba et al., 2010).

For the model architecture, this study applied the architecture with only one hidden layer. The application of more than one hidden layers has also been tried but it much slowed down the computation and did not bring significant improvement to the model performance. In terms of the number of the hidden neurons, the trial and error method was applied and the number of the hidden neurons was changed from 10 to 70. It was

found that with the increase of the hidden neurons the mean squared error between the estimates and the targets decreased. Hence, the number of the hidden neurons was made as large as possible. Because of the limited computer memory, 70 hidden neurons were selected to use for all the models developed in this study. The more hidden neurons may lead to smaller errors. However, the improvement would be small and there will be a risk of overtraining as well. The results in sensitivity analysis (to be shown in **section 4.1**) also indicated this selection was reasonable. The number of hidden neurons used in this study is large compared with the value of 20 used in Kornelsen & Coulibaly (2014b). This may result from the large training dataset as well as the complexity and the high spatial heterogeneity of vadose zone hydrology in the large region considered. One output neuron was used in the model configuration to estimate RZSM at either 20cm or 50cm.

Turning to the training algorithm, the Levenberg-Marquardt algorithm was applied to train the MLPs. Compared with other algorithms, the Levenberg-Marquardt algorithm is more effective because it balances the speed of Newton's method and the convergence with quickest descent (Hagan et al., 2002). The cost function in the back propagation algorithm was selected to be mean squared error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2 \tag{16}$$

where N is the number of the examples; y_i and t_i are the outputs and the corresponding targets respectively. In order to avoid the MLPs overfitting the target data, the dataset used to train the model was divided into three parts: a training subset (70%), a testing subset (15%) and a cross-validation subset (15%) which was used for early stopping.

As each training session starts with different initial weights and biases as well as different divisions (because of the random order) of data into the three subsets, the outputs for each training process are different. To gain more stable estimates, each MLP was trained 20 times and the average of the ensemble 20 output sets was treated as the final estimation.

Note that all these settings described above were applied to all the MLPs developed in this study.

3.3.3 The Design of the Two Experiments

In this study, two experiments including four MLPs were developed in order to find out the important inputs for MLPs to estimate RZSM in a large area and gain insight into the effects brought by errors in the observations. The first experiment focused on identifying the state variables that are generally significant for RZSM estimation in a large region with various climatic patterns and land surface conditions as well as exploring the benefit brought by introducing soil texture data into the model. Two models (denoted as MLP1-1 and MLP1-2) are included in this experiment to compare the situations with and without soil texture data.

The second experiment focused on the impacts of potential errors in the datasets. As mentioned before, the forcing data provided by GLDAS/Noah products have been proven to be relatively reliable in previous studies (Dorigo et al., 2013; Wang et al., 2011) and do not have missing values. Here the author mainly focused on the effects of potential errors in the in-situ soil moisture observations. It should be noted that the soil moisture observation networks around the world integrated by the ISMN are

heterogeneous with respect to measurement techniques. The quality of the soil moisture data may influence not only the development but also the evaluation of the models. As described in Dorigo et al. (2013), various types of errors exist in the soil moisture observations and an automated quality control system is necessary as the visual inspection is mostly not possible for a great volume of data. To meet the increasing concerns about the data quality, Dorigo et al. (2011) first proposed a simple quality control system for the ISMN dataset to provide data users with flags reflecting the data quality. However, some subtle outliers were failed to be flagged and this system was further improved and demonstrated in Dorigo et al. (2013). Because of various conditions and the different instruments used at the stations of a variety of networks, the automated quality control system may face challenges to capture all the potential errors and some data may be overflagged. For example, some data may be out of the selected range of the soil moisture and flagged but their time series can look realistic and some spurious observations may fail to be flagged (Dorigo et al., 2013). Therefore, it may be difficult to avoid all the potential errors in soil moisture datasets while overflagging the data may decrease the data of special cases affecting the generalization of MLPs. It is necessary to test whether the MLPs were robust when potential errors were involved. In this study, the quality flags provided by ISMN were used. Though potential errors may still exist these quality flags can be used to reduce the uncertainty from some of the errors in the data. The differences in the model performance of MLPs with and without these flagged data can explore the robustness of the MLPs. The detailed definition of each data flag in the automated quality control system (Dorigo et al., 2013) can be found at http://ismn.geo.tuwien.ac.at/data-

access/quality-flags/. In the second experiment, data with flag “C” —values exceeding threshold and “D”— questionable/dubious are considered as poor data. Two models, namely MLP2-1 and MLP2-2, were established with the same architecture and input variables as MLP1-1 and MLP1-2 respectively, except that the examples with flag “C” and “D” are removed. Note that linear interpolation was implemented to fill gaps in datasets before transforming the hourly data to daily scale. The infilled values lack of quality flag. The infilled values generated using the poor data through interpolation were also identified and removed. The differences of the four MLPs are summarized below:

- MLP1-1: the nine state variables selected in **section 3.3.1** were used as inputs
- MLP1-2: the two state variables with the least significance shown in MLP1-1 were replaced by soil texture information (fractions of sand and clay)
- MLP2-1: have the same configuration and input variables with MLP1-1 but the data with flag “C” and “D” were removed
- MLP2-2: have the same configuration and input variables with MLP1-2 but the data with flag “C” and “D” were removed

Note that each of the four models actually consists of two MLPs to simulate RZSM at the depth of either 20cm or 50cm, since only one output neuron was used in the model configuration.

3.3.4 Evaluation Criteria

Three criteria, namely root mean squared error (RMSE), relative error (RE) and correlation coefficient (R), were used to evaluate the model performance. RMSE was used to indicate the accuracy of the estimation. Considering that soil characteristics

vary spatially, the same RMSE may lead to different RE, given soil moisture varying within different ranges. This criterion can explore whether the models were able to capture the soil characteristics of each station. RE was calculated based on the RMSE and the mean soil moisture of the corresponding station. R was used to test the ability of MLPs to capture the soil moisture temporal variations. Their equations are given below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2} \tag{17}$$

$$RE = \frac{RMSE}{\frac{1}{N}\sum_{i=1}^{N} t_i} \tag{18}$$

$$R = \frac{\sum_{i=1}^{N}(y_i - \overline{y_i})(t_i - \overline{t_i})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y_i})^2 \sum_{i=1}^{N}(t_i - \overline{t_i})^2}} \tag{19}$$

3.4 Chapter Summary

This chapter presented the basic theory of ANNs. The architecture and the training process of the feedforward MLPs were introduced. MLPs are made up of the interconnected neuron models and characterized by one or more hidden layers between the input and output layers. MLPs use back propagation algorithm to complete the training process. The cross-validation method and data pre- and post-processing are usually applied to avoid overfitting and to accelerate learning process, respectively.

To achieve the objectives in this study, the study area is selected to be the United States for the various climate types and the large number of available data provided

there. Soil moisture data for 557 stations were collected. The small gaps within one day were filled. The records of large gaps were discarded. The corresponding forcing data and soil texture information were acquired for the stations. The forcing data and the soil moisture data were converted to daily values.

The development of the MLPs and the two experiments was also fully described in this chapter. Nine potential variables were selected as inputs according to the literature review in **Chapter 2**. The correlation analysis and sensitivity analysis revealed that the cumulative precipitation and temperature were effective inputs and the accumulation of the previous 30days was generally better than 15 days. The final selected input variables included ET, SH, R30, S30, Ta30, Ts30, WS and SM05. The MLPs were designed to have 70 hidden neurons, tangent sigmoid function and linear function as activation function in hidden layer and output layer respectively. The Levenberg-Marquardt algorithm was selected for model training. To gain more stable estimates, each MLP was trained 20 times and the average of the ensemble 20 output sets was treated as the final estimation. Two experiments were designed. The first experiment consisted of two MLPs with and without soil texture information to explore the effects caused by the inclusion of land surface property data (e.g. soil texture). The second experiment included two MLPs similar with those two in the first experiment except for the exclusion of poor data. It was designed to explore the impacts from potential errors in soil moisture datasets. The MLPs were evaluated by three criteria, namely RMSE, RE and R. The results of the experiments will be presented in the next chapter.

Chapter 4

Results

In this chapter, the results of the two experiments will be presented. A detailed sensitivity analysis using the same method described in **Chapter 3** will be given for the MLPs in the first experiment, to explore the importance of the input variables for RZSM estimation and verify the appropriateness of the model design. The performance of the four MLPs will then be analyzed to achieve the objectives of this study. A discussion will also be conduct to compare the findings in this study with the others.

4.1 Sensitivity Analysis

The sensitivity analysis was conduct on the two MLPs in the first experiment with the training dataset for the 448 stations. The method used here is the same with that mentioned in **Chapter 3**.The absolute values of sensitivity index of the selected variables for the models built in the first experiment are summarized in **Table 2**. All the sensitivity indices are negative (not shown) indicating that the models were able to capture the information provided by the input variables. Therefore, the selected configuration was reasonable and the two models were well-trained. For MLP1-1 which only included state variables, SM05 was the most important input to estimate SM20, with the sensitivity index of 0.523 indicating that the lack of SM05 data can increase RMSE by 52%. Ta30 ranks the second with the value of 0.415, followed by Ts30, S30 and R30. On the contrary, ET, SR, SH and WS made much smaller differences in the outputs. The influences of these variables on the model performance are less than 5%. For the estimation of SM50, Ta30 keeps a similar value ranking the first while the contribution of SM05 became smaller compared to that of SM20. The sensitivity indices

for most of the other variables also declined except for a slight increase for evapotranspiration. This can be explained by the fact that the connection between the surface forcing and RZSM is weakened with the increase of soil depth for most variables, while the root uptake from the deeper soil layer increases the effects from evapotranspiration. Since SH and WS are the least influential state variables in MLP1-1, they were replaced with the fractions of sand and clay in MLP1-2.

In terms of MLP1-2, the great differences between sensitivity indices of soil texture and the other state variables signify the importance of soil texture data. The indices of soil texture are larger than one indicating that clamping them led to errors a few times larger, according to equation (15). MLP1-2 was especially sensitive to the fraction of sand, as its sensitivity index was twice that of the fraction of clay. The proportion of sand may be more representative in reflecting the soil hydraulic properties. In the study conducted by Twarakavi et al. (2010), high similarity was found in classification between the soil texture triangle and the soil hydraulic triangle given the soils with a large proportion of sand. For the state variables, SM05 remains the most important input for the estimation of RZSM. Its sensitivity indices for estimates of 20cm and 50cm were similar with those of MLP1-1, suggesting that it has a stable effect on model performance. Unlike SM05, the sensitivity indices for most of the other state variables become smaller when the soil texture information was introduced into the model. MLP1-2 may better represent the subsurface hydrological processes with the provided soil texture information and give smaller weights to the surface meteorological forcing. Consequently, the influences of these meteorological variables on estimation of SM20 and SM50 were similar for MLP1-2, unlike the differences found in MLP1-1.

Table 2 Absolute values of sensitivity indices of input variables for models in the first experiment

Model	Estimation	SM05	Ta30	Ts30	S30	R30	ET	SR	SH	WS	Sand	Clay
MLP1-1	SM20	0.523	0.415	0.235	0.091	0.077	0.038	0.036	0.031	0.003	-	-
	SM50	0.290	0.426	0.166	0.079	0.042	0.059	0.032	0.021	0.004	-	-
MLP1-2	SM20	0.539	0.166	0.167	0.023	0.028	0.030	0.030	-	-	5.465	2.545
	SM50	0.317	0.141	0.251	0.029	0.025	0.030	0.020	-	-	5.247	2.515

Table 3 Averages of RMSE, RE and R for the estimation of soil moisture at 448 training stations and 109 validation stations

Estimation	Model	Training			Validation		
		RMSE (m^3m^{-3})	RE	R	RMSE (m^3m^{-3})	RE	R
SM20	MLP1-1	0.069	0.541	0.745	0.066	0.360	0.778
	MLP1-2	0.065	0.513	0.760	0.068	0.371	0.770
	MLP2-1	0.068	0.535	0.750	0.066	0.374	0.775
	MLP2-2	0.062	0.495	0.767	0.069	0.384	0.768
SM50	MLP1-1	0.085	0.619	0.611	0.090	0.550	0.607
	MLP1-2	0.078	0.567	0.626	0.093	0.564	0.592
	MLP2-1	0.083	0.623	0.633	0.091	0.568	0.619
	MLP2-2	0.074	0.556	0.644	0.094	0.575	0.593

Considering the results of the first experiment, it was interesting that S30 had similar influence on the estimates as R30. As mentioned before, many of the selected stations are from SNOTEL whose stations mostly concentrate in the cold mountainous regions in the western United States. Hence, many examples provided to train the MLPs have the precipitation in the form of snowfall. The MLPs may capture the infiltration into frozen soil during the freezing-thawing process using the surface soil temperature and accumulative snowfall. In addition, the snow accumulation and melt can exert significant controls on soil moisture in wet period (spring) in a snow-dominated mountain region (Williams et al., 2009). It can also be noticed that the temperature information, including Ta30 and Ts30, played a more important role than the precipitation (R30 and S30). The water and energy dynamics in the root zone may be mainly reflected by the surface soil moisture and the temperature information. Therefore the effects of precipitation were trimmed.

The sensitivity analysis reveals that MLPs have the ability to identify variables and parameters that directly affects the water balance in root zone. Surface soil moisture, cumulative air temperature, cumulative surface soil temperature, cumulative rainfall and cumulative snowfall as well as soil texture are generally significant for the retrieval of RZSM in a large region with various climatic patterns and land surface properties.

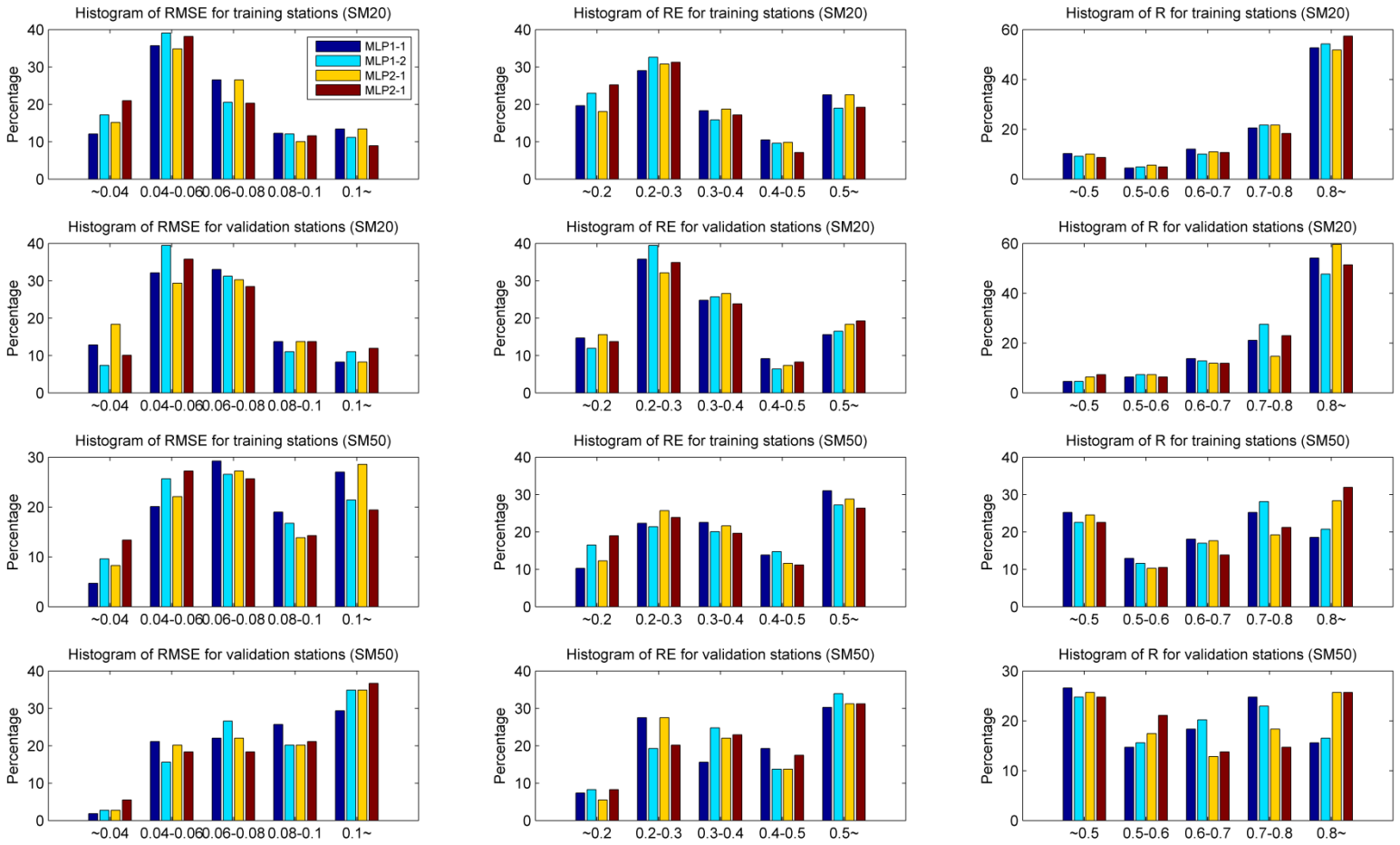

Figure 10 Histograms of RMSE, RE and R for estimation of SM20 and SM50

Note: In terms of RMSE, "~0.04" means values less than 0.04 m^3m^{-3}, "0.1~" means values greater than 0.1 m^3m^{-3} and so on.

4.2 Model Performance

4.2.1 General Performance

The performance of the models in terms of RMSE, RE and R for training and validation is given in **Table 3**. The RMSE, RE and R are the average values of the selected 448 training stations and 109 validation stations. Since the average performance statistic may be affected by extreme values, especially RE, the histograms of the three criteria for the stations are also shown in **Figure 10**. In general, the four MLPs have better performance for the estimation of SM20 with averaged values (for the four MLPs including training and validation stations) of 0.067 m^3m^{-3}, 0.447, and 0.764 for RMSE, RE and R respectively, while those of SM50 are 0.086 m^3m^{-3}, 0.578, 0.616 respectively.

In terms of RMSE, although the accuracy of the simulated SM20 was not very high compared to the usual targeted value (0.04 m^3m^{-3}) in the SMAP (Entekhabi et al., 2010) and SMOS (Kerr et al., 2001) missions, this accuracy can still be considered reasonable as it is comparable to the average sensor accuracy of 0.05 m^3m^{-3} for the soil moisture observation networks in ISMN (Dorigo et al., 2013). According to the histograms shown in **Figure 10**, more than half of the stations have RMSE of SM20 less than 0.06 m^3m^{-3} and about 75% stations have RMSE of SM20 less than 0.08 m^3m^{-3}. RE seems to be a critical criterion to evaluate the model performance. About 50% and 70% of the stations have RE of SM20 less than 0.3 and 0.4 respectively. This reveals the difficulties for MLPs to capture the characteristics associated with the land surface properties for the stations. This will be further discussed in the comparison of the two models in the first experiment (**section 4.2.2**). In the histograms of R, the models can successfully

characterize the soil moisture dynamics at the depth of 20cm since about 75% stations have R above 0.7. The histograms also clearly show the challenges for the MLPs to estimate SM50. Overall, the developed models have high skill in soil moisture estimation at the depth of 20cm with correlation coefficient of above 0.7 in most cases and reasonable accuracy. However, the estimation at the depth of 50cm faces some challenges. Comparing the results of training and validation, the three criteria for the training stations do not have large differences from those for the validation stations, as shown in **Table 3**. Although the RE of SM20 for the training stations differs greatly from that for the validation stations, the corresponding two histograms still look similar. The large differences may be caused by the extreme values for some of the training stations. For the estimated SM20 by MLP1-1, 6.5% training stations have RE above 100% while only 1.8% validation stations have RE above 100%. In addition, the largest RE in validation is 156% while some training stations have RE larger than this value.

4.2.2 The First Experiment

Comparing the performance of the models with and without soil texture data, slight improvements can be detected in the training process after introducing the fractions of sand and clay, especially for the estimation of SM50. MLP1-2 has lower RMSE and RE as well as higher R, compared with the corresponding values of MLP1-1 (**Table 3**). It is more obvious in the histograms for the training stations that MLP1-2 had more stations with RMSE in the range of '~0.04' and '0.04—0.06' than those of MLP1-1 (**Figure 10**). This decreased the overall RMSE of MLP1-2 for training stations, as shown in **Table 3**. Similar phenomena can also be detected in the histograms of RE and R for the training stations. This confirms the finding mentioned in the sensitivity analysis that the inclusion

of surface soil texture information can help the MLPs in inferring the soil hydrological processes. In the training processes, the MLPs were provided with the data for all the training stations. Therefore the training dataset had a larger arrange compared to a specific station. The MLPs may encounter difficulties to capture the characteristics of land surface properties for each station. Consequently, RE become a critical criterion as the training processes were to minimize the overall MSE for all the examples. Though the soil texture cannot fully reflect the characteristics of the soil profiles, it still provided some help for the MLPs to capture the differences. As shown in **Figure 11**, without soil texture data the estimates of MLP1-1for both SM20 and SM50 have a positive bias, resulting in relatively worse performance with RE of 0.677 for SM20 and 0.305 for SM50, respectively. When the soil texture data were included, the estimates of MLP1-2 were improved with RE of 0.153 for SM20 and 0.161 for SM50, respectively.

During validation MLP1-2 performed on par or worse than MLP1-1. This may be attributed to the limited soil texture information. Frist, as a terrestrial property, soil texture for each station does not change, resulting in low diversity of this input for the models in comparison with the other input variables. Second, the soil texture data used here were decided according to the classification of soil types (Webb et al., 2000). For stations with the same soil type, the proportions of sand and clay are the same correspondingly. In fact, the selected 448 training stations only included 44 soil types. This further decreased the diversity of these two input variables. Third, the 448 examples were not evenly divided amongst these 44 soil types. As SNOTEL provided a large number of stations concentrated in the western United States, some unique soil types in the central and eastern United States were represented by limited number of

stations. The MLPs may encounter difficulties in presenting the interaction between various climatic patterns and these soil types with the limited training examples. The proportions of stations for each soil type are shown in **Table 4**.

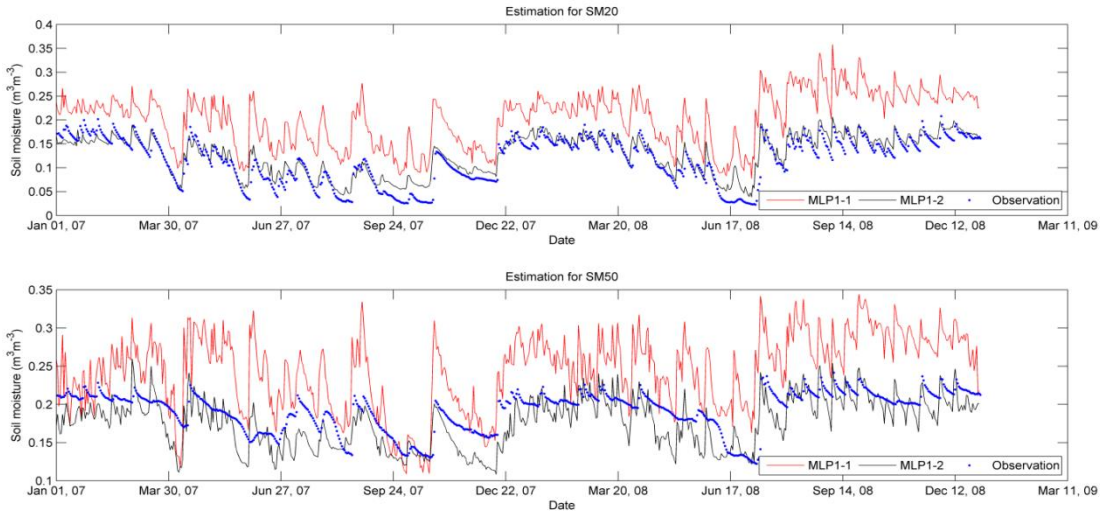

Figure 11 The time series of the simulated and observed soil moisture for station Pee Dee (from SCAN)

4.2.3 The Second Experiment

Comparing the performance of MLP1-1 and MLP2-1 as well as MLP1-2 and MLP2-2, models with poor data performed on par with models without poor data, as the differences of the corresponding three criteria are less than 5% (**Table 3**). In the histograms of SM50 in terms of R, MLP2-1 and MLP2-2 have more stations with R larger than 0.8. Some of the poor data are values exceeding the predefined threshold and the removal of these data may decrease the number of potential errors, generating higher R in evaluation. However, the removal of poor data does not largely decrease the number of the stations with R lower than 0.5. The other histograms also indicate that MLP2-1 and MLP2-2 perform similarly with MLP1-1 and MLP1-2 respectively. Therefore, The MLPs were not sensitive to the potential errors in the observations.

Table 4 The proportions (%) of stations for each type of soil

Soil Type	Training	Validation	Total Satations
Ferric Acrisol	3.8	4.6	3.9
Gleyic Acrisol	0.9	0.0	0.7
Orthic Acrisol	9.6	14.7	10.6
Plinthic Acrisol	0.4	0.0	0.4
Dystric Cambisol	1.1	0.9	1.1
Eutric Cambisol	0.7	0.0	0.5
Humic Cambisol	0.4	0.9	0.5
Calcic Cambisol	0.2	0.0	0.2
Luvic Chernozem	0.2	0.0	0.2
Eutric Podzoluvisol	0.2	0.0	0.2
Dystric Gleysol	2.0	1.8	2.0
Eutric Gleysol	2.0	0.9	1.8
Mollic Gleysol	0.2	0.0	0.2
Gelic Gleysol	0.9	0.0	0.7
Gleyic Phaeozem	0.2	0.0	0.2
Haplic Phaeozem	1.3	0.9	1.3
Luvic Phaeozem	18.5	15.6	18.0
Lithosol	2.0	0.0	1.6
Calcaric Fluvisol	2.2	0.0	1.8
Haplic Kastanozem	6.3	8.3	6.6
Calcic Kastanozem	0.2	0.0	0.2
Luvic Kastanozem	7.1	13.8	8.4
Albic Luvisol	8.0	11.9	8.8
Chromic Luvisol	1.1	0.0	0.9
Gleyic Luvisol	0.9	0.0	0.7
Calcic Luvisol	1.6	0.0	1.3
Orthic Luvisol	0.7	0.9	0.7
Eutric Histosol	0.2	0.0	0.2
Gleyic Podzol	0.7	0.0	0.5
Leptic Podzol	0.9	0.0	0.7
Orthic Podzol	1.6	0.9	1.4
Calcaric Regosol	3.3	2.8	3.2
Dystric Regosol	0.2	0.0	0.2
Eutric Regosol	1.3	1.8	1.4
Mollic Solonetz	0.2	0.0	0.2
Orthic Solonetz	1.3	0.0	1.1
Vitric Andosol	3.6	5.5	3.9
Pellic Vertisol	0.2	0.0	0.2
Haplic Xerosol	0.4	0.0	0.4
Calcic Xerosol	0.4	0.0	0.4
Luvic Xerosol	2.7	1.8	2.5
Haplic Yermosol	1.8	0.9	1.6
Calcic Yermosol	1.8	0.9	1.6
Luvic Yermosol	6.3	10.1	7.0

4.2.3 The Flexibility of the MLPs

The similar performance for the training and validation stations has indicated the flexibility and good generalization of the MLPs. To gain further insight into the ability of the MLPs to represent the interactions between the land surface and various climatic patterns, the model performance is depicted spatially in **Figure 12** with respect to the climate classification (Peel et al., 2007). As the spatial results for the different models did not have great differences, for brevity only the results (RMSE and R) of MLP1-2 are presented. It can be detected that the climatic patterns in the western United States tend to be longitudinally distributed and in the eastern part the climatic patterns have a latitudinal distribution. The performance of MLP1-2 shown in **Figure 12** does not reveal any trends except that the stations in the southern Great Plains tend to have worse results for the estimation at 50cm. This result is not consistent with the high strength of land-atmosphere coupling (Koster et al., 2004) and surface-subsurface coupling (Ford et al., 2014) found in the Great Plains of North America. As there are a limited number of stations to provide training and validation data in this area, this finding needs further study to confirm. Similar results and phenomena can also be found in the estimation of MLP1-1 (not shown). Generally, the developed MLPs were flexible and could be applied in different geographical regions.

The results are also spatially depicted in **Figure 13** with respect to the various soil types (Webb et al., 2000). No obvious trend can be detected. Though the soil texture data of the top horizon provide useful information to the models as shown in the sensitivity analysis, they may not reflect all the major characteristics of soil profiles which directly affect the dynamics of RZSM.

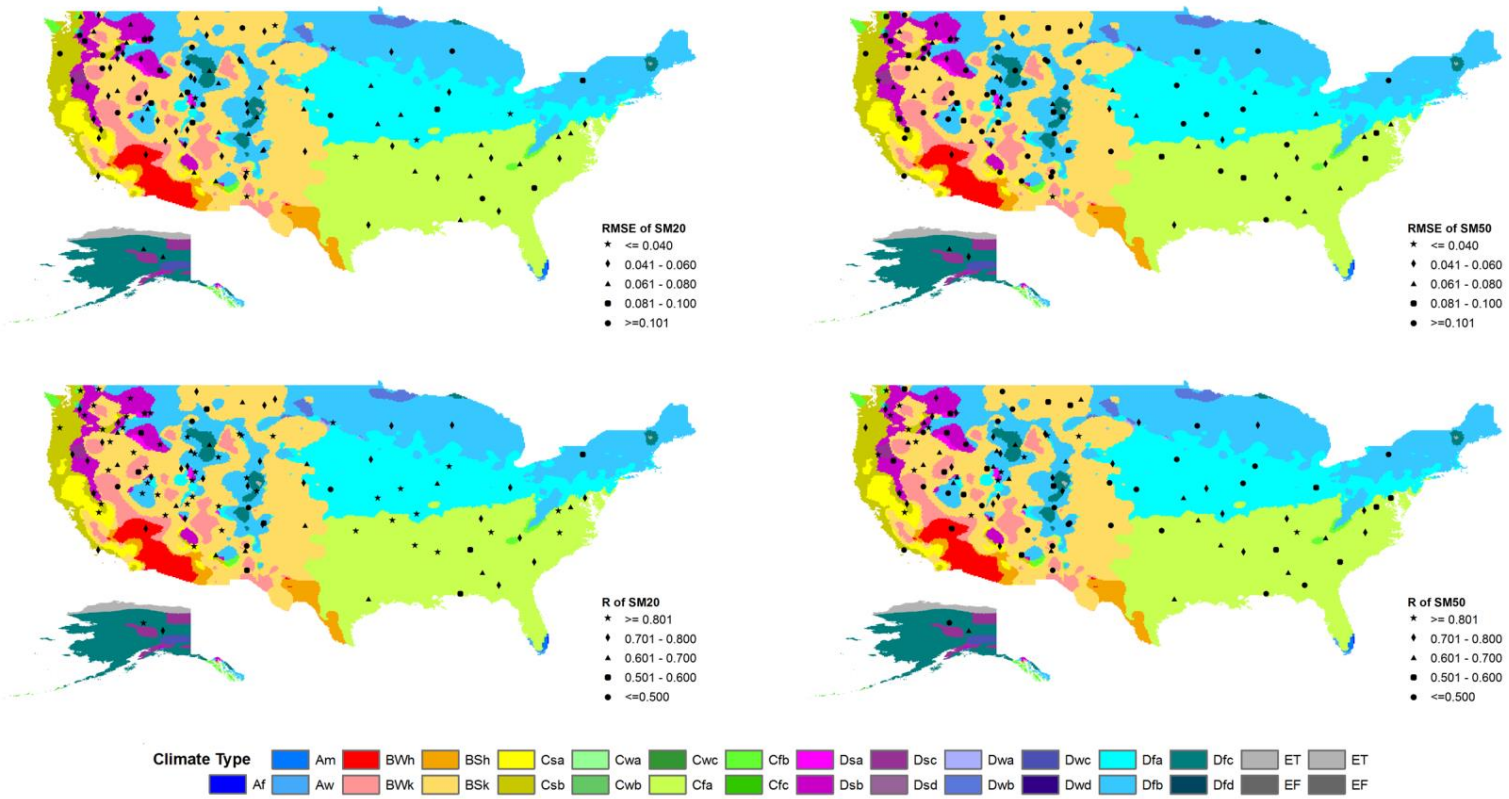

Figure 12 Spatial distribution of RMSE and R of estimation for MLP1-2 with respect to Köppen-Geiger climate classification (Peel et al., 2007)

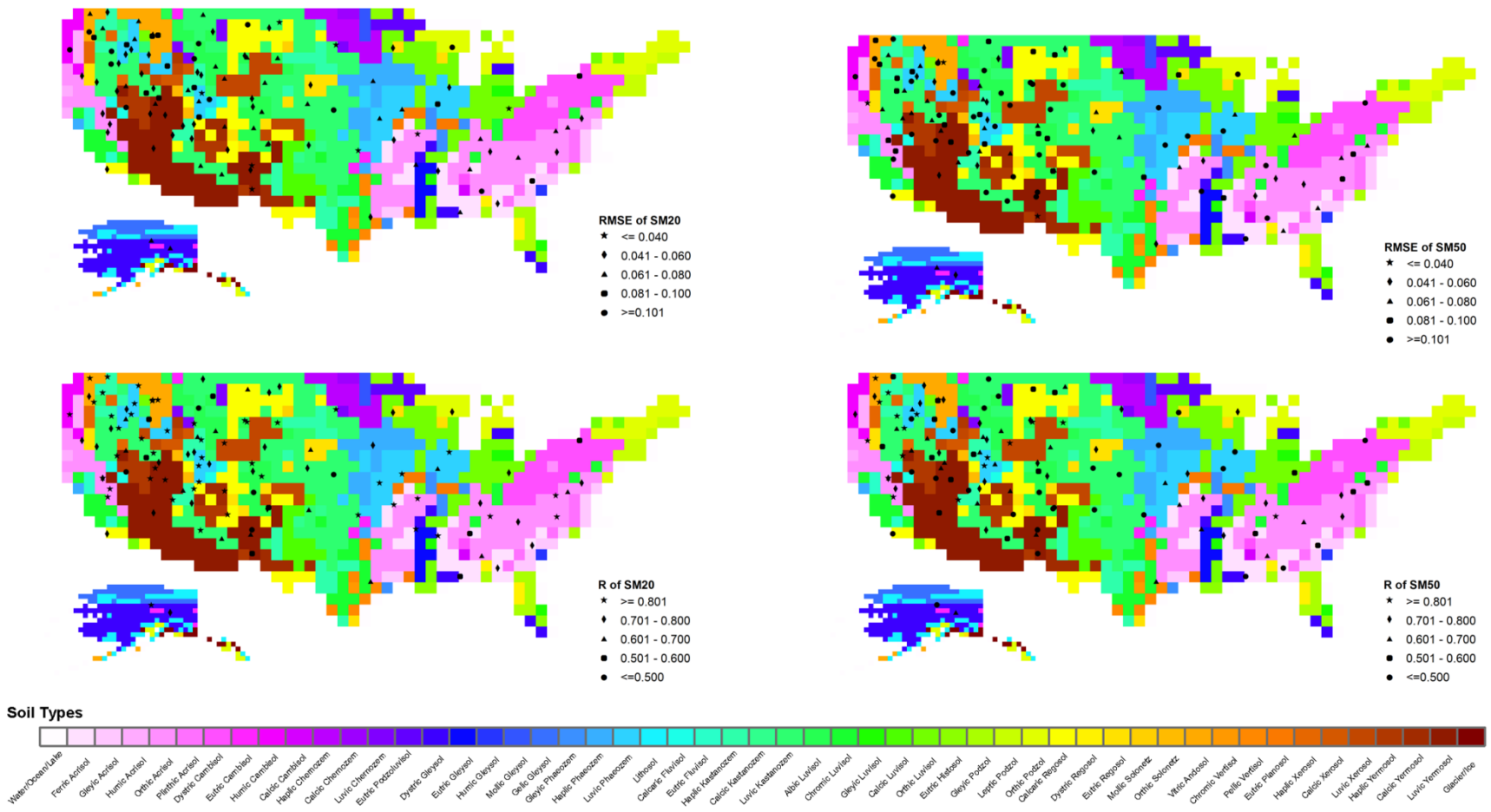

Figure 13 Spatial distribution of RMSE and R of estimation for MLP1-2 with respect to various soil types (Webb et al., 2000)

4.3 Discussion

Comparing with the previous studies (Deng et al., 2011; Elshorbagy et al., 2010b; Ramírez-Beltran et al., 2008), the MLPs in this study did not achieve the same accuracy for RZSM estimation. However, it should be noted that in these studies MLPs were applied in much smaller regions. The estimation results may vary over the locations as found in (De Lannoy et al., 2006; Heathman et al., 2003; Mertens et al., 2005). De Lannoy et al. (2006) reported RMSE ranging from 0.022 to 0.090 m^3m^{-3} for the RZSM estimation of the selected 12 probes, when applying CLM 2.0 model. Heathman et al. (2003) found RMSE between 0.015 and 0.083 m^3m^{-3} for the simulated soil moisture by RZWQM model at the layer 15-30cm of the 4 selected sites. Mertens et al. (2005) simulated soil moisture at the surface and the depths of 30cm and 60cm with MIKE-SHE model for 25 locations and reported RMSE ranging from 0.041 to 0.089m^3m^{-3}. In this study, despite the large scale and high number of stations considered, most of the stations had RMSE for SM20 less than 0.080m^3m^{-3}, which is within the range of the RMSE reported in the previous studies. This accuracy can be considered reasonable as the study area in this research is much larger.

The MLPs were less effective to achieve accurate estimates for SM50, which is consistent with the findings in Kornelsen & Coulibaly (2014b). The MLPs for RZSM retrieval using surface observations relied on the assumption that the surface information was sufficient to infer subsurface soil moisture dynamics (Kornelsen & Coulibaly, 2014b). However, the assumption is invalid when the upper layers and deeper layers decouple. The importance of surface-subsurface coupling has also been

reported in several studies using physical models for the estimation of soil moisture profiles from surface measurements (De Lannoy et al., 2007a; Kumar et al., 2009; Walker et al., 2002). The surface-subsurface decoupling controlled by the soil hydraulic properties may occur in coarse textured and stratified soils (Vereecken et al., 2008) as well as dry condition (Hirschi et al., 2014; Walker et al., 2002). A shallow water table may also cause difficulties to retrieve deeper RZSM using surface observations.

In the previous study (Kornelsen & Coulibaly, 2014b), the potential of MLPs in the application for RZSM estimation in a smaller geographic region was demonstrated. Kornelsen & Coulibaly (2014b) also suggested the feasibility of MLPs to be more general and flexible tools. In contrast, the flexibility of hydrological models may be limited by the correctness of the model physics and the appropriateness of parameter selection. The study of Xia et al. (2014) revealed that land surface models may have large biases in specific regions. In this study, the potential of MLPs for large-scale RZSM estimation is demonstrated. It is shown that the performance of the MLPs is not constrained to specific climate types. A few other advantages of MLPs were also shown in this study. Given that the training data were not required to be temporally continuous, the MLPs avoided the uncertainty from the methods for infilling large gaps. Furthermore, the MLPs were insensitive to potential errors in the soil moisture data. In addition, once the MLPs were trained it cost little computation to simulate the RZSM. The much less demand of the data for land surface properties also makes MLPs more practical than some hydrological models. Considering these advantages, MLPs can be effective and flexible tools for RZSM estimation for the shallow depths (SM20). For the deeper soil layer (SM50), further improvement to the MLP based models is needed.

4.4 Conclusions

With the increasing soil moisture measurements provided by large observation networks and the successful applications of ANNs in modeling various hydrological processes, the potential of MLPs for the retrieval of RZSM at different points across the entire United States was evaluated in this study. The selected state variables as inputs include soil moisture at the depth of 5cm (SM05), evapotranspiration (ET), specific humidity (SH), shortwave radiation (SR), wind speed (WS), accumulative previous 30days rainfall (R30), snowfall (S30), surface air temperature (Ta30) and soil temperature (Ts30). Soil texture information (proportions of sand and clay) were also acquired to reflect terrestrial characteristics. In this study, two experiments consisting of four MLPs were conducted to estimate daily soil moisture at the depths of 20cm (SM20) and 50cm (SM50). A sensitivity analysis indicated that the MLPs were sensitive to soil texture, and variables SM05, Ta30, Ts30, S30 and R30 are generally important for RZSM estimation across a large region. The MLPs had the ability to identify input variables that directly affect the water balance in root zone. The results of training and independent validation reveal that the developed MLPs were valid to generate reasonably accurate estimates of soil moisture at 20cm. The various climatic patterns in the United States caused little impact on the model performance. However, the models were less effective to achieve the desired accuracy of SM50 estimates. The comparison of models in the first experiment shows the advantage of including soil texture data to help the MLP to infer RZSM dynamics, especially SM50. The inclusion of other surface parameters may enhance the ability of MLPs to present the interaction between surface and subsurface, which requires further study. Although the soil moisture data were collected from

stations with various instruments, the MLPs were not sensitive to the potential errors of the datasets, suggesting its robustness. In general, once the MLPs are developed, they can be computationally effective tools to estimate RZSM at 20cm under various types of climate**.**

Chapter 5

Conclusions

5.1 Conclusions

Root zone soil moisture is an important variable for many environmental studies. However, the current measurement techniques are not sufficient to acquire accurate large-scale RZSM data at the spatial resolution of interest. Though many models have been successfully applied in relatively small areas, the large-scale RZSM estimation still faces many difficulties as it requires the high flexibility and practicality of soil moisture models. This study aims to develop feasible models for large-scale RZSM estimation using the surface observations.

Firstly, a literature review was conduct to identify potential methods and it was presented in **Chapter 2**. With respect to the degree of representation of the involved physical processes, the soil moisture models were classified into two categories, namely physically based models as well as statistical and data-driven models. It was found that physically based models were more widely used. As data assimilation was found to have the ability to improve model performance, integrating physically based models with data assimilation methods has been considered as a promising method for soil moisture estimation. However, this method may hard to avoid the errors in model physics affecting the model flexibility. In addition, it requires lots of soil properties data and computational resources, reducing the practicality of this method. On the contrary, statistical and data-driven models have high potential for large-scale RZSM estimation but have not been fully explored. This study applied artificial neural networks, specifically multilayer perceptrons. They are data-driven methods which have been

widely used in hydrology for the high ability of nonlinear input-output mapping (Abrahart et al., 2012; Abrahart & See, 2007; ASCE, 2000a, 2000b) Their potential for large-scale RZSM estimation has also been suggested (Gill et al., 2006; Kornelsen & Coulibaly, 2014b).

Two experiments including four models were developed to estimate RZSM in the United States where there are various climate types and a large number of available soil moisture data. The input variables for the models were carefully selected. The analysis presented in **Chapter 3** showed that the cumulative precipitation and temperature were effective inputs and the accumulation of the previous 30days was generally better than 15 days. The evaluation of the models was fully described in **Chapter 4**. The sensitivity analysis found that MLPs were able to identify the influential forcing variables including SM05, Ta30, Ts30, R30 and S30. The MLPs were extremely sensitive to soil texture information if provided. The evaluation of model performance using the three criteria showed that: (1) the MLPs were effective for the estimation of soil moisture at the depth of 20cm; (2) the model performance was not affected by the local climate; (3) the MLPs were less effective to achieve the desired accuracy for the estimation at the depth of 50cm; (4) the inclusion of soil texture data as inputs can help the MLPs to infer RZSM dynamics, especially for the estimation at 50cm; (5) the MLPs were not sensitive to the potential errors of the soil moisture datasets.

5.2 Future Work

Although this study found that the MLPs were effective to estimate soil moisture at 20cm for a large area, the performance for the estimation at 50cm was not satisfactory.

Such limitation may be explained by the limited information for the MLPs to infer the soil moisture dynamics at the deeper soil layer. In this study, it was shown that the inclusion of soil texture data can improve the estimation at 50cm. The application of other land surface information, such as topographic wetness index and leaf area index, may also help to capture the dynamics in subsurface. Further study is required to improve the estimation at the deeper layer. Furthermore, there are also other statistical and data-driven methods with potential for the applications in large-scale RZSM estimation. SVMs are good candidates and have also been used to estimate RZSM. In some cases (Deng et al., 2011; Gill et al., 2006; Wu et al., 2008; Zhao et al., 2014) they achieved better performance than MLPs for soil moisture estimation. However, it was also found that MLPs had better ability to handle the nonlinearity in the subsurface systems (Elshorbagy et al., 2010a, 2010b). It may be necessary to explore the differences and feasibility of these two methods for large-scale RZSM estimation.

References

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., Wilby, R. L. 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography, 36*(4), 480-513. doi: 10.1177/0309133312444943

Abrahart, R. J., See, L. M. 2007. Neural network modelling of non-linear hydrological relationships. *Hydrology and Earth System Sciences, 11*(5), 1563-1579. doi: 10.5194/hess-11-1563-2007

Albergel, C., Rudiger, C., Pellarin, T., Calvet, J. C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., Martin, E. 2008. From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations. *Hydrology and Earth System Sciences, 12*(6), 1323-1337.

ASCE. 2000a. Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123. doi: 10.1061/(ASCE)1084-0699(2000)5:2(115)

ASCE. 2000b. Artificial Neural Networks in Hydrology. II: Hydrologic Applications. *Journal of Hydrologic Engineering, 5*(2), 124-137. doi: 10.1061/(ASCE)1084-0699(2000)5:2(124)

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., Vitart, F. 2015. ERA-Interim/Land: a global land surface reanalysis data set. *Hydrology and Earth System Sciences, 19*(1), 389-407. doi: 10.5194/hess-19-389-2015

Baroni, G., Tarantola, S. 2014. A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environmental Modelling & Software, 51*, 26-34. doi: 10.1016/j.envsoft.2013.09.022

Bell, J. E., Palecki, M. A., Baker, C. B., Collins, W. G., Lawrimore, J. H., Leeper, R. D., Hall, M. E., Kochendorfer, J., Meyers, T. P., Wilson, T., Diamond, H. J. 2013. U.S. Climate Reference Network Soil Moisture and Temperature Observations. *Journal of Hydrometeorology, 14*(3), 977-988. doi: 10.1175/jhm-d-12-0146.1

Berg, A. A., Famiglietti, J. S., Rodell, M., Reichle, R. H., Jambor, U., Holl, S. L., Houser, P. R. 2005. Development of a hydrometeorological forcing data set for global soil moisture estimation. *International Journal of Climatology, 25*(13), 1697-1714. doi: 10.1002/joc.1203

Braun, F. J., Schadler, G. 2005. Comparison of soil hydraulic parameterizations for mesoscale meteorological models. *Journal of Applied Meteorology, 44*(7), 1116-1132. doi: 10.1175/jam2259.1

Brimelow, J. C., Hanesiak, J. M., Raddatz, R. 2010. Validation of soil moisture simulations from the PAMII model, and an assessment of their sensitivity to uncertainties in soil hydraulic parameters. *Agricultural and Forest Meteorology, 150*(1), 100-114. doi: 10.1016/j.agrformet.2009.09.006

Calvet, J. C., Noilhan, J. 2000. From near-surface to root-zone soil moisture using year-round data. *Journal of Hydrometeorology, 1*(5), 393-411. doi: 10.1175/1525-7541(2000)001<0393:fnstrz>2.0.co;2

Chen, F., Crow, W. T., Starks, P. J., Moriasi, D. N. 2011. Improving hydrologic predictions of a catchment model via assimilation of surface soil moisture. *Advances in Water Resources, 34*(4), 526-536. doi: 10.1016/j.advwatres.2011.01.011

Chirico, G. B., Medina, H., Romano, N. 2014. Kalman filters for assimilating near-surface observations into the Richards equation - Part 1: Retrieving state profiles with linear and nonlinear numerical schemes. *Hydrology and Earth System Sciences, 18*(7), 2503-2520. doi: 10.5194/hess-18-2503-2014

Cornelissen, T., Diekkruger, B., Bogena, H. R. 2014. Significance of scale and lower boundary condition in the 3D simulation of hydrological processes and soil moisture variability in a forested headwater catchment. *Journal of Hydrology, 516*, 140-153. doi: 10.1016/j.jhydrol.2014.01.060

Cosh, M. H., Jackson, T.J., Bindlish, R., Prueger, J.H. 2002. Estimation of watershed scale soil moisture from point measurements in SMEX02. *American Geophysical Union. EOS Trans. of AGU, 83:F507*.

Crawford, T. M., Stensrud, D. J., Carlson, T. N., Capehart, W. J. 2000. Using a soil hydrology model to obtain regionally averaged soil moisture values. *Journal of Hydrometeorology, 1*(4), 353-363. doi: 10.1175/1525-7541(2000)001<0353:uashmt>2.0.co;2

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., Walker, J. P. 2012. Upscaling Sparse Ground-Based Soil Moisture Observations for the Validation of Coarse-Resolution Satellite Soil Moisture Products. *Reviews of Geophysics, 50*, 20. doi: 10.1029/2011rg000372

De Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N., Verhoest, N. E. C. 2006. Assessment of model uncertainty for soil moisture through ensemble verification. *Journal of Geophysical Research-Atmospheres, 111*(D10), 18. doi: 10.1029/2005jd006367

De Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N., Verhoest, N. E. C. 2007a. State and bias estimation for soil moisture profiles by an ensemble Kalman filter: Effect of assimilation depth and frequency. *Water Resources Research, 43*(6), 15. doi: 10.1029/2006wr005100

De Lannoy, G. J. M., Houser, P. R., Verhoest, N. E. C., Pauwels, V. R. N., Gish, T. J. 2007b. Upscaling of point soil moisture measurements to field averages at the OPE3 test site. *Journal of Hydrology, 343*(1-2), 1-11. doi: 10.1016/j.jhydrol.2007.06.004

De Lannoy, G. J. M., Reichle, R. H., Houser, P. R., Pauwels, V. R. N., Verhoest, N. E. C. 2007c. Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter. *Water Resources Research, 43*(9), 14. doi: 10.1029/2006wr005449

Deng, J. Q., Chen, X. M., Du, Z. J., Zhang, Y. 2011. Soil Water Simulation and Predication Using Stochastic Models Based on LS-SVM for Red Soil Region of China. *Water Resources Management, 25*(11), 2823-2836. doi: 10.1007/s11269-011-9840-z

Diamond, H. J., Karl, T. R., Palecki, M. A., Baker, C. B., Bell, J. E., Leeper, R. D., Easterling, D. R., Lawrimore, J. H., Meyers, T. P., Helfert, M. R., Goodge, G., Thorne, P. W. 2013. U.S. Climate Reference Network after One Decade of Operations: Status and Assessment. *Bulletin of the American Meteorological Society, 94*(4), 485-498. doi: 10.1175/bams-d-12-00170.1

Dingman, S. L. 2008. *Physical Hydrology* (Second ed.). America: Waveland Press Inc.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., Jackson, T. 2011. The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences, 15*(5), 1675-1698. doi: 10.5194/hess-15-1675-2011

Dorigo, W. A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A. D., Zamojski, D., Cordes, C., Wagner, W., Drusch, M. 2013. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. *Vadose Zone Journal, 12*(3), 0. doi: 10.2136/vzj2012.0097

Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D. P. 2010a. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology. *Hydrology and Earth System Sciences, 14*(10), 1931-1941. doi: 10.5194/hess-14-1931-2010

Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D. P. 2010b. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application. *Hydrology and Earth System Sciences, 14*(10), 1943-1961. doi: 10.5194/hess-14-1943-2010

Elshorbagy, A., El-Baroudy, I. 2009. Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *Journal of Hydroinformatics, 11*(3-4), 237-251. doi: 10.2166/hydro.2009.032

Elshorbagy, A., Parasuraman, K. 2008. On the relevance of using artificial neural networks for estimating soil moisture content. *Journal of Hydrology, 362*(1-2), 1-18. doi: 10.1016/j.jhydrol.2008.08.012

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L., Van Zyl, J. 2010. The Soil Moisture Active Passive (SMAP) Mission. *Proceedings of the Ieee, 98*(5), 704-716. doi: 10.1109/jproc.2010.2043918

Entekhabi, D., Rodriguez-iturbe, I., Castelli, F. 1996. Mutual interaction of soil moisture state and atmospheric processes. *Journal of Hydrology, 184*(1-2), 3-17.

Evensen, G. 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics, 53*(4), 343-367. doi: 10.1007/s10236-003-0036-9

Evensen, G. 2009. *Data Assimilation The Ensemble Kalman Filter* (2nd ed.). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., Lüthi, D., Schär, C. 2007. Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave. *Journal of Climate, 20*(20), 5081-5099. doi: 10.1175/jcli4288.1

Ford, T. W., Harris, E., Quiring, S. M. 2014. Estimating root zone soil moisture using near-surface observations from SMOS. *Hydrology and Earth System Sciences, 18*(1), 139-154. doi: 10.5194/hess-18-139-2014

Ghedira, H., Lakhankar, T., Jahan, N., Khanbilvard, R., ieee. 2004. *Combination of passive and active microwave data for soil moisture estimates*. New York: Ieee.

Gill, M. K., Asefa, T., Kemblowski, M. W., McKee, M. 2006. Soil moisture prediction using support vector machines. *Journal of the American Water Resources Association, 42*(4), 1033-1046. doi: 10.1111/j.1752-1688.2006.tb04512.x

Gill, M. K., Kemblowski, M. W., McKee, M. 2007. Soil Moisture Data Assimilation Using Support Vector Machines and Ensemble Kalman Filter. *Journal of the American Water Resources Association, 43*(4), 1004-1015. doi: 10.1111/j.1752-1688.2007.00082.x

Greve, P., Warrach-Sagi, K., Wulfmeyer, V. 2013. Evaluating Soil Water Content in a WRF-Noah Downscaling Experiment. *Journal of Applied Meteorology and Climatology, 52*(10), 2312-2327. doi: 10.1175/jamc-d-12-0239.1

Guswa, A. J., Celia, M. A., Rodriguez-Iturbe, I. 2002. Models of soil moisture dynamics in ecohydrology: A comparative study. *Water Resources Research, 38*(9), 15. doi: 10.1029/2001wr000826

Hagan, M. T., Demuth, H. B., Beale, M. 2002. *Neural Network Design*. Boston, MA 02116: PWS Publishing Co.

Han, E. J., Merwade, V., Heathman, G. C. 2012. Application of data assimilation with the Root Zone Water Quality Model for soil moisture profile estimation in the upper Cedar Creek, Indiana. *Hydrological Processes, 26*(11), 1707-1719. doi: 10.1002/hyp.8292

Haykin, S. 1999. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, New Jersey 07458: Prentice Hall.

Heathman, G. C., Starks, P. J., Ahuja, L. R., Jackson, T. J. 2003. Assimilation of surface soil moisture to estimate profile soil water content. *Journal of Hydrology, 279*(1-4), 1-17. doi: 10.1016/s0022-1694(03)00088-x

Hirschi, M., Mueller, B., Dorigo, W., Seneviratne, S. I. 2014. Using remotely sensed soil moisture for land-atmosphere coupling diagnostics: The role of surface vs. root-zone soil moisture variability. *Remote Sensing of Environment, 154*, 246-252. doi: 10.1016/j.rse.2014.08.030

Huang, C. J., Li, L., Ren, S. H., Zhou, Z. S. 2011. Research of Soil Moisture Content Forecast Model Based on Genetic Algorithm BP Neural Network. In D. L. Li, Y. Liu, Y. Y. Chen (Eds.), *Computer and Computing Technologies in Agriculture Iv, Pt 2* (Vol. 345, pp. 309-316). Berlin: Springer-Verlag Berlin.

Huang, J., vandenDool, H. M., Georgakakos, K. P. 1996. Analysis of model-calculated soil moisture over the United States (1931-1993) and applications to long-range temperature forecasts. *Journal of Climate, 9*(6), 1350-1362. doi: 10.1175/1520-0442(1996)009<1350:aomcsm>2.0.co;2

Hurkmans, R., Paniconi, C., Troch, P. A. 2006. Numerical assessment of a dynamical relaxation data assimilation scheme for a catchment hydrological model. *Hydrological Processes, 20*(3), 549-563. doi: 10.1002/hyp.5921

Jacques, D., Simunek, J., Timmerman, A., Feyen, J. 2002. Calibration of Richards' and convection-dispersion equations to field-scale water flow and solute transport under rainfall conditions. *Journal of Hydrology, 259*(1-4), 15-31. doi: 10.1016/s0022-1694(01)00591-1

Jajarmizadeh, M., Harun, M., Salarpour, M. 2012. A Review on Theoretical Consideration and Types of Models in Hydrology. *Journal of Environmental Science and Technology, 5*(5), 249-261. doi: 10.3923/jest.2012.249.261

Jiang, H. L., Cotton, W. R. 2004. Soil moisture estimation using an artificial neural network: a feasibility study. *Canadian Journal of Remote Sensing, 30*(5), 827-839.

Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering, 82*(D), 35-45.

Kerr, Y. H., Waldteufel, P., Wigneron, J. P., Martinuzzi, J. M., Font, J., Berger, M. 2001. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission. *Ieee Transactions on Geoscience and Remote Sensing, 39*(8), 1729-1735. doi: 10.1109/36.942551

Khan, M. S., Coulibaly, P. 2006. Bayesian neural network for rainfall-runoff modeling. *Water Resources Research, 42*(7), 18. doi: 10.1029/2005wr003971

Kornelsen, K., Coulibaly, P. 2013. Advances in soil moisture retrieval from synthetic aperture radar and hydrological applications. *Journal of Hydrology, 476*, 460-489. doi: 10.1016/j.jhydrol.2012.10.044

Kornelsen, K., Coulibaly, P. 2014a. Comparison of Interpolation, Statistical, and Data-Driven Methods for Imputation of Missing Values in a Distributed Soil Moisture Dataset. *Journal of Hydrologic Engineering, 19*(1), 26-43. doi: 10.1061/(asce)he.1943-5584.0000767

Kornelsen, K., Coulibaly, P. 2014b. Root-zone soil moisture estimation using data-driven methods. *Water Resources Research, 40*(4), 2946-2962. doi: 10.1002/2013WR014127

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C. T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C. H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., Yamada, T., Team, G. 2004. Regions of strong coupling between soil moisture and precipitation. *Science, 305*(5687), 1138-1140. doi: 10.1126/science.1100217

Kumar, S. V., Reichle, R. H., Koster, R. D., Crow, W. T., Peters-Lidard, C. D. 2009. Role of Subsurface Physics in the Assimilation of Surface Soil Moisture Observations. *Journal of Hydrometeorology, 10*(6), 1534-1547. doi: 10.1175/2009jhm1134.1

Lamorski, K., Pastuszka, T., Krzyszczak, J., Slawinski, C., Witkowska-Walczak, B. 2013. Soil Water Dynamic Modeling Using the Physical and Support Vector Machine Methods. *Vadose Zone Journal, 12*(4), 12. doi: 10.2136/vzj2013.05.0085

Liu, H. B., Xie, D., Wu, W. 2008. Soil water content forecasting by ANN and SVM hybrid architecture. *Environmental Monitoring and Assessment, 143*(1-3), 187-193. doi: 10.1007/s10661-007-9967-9

Lü, H., Yu, Z., Zhu, Y., Drake, S., Hao, Z., Sudicky, E. A. 2011a. Dual state-parameter estimation of root zone soil moisture by optimal parameter estimation and extended Kalman filter data assimilation. *Advances in Water Resources, 34*(3), 395-406. doi: 10.1016/j.advwatres.2010.12.005

Lü, H. S., Li, X. L., Yu, Z. B., Horton, R., Zhu, Y. H., Hao, Z. C., Xiang, L. 2010. Using a H-infinity filter assimilation procedure to estimate root zone soil water content. *Hydrological Processes, 24*(25), 3648-3660. doi: 10.1002/hyp.7778

Lü, H. S., Yu, Z. B., Horton, R., Zhu, Y. H., Wang, Z. L., Hao, Z. C., Xiang, L. 2011b. Multi-scale assimilation of root zone soil water predictions. *Hydrological Processes, 25*(20), 3158-3172. doi: 10.1002/hyp.8034

Malekian, R., Gordon, R., Madani, A., Robertson, S. 2014. Evaluation of the Versatile Soil Moisture Budget model for a humid region in Atlantic Canada. *Canadian Water Resources Journal, 39*(1), 73-82. doi: 10.1080/07011784.2014.888891

Manfreda, S., Brocca, L., Moramarco, T., Melone, F., Sheffield, J. 2014. A physically based approach for the estimation of root-zone soil moisture from surface measurements. *Hydrology and Earth System Sciences, 18*(3), 1199-1212. doi: 10.5194/hess-18-1199-2014

Maxwell, R. M., Miller, N. L. 2005. Development of a coupled land surface and groundwater model. *Journal of Hydrometeorology, 6*(3), 233-247. doi: 10.1175/jhm422.1

Medina, H., Romano, N., Chirico, G. B. 2014a. Kalman filters for assimilating near-surface observations into the Richards equation - Part 2: A dual filter approach for simultaneous retrieval of states and parameters. *Hydrology and Earth System Sciences, 18*(7), 2521-2541. doi: 10.5194/hess-18-2521-2014

Medina, H., Romano, N., Chirico, G. B. 2014b. Kalman filters for assimilating near-surface observations into the Richards equation - Part 3: Retrieving states and parameters from laboratory evaporation experiments. *Hydrology and Earth System Sciences, 18*(7), 2543-2557. doi: 10.5194/hess-18-2543-2014

Mehta, V. K., Walter, M. T., Brooks, E. S., Steenhuis, T. S., Walter, M. F., Johnson, M., Boll, J., Thongs, D. 2004. Application of SMR to modeling watersheds in the Catskill Mountains. *Environmental Modeling & Assessment, 9*(2), 77-89. doi: 10.1023/b:enmo.0000032096.13649.92

Mertens, J., Madsen, H., Kristensen, M., Jacques, D., Feyen, J. 2005. Sensitivity of soil parameters in unsaturated zone modelling and the relation between effective, laboratory and in situ estimates. *Hydrological Processes, 19*(8), 1611-1633. doi: 10.1002/hyp.5591

Mertens, J., Stenger, R., Barkle, G. F. 2006. Multiobjective inverse modeling for soil parameter estimation and model verification. *Vadose Zone Journal, 5*(3), 917-933. doi: 10.2136/vzj2005.0117

Monsivais-Huertero, A., Graham, W. D., Judge, J., Agrawal, D. 2010. Effect of simultaneous state-parameter estimation and forcing uncertainties on root-zone soil moisture for dynamic vegetation using EnKF. *Advances in Water Resources, 33*(4), 468-484. doi: 10.1016/j.advwatres.2010.01.011

Montaldo, N., Albertson, J. D. 2001. On the use of the force-restore SVAT model formulation for stratified soils. *Journal of Hydrometeorology, 2*(6), 571-578. doi: 10.1175/1525-7541(2001)002<0571:otuotf>2.0.co;2

Montaldo, N., Albertson, J. D. 2003. Multi-scale assimilation of surface soil moisture data for robust root zone moisture predictions. *Advances in Water Resources, 26*(1), 33-44. doi: 10.1016/s0309-1708(02)00103-3

Montaldo, N., Albertson, J. D., Mancini, M. 2007. Dynamic calibration with an ensemble kalman filter based data assimilation approach for root-zone moisture predictions. *Journal of Hydrometeorology, 8*(4), 910-921. doi: 10.1175/jhm582.1

Montaldo, N., Albertson, J. D., Mancini, M., Kiely, G. 2001. Robust simulation of root zone soil moisture with assimilation of surface soil moisture data. *Water Resources Research, 37*(12), 2889-2900. doi: 10.1029/2000wr000209

Nagarajan, K., Judge, J., Graham, W. D., Monsivais-Huertero, A. 2011. Particle Filter-based assimilation algorithms for improved estimation of root-zone soil moisture under dynamic vegetation conditions. *Advances in Water Resources, 34*(4), 433-447. doi: 10.1016/j.advwatres.2010.09.019

Nearing, G. S., Gupta, H. V., Crow, W. T., Gong, W. 2013. An approach to quantifying the efficiency of a Bayesian filter. *Water Resources Research, 49*(4), 2164-2173. doi: 10.1002/wrcr.20177

Nguyen, D., Widrow, B., Ieee. 1990. IMPROVING THE LEARNING SPEED OF 2-LAYER NEURAL NETWORKS BY CHOOSING INITIAL VALUES OF THE ADAPTIVE WEIGHTS. *Ijcnn International Joint Conference on Neural Networks, Vols 1-3*, C21-C26.

Nishat, S., Guo, Y., Baetz, B. W. 2007. Development of a simplified continuous simulation model for investigating long-term soil moisture fluctuations. *Agricultural Water Management, 92*(ayer1-2), 53-63. doi: 10.1016/j.agwat.2007.04.012

Ochsner, T. E., Cosh, M. H., Cuenca, R. H., Dorigo, W. A., Draper, C. S., Hagimoto, Y., Kerr, Y. H., Njoku, E. G., Small, E. E., Zreda, M. 2013. State of the Art in Large-Scale Soil Moisture Monitoring. *Soil Science Society of America Journal, 77*(6), 1888-1919. doi: 10.2136/sssaj2013.03.0093

Pan, H. L., Mahrt, L. 1987. Interaction between Soil Hydrology and Boundary-Layer Development. *Boundary-Layer Meteorology, 38*(1-2), 185-202. doi: 10.1007/bf00121563

Paniconi, C., Marrocu, M., Putti, M., Verbunt, M. 2003. Newtonian nudging for a Richards equation-based distributed hydrological model. *Advances in Water Resources, 26*(2), 161-178. doi: 10.1016/s0309-1708(02)00099-4

Panigrahi, B., Panda, S. N. 2003. Field test of a soil water balance simulation model. *Agricultural Water Management, 58*(3), 223-240. doi: 10.1016/s0378-3774(02)00082-3

Peel, M. C., Finlayson, B. L., McMahon, T. A. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences, 11*(5), 1633-1644. doi: 10.5194/hess-11-1633-2007

Qin, S. X., Ma, J. W., Wang, X. J. 2013. Development of a hierarchical Bayesian network algorithm for land surface data assimilation. *International Journal of Remote Sensing, 34*(6), 1905-1927. doi: 10.1080/01431161.2012.727495

Qiu, Y., Fu, B. J., Wang, J., Chen, L. D. 2003. Spatiotemporal prediction of soil moisture content using multiple-linear regression in a small catchment of the Loess Plateau, China. *Catena, 54*(1-2), 173-195. doi: 10.1016/s0341-8162(03)00067-x

Ramírez-Beltran, N. D., Castro, J. M., Harmsen, E., Vásquez, R. 2008. Stochastic Transfer Function Model and Neural Networks to Estimate Soil Moisture1. *Journal of the American Water Resources Association, 44*(4), 847-865. doi: 10.1111/j.1752-1688.2008.00208.x

Reichle, R. H. 2008. Data assimilation methods in the Earth sciences. *Advances in Water Resources, 31*(11), 1411-1418. doi: 10.1016/j.advwatres.2008.01.001

Richards, L. A. 1931. Capillary Conduction of Liquids Through Porous Mediums. *Physics, 1*(5), 318-333. doi: 10.1063/1.1745010

Robinson, D. A., Campbell, C. S., Hopmans, J. W., Hornbuckle, B. K., Jones, S. B., Knight, R., Ogden, F., Selker, J., Wendroth, O. 2008. Soil Moisture Measurement for Ecological and Hydrological Watershed-Scale Observatories: A Review. *Vadose Zone Journal, 7*(1), 358. doi: 10.2136/vzj2007.0143

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C. J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin*, J. K., Walker, J. P., Lohmann, D., Toll, D. 2004. The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society, 85*(3), 381-394. doi: 10.1175/bams-85-3-381

Romano, N. 2014. Soil moisture at local scale: Measurements and simulations. *Journal of Hydrology, 516*, 6-20. doi: 10.1016/j.jhydrol.2014.01.026

Romano, N., Palladino, M., Chirico, G. B. 2011. Parameterization of a bucket model for soil-vegetation-atmosphere modeling under seasonal climatic regimes. *Hydrology and Earth System Sciences, 15*(12), 3877-3893. doi: 10.5194/hess-15-3877-2011

Rossler, O., Loffler, J. 2010. Potentials and limitations of modelling spatio-temporal patterns of soil moisture in a high mountain catchment using WaSiM-ETH. *Hydrological Processes, 24*(15), 2182-2196. doi: 10.1002/hyp.7663

Sabater, J. M., Jarlan, L., Calvet, J. C., Bouyssel, F., De Rosnay, P. 2007. From near-surface to root-zone soil moisture using different assimilation techniques. *Journal of Hydrometeorology, 8*(2), 194-206. doi: 10.1175/jhm571.1

Sabater, J. M., Rudiger, C., Calvet, J. C., Fritz, N., Jarlan, L., Kerr, Y. 2008. Joint assimilation of surface soil moisture and LAI observations into a land surface model. *Agricultural and Forest Meteorology, 148*(8-9), 1362-1373. doi: 10.1016/j.agrformet.2008.04.003

Schaap, M. G., Leij, F. J., van Genuchten, M. T. 2001. ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology, 251*(3-4), 163-176. doi: 10.1016/s0022-1694(01)00466-8

Schaedler, G. 2007. A comparison of continuous soil moisture simulations using different soil hydraulic parameterizations for a site in Germany. *Journal of Applied Meteorology and Climatology, 46*(8), 1275-1289. doi: 10.1175/jam2528.1

Schaefer, G. L., Cosh, M. H., Jackson, T. J. 2007. The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *Journal of Atmospheric and Oceanic Technology, 24*(12), 2073-2077. doi: 10.1175/2007jtecha930.1

Sheikh, V., van Loon, E. E. 2007. Comparing performance and parameterization of a one-dimensional unsaturated zone model across scales. *Vadose Zone Journal, 6*(3), 638-650. doi: 10.2136/vzj2006.0077

Sheikh, V., Visser, S., Stroosnijder, L. 2009. A simple model to predict soil moisture: Bridging Event and Continuous Hydrological (BEACH) modelling. *Environmental Modelling & Software, 24*(4), 542-556. doi: 10.1016/j.envsoft.2008.10.005

Starks, P. J., Heathman, G. C., Ahuja, L. R., Ma, L. W. 2003. Use of limited soil property data and modeling to estimate root zone soil water content. *Journal of Hydrology, 272*(1-4), 131-147. doi: 10.1016/s0022-1694(02)00260-3

Subbaiah, R. 2013. A review of models for predicting soil water dynamics during trickle irrigation. *Irrigation Science, 31*(3), 225-258. doi: 10.1007/s00271-011-0309-x

Tavakoli, M., De Smedt, F. 2013. Validation of soil moisture simulation with a distributed hydrologic model (WetSpa). *Environmental Earth Sciences, 69*(3), 739-747. doi: 10.1007/s12665-012-1957-8

Teuling, A. J., Uijlenhoet, R., Hupet, F., van Loon, E. E., Troch, P. A. 2006. Estimating spatial mean root-zone soil moisture from point-scale observations. *Hydrology and Earth System Sciences, 10*(5), 755-767.

Turchin, V. F. 1977. *The Phenomenon of Science: A Cybernetic Approach to Human Evolution*. New York: Columbia University Press.

Twarakavi, N. K. C., Simunek, J., Schaap, M. G. 2010. Can texture-based classification optimally classify soils with respect to soil hydraulics? *Water Resources Research, 46*, 11. doi: 10.1029/2009wr007939

USDA, NRCS. 2010. Snow Survey and Water Supply Forecasting *National Engineering Handbook, Part 622*. Washington, DC.

USDA, NRCS. 2012. Snow Survey and Water Supply Forecasting *National Engineering Handbook, Part 622*. Washington, DC.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer, New York.

Vapnik, V. 1998. *Statistical learning theory*. New York: Wiley, New York

Vereecken, H., Huisman, J. A., Bogena, H., Vanderborght, J., Vrugt, J. A., Hopmans, J. W. 2008. On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research, 44*(4), n/a-n/a. doi: 10.1029/2008wr006829

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G., van Genuchten, M. T. 2010. Using Pedotransfer Functions to Estimate the van Genuchten-Mualem Soil Hydraulic Properties: A Review. *Vadose Zone Journal, 9*(4), 795-820. doi: 10.2136/vzj2010.0045

Walker, J. P., Willgoose, G. R., Kalma, J. D. 2001a. One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application. *Journal of Hydrometeorology, 2*(4), 356-373. doi: 10.1175/1525-7541(2001)002<0356:odsmpr>2.0.co;2

Walker, J. P., Willgoose, G. R., Kalma, J. D. 2001b. One-dimensional soil moisture profile retrieval by assimilation of near-surface observations: a comparison of retrieval algorithms. *Advances in Water Resources, 24*(6), 631-650. doi: 10.1016/s0309-1708(00)00043-9

Walker, J. P., Willgoose, G. R., Kalma, J. D. 2002. Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: Simplified Kalman

filter covariance forecasting and field application. *Water Resources Research, 38*(12), 37-31-37-13. doi: 10.1029/2002wr001545

Wang, F., Wang, L., Koike, T., Zhou, H., Yang, K., Wang, A., Li, W. 2011. Evaluation and application of a fine-resolution global data set in a semiarid mesoscale river basin with a distributed biosphere hydrological model. *Journal of Geophysical Research, 116*(D21). doi: 10.1029/2011jd015990

Wang, W., Jones, P., Partridge, D. 2000. Assessing the impact of input features in a feedforward neural network. *Neural Computing & Applications, 9*(2), 101-112. doi: 10.1007/pl00009895

Webb, R. S., Rosenzweig, C. E., Levine, E. R. 2000. *Global Soil Texture and Derived Water-Holding Capacities (Webb et al.)*. Oak Ridge, Tennessee, U.S.A. : Oak Ridge National Laboratory Distributed Active Archive Center.

Wegehenkel, M. 2005. Validation of a soil water balance model using soil water content and pressure head data. *Hydrological Processes, 19*(6), 1139-1164. doi: 10.1002/hyp.5557

Williams, C. J., McNamara, J. P., Chandler, D. G. 2009. Controls on the temporal and spatial variability of soil moisture in a mountainous landscape: the signature of snow and complex terrain. *Hydrology and Earth System Sciences, 13*(7), 1325-1336.

Wu, W., Wang, X., Xie, D. T., Liu, H. B. 2008. Soil water content forecasting by support vector machine in purple hilly region. In D. L. Li (Ed.), *Computer and Computing Technologies in Agriculture, Vol 1* (Vol. 258, pp. 223-230). New York: Springer.

Xia, Y. L., Sheffield, J., Ek, M. B., Dong, J. R., Chaney, N., Wei, H. L., Meng, J., Wood, E. F. 2014. Evaluation of multi-model simulated soil moisture in NLDAS-2. *Journal of Hydrology, 512*, 107-125. doi: 10.1016/j.jhydrol.2014.02.027

Yamaguchi, Y., Shinoda, M. 2002. Soil moisture modeling based on multiyear observations in the Sahel. *Journal of Applied Meteorology, 41*(11), 1140-1146. doi: 10.1175/1520-0450(2002)041<1140:smmbom>2.0.co;2

Yang, N., Liu, L. M., Xiang, F. 2009. *Use BP Network to Retrieve Soil Moisture with Multiple Meterological Paramenters*. New York: Ieee.

Yonaba, H., Anctil, F., Fortin, V. 2010. Comparing Sigmoid Transfer Functions for Neural Network Multistep Ahead Streamflow Forecasting. *Journal of Hydrologic Engineering, 15*(4), 275-283. doi: 10.1061//ASCE/HE.1943-5584.0000188

Yu, Z. B., Liu, D., Lu, H. S., Fu, X. L., Xiang, L., Zhu, Y. H. 2012. A multi-layer soil moisture data assimilation using support vector machines and ensemble particle filter. *Journal of Hydrology, 475*, 53-64. doi: 10.1016/j.jhydrol.2012.08.034

Yuan, X., Liang, X. Z. 2011. Evaluation of a Conjunctive Surface-Subsurface Process Model (CSSP) over the Contiguous United States at Regional-Local Scales. *Journal of Hydrometeorology, 12*(4), 579-599. doi: 10.1175/2010jhm1302.1

Zeng, X. B., Decker, M. 2009. Improving the Numerical Solution of Soil Moisture-Based Richards Equation for Land Models with a Deep or Shallow Water Table. *Journal of Hydrometeorology, 10*(1), 308-319. doi: 10.1175/2008jhm1011.1

Zhang, S. W., Li, H. R., Zhang, W. D., Qiu, C. J., Li, X. 2005. Estimating the soil moisture profile by assimilating near-surface observations with the ensemble Kalman filter (EnKF). *Advances in Atmospheric Sciences, 22*(6), 936-945.

Zhao, L. X., Shui, P. B., Jiang, F., Qiu, H. Q., Ren, S. M., Li, Y. K., Zhang, Y. 2014. Using monitoring data of surface soil to predict whole crop-root zone soil water content with PSO-LSSVM, GRNN and WNN. *Earth Science Informatics, 7*(1), 59-68. doi: 10.1007/s12145-013-0130-6

Zou, P., Yang, J., Fu, J., Liu, G., Li, D. 2010. Artificial neural network and time series models for predicting soil salt and water content. *Agricultural Water Management, 97*(12), 2009-2019. doi: 10.1016/j.agwat.2010.02.011

Appendix 1: Studies using physically based models for root zone soil moisture estimation (2000-2014)

Reference	Study area/ Synthetic test	Models and methods	Major results and findings
Calvet & Noilhan (2000)	MUREX fallow land, France (0.175 km^2)	- ISBA model - quasi-Newton optimization algorithm - VAR	- Well-calibrated model performance: NSE: 89% RMSE: 0.015 m^3m^{-3} - Best assimilation: NSE: 54%, RMSE: 0.030 m^3m^{-3}
Crawford et al. (2000)	38 sites from Mesonet, Oklahoma, US	- Soil Hydrology Model	- The model performed better for the upper soil layer than deeper layers, where the variation was sluggish. (RMSE:0.09-0.13m^3m^{-3}, R^2:0.27-0.64)
Montaldo & Albertson (2001)	A field in Durham, North Carolina, US	- ISBA model	- The modification used to deal with the vertical discontinuity of soil hydraulic properties for stratified soils provided great improvements, reducing RMSE from0.048 to 0.017m^3m^{-3}.
Montaldo et al. (2001)	A site in Cork, Ireland (0.15 km^2)	- ISBA model - Direct insertion assimilating the inverse soil moisture using surface observations	- The method performed well over 3 orders of magnitude of misspecification of saturated hydraulic conductivity and improved the model skill under uncertain initial conditions.
(Walker et al., 2001a)	Nerrigundah catchment, New South Wales, Australia (0.06 km^2)	- ABDOMEN model - "NLFIT" Bayesian nonlinear regression program - SKF	- The model was an excellent approximation to the Richards equation. - Using SKF was only as good as the calibrated model. - The assimilation method could correct errors in initial conditions and forcing data but could not correct errors in the model physics.
Guswa et al. (2002)	Nylsvley, Africa	- A budget model - Richards equation	- In African savannah, the two models performed similarly only if the plant can extract water vertically from the wet areas to make up for roots in the dry areas.
Jacques et al. (2002)	A field in Bekkevoort, Belgium	- Richards equation - Levenberg-Marquardt optimization	- The method with various soil parameters at different layers was better than the method assuming a homogeneous profile.
Walker et al. (2002)	Nerrigundah catchment, New South Wales, Australia (0.06 km^2)	- Quasi 3D ABDOMEN model - "NLFIT" Bayesian nonlinear regression program - Modified KF	- Soil moisture profile cannot be retrieved when surface and subsurface soil layers decouple. - The assimilation provide improvement for poor stimulation by reducing RMSE of average 0.07 m^3m^{-3} and it was insensitive to poor initialization but it caused slight degradation for good stimulation. - Update frequency depended on errors in model physics and forcing data.

Yamaguchi & Shinoda (2002)	Two sites in the Sahel, Niger	- A water balance model - A exponential function using antecedent precipitation index	- Both models simulated the seasonal and interannual variability of soil moisture reasonably well in the semiarid region. R: 0.93-0.96.
Montaldo & Albertson (2003)	A site in Cork, Ireland ($0.15\ km^2$)	- ISBA model - Multi-scale data assimilation with state and parameter updated	- The method provided great improvement and outperformed the method with only state updated. RMSE:0.037-0.080m^3m^{-3}
Paniconi et al. (2003)	A hypothetical test catchment ($0.165\ km^2$)	- 3D CATHY model - Newtown nudging for 4D data assimilation	- Nudging improved the stimulation, decreasing errors by about 30% when perturbation was introduced in forcing data and initial conditions. - The assimilation method t introduced little computational cost.
Panigrahi & Panda (2003)	A farm in Kharagpur, India. ($0.003\ km^2$)	- A budget model	- The model successfully estimated the soil moisture in the active root zone of the crop. R^2: 0.92-0.95
Starks et al. (2003)	Little Washita River Experimental Watershed, Oklahoma, US ($610\ km^2$)	- RZWQM model	- Scenarios with soil property information of textural class only or hydraulic properties determined in situ achieved the smallest errors. - For textural class only: RMSE:0.01-0.02m^3m^{-3}, R: 0.51-0.92
Mehta et al. (2004)	Town Brook (37 km^2) and Biscuit Brook (9.6 km^2), New York, US	- SMR model	- The model well simulated the streamflow in snowmelt and non-snowmelt periods, and can well depicted spatial distribution of soil moisture with R^2:0.63-0.79
Berg et al. (2005)	Data from Iowa, Illinois, Mongolia, Russia and China	- Mosaic LSM	- The simulations had good agreement with the field observations, but were relatively worse for the dry regions. (Average anomaly correlation: 0.38)
Braun & Schadler (2005)	3 sites in Upper Rhine Valley region, Germany	- VEG3D model - A few soil hydraulic functions and parameter sets	- The van Genuchten/Rawls–Brakensiek method gives the best results, outperforming the Campbell/Clapp–Hornberger method especially in the cases of medium and high soil moisture.
Maxwell & Miller (2005)	Usadievskiy Watershed, Valdai, Russia	-CoLM coupled with the groundwater model ParFlow	- The shallow simulations (20cm) are similar for the coupled and uncoupled model, but the coupled model had better results for deeper layers (below 40cm). - The dynamic water table affected watershed flow simulation.
Mertens et al. (2005)	A field in Belgium ($0.004\ km^2$)	-MIKE-SHE model - SCE algorithm - Monte Carlo analysis - Screening method	- The model was mostly sensitive to saturated hydraulic conductivity, saturated soil moisture for the upper layer and the depth to the lower layer. - The calibrated model achieved good results (RMSE: 0.041-0.089m^3m^{-3}), but many calibrated parameters were different from the measured values.

Wegehenkel (2005)	A field in Müncheberg, Germany (0.009 km^2)	- SAWAH model	- The good simulation of soil moisture (RMSE: 0.026-0.049m^3m^{-3}) did not necessarily lead to good simulation of pressure heads.
Zhang et al. (2005)	Synthetic data	- Richards equation - EnKF - Direct insertion	- The EnKF outperformed the direct insertion and was insensitive to the observation depth. - The nonlinearities had negative influence on the optimal estimates but not very seriously.
De Lannoy et al. (2006)	Maryland, US (0.21 km^2)	- CLM 2.0 model - Monte Carlo search	- RMSE:0.022-0.090m^3m^{-3} R:0.47-0.89 - Parameter uncertainty was majorly important and the realistic perturbation of parameters and forcings was most effective to generate ensemble members. - In extreme events the model behaved very nonlinearly.
Hurkmans et al. (2006)	Brisy subcatchment, Belgium (4.64 km^2)	- 3D CATHY model - Newtown nudging for 4D data assimilation	- The assimilation improved the estimation but the numerical cost increased with the increase of the parameters, except for the vertical influence radius. - Intermediate update frequency achieved the lowest errors. - The method was computationally efficient but its results were sensitive to its parameters.
Mertens et al. (2006)	A lysimeter experiment in New Zealand	- HYDRUS-1D - SCE algorithm - Monte Carlo analysis	- SCE algorithm was capable to estimate model parameters in a multiobjective context. - Pareto front can be used to identify model errors.
De Lannoy et al. (2007a)	Maryland, US (0.21 km^2)	- CLM 2.0 model - Friedland's state and bias estimation - Monte Carlo search - EnKF	- Assimilation of complete profiles had the largest effect on the deeper soil layers and the optimal assimilation depth depended on calibration results. - Surface layer assimilation had less impact than the assimilation of the other layers and the propagation of the innovations to the other layers was insufficient. - The bias correction improved the estimation, with optimal update frequency of about 1-2 weeks.
De Lannoy et al. (2007c)	Maryland, US (0.21 km^2)	- CLM 2.0 model - Friedland's state and bias estimation - Monte Carlo search - EnKF	- The bias estimation and correction reduced the RMSE by about 60% but limited to the variables with available observations. - The best algorithm for state and bias estimation depends on one's needs and the selected model. - It was better to post-process the estimates without updating the model states to avoid water imbalance.

Montaldo et al. (2007)	A site in Cork, Ireland (0.15 km^2)	- ISBA model - EnKF with dynamic parameter calibration	- EnKF with only state update failed to reduced effects from the incorrect model parameter, which was improved by state-parameter estimation.
Nishat et al. (2007)	A site in Guelph, Ontario, Canada	- A budget model	- The model can be used to assess the general soil moisture dynamics, but it had overestimation at 37.5cm and underestimation at 80cm. (RMSE: 0.045-0.115m^3m^{-3})
Sabater et al. (2007)	SMOSREX site, France	- ISBA-A-gs model - EKF - EnKF - Simplified 1DVAR - T-VAR	- The 1DVAR method outperformed the other methods with higher accuracy (RMSE: 0.02m^3m^{-3}), less computation than EnKF and T-VAR, and the robustness under a wide range of background or observation errors.
Schaedler (2007)	A site in Upper Rhine Valley region, Germany	- VEG3D model - A few soil hydraulic functions and parameter sets	- The combinations of van Genuchten function with parameter sets Clapp–Hornberger and Rawls–Brakensiek, and Campbell function with Cosby parameter set had the best overall performance. (For all schemes RMSE:0.013-0.078m^3m^{-3}, R:0.74-0.94)
Sheikh & van Loon (2007)	Catsop catchment in the Netherlands (0.42 km^2)	- SWAP model - Levenberg–Marquard optimization method	- The Mualem–van Genuchten parameters α and n calibrated at large-scale were close to the mean of the parameters at point scale. - The model had better results for spatial averaged output than the point scale. (RMSE:0.011-0.051m^3m^{-3})
Albergel et al. (2008)	Southwestern France (a 400km transect)	- Exponential Filter based on water balance with a surface layer and a root zone as reservoir - SIM model suite	- 7 out of the 12 stations had NSE above 0.7, R:0.495-0.958, averaged RMSE:0.031 m^3m^{-3} - The parameter T was affected by soil depth and a climatic effect may exist but not significantly related to soil properties. - The method was insensitive with parameter T.
Sabater et al. (2008)	The SMOSREX site, France	- ISBA-A-gs model - simplified 1DVAR	- The joint assimilation of surface soil moisture and LAI and dynamical correction of the wilting point improved the estimation of root zone soil moisture (RMSE: 0.03m^3m^{-3} NSE: 0.86) and vegetation biomass.
Kumar et al. (2009)	Synthetic data for the contiguous United States (from 30.58N, 124.58W to 50.58N, 75.58W)	- Catchment LSM - Mosaic LSM - Noah LSM - CLM 2.0 - EnKF	- The Catchment LSM had a relatively strong surface-subsurface coupling. - A LSM with strong surface-subsurface coupling may be a more robust choice for assimilation. - The improvements through assimilation were sensitive to the local climate and the soil types. - The improvements were higher if the true subsurface physics had a strong surface-subsurface correlation, especially if the assimilation model also has a strong correlation.

Sheikh et al. (2009)	Catsop Basin, Netherlands (0.42 km^2)	- BEACH model - BUDGET model	- BEACH model could estimate spatially distributed soil moisture with acceptable accuracy. RMSE:0.011-0.065,R:0.46-0.96
Zeng & Decker (2009)	A station in Illinois, US	- CLM3 with a modified bottom boundary	-The new method reduced deficiency in the numerical solution of θ-based Richards equation caused by the free drainage bottom boundary.
Brimelow et al. (2010)	3 DroughtNet sites in Alberta, Canada	- PAMII model	- The model performed well R^2:0.65-0.96. - Using the ensemble pedotransfer functions improved the soil hydraulic parameters estimation.
Lü et al. (2010)	Meilin Experiment Station, China (0.7 km^2)	- Richards equation - H_∞ Filter	-HF was sensitive to soil hydraulic parameters but insensitive to initial soil moisture condition. - HF assimilation improved the simulation: for the best scenario,RMSE:0.02-0.035 m^3m^{-3}, R:0.72-0.92
Monsivais-Huertero et al. (2010)	MicroWEX-2, North Central Florida, US (0.036 km^2)	- LSP-DSSAT model - EnKF	- Simultaneous state–parameter estimation was superior to the open-loop and only-state estimation. RMSE: 0.015-0.023 m^3m^{-3} - The every 3-day assimilation achieved better results. - The effects of forcing uncertainty were insignificant. - Errors in model physics caused errors in estimates.
Rossler & Loffler (2010)	Lötschen valley, Bernese Alps, Switzerland (160 km^2)	- WaSiM-ETH model	- The model reproduced the general soil moisture patterns in the high mountain area with limited accuracy R: 0.47-0.70, due to the coarse weather data and the sensitivity to skeleton fraction.
Chen et al. (2011)	Cobb Creek watershed, Oklahoma, US (341 km^2)	- SWAT - EnKF - SCE-UA	- In EnKF improved the simulation of the upper layer, but the insufficient vertical coupling in the model impedes its ability to update deep soil moisture, ground water flow and surface runoff.
Lü et al. (2011a)	Meilin Experiment Station, China (0.7 km^2)	- Richards equation - Particle swarm optimization algorithm - EKF	-The method with dual state-parameter estimation was superior to the open-loop and the only-state estimation (RMSE: 0.019-0.110 m^3m^{-3}, R: 0.629-0.939). - The assumption of a homogeneous soil profile caused inaccurate estimates at the depth of 100cm.
Lü et al. (2011b)	A station in Anhui, China	- Richards equation - Particle swarm optimization algorithm - Direct insertion	- The dual state-parameter estimation method outperformed direct insertion but the errors were still relatively large. - The wrong bottom boundary affected the results.
Nagarajan et al. (2011)	MicroWEX-2, North Central Florida, US (0.036 km^2)	- LSP-DSSAT model - EnKF -PF	- EnKF outperformed (RMSE:0.019 m^3m^{-3}) to PF (RMSE:0.021-0.029 m^3m^{-3}) and the open-loop (RMSE:0.032 m^3m^{-3}) -Errors in model physics affected the simulation.

Romano et al. (2011)	Synthetic data representing the climate type in Mediterranean	- SWAP model - A budget model	- The field capacity of the budget model decided by a specific point of water retention function may produce differences compared the predictions of SWAP model, especially for coarser soils and in the rainiest season of a Mediterranean climate area.
Yuan & Liang (2011)	The contiguous United States	- CSSP model - CLM 3.5 - CoLM	- CSSP could capture the seasonal and interannual variations in soil moisture and outperformed the other models, but had low amplitude for the annual cycle of surface soil moisture and less credible soil moisture anomaly for semi-humid regions.
Han et al. (2012)	2 sites in Matson Ditch subcatchment , Indiana ($0.049\ km^2$)	- RZWQM model - EnKF - Direct insertion	- EnKF outperformed the other methods (R: 0.38-0.91 RMSE: 0.028-0.103 m^3m^{-3}), especially the upper layers. -The larger update frequency and ensemble size increased the improvements but the results became stable when they were increased to certain values.
Greve et al. (2013)	Southwestern France (a 400km transect)	- WRF-Noah model	- The simulation reproduced the annual cycle but was wetter in winter-spring and drier in summer because of model errors. (RMSE:0.048 m^3m^{-3}, R^2:0.83)
Nearing et al. (2013)	Synthetic data	- A dynamic soil moisture accounting model -EnKF - Shannon entropy	- During assimilation, EnKF did not use all the information available in the surface soil moisture observations.
Qin et al. (2013)	Northern Alabama, US ($14{,}400\ km^2$)	- VIC model - Hierarchical Bayesian network	-The assimilation with Hierarchical Bayesian network improved the simulation and provided spatial and temporal distribution information of soil moisture with good agreement with the observations. R:0.96-0.99, MSE:0.001-0.039 $(m^3m^{-3})^2$
Tavakoli & De Smedt (2013)	A site in Baron Fork river basin, Oklahoma, US	- WetSpa model	- The model achieved good simulation but showed somewhat more abrupt temporal fluctuations in response to rainfall events (RMSE: 0.025 m^3m^{-3}).
Baroni & Tarantola (2014)	A site in Bornim, Brandenburg, Germany ($0.3\ km^2$)	- General Probabilities Framework - SWAP model	- The output soil moisture was more sensitive to the uncertainty in observations, soil properties, than the number of nodes in the model and errors in forcing data. - The uncertainty sources differed for each output variable and evaluation of a model should consider multiple output variables.
Chirico et al. (2014)	Synthetic data	- Different numerical solutions of Richards equation: EX,CN and NL - Different assimilation method: SKF, EKF, UKF, EnKF	-SKF-CN assimilated matric heads more effectively than UKF-NL or EnKF-NL to retrieve matric heads and EX was very computationally inefficient. -When assimilating soil moisture to retrieve matric heads, EKF-CN was more effective than UKF-NL and EnKF-NL, but may not ensure accuracy for all cases. -UKF was as feasible as EnKF for systems of small dimensionality.

Cornelissen et al. (2014)	Wüstebach test site, Germany (0.27 km^2)	- 3D HydroloGeoSphere model	- The model well captured the long-term soil moisture dynamics but poorly for the short-term, because of the neglect of macropore flow. (R^2:above 0.41) - The topsoil spatial pattern was well reproduced but the topographic effect was overestimated. - Considering bedrock slightly improved the results.
Malekian et al. (2014)	A site near Truro, Nova Scotia, Canada	- VSMB model	- The model was able to predict soil moisture with some accuracy (RMSE:0.028-0.036m^3m^{-3}, R^2:0.75-0.90, NSE:0.47-0.79)
Manfreda et al. (2014)	5 stations in West Africa and 3 stations in New Mexico, US	- A budget model - Exponential Filter	- The model was superior to Exponential Filter in semiarid regions though it had more parameters to simulate soil water index (RMSE: 0.019-0.066, R: 0.71-0.97)
Medina et al. (2014a)	Synthetic data	- DSUKF coupled with CN scheme - dual EnKF coupled with NL scheme	- DSUKF with 7 sigma points reduced the uncertainty of the states for any wrong initial parameterization and got similar accuracy but less computational time than DEnKF with 25 members. - The soil hydraulic parameters retrieval was strongly affected by parameter initialization, the range of the states, the boundary conditions and the form of the state-space formulation. -The saturated hydraulic conductivity was hard to identify due to the correlation with other parameters
Medina et al. (2014b)	A laboratory experiment of two soil profiles	- DSUKF with CN scheme	-When assimilating observations at 1cm and 2cm, the retrieved parameters well agreed with that of assimilating the entire observed profiles. - The method had good performance to retrieve soil moisture. RMSE:0.005-0.059m^3m^{-3} -The method was more sensitive to the observation depths than to the update frequency.
Xia et al. (2014)	The contiguous United States	- Noah model - Mosaic model - SAC model - VIC model	- The models have high stimulation skill with averaged anomaly correlation above 0.7, but had large biases which may be caused by model errors or data errors. - Anomaly correlation:-0.06-0.91, relative bias:-40%-100%

Note: RMSE—root mean square error, NSE: Nash-Sutcliffe efficiency

Appendix 2: Studies using statistical and data-driven models for root zone soil moisture estimation (2000-2014)

Reference	Study area	Input data	Method	Major results and findings
Qiu et al. (2003)	Danangou catchment, Loess Plateau, China (3.5 km^2)	Land use, profile slope shape, plan slope shape, aspect, slope gradient, elevation, ground temperature, TA, P, RH, WS, ST	- MLR	- The regression model was capable to predict SM (R^2:0.43-0.79), but the capability decreases with the increase of the soil depth.
Gill et al. (2006)	Little Washita River Experimental Watershed, Oklahoma, US (610 km^2)	The previous and current step of AT, RH, SR, ST at 5 and 10cm, SM	- SVMs - MLPs	- The prediction of SVMs had good agreement with observations and were better than MLPs (MAE: 0.034-0.037 vs. 0.042-0.050m^3m^{-3}, RMSE: 0.041-0.042 vs. 0.060-0.061m^3m^{-3} R: 0.87-0.89 vs. 0.73-0.74).
Gill et al. (2007)	A site at Ames, Iowa, US	The previous and current step of AT, RH, SR, ST at 5 and 10cm, SM	- SVMs - EnKF	- The assimilation improves the prediction accuracy by SVMs. (RMSE: 0.004-0.021 vs. 0.024-0.028m^3m^{-3}, R: 0.93-0.99 vs. 0.81-0.89.)
Elshorbagy & Parasuraman (2008)	Three sites at Athabasca basin, Alberta, Canada (0.03 km^2)	NR, (previous or accumulated) P and AT, (squared) ST	- HONNs - MLPs - SDW	- Using ST and accumulated inputs improved the simulation. - HONNs outperformed MLPs and SDW. - The simulation was affected by the structure and formation of the soil covers. (The best model, R :0.59-0.84)
Liu et al. (2008)	A site at Chongqing, China (0.001 km^2)	SM of previous time steps	- SOMs - SVMs - Coupling SOMs and SVMs	- The hybrid model required less training data and performed better. (RMSE:0.001 vs. 0.030 (SOMs) and 0.033 (SVMs) m^3m^{-3})
Wu et al. (2008)	A site at Chongqing, China (0.001km^2)	SM of previous time steps	- SVMs - MLPs	- SVMs outperformed MLPs but were sensitive to the kernel choice. (RMSE: 0.002-0.033 m^3m^{-3})
Elshorbagy & El-Baroudy (2009)	Athabasca basin, Alberta, Canada (0.03 km^2)	The cumulative or time-lagged NR, AT, P, ST	- EPR - GP	- Using cumulative inputs was better than time-lagged inputs for the storage effect of soil moisture, especially thick soil layer. (RMSE:0.01-0.06 m^3m^{-3}, R:0.53-0.78) - None of them was superior.

Yang et al. (2009)	Observations from Gansu, China	Air pressure, AT, wet bulb temperature, vapour pressure, RH, total clouds amount, low cloud amount, P, WS, sunshine hours, E, ST	- MLPs	- MLPs were feasible to retrieve soil moisture with the meteorological variables if given reasonable parameters and sample size. (RMSE: 0.143 m^3m^{-3}, R:0.70) - Levenberg-Marquardt training algorithm was superior.
Elshorbagy et al. (2010a) Elshorbagy et al. (2010b)	Five datasets from Alberta, Canada and Ourthe subcatchment around the border of France and Netherlands	NR, accumulated P and AT, ST	- MLPs - GP - EPR - SVMs - M5 model trees - KNN - MLR - naïve model	- MLPs outperformed the other methods to simulate soil moisture. (R:0.55-0.6, RMSE: 0.015-0.046m^3m^{-3}) - GP was good for various hydrological fluxes simulation and EPR was close to GP in the less nonlinear case. -SVMs were sensitive to kernel choice. - MLPs, GP, K-nn were suitable for highly nonlinear cases, M5 model trees and K-nn again could succeed in linear cases.
Zou et al. (2010)	A site in North China Plain	SM of previous time steps	- MLPs - ARIMA	- MLPs outperformed ARIMA method for averaged soil moisture profile prediction (R^2=0.90), but ARIMA got better results for prediction at 20cm.
Deng et al. (2011)	A site in Qiyang County, Hunan, China	Cumulative P, AT, E, and daily maximum and minimum temperature	- Simulation syste LSSVMs, MLPs, ANFIS - Prediction system: Chaos Theory, Wavelet Transform, ARMA	- LSSVMs had more stabilities and advantages in soil moisture simulation over MLPS and ANFIS (R: 0.884, RMSE: 0.007 m^3m^{-3}). - The de-noising methods may ignore the details while appropriate wavelet transformation improved the results.
Huang et al. (2011)	Hongxing farm, Heilongjiang, China	P, AT, RH, E, sunshine hours, SM at different depths	- MLPs trained with genetic algorithm	- MLP with genetic algorithm had better simulation capabilities and better generalization ability than usual MLPs. (MAE: 0.019 vs. 0.029 m^3m^{-3})
Yu et al. (2012)	Meilin Experiment Station, China (0.7 km^2)	The previous and current step of AT, RH, SR, ST at 5 and 20cm, SM	- SVMs - Combining EnKF and PF - EnKF - PF	- SVMs were robust and larger training dataset helped to get better results. - Data assimilation improved the prediction and EnPF outperformed EnKF and PF. - The large resampling size exceeding a certain amount degraded the performance of EnPF and PF.

Lamorski et al. (2013)	A site in Lublin, Poland	Maximum daily temperature, P, WS, humidity, SR, antecedent SM	- HYDRUS-1D - SVMs	- SVMs outperformed HYDRUS-1D to get higher accuracy and capture the soil moisture variation. (R^2:0.44-0.86, RMSE: 0.04-0.035 m^3m^{-3}) - The results of upper layers were better.
Kornelsen & Coulibaly (2014a)	4 sites in Halton-Hamilton Watershed, Ontario, Canada	AT, RH, incident SR, potential ET , API, antecedent SM, SM of the upper layer	- MLPs - EPR - other a few statistics methods	- MLPs and interpolation methods outperformed the other methods to infill random missing values and large gaps in soil moisture datasets. (RMSE:0.002-0.117 m^3m^{-3})
Kornelsen & Coulibaly (2014b)	6 Sites in the Lower Great Lakes Area	Surface SM, RH, AT, SR, WS, API, LAI, soil texture , potential ET	- MLPs - HYDRUS-1D - EnKF	- MLPs well estimated the synthetic soil moisture data, but the accuracy was reduced when using field data outside the training condition. (RMSE:0.01-0.07 m^3m^{-3}, R:0.68-0.99) - The transferability of the model was limited to the same geographic region.
(Zhao et al. (2014))	85 stations in Beijing, China	SM at top 10cm, organic matters, saturated water contents	- PCA - PSO-LSSVMs - MLPs - GRNNs - WNNs	- PSO-LSSVMs performed better than the other methods. R^2:0.875

Note: API— antecedent precipitation index, AT—air temperature, E—evaporation, ET—evapotranspiration, P—precipitation, RH— relative humidity, SM—soil moisture, SR—solar radiation, NR—net radiation, ST—soil temperature, WS—wind speed, RMSE—root mean square error, MAE—mean absolute error