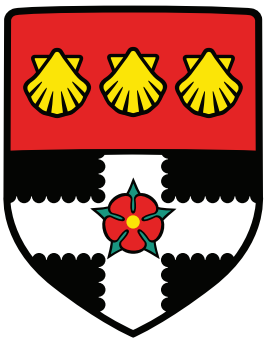


University of Reading

Department of Meteorology

Ensemble generation in land surface models for soil moisture data assimilation

Amsalework Ayele Ejigu

A thesis submitted for the degree of Doctor of Philosophy

May 2020

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Amsalework Ayele Ejigu

Abstract

Soil moisture is a crucial meteorological variable to understand land surface and atmospheric processes like the water cycle, the carbon cycle and the energy balance. However, its link with those processes makes the measurement and modelling difficult. Data assimilation is a mechanism to combine observations with modelled estimates and uncertainties to provide the best prediction for the state or parameter of a system. Proper uncertainty representation is an essential procedure to get a skilful result from the data assimilation.

In this thesis, we demonstrate uncertainty representation techniques in the forcing data, numerical models and the parameters for soil moisture data assimilation. We use the Joint UK Land Environment Simulator land surface model to estimate soil moisture and the Ensemble Transform Kalman Filter and the four-Dimensional Ensemble Variational data assimilation methods to combine soil moisture estimates with observations. Satellite observed, synthetic and *in-situ* soil moisture data are assimilated.

When *in-situ* soil moisture observations for three soil layers are assimilated, employing stochastic forcing via generated rainfall to account for errors in observed rainfall has shown substantial improvement for ensemble spread as well as forecast skill of posterior surface soil moisture. However, additional stochastic forcing via model error is needed to improve forecast skills for the deeper layers. For parameter estimation, prior soil texture parameter errors are represented by the Dirichlet distribution where both share positivity and boundedness. Synthetic data assimilation results show that truth parameters can be recovered even though prior parameters are less informed. The advantage over the Gaussian distribution is that the Dirichlet distribution automatically assigns correlations for the prior covariance matrix. The robustness of the method is tested for different soil types. Posterior parameters obtained from assimilating *in-situ* and satellite observations showed improvement in soil moisture forecast skills beyond the assimilation window. It is also shown that satellite observations are representative of the state of soil moisture for areas with no or less woody vegetation cover.

Acknowledgements

Firstly, I would like to thank my supervisors Dr Tristian Quaife and Dr Amos Lawless, for their invaluable guidance, advice and support throughout the PhD journey. I have learnt not only science but also a different view of the broader world. I also thank Dr Phil Browne and Dr Gernot Geppert for their inputs, who were co-supervisors at different times. I greatly value the efforts of my monitoring committee members Professor Geoff Wadge, Professor Robert Gerney and Dr Ross Banister, for their contributions and follow-ups. Thanks for the DARC research group especially to Dr Ewan Pinnington, Professor Peter Jan Van Leeuwen and Dr Javier Amezcua, for the discussions. Thanks to Dr Peter Craufurd and Mrs Rahel Assefa from CIMMYT and Dr Emily Black for linking my work in context within the TAMASA project, Professor Roger Stern who helped me during the start of the PhD project and Professor Neil Turok, the founder of AIMS, for giving me a chance to pursue my dreams in applied Mathematics. I acknowledge the financial support from the Bill and Melinda Gates Foundation via CIMMYT and the extra funding from the meteorology department.

Special thanks and appreciation for my husband and best friend Melaku, who always supports me whenever I need a hand and for the sacrifices on his professional career to be there for the family. Thanks for Amha for giving up on his holidays due to my working pattern. I extend my thanks to my mum, dad, in-laws and all the family and friends back home in Ethiopia and beyond for the prayers and support. I also thank my niece, Hiwot, for her strength and support especially during the difficult time when we lose dearest sister Yeshi. Thanks for the Ethiopian community in the UK for creating the home environment away from home. Thank you to all colleagues in the department, especially to Dagmawi and Sifan, for checking on me and for discussions, whether it is about meteorology, future career or Ethiopian politics. I appreciate the support from our landlord Richard Coe for availing his house as our own and the Little Learners Nursery staffs for looking after Elda since she was seven months old. *Say to God, "How awesome are Your works! Through the greatness of Your power Your enemies shall submit themselves to You. All the earth shall worship You And sing praises to You; They shall sing praises to Your name." Psalm 66:3-4.*

List of Acronyms

Acronym	Definition
DA	Data Assimilation
DJF	December January February
DRBC	The Darcy Richard Brooks and Corey
Dec	December
ECMWF	European Center for Medium Range Weather Forecast
ES	Error-Spread score
EMPIRE	Employing Message Passing Interface for Researching Ensemble
EnOL	Ensemble Open Loop
ERA - Interim	ECMWF Re-Analysis
ETKF	Ensemble Transform Kalman Filter
IQR	Inter Quartile Range
JJA	June July August
JULES	The Joint UK Land Environment Simulator
LETKF	Localised Ensembled Transform Kalman Filter
RMSD	Root Mean Squared Difference
RMSE	Root Mean Squared Error
Sep	September
WFDEI	WATCH Forcing Data methodology applied to ERA-Interim data
4DEnVar	Four Dimensional Ensemble Variational

Contents

Chapter 1

Introduction

1.1 Motivation

Surface soil moisture is a crucial variable which plays a vital role in land surface hydrology and as a result controls the land-atmosphere interactions. It links fundamental earth system processes: the water cycle, energy balance and carbon cycles (Legates et al., 2011). Figure 1.1 shows a simple representation of the water cycle, linking the land surface and the atmosphere. Root zone soil moisture determines the amount of evaporation and transpiration and related latent heat flux. It also determines how much of the thermal infra-red radiation is reflected into the atmosphere, determines the sensible heat flux and ground heat fluxes (Reichle et al., 2001; Margulis et al., 2002; Reichle et al., 2002; O'Neill et al., 2017). Furthermore, soil moisture is an important variable to forecast extreme weather events like drought and flooding and to suggest solutions for some of the pressing issues of the globe: food insecurity, displacement and disease outbreaks.

In the years between 1900 and 2013, there were 642 drought events reported across the world, and 291 (45%) were from Africa where agriculture is mainly based on rain-fed farming. Particularly in Sub-Saharan Africa, 90% of the staple food production comes from rain-fed farming (Cooper et al., 2008) and a slight rainfall variability jeopardises food security which affects millions in the region (Boyd et al., 2013). Soil moisture deficit, drought, is responsible for up to 60% of the losses for Africa's staple food produce (Tadele, 2017).

To mitigate the effects of drought, different early drought warning and monitoring mechanisms have been developed. The Standard Precipitation Index (SPI) uses precipitation as a

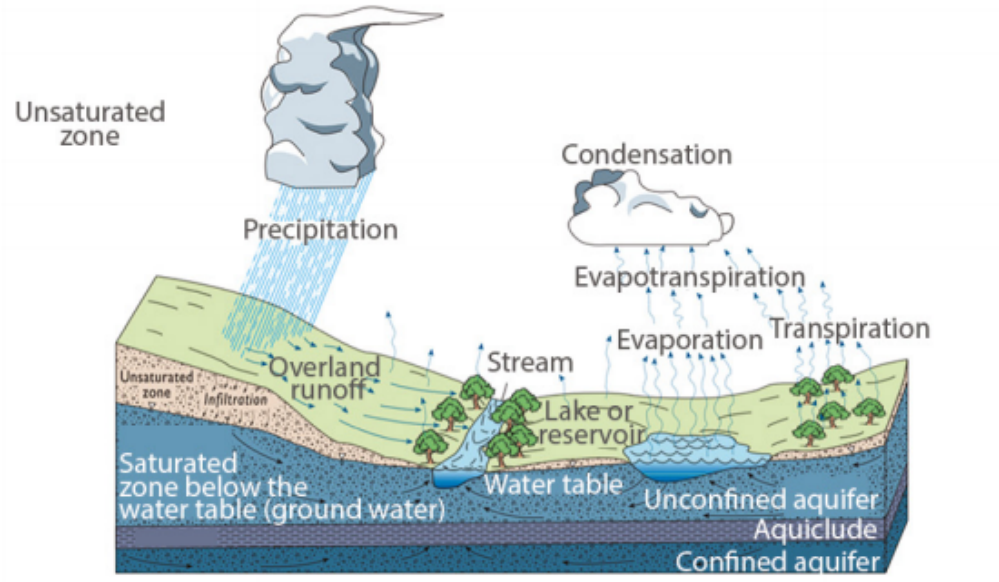


Figure 1.1: Schematic representation of water cycle reproduced from Fandel et al. (2018).

drought indicator. However, precipitation is highly variable and using soil moisture (together with groundwater) is preferable to precipitation since it can indicate the effects of natural water demand and balance (Houborg et al., 2012). Luo et al. (2008) developed a drought index: the percentile of the current soil moisture with the climatology distribution. Sheffield and Wood (2008) used soil moisture-based drought index to characterised the duration, intensity and severity of a drought for the years 1950 – 2000. Narasimhan and Srinivasan (2005); Engda and Kelleners (2016); Steinemann and Cavalcanti (2006); Luo et al. (2008) and Velpuri et al. (2015) also used soil moisture for drought characterisation.

Soil moisture-based drought indices are better than precipitation-based indexes since soil moisture responds to both precipitation and evapotranspiration (Engda and Kelleners, 2016). On average, around 14% of the precipitation fall on land remains after three days in the top 5 cm of the surface soil layer (McColl et al., 2017). From the study by Black et al. (2016) we can see that water availability to plants can be described by the variable β defined as

$$\beta = \begin{cases} 1 \text{ if } \theta \geq \theta_c \\ \frac{\theta - \theta_w}{\theta_c - \theta_w} \text{ if } \theta_w < \theta < \theta_c \\ 0 \text{ if } \theta \leq \theta_w \end{cases} \tag{1.1}$$

where θ is volumetric water content, θ_c is the critical water content, and θ_w is the wilting point. β ranges between zero and one. When it is one, water stress does not affect plant growth (Clark

et al., 2011).

Figure 1.2 shows the correlation of rainfall with soil moisture on the surface and beta as well as the correlation between soil moisture and beta for three time periods for a region in Zambia, the variable β has a better correlation with soil moisture than rainfall. . The main objective of the study is to supply a weather insurance index which characterises agricultural drought based on ensembles of satellite observed rainfall from Tropical Application of Meteorology Using Satellite Data and Ground-Based Observations (TAMSAT) (Maidment et al., 2014). The figure suggested that soil moisture is a better indicator of *beta* than rainfall is, i.e. how stressed plants are can be better determined by soil moisture than by rainfall.

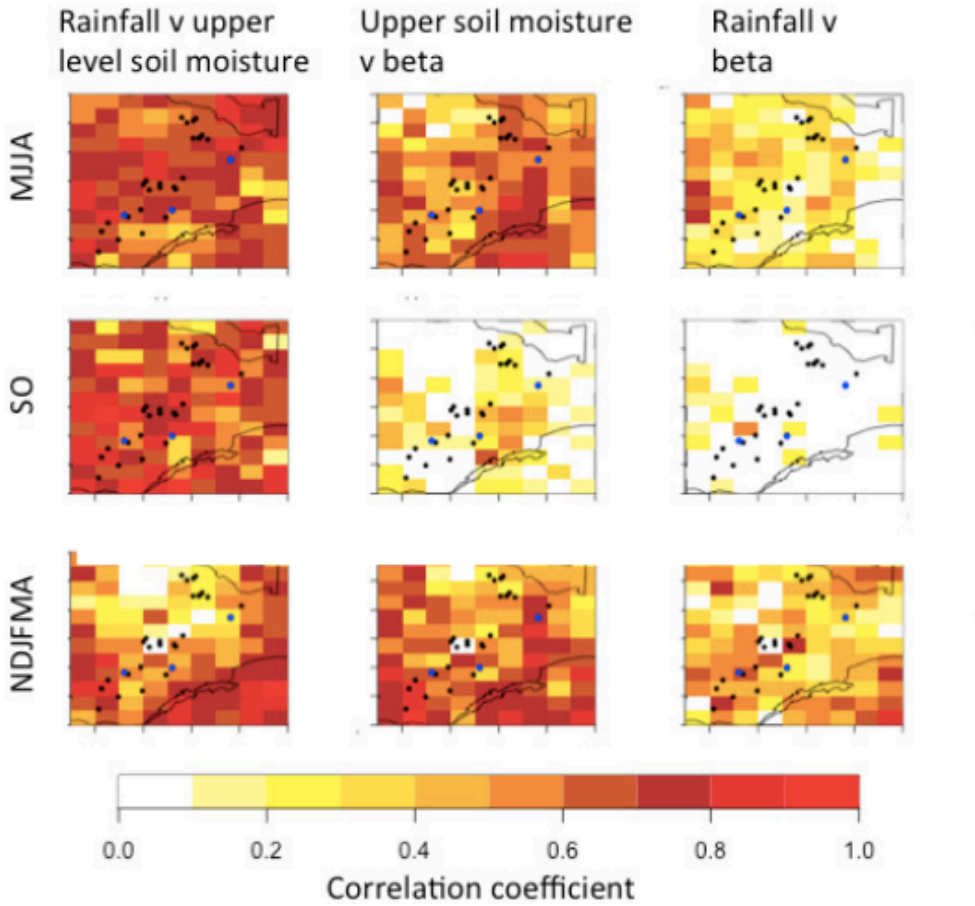


Figure 1.2: correlation between rainfall and upper-level soil moisture (left); upper-level soil moisture and beta (middle); rainfall and beta (right) for (from top to bottom) May-August, September-October, NovemberApril. Black circles are the localities for which loss data are available; blue circles represent the locations in Zambia considered in the study (from west to east: Chikanta, Magoye, Makafu). The rainfall in the figure is from an example TAMSAT ensemble member where top-level soil moisture and beta are outputs of the Joint UK Land and Environment Simulator (JULES) processed-based land surface model by forcing with the TAMSAT rainfall. This figure is adapted from Black et al. (2016).

Flooding affects the lives of individuals and as well as the economy (Silvestro et al., 2019).

For example, in 2007, 8349 people were killed, more than 21 billion $US economic damage was recorded, and 164 million people were affected by flooding (Pappenberger et al., 2008). Soil moisture is useful for flood prediction systems. Massari et al. (2018) used soil moisture to enhance the skill of predicting and characterisation of flooding. In their work, the estimation of rainfall and soil moisture estimation from a river flow model is improved by the help of surface soil moisture observations.

Soil moisture is vital to investigate the effects of changing climate for carbon uptake (Zhao and Li, 2015; Green et al., 2019). A prolonged dry condition impacts the plant transpiration and photosynthesis, and hence it degrades the carbon uptake of the land. As a result, the ecosystem could change from dense vegetation to grassland, for example, due to extended moisture stress (Zeng, 1999; Seneviratne et al., 2010). Vegetation loss is directly related to the available soil moisture which can change land surface properties like surface albedo, evapotranspiration and then precipitation as a result of the soil moisture feedback (Koster et al., 2014; Zhang et al., 2008; Liu et al., 2010). For the regions of moderate soil wetness where available soil moisture limits evapotranspiration, the land-atmosphere coupling is strong. When soil moisture is between the wilting point (below which there is no evaporation) and critical soil moisture (above which is the maximum evapotranspiration), soil moisture is a determining factor for evapotranspiration (Seneviratne et al., 2010).

Extreme weather events need more attention with the growing population, and climate change will most likely exacerbate their occurrence and impact. Hence, accurate knowledge of the state of the soil moisture can save lives and reduce suffering. As such, soil moisture observations are a valuable source of information to help characterise the occurrences and related impacts of extreme events. The following section discusses the state of global soil moisture observations.

1.2 Soil moisture observations

Soil moisture content can be expressed in terms of mass (gravimetric) or volume. The gravimetric representation is the ratio of the mass of water to the mass of dry soil ($kgkg^{-1}$) whereas the volumetric water content (m^3m^{-3}) is a ratio of the volume of water to the sum of the volume of water, gas and soil.

The gravimetric methods require soil sample collection, weighing before and after drying the samples and calculating the soil moisture content. While known to be accurate and mostly used for a one-time soil moisture measurement, these methods are labour intensive and disruptive as they involve removing the soil sample physically and sending samples to the labs. Hence, using gravimetric methods for continuous-time record of soil moisture is not feasible. An alternative approach is an indirect method by measuring one variable and converting it into water content using known relationships between water content and the measured variable. Indirect methods can be implemented from very small spatial coverage (point measurement) up to large area coverage. Satellite observations are remotely sensed, indirect measurements where the measuring resolution could be up to 50 km by 50 km.

In-situ measurements which use indirect methods with automated networks for data entry are a very efficient way of obtaining soil moisture observations, example the Mesoscale Networks soil moisture data: see subsection 3.2.2. The advantage of such measurements is the ability to be obtained at different soil depths, but it is not feasible to have such networks globally. In fact, such measuring networks are limited to the developed world while soil moisture observations are necessary for Africa's agriculture where most of the farming is reliant on rainfall and highly susceptible to drought. Mohanty et al. (2017) showed that only three *in situ* soil moisture sites are present in Africa to validate satellite observed soil moisture data.

The International Soil Moisture Network, (ISMN) is an international effort to provide open access to a homogeneous and quality-controlled data archive of in-situ soil moisture measurements at several depths, Dorigo et al. (2011a), Dorigo et al. (2011b). This is a valuable resource for scientific research, model calibration and validation. In addition to soil moisture, related variables like temperature and precipitation are included in the ISMN dataset. However, despite the effort, the contributors to the network are mainly from the developed world. Top layer observations from *in-situ* measurements are important to validate observations from the satellites despite the difference in spatial resolution.

Satellites, on the other hand, can observe the whole globe within few days but limited to the top surface of the soil, up to 5 cm. Figure 1.3 is global volumetric soil moisture content from the Soil Moisture Active Passive (SMAP) satellite, observed between June 1 - 7, 2015. The SMAP satellite was launched in January 2015 to provide global soil moisture observations with coverage every two to three days (O'Neill et al., 2017) to retrieve soil moisture in addition to

existing satellite missions. For example, Scanning Multichannel Microwave Radiometer (SMMR) and the Special Sensor Microwave - Imager (SSM/I) have been providing soil moisture retrievals in the duration between 1978–1987 and 1987–2007 respectively. Karthikeyan et al. (2017) have listed such other satellite missions.

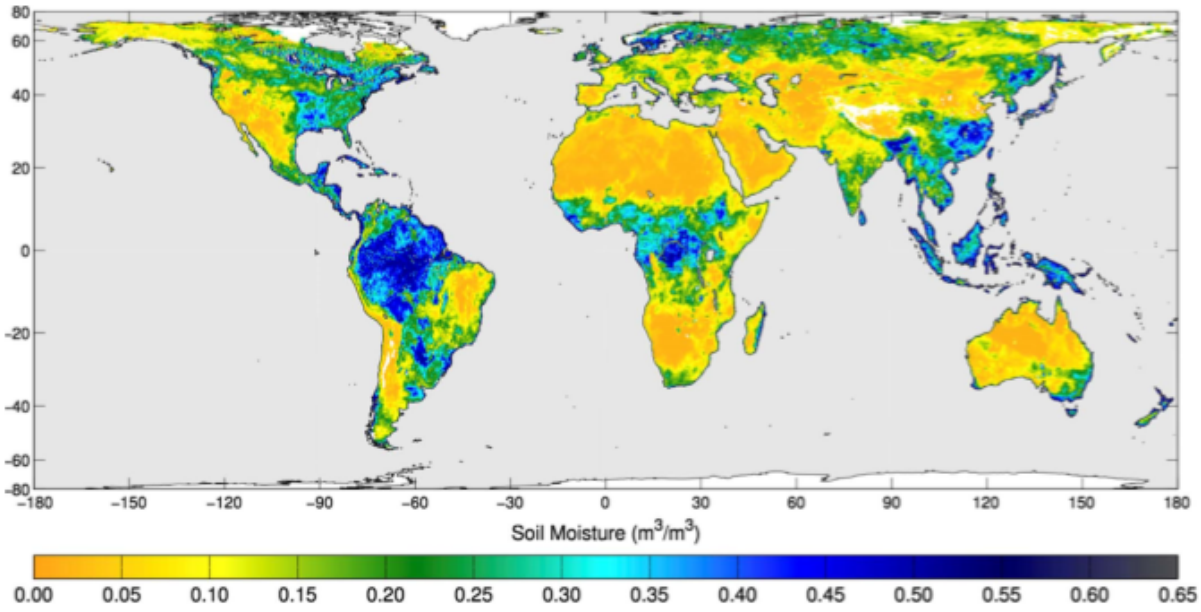


Figure 1.3: SMAP volumetric soil moisture, June 1 - 7 2015 (O'Neill et al., 2017)

The availability of satellite soil moisture observations has made it possible to fill the gaps due to lack of *in-situ* soil moisture data in a variety of applications. Lu et al. (2011) characterised climate change and showed the changing pattern of climate in Africa using satellite soil moisture observations. They retrieved 20 years (1988 - 2007) daily soil moisture data from Special Sensor Microwave/Imagery (SSM/I) using a radiative transfer model. They computed the difference between the second and first decade-average summer soil moisture, JJA for the Northern region and DJF for the Southern region. It was indicated that the southern and central Africa is getting wetter and Northern Africa is drier.

Wagner et al. (2013) highlighted the importance of satellite soil moisture data, from Advanced Scatterometer ASCAT, for several applications. They demonstrated applications in numerical weather predictions, runoff forecasting, vegetation and crop growth monitoring, epidemic risk assessment and societal risk assessment for different parts of the world. Baugh et al. (2020) analysed the impact of soil moisture data from Soil Moisture and Ocean Salinity (SMOS) for streamflow predictions. Zhao et al. (2016) estimated global soil moisture using brightness

temperature observations (highly correlated to soil moisture) from AMSRE.

In the cases where both direct and indirect soil moisture observations are absent, numerical models can be used as a source of information about the state of the soil moisture based on the known physics and meteorology at hand. The following section discusses the role of land surface models for soil moisture prediction.

1.3 Land surface models

Land surface models (LSMs) play a major role in understanding the interaction between the land surface and the atmosphere. They are essential parts of climate models to provide land surface information on climate projections (Pitman, 2003). It is crucial to represent land surface processes as accurate as possible in climate models to understand the changing climate. This is because land surface variables like soil moisture, for example, as explained above, play a major role in climate projections.

The ultimate power of land surface models other than helping to understand the dynamical system of interest is that, they can predict the state of the system at the future time which observations can not tell and with a required spatio-temporal resolution. In the case of soil moisture, land surface models provide estimates at various depths of the soil and without time gaps, as long as the meteorological forcing variables are available. Based on the fundamental relationship of variables, land surface models are also able to output variables or parameters which can not be directly observed.

However, the limitation of land surface models, as any other numerical model, is the uncertainty in the predicted values. Either due to the model uncertainty representing the physical process, uncertainties in the meteorological data used to force the model and/or errors in parameter values used, the prediction could be far from the observed data. Mathematical simplifications to avoid over-parametrisation of the model or missing processes are a source of uncertainty in land surface models (Notaro, 2008). Observations of meteorological variables like rainfall and temperature are subject to measurement errors. And also a representation of highly complex processes in land surface models is a rough approximation given heterogeneous land surfaces in reality. Therefore, soil moisture estimates from land surface models are subject to errors due to the errors in the model formulation, parameters used and uncertainties in the meteorological

forcing data used to integrate the model.

1.4 Data assimilation

Independently, neither observations nor numerical models are perfect to represent the state of the system. In-situ observations are better in accuracy but sparse in spatial coverage. Whereas satellite observations are great in spatial coverage but limited to the top surface of the earth. On the other hand, numerical models give estimates without time and space resolution or coverage limitations provided that meteorological forcing data are available but could provide forecasts which are far from the observed reality. Hence, combining observations with numerical models can optimise the quality of estimates from numerical models. The technique called data assimilation is a mechanism which enables us to combine numerical model outputs with observed data and its uncertainty to obtain the most likely state of the system. Data assimilation is an important procedure for different purposes in the environmental forecasting - meteorology, oceanography and hydrology (Reichle, 2000). Especially in meteorology, weather forecasts are updated multiple times a day by incorporating observations from different sources. In many operational weather forecast centres, the forecast skill has improved as a result of data assimilation in addition to advancements in the modelling and computing resources (Bauer et al., 2015; Dee et al., 2011).

In hydrology data assimilation is used for several purposes: to improve the estimation of the root zone water content (Heathman et al., 2003), to improve the estimation of soil moisture and streamflow (Lievens et al., 2015), for flood forecasting (Silvestro and Rebora, 2014), to study the effect of assimilating the surface soil moisture on hydrological processes (Han et al., 2012) and to estimate parameters for numerical models (Pinnington et al., 2018), to mention a few.

Ensemble-based data assimilation methods use ensembles of model runs to represent uncertainties in the model state. In doing so, an appropriate spread among ensemble members is an important step. However, due to small sample size or unaccounted for model error, underestimation of ensemble spread, which is known as ensemble collapse, will occur. The ensemble spread being small is associated with high certainty, hence the posterior estimate from the data assimilation will rely on the prior model state and will disregard observations. This leads to a long-standing problem in ensemble data assimilation, which is called filter divergence (Wu et al., 2013). In such a case, the growth of ensemble spread over time is smaller than the growth of root

mean square error of the ensemble mean. Covariance inflation, additive, multiplicative or a combination of them, are techniques which have been implemented to mitigate the issue of ensemble collapse (Hamill and Whitaker, 2011; Luo and Hoteit, 2013; Wu et al., 2013). A special case where ensemble collapse is acceptable is when repeated wetting and drying events eventually draw all ensemble members to the same upper and lower bounds. In this case, observations and model estimates will match, and ensemble collapse will not be a problem. This implies that there is no need for the data assimilation as observations and model estimates agree within the observation error.

In this thesis, we deploy stochastic forcing based on uncertainties of the forcing data, parameters and also in the model itself to enhance the ensemble spread of the prior model state. By doing so, we aim to alleviate the filter divergence, and posterior soil moisture will benefit from observations. As such, ensemble spread of posterior soil moisture will be a better predictor of the accuracy of the ensemble mean of the posterior soil moisture. The following section discusses the details of the aim of this thesis.

1.5 Thesis Aims

The aim of the thesis is to address the fundamental science of how data assimilation systems should be developed for soil moisture assimilation, improving uncertainty representation for a prior distribution, with physically meaningful mechanisms. This will enable us to improve the skill of soil moisture forecast beyond the assimilation period. The JULES model, both in its full implementation and only the water balance part (DRBC model: abbreviated for Darcy Richard Brooks and Corey) is used to implement the data assimilation experiments with the following objectives:

1. **Apply stochastic forcing to generate ensemble spread for the ETKF**

   Ensemble forecasts of the DRBC model by initial condition perturbation have small spread. This leads to ensemble collapse, where the model forecast statistics is represented by a single ensemble member. Having small spread means the forecast has high certainty and observations will not have a substantial impact when assimilated. Here we are addressing the question:

   Can we use stochastic forcing to generate ensemble spread for soil moisture data assimi-

lation with ETKF?

2. **Determine whether or not non-Gaussian distributions can be used to initialise model ensembles for 4DEnVar**
   The four Dimensional Ensemble Variational (4DEnVar) uses ensemble perturbations from the full nonlinear model trajectories to approximate the tangent linear and adjoint models. The initialisation of those ensemble members which give rise to ensemble perturbations for the 4DEnVar is usually done by sampling the parameter or state of interest from a prior covariance matrix, assuming a Gaussian error. However, the parameters we consider here, soil texture parameters, do not have Gaussian distributed errors but share fundamental properties of other distributions, positivity and boundedness. Hence, we aim to answer:
   Can the Dirichlet distribution be used to initialise model ensembles for 4DEnVar?

3. **Investigate the improvement of soil moisture forecast skill as a result of posterior parameters**
   Assimilating soil moisture observations has been a way to improve estimates of soil moisture from land surface models. As soil texture parameters are related to soil moisture estimates, parameter estimation using observed soil moisture data is essential. Hence, the question we are asking is:
   Does parameter estimation improve soil moisture forecast skill beyond the assimilation window?

1.6 Thesis outline

The rest of the thesis is organised as follows.

- **Chapter 2** reviews different types of data assimilation methods and soil moisture data assimilation studies. Challenges of existing data assimilation methods and improvement strategies are discussed.
- **Chapter 3** presents general information about the sites considered in the experiments. Furthermore, the JULES land surface model and the hydrological part of it (called DRBC), meteorological forcing data and soil moisture data used for the data assimilation experiments are described.

- **Chapter 4** shows techniques of ensemble spread generation in the DRBC model for ETKF. Stochastic forcing is imposed into the model based on errors in the rainfall data and model error. Data assimilation results show that stochastic forcing helps the model to gain spread and improve soil moisture estimates.

- **Chapter 5** introduces new sampling techniques for non-Gaussian prior errors. Soil texture parameters are positive and bounded and are sampled from the Dirichlet distribution. Data assimilation results obtained from assimilating synthetic observations showed that it is possible to restore the true parameters even if the prior is wrongly specified.

- **Chapter 6** uses the sampling technique introduced in chapter 5 and presents the implementation of soil moisture data assimilation for parameter estimation. In-situ and satellite observations assimilated into the JULES land surface model shows that posterior soil moisture is more skilful compared to the prior estimate as a result of posterior parameters. Retrospective forecast of soil moisture estimates with posterior parameters shows a great improvement in accuracy compared to an open-loop model run which uses the prior parameters.

- **Chapter 7** summarises the findings of the thesis, presents the scope and limitations of the work and suggests further work opportunities.

Chapter 2

Review of data assimilation methods and soil moisture data assimilation

In this chapter, we introduce data assimilation and review the literature on soil moisture data assimilation. Section 2.1 introduces different data assimilation methods where subsection 2.1.1 and subsection 2.1.2 gives description of the methods used in this thesis. Section 2.2 discusses studies on soil moisture data assimilation which are relevant for the work in this thesis.

2.1 Introduction to data assimilation

Data assimilation (DA) is a mathematical technique which aims to find the most likely estimate of the true state or parameter, known as analysis or posterior, by optimally combining the prior estimate from a numerical model with available observations and associated uncertainties. Providing the prior estimate involves finding the best guess for the initial state of the model and its uncertainty. DA methods can generally be classified into variational and sequential. Sequential methods solve the system of equations needed to find an optimal solution explicitly whenever observations are incorporated. In conjunction with the optimal solution, the associated uncertainty is also obtained at each assimilation step. Variational methods solve the system of equations needed to find an optimal solution implicitly by minimising the cost function, i.e. the misfit between observations and model estimate, by considering all available observations in the assimilation period. Then the optimal solution serves as an initial condition to obtain an optimal trajectory of the model state. In both approaches, the uncertainties for the prior and observation

are assumed to be Gaussian.

For Numerical Weather Prediction (NWP), for example, models are chaotic (small perturbation of initial conditions can lead to a significant deviation in the model integration at later times), data assimilation mostly focuses on finding better initial conditions. Whereas land surface models are not chaotic, hence determining the error statistics in the forcing data and model parametrisation are more important for land surface data assimilation than improving initial conditions (Darvishi and Ahmadi, 2014). In this thesis, we use variants of both sequential and variational methods and subsection 2.1.1 and subsection 2.1.2 discuss those methods in details.

Consider a dynamical system which describes the true state of the system by:

$$\mathbf{x}_{t+1} = \mathcal{M}_{t+1,t}(\mathbf{x}_t) \tag{2.1}$$

were $\mathbf{x} \in \mathcal{R}^n$ is a state vector and $\mathcal{M}_{t+1,t}$ is a non-linear model which updates the state at time $t+1$ from time t. As the true state is unknown, the initial guess ($\mathbf{x}^b \in \mathcal{R}^n$) and observations ($\mathbf{y} \in \mathcal{R}^m$) are only approximations of the true state such that

$$\mathbf{x}^b = \mathbf{x} + \epsilon^b \tag{2.2}$$

$$\mathbf{y} = \mathcal{H}(\mathbf{x}) + \epsilon^o \tag{2.3}$$

where ϵ^b and ϵ^o are background and observation errors respectively which are assumed to be uncorrelated, unbiased and Gaussian with known covariances $\mathbf{B} = \mathbb{E}[\epsilon^b(\epsilon^b)^T], \mathbf{R} = \mathbb{E}[\epsilon^o(\epsilon^o)^T]$, $\mathcal{H} : \mathcal{R}^n \rightarrow \mathcal{R}^m$ is the observation operator which maps the state space into the observation space, can be linear or non-linear. The best estimate of $\mathbf{x}$, called analysis or posterior $\mathbf{x}^a$, is the most likely representation of the observations given the prior. In the following sections, we describe the process of finding $\mathbf{x}^a$ for sequential and variational methods.

For a probability density functions, pdf, the Bayes' theorem gives us a basis for finding the posterior such that:

$$p^a(\mathbf{x}|\mathbf{y}) = \frac{p^b(\mathbf{x})p^o(\mathbf{y}|\mathbf{x})}{p^o(\mathbf{y})} \tag{2.4}$$

where $p^a(\mathbf{x}|\mathbf{y})$ is the pdf for the posterior, $p^b(\mathbf{x})$ is the pdf for the prior, $p^o(\mathbf{y}|\mathbf{x})$ is pdf for the observations given the prior and $p^o(\mathbf{y})$ is pdf for the observations (normalising factor).

As we are aiming to maximise the probability of $p^a(\mathbf{x}|\mathbf{y})$, the normalising factor which is constant wrt $\mathbf{x}$ can be omitted, hence

$$p^a(\mathbf{x}|\mathbf{y}) \propto p^b(\mathbf{x})p^o(\mathbf{y}|\mathbf{x}). \tag{2.5}$$

Assuming Gaussian probability density functions for $p^b(\mathbf{x})$ and $p^o(\mathbf{y}|\mathbf{x})$, applying the Bayes' theorem and dropping constants results Equation 2.6:

$$p^a(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) - \frac{1}{2}(\mathbf{y}-\mathcal{H}(\mathbf{x}))^T\mathbf{R}^{-1}(\mathbf{y}-\mathcal{H}(\mathbf{x}))\right) \tag{2.6}$$

While the objective is to maximize $p^a(\mathbf{x}|\mathbf{y})$ in Equation 2.6, sequential and variational data assimilation methods differ in their approach. The following two sections describe how each method solves Equation 2.6.

2.1.1 Sequential methods

The basis for sequential data assimilation is the Kalman filter (KF) which was introduced by Kalman (Kalman, 1960). The forecast and analysis steps are obtained sequentially together with the error statistics. The method assumes that the model representing the dynamical system and observation operator are linear and error statistics follow the Gaussian distribution.

The Extended Kalman Filter (EKF) is an extension of the KF when the system is non-linear. It can give a good approximation of the posterior for a weakly non-linear observation system; however, as the nonlinearity increases, the accuracy of the approximation declines (Hunt et al., 2007). Incorporation of ensemble techniques in sequential data assimilation methods become necessary to approximate the covariance evolution with a few ensemble members than the number of model runs required by EKF; hence the Ensemble Kalman Filter arises.

The Ensemble Kalman Filter (EnKF) is a sequential data assimilation method which uses a

Monte-Carlo approach to represent the error covariance matrix and use Bayes' theorem to determine the posterior pdf with a Gaussian assumption (Evensen, 1994; Margulis et al., 2002). Compared to EKF, it is easier to calculate the forecast error covariance matrix. The downside of EnKF and its variants is ensemble collapse where all ensemble members converge to a single estimate, i.e. ensemble spread becomes very small, and it does not have proper representation of the model uncertainty. In such a case, the posterior estimates will be nearly identical to the prior as observations will be disregarded due to the system being overconfident on the model.

The Ensemble Transform Kalman Filter (ETKF) is one of the variants of the EnKF which belongs to the family of Square Root Filters (SRFs) methods. Bishop et al. (2001) first introduced it. Here we give a mathematical description of the ETKF by using the notations from Bishop et al. (2001) and illustrations from Fairbairn (2009) and Hunt et al. (2007). The equation for the analysis is similar to the original Kalman filter when the perturbation matrix replaces the background error.

Consider N_e ensemble members of n dimensional model state vector and non-linear model operator $\mathcal{M}$ introduced in Equation 2.1 with the notation superscript a representing the analysis (posterior) and superscript b representing background (prior) state,

$$\{\mathbf{x}_t^{a,i} : i = 1, 2, ..., N_e\}. \tag{2.7}$$

The following equation describes the evolution of the analysis state to give a new background with a non-linear model $\mathcal{M}$:

$$\mathbf{x}_{t+1}^{b,i} = \mathcal{M}_{t+1,t}\left(\mathbf{x}_t^{a,i}\right) + \epsilon_t \tag{2.8}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{Q})$ is a stochastic model error drawn from a Gaussian distribution with error covariance matrix $\mathbf{Q}$.

The background ensemble perturbations matrix, scaled by number of ensemble members N_e is given by

$$\mathbf{X}^b (\in \mathbb{R}^{n \times N_e}) = \frac{1}{\sqrt{N_e - 1}} (\mathbf{x}^{b,1} - \overline{\mathbf{x}}^b, \mathbf{x}^{b,2} - \overline{\mathbf{x}}^b, ..., \mathbf{x}^{b,N_e} - \overline{\mathbf{x}}^b) \tag{2.9}$$

where $\bar{\mathbf{x}}^b$ is ensemble mean

$$\bar{\mathbf{x}}^b = \frac{1}{Ne}\sum_{i=1}^{N_e} \mathbf{x}^{b,i}.$$

Here the time subscript is omitted. The background error covariance matrix becomes

$$\mathbf{P}^b = \mathbf{X^b}(\mathbf{X^b})^T. \tag{2.10}$$

The state ensemble members can be mapped to the observation space by

$$\mathbf{y}^{b,i} = \mathbf{H}\mathbf{x}^{b,i} \tag{2.11}$$

The forecast ensemble perturbations become

$$\mathbf{Y}^b = \mathbf{H}\mathbf{X}^b \tag{2.12}$$

Then substituting equation 2.10 into the Kalman gain (of the Kalman Filter) equation becomes

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1} \tag{2.13}$$

$$= \mathbf{X}^b(\mathbf{X}^b)^T\mathbf{H}^T(\mathbf{H}\mathbf{X}^b(\mathbf{X}^b)^T\mathbf{H}^T + \mathbf{R})^{-1} \tag{2.14}$$

$$= \mathbf{X}^b(\mathbf{Y}^b)^T\mathbf{S}^{-1}, \mathbf{S} = \mathbf{H}\mathbf{X}^b(\mathbf{X}^b)^T\mathbf{H}^T + \mathbf{R} \tag{2.15}$$

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^b + \mathbf{K}(\mathbf{y} - \bar{\mathbf{y}}^b). \tag{2.16}$$

The analysis error covariance matrix is given by

$$\mathbf{P}^a = (N_e - 1)^{-1}\sum_{i=1}^{N_e}(\mathbf{x}^{a,i} - \bar{\mathbf{x}}^a)(\mathbf{x}^{a,i} - \bar{\mathbf{x}}^a)^T \tag{2.17}$$

$$= (\mathbf{X^a})(\mathbf{X^a})^T \tag{2.18}$$

$$= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b \tag{2.19}$$

$$= (\mathbf{I} - \mathbf{X}^b(\mathbf{Y}^b)^T\mathbf{S}^{-1}\mathbf{H})\mathbf{X}^b(\mathbf{X}^b)^T \tag{2.20}$$

$$= \mathbf{X}^b(\mathbf{I} - (\mathbf{Y}^b)^T\mathbf{S}^{-1}\mathbf{Y}^b)\mathbf{X}^{bT}. \tag{2.21}$$

The analysis ensemble perturbation matrix $\mathbf{X}^a$ is updated according to

$$\mathbf{X}^a = \mathbf{X}^b\mathbf{T} \tag{2.22}$$

such that

$$\mathbf{T}\mathbf{T}^T = \mathbf{I} - (\mathbf{Y}^b)^T\mathbf{S}^{-1}\mathbf{Y}^b. \tag{2.23}$$

Then the analysis step which maximises Equation 2.6 is updated by

$$\mathbf{x}^a = \bar{\mathbf{x}}^a + \mathbf{X}^a \tag{2.24}$$

The ETKF is different from other Square Root Filters in finding the $\mathbf{T}$ matrix using the identity

$$\mathbf{I} - (\mathbf{Y}^b)^T\mathbf{S}^{-1}\mathbf{Y}^b = (\mathbf{I} + (\mathbf{Y}^b)^T\mathbf{R}^{-1}\mathbf{Y}^b)^{-1}. \tag{2.25}$$

Using the above method is preferable because inverting the matrix $\mathbf{R}$ is much simpler than inverting $\mathbf{S}$. Then using the eigenvalue decomposition we get

$$(\mathbf{Y}^b)^T\mathbf{R}^{-1}\mathbf{Y}^b = \mathbf{U}\Lambda\mathbf{U}^T \tag{2.26}$$

$$\Rightarrow \mathbf{T}\mathbf{T}^T = \mathbf{U}(\mathbf{I} + \Lambda)^{-1}\mathbf{U}^T \tag{2.27}$$

$$\Rightarrow \mathbf{T} = \mathbf{U}(\mathbf{I} + \Lambda)^{-1/2} \tag{2.28}$$

where $\mathbf{U}$ is orthogonal and Λ is diagonal matrices. Once we get the analysis ensemble mean and the analysis ensemble perturbation matrix, the analysis update evolves according to equation 2.24. The main advantage of ETKF is the easy in calculating $(\mathbf{I} + \Lambda)^{-1/2}$ since both matrices are diagonal.

When the dynamical model is not chaotic (as in the case of DRBC model), ETKF suffers from ensemble collapse. From equations 2.13 and 2.16, ensemble collapse refers to minimal value for $\mathbf{P}^b$, which makes the analysis mean mainly dependant on the background mean. As a result, observations will have a very small influence and will not impact the analysis mean.

2.1.2 Variational methods

Variational data assimilation methods are widely used in many meteorological services and have been successful in forecasting the weather for a long time. Specifically, the four-dimensional variational (4DVar) method is among the most commonly used (Lawless, 2013). However, one of the caveats of using the method is that calculating the tangent linear and adjoint models is costly for complex models and large dimensional problems. Since its first implementation in 1997 at the European Centre for Medium-Range Weather Forecasts (ECMWF), the four-dimensional variational (4DVar) data assimilation method becomes a preferred data assimilation method at most leading operational Numerical Weather Prediction (NWP) centres (Köpken et al., 2004; Milan et al., 2019). Its capability to assimilate observations at different times and being highly constrained to the model dynamics makes it preferable among other data assimilation methods.

The objective in using the 4DVar is to minimise the misfit between the background and model predicted state of the modelled dynamics and between observations and model predicted observations. Mathematically, the error or misfit distance is represented as:

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\sum_{t=0}^{N}\left(\mathbf{y}_t - \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^b))\right)^T \mathbf{R}^{-1}\left(\mathbf{y}_t - \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^b))\right), \tag{2.29}$$

$$\nabla_{\mathbf{x}} J = \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \sum_{t=0}^{N} \mathbf{M}_{t,0}^T \mathbf{H}_t^T \mathbf{R}^{-1}\left(\mathbf{y}_t - \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^b))\right), \tag{2.30}$$

where $\mathbf{B}$ is the background error covariance matrix, $\mathbf{M}_{t,0} = \mathbf{M}_{t-1}\mathbf{M}_{t-2},...,\mathbf{M}_0$ is tangent linear model such that $\mathbf{M}_t = \frac{\partial \mathcal{M}_{t,t-1}(\mathbf{x}_t)}{\partial \mathbf{x}_t}$, $\mathbf{M}_{t,0}^T$ is the transpose of the tangent linear model or adjoint model. $\mathbf{H}_t = \frac{\partial \mathcal{H}_t(\mathbf{x}_t)}{\partial \mathbf{x}_t}$ is the tangent linear of the nonlinear observation operator. Here, the background error covariance matrix $\mathbf{B}$ is constant with time even though it evolves with the model dynamics implicitly. Also, for some application like in meteorology, for example, the background error covariance matrix $\mathbf{B}$ could become large or ill-conditioned and difficult to find the inverse as a result. Then minimisation of Equation 2.29 could be slow and may not converge efficiently. To avoid the computation of $\mathbf{B}$ and it's inverse explicitly and to make sure Equation 2.29 converges efficiently, preconditioning has been used (Bannister, 2017; Tian et al., 2008)

Define $\mathbf{U}$, a preconditioning matrix as:

$$\mathbf{B} = \mathbf{U}\mathbf{U}^T \tag{2.31}$$

and

$$\mathbf{x} = \mathbf{x}^b + \mathbf{U}\mathbf{w} \tag{2.32}$$

Substituting Equations 2.31 and 2.32, Equation 2.29 becomes

$$\mathbf{J}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}\sum_{t=0}^{N}(\mathbf{H}_t\mathbf{M}_{t,0}\mathbf{U}\mathbf{w} + \mathbf{d}_t)^T\mathbf{R}^{-1}(\mathbf{H}_t\mathbf{M}_t\mathbf{U}\mathbf{w} + \mathbf{d}_t) \tag{2.33}$$

with tangent linear approximation

$$\mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}_b + \mathbf{U}\mathbf{w})) \approx \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}_b)) + \mathbf{H}_t\mathbf{M}_{t,0}\mathbf{U}\mathbf{w}, \tag{2.34}$$

and

$$\mathbf{d}_t = \mathbf{y}_t - \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}_b)). \tag{2.35}$$

Now the explicit computation of the $\mathbf{B}$ is avoided so long as $\mathbf{U}$, and $\mathbf{w}$ are known, but still the tangent linear and adjoint models are needed for Equation 2.33. To ease those difficulties in using the 4DVar but still keeping its non-sequential nature, there have been lots of efforts from researchers (Tian et al., 2008; Liu et al., 2008; Bannister, 2017). The two main improved variants of 4DVar are hybrid four-dimensional variational (hybrid-4DVar) and four Dimensional Ensemble Variational (4DEnVar). In hybrid-4DVar, the background error covariance matrix is obtained by blending the climatology error covariance matrix and the forecast error covariance matrix from ensemble perturbations, tangent linear and adjoint models still required. Whereas in 4DEnVar, the tangent linear and the adjoint models are not needed, and forecast error covariance matrix is used from ensemble perturbations (Liu et al., 2008; Bannister, 2017).

The 4DEnVar uses ensembles of the full non-linear model trajectories to calculate the tangent linear and adjoint models. As a result, approximations of the tangent linear and adjoint

models become obsolete in the 4DEnVar. Here we give formulation of the 4DEnVar using the notations from Liu et al. (2008).

Using a background perturbation $\mathbf{X^b}$ as described in Equation 2.9, the background error can be approximated as

$$\mathbf{B} \approx \mathbf{X^b}\mathbf{X^b}^T. \tag{2.36}$$

The analysis is described by

$$\mathbf{x} = \mathbf{x}^b + \mathbf{X^b}\mathbf{w}, \tag{2.37}$$

$$\tag{2.38}$$

where the vector $\mathbf{w}(\in \mathbb{R}^{n\times 1})$ provides the weights of each ensemble perturbations. Then the cost function of the 4DEnVar becomes:

$$\mathbf{J}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}\sum_{t=0}^{N}\left(\mathbf{H}_t\mathbf{M}_{t,0}\mathbf{X}^b\mathbf{w} + \mathbf{d}_t\right)^T\mathbf{R}^{-1}\left(\mathbf{H}_t\mathbf{M}_t\mathbf{X}^b\mathbf{w} + \mathbf{d}_t\right), \tag{2.39}$$

and the gradient of the cost function is

$$\nabla_{\mathbf{w}}\mathbf{J} = \mathbf{w} + \sum_{t=0}^{N}(\mathbf{HMX}^b)^T\mathbf{R}^{-1}(\mathbf{HMX}^b\mathbf{w} + \mathbf{d}_t). \tag{2.40}$$

From the idea of the EnKF, the perturbation in the observation space becomes:

$$\mathbf{HMX}^b \approx \frac{1}{\sqrt{N_e - 1}}(\mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^{b,1})) - \mathcal{H}_t(\mathcal{M}_{t,0}(\overline{\mathbf{x}^b})), \mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^{b,2})) - \mathcal{H}(\mathcal{M}_{t,0}(\overline{\mathbf{x}^b})), \ldots,$$
$$\mathcal{H}_t(\mathcal{M}_{t,0}(\mathbf{x}^{b,N})) - \mathcal{H}_t(\mathcal{M}_{t,0}(\overline{\mathbf{x}^b}))). \tag{2.41}$$

As a result of the approximation in equation 2.41, the 4DEnVar avoids approximations of the tangent linear model ($\mathbf{M}$), adjoint model ($\mathbf{M}^T$) as well as tangent linear approximation of the observation operator $\mathbf{H}$ which are needed in equations 2.39 and 2.40.

Liu et al. (2008) tested their newly formulated data assimilation method, 4DEnVar for one-dimensional shallow water model and compared experimental results with 4DVar and EnKF. 4DEn-

Var was able to give analysis which is as good as the analysis from 4DVar and EnKF but with less computational cost. Liu et al. (2009) implemented with the Advanced Research Weather Research and Forecasting (ARW-WRF) model. Experimental results of 4DEnVar were compared with three-dimensional Ensemble Variational (3DEnVar), and they reported that 4DEnVar has overall better analysis. Buehner et al. (2010) Compared the performance of 4DEnVar with EnKF and 4DVar (using static and flow-dependent **B** matrices) using the Environment Canada operational model. The meteorological data they assimilated are from different sources, radiosondes, aircraft and satellites. They have shown that 4DVar with evolving **B** matrix (4DVar-Ben) outperformed all other data assimilation methods, whereas 4DEnVar has similar performance as EnKF but better than the 4DVar with static **B** matrix.

Similarly, Fairbairn et al. (2014) compared ensembles of 4DEnVar with 4DVar and EnKF for the best analysis. A toy model (Lorenz 2005, which represents waves propagation of unspecified atmospheric quantity) was used for the data assimilation experiments. They demonstrated that 4DVar with perfect climatology **B** matrix performed less than the other two methods with flow-dependent **B** matrix. When a similar flow-dependent background error covariance matrix is used for all the three methods, similar analysis error was obtained. It was also shown that for large ensemble size, ensemble-based methods perform significantly better than the 4DVar as they use flow-dependent background-error covariance matrix. Whenever considering large ensemble size is possible, ensemble-based methods are preferable as they avoid calculating the tangent linear and adjoint models approximations with a reasonable high rank and flow-dependent representation of the background error covariance matrix.

So far, the implementation of the 4DEnVar uses a background error covariance matrix from existing data assimilation systems and with a Gaussian assumption. The problem we are dealing with here is, soil texture parameters (discussed in chapter 5) are positive and have bounded sum, errors do not follow a Gaussian distribution. However, there are cases where the problem under consideration is non-Gaussian, for example, in the case of soil texture parameters. In this thesis, we demonstrate the use of non-Gaussian distributions for the 4DEnVar. Chapter 5 presents the statement of the problem and experimental results.

2.2 Soil moisture data assimilation

In this section, we review soil moisture data assimilation studies and set a frame of reference for the work done in this thesis. Sequential and ensemble-based methods have been a popular technique for soil moisture data assimilation. On the other hand, variational data assimilation methods are also used. In both cases, soil moisture data from satellite observations, synthetic observations and in-situ measurements are assimilated. Parameters in land surface processes are crucial for modelling the system and yet less known, compared to atmospheric model parameters (Reichle, 2000). For soil moisture estimation, the role of parameters is paramount, see section 5.5. Here we present a review of studies on soil moisture data assimilation both for parameter and state estimation, in some cases, joint parameter and state.

Several researchers implemented the EnKF and its variants for soil moisture data assimilation. Reichle et al. (2002) assimilated synthetic near-surface soil moisture data with the EnKF into a hydrological model at a catchment level as opposed to the grid level. They compared the performance of the EnKF with the EKF and found that estimation error for volumetric soil moisture has shown a slight improvement for the EnKF by increasing ensemble size. Margulis et al. (2002) used the EnKF to assimilate radio brightness observations from the Southern Great Plains 1997 (SGP97) field experiment by using a radiative transfer model. They showed that the analyses are better than the model estimate, which does not incorporate brightness measurements and with good agreement when it is compared to ground measurement soil moisture data.

Similarly, Zheng et al. (2018), Han et al. (2012) and Liu et al. (2016) used the EnKF for soil moisture data assimilation. Zheng et al. (2018) assimilated soil moisture observations from NOAA-NESDIS Soil Moisture Operational Product System (SMOPS) into the NOAA-NCEP Global Forecast System (GFS). Han et al. (2012) assimilated synthetic surface soil moisture observations into the Soil and Water Assessment Tool (SWAT) hydrological model. They showed that surface soil moisture data assimilation had improved the soil moisture updates. Rainfall was artificially differed to investigate how surface soil moisture data assimilation compensates for the errors in rainfall and effects on root zone soil moisture. However, verifying with real data is needed to fully rely on their results. Liu et al. (2016) have shown that assimilating Active and Passive microwave observations (radar backscattering and brightness temperature) to estimate soil moisture at two layers and related dry biomass and Leaf Area Index LAI. The decision support system for agrotechnology transfer (DSSAT) crop growth model was used. The radar backscattering is sensitive

to surface roughness and used to update vegetation biomass, whereas brightness temperature is sensitive to soil moisture. In their study, they have shown that active and passive microwave observations can help to improve soil moisture and monitor crop growth estimates compared to the model only estimates.

Although EnKF and its variants are widely used for soil moisture DA method, there is a long-standing associated downside of using the method. When small ensemble size is used or due to model error, ensemble spread will be underestimated, and ensemble collapse or filter degeneracy occurs (Wu and Zheng, 2018). Then it will lead to a scenario that the data assimilation result is reliant on the dynamical model and ignores the observations. To overcome the filter degeneracy, covariance inflation, additive, multiplicative or both, could be used (Wu and Zheng, 2018). Han et al. (2014) implemented multiplicative covariance inflation when they assimilated satellite-based soil moisture observations for drought monitoring mechanism using EnKF. Wu et al. (2016) deployed multiplicative covariance inflation. They assimilated top layer soil moisture observations into a land surface model using ETKF. Wu and Zheng (2018) has more literature on the different types of inflation.

Reichle et al. (2001) assimilated radio brightness into a hydrological model which considers the vertical flow of water using 4DVar to update soil moisture estimates at a finer scale by downscaling the radio brightness using a radiative transfer model as an observation operator. In their work, precipitation is withheld from the hydrological model and instead, model error is accounted whenever precipitation is observed in the data assimilation system where the model error is estimated from the brightness temperature. When a model error is used, area-averaged temporal mean Root Mean Square Error (RMSE) of $0.034 m^3 m^{-3}$ is obtained for the analysis soil moisture compared to the reference experiment (with observed rainfall) with RMS error of $0.014 m^3 m^{-3}$ where RMS error of $0.19 m^3 m^{-3}$ is for a prior estimate when rainfall is withheld. In hydrological models, land surface fluxes enter into the process additively, and errors in the fluxes can be represented as a model error. They have shown that accounting for a model error can compensate for the absence of quantitative rainfall forcing data.

Heathman et al. (2003) estimated soil hydraulic parameters using a direct insertion data assimilation method to a hydrology model. The study considered four sites in south-central Oklahoma. A gravimetric surface soil water content which was measured from June 18 - July 16, 1997, for an experiment (Southern Great Plains 1997 SGP97) was used as a surrogate for re-

motely sensed surface moisture data. Based on the estimated parameters, they showed that it is possible to improve the soil water estimates of the root zone up-to 30cm deep by assimilating the surface soil moisture. The work by Pinnington et al. (2018) also showed that assimilating soil moisture data has improved estimate soil texture parameters in the JULES model. They further updated soil moisture based on improved parameters; as a result, RMSE reduction for the posterior soil moisture estimates compared to the open-loop model runs was observed.

Pinnington et al. (2018) also showed that improvement in rainfall data uncertainty improves soil moisture estimation, both with and without data assimilation. The maximum RMSE reduction was obtained when data assimilation is used while the JULES model was forced with TAMSAT more accurate rainfall data. This result shows that errors in the forcing data, rainfall, need to be taken into account. Dunne and Entekhabi (2005) also showed the importance of correct characterisation of errors in rainfall. In their work, a twin experiment, using EnKF and Ensemble smoother, with rainfall ensembles from a different site did not recover the truth soil moisture estimate with rainfall forcing from the site. Reichle et al. (2001); Ettema and Viterbo (2001); Zheng and Eltahir (1998) have shown that precipitation is the main component of the forcing data in hydrological models and yet prone to errors. Due to the direct relation between rainfall and soil moisture, improving rainfall estimation will certainly improve soil moisture estimation.

From the above property of land surface models, it is difficult to maintain the spread of the ensemble when ETKF is implemented, which is an inherent problem for ensemble-based data assimilation methods. That means we encounter ensemble collapse or filter degeneracy (Morzfeld et al., 2016; Wu and Zheng, 2018), and observations will not impact any more for the data assimilation. In such a case, there are ad hoc methods as a rescue mechanism. Depending on the particular situation, covariance inflation (additive or multiplicative), perturbing model parameters, perturbing forcing data of the model or model error can be considered to mitigate ensemble collapse. Here, we used stochastic forcing via generated rainfall (instead of the observed rainfall) and by allowing for a model to be imperfect. The full description of stochastic rainfall generation is given in subsection 4.3.1. The characterisation of the model error is addressed by including model error covariance matrix to inflate the background error covariance matrix $\mathbf{P}^b$. The magnitude and shape of the model error covariance matrix are given in section 4.4.

2.3 Summary

Soil moisture data assimilation has been explored and advancements been shown on addressing key questions: how soil moisture data is used to improve state and parameter estimation? However there are still areas where improvement is needed: considering errors of the forcing data, how to limit the amount of stochastic forcing to respect the model dynamics but having appropriate spread and how parameters with a non-Gaussian error are sampled? The following challenges have been identified:

- As any other numerical model, land surface models are not perfect. Besides, the main meteorological forcing data, rainfall, is prone to errors. However, for the sake of simplicity, both models and rainfall observations are considered to be perfect in most cases. Accounting for model and forcing data uncertainty needs to be addressed in order the modelled state represents what is observed in reality.

- In order to tackle the inherent problem of ensemble data assimilation methods; ensemble collapse or ensemble degeneracy, different covariance inflation techniques have been implemented. While it solves the problem, there is not much physical meaning for inflating the covariance apart from making the model less certain. More meaningful approaches for inflating the covariance would be intuitive to include in the data assimilation scheme.

- In chapter 4, errors in rainfall observations and in the numerical model are considered. Reichle et al. (2001) has done similar work for satellite soil moisture data assimilation by considering errors in rainfall where rainfall is estimated from brightness temperature. The advantage here is that rainfall is a forcing data for all hydrological models and representing errors by generating stochastic rainfall generation is easily accessible.

- Gaussian distributed errors are assumed in data assimilation. However, there are cases where errors do not follow a Gaussian distribution. For example, soil texture parameters are positive, and they sum up to a hundred.

In this thesis, both ETKF and 4DEnVar methods are used for the data assimilation experiments. Stochastic forcing techniques and parameter sampling techniques are used for the JULES model; however, the methods can be implemented for any land surface model.

Chapter 3

Models and Data

In this chapter, we describe sites considered in the experiments; we present the JULES land surface model used in chapters 5 and 6 of this thesis. We also describe the DRBC model, which is a part of the JULES hydrology, used in chapter 4 of this thesis. Then we describe different soil moisture data products used for the data assimilation experiments.

3.1 Models

3.1.1 The JULES model

The Joint UK Land Environment Simulator (JULES) model is a community land surface model which simulates land surface processes such as surface energy balance, hydrological cycle, carbon cycle and vegetation dynamics. The JULES model solves surface energy balance, with the general formulation given by:

$$\frac{dT}{dt} = R_n - G - H - \tau E, \tag{3.1}$$

where T is surface temperature, R is net radiation, G is ground heat flux, H is sensible heat flux, τ is the specific latent heat of evaporation and E is the rate of evaporation.

In doing so, the hydrology part of it includes a vertical flow of water in a soil column which uses 1-Dimensional Darcy's law and moisture content of each soil layer by Richards equation. Hydraulic characteristics can be represented by two alternative equations, Van Genuchten's equations, scientifically robust but more complex, and Brooks and Corey equations. Both alterna-

tives are implemented in the JULES model, where the user chooses between the two. In subsection 3.1.2 we describe part of the hydrology in the JULES model which uses Darcy's law, Richards equation and Brooks and Corey for hydraulic characteristics, which we call it DRBC.

The JULES model is developed from the UK Met Office Surface Exchange Scheme (MOSES). It can be used as a standalone model or coupled with the Unified Model, (Best et al., 2011). Figure 3.1 shows a schematic representation of the JULES model for a single grid cell. It shows the nine surface types. The five plant functional types are broad-leaf trees, needle-leaf trees, temperate (C3) grasses, tropical (C4) grasses, shrubs and the four non-vegetated types are urban, inland water, bare-soil and land-ice. Based on the specifications of a particular location, a proportion of the nine surface types are assigned. The JULES model can be configured to run a single point or a gird scale.

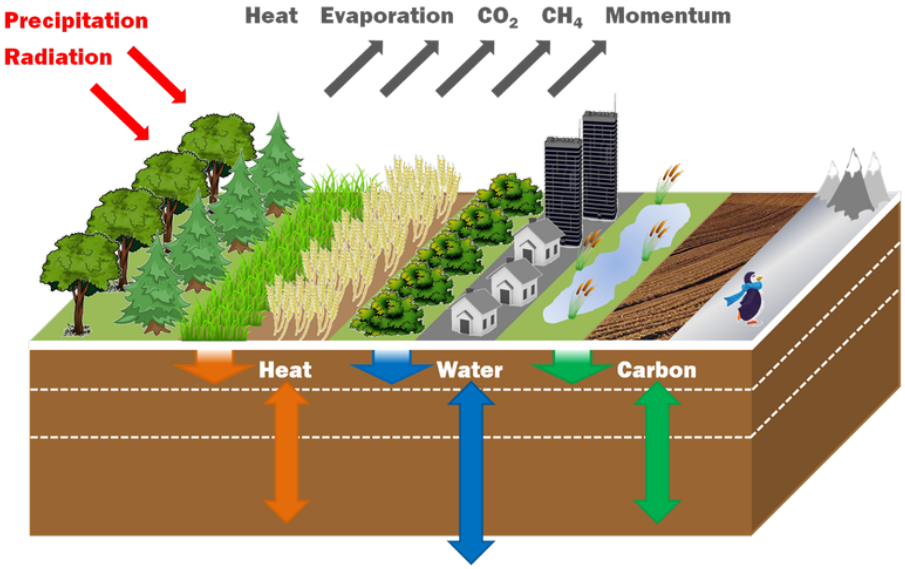


Figure 3.1: Illustration of the main features of JULES, `https://jules.jchmr.org/content/about`.

In this thesis, the JULES model is used to estimate soil moisture in its full implementation as well as by considering only part of the water balance, the DRBC model, described in the next subsection.

3.1.2 The DRBC model

To enable us to do the science by concentrating on the important part of the model for soil moisture, we separated the water balance part of the JULES model and study its dynamical properties

when soil moisture is assimilated. Soil moisture and evaporation for the top layer from JULES and DRBC were compared, and a good agreement was observed (was observed).

A column of soil discretization showing the vertical flow of water and moisture content in each layer, as explained by Essery et al. (2009) and Best et al. (2011), is depicted in Figure 3.2

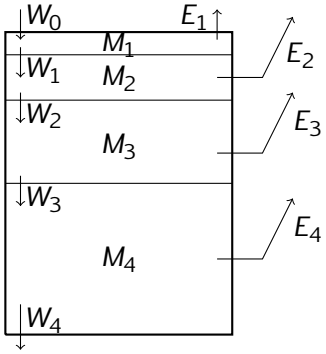


Figure 3.2: Soil layers with water flux and available water in each layer.

where M_k $(\mathrm{kgm^{-2}})$ is mass of soil moisture per unit area in layer k, W_k $(\mathrm{kgm^{-2}s^{-1}})$ is the water flux from soil layer k into the layer $k+1$ for natural number $k \leq 4$, W_0 $(\mathrm{kgm^{-2}s^{-1}})$ is precipitation and layer W_4 is drainage and E_k $(\mathrm{kgm^{-2}s^{-1}})$ is moisture flux due to evapo-transpiration from each layer (Essery et al., 2009). Here we only considered bare-soil evaporation.

Mathematical expressions for Darcy's law and Richards equation coupled with the Brooks and Corey relationship are given as:

$$\frac{dM_k}{dt} = W_{k-1} - W_k - E_k, \tag{3.2}$$

$$W_k = \rho_w k_h \left(\frac{\partial \psi}{\partial z} + 1 \right), \tag{3.3}$$

$$\psi = \psi_s \left(\frac{\theta}{\theta_s} \right)^{-b}, \tag{3.4}$$

$$k_h = k_{hs} \left(\frac{\theta}{\theta_s} \right)^{2b+3}, \tag{3.5}$$

$$M_k = \rho_w \Delta z \theta_k, \tag{3.6}$$

where ψ (m) is water suction, Δz is layer depth, ρ_w is density of water, k_h $(\mathrm{ms^{-1}})$ is hydraulic conductivity and θ $(\mathrm{m^3m^{-3}})$ is volumetric soil moisture.

From the above equations, Equation 3.2 is the Richards equation which describes the change of mass of water per unit time in each layer. The mass of water per unit area for each soil layer

is a result of water coming into the layer from rainfall and water, leaving the layer due to evapotranspiration, infiltration in a given time. Equation 3.3 is Darcy's law which describes the vertical flow of water due to gravity and Equation 3.4 and Equation 3.5 Brooks and Corey relationships relating soil water content, water suction and hydraulic conductivity. The rate of evaporation from the top layer is given in Appendix A; details are given in Essery et al. (2009) for each soil layer. Parameters ψ_s, k_{hs}, θ_s and b are empirical constants which can be calculated from the soil texture, sand, silt and clay proportions of the soil using set of equations known by the name pedo-transfer functions. For the DRBC model, the meteorological forcing variables are rainfall, temperature, pressure, wind speed, and humidity and the JULES model takes additional forcing variables longwave radiation, short wave radiation and snowfall.

3.2 Data

3.2.1 Site information

Sites considered in this thesis are from the observing networks of mesoscale weather events (Mesonet), distributed over Oklahoma, south-central of the United States. At the moment there are 120 stations, operational from 1991 to date and provide meteorological data with every five minutes and soil moisture every 30 minutes. One of the reasons for choosing Oklahoma Mesonet sites is the completeness of the data. For the year 2016, out of the 17520 data points of 30 min time interval, only 14 data points are missed for one of the stations (Antlers station), for example. The forcing data to be used in the JULES and DRBC models need processing so that the units are in agreement with those expected by the models and the rainfall data needs to be in rates than cumulative rainfall.

Oklahoma's climate ranges from humid subtropical, hot-humid summer and mild to cold winters in the east to semiarid with extreme temperatures in the west. The annual average temperature varies from 52 F to 19 F, East to West. The annual average rainfall also ranges from 63 to 55 inches east to west. Figures 3.3, 3.4 and 3.5 show annual rainfall, temperature and elevation respectively.

The land cover is mostly tall-grass in the east, and in some cases, woody vegetation is present (Fuhlendorf and Engle, 2004). The topography varies from prairie, plains, and hills where forests dominate the north-east. Figure 3.5 shows a topographic map with Mesonet sites where

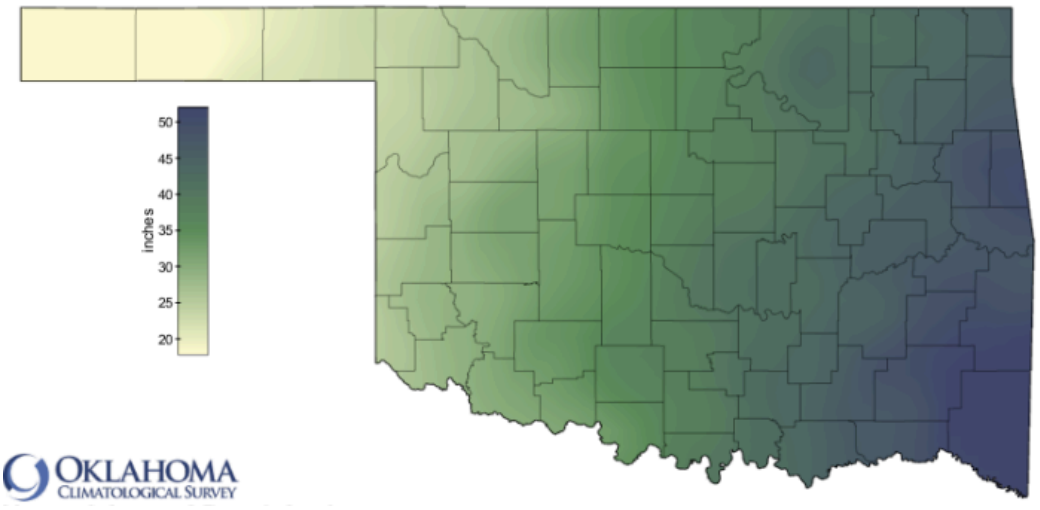


Figure 3.3: Oklahoma annual mean precipitation for a 30-year period (1981 - 2010) using quality assured observations from the National Weather Service cooperative observer network, taken from http://climate.ok.gov/index.php/climate/map/normal_annual_precipitation/oklahoma_climate, accessed on 12th Aug 2019.

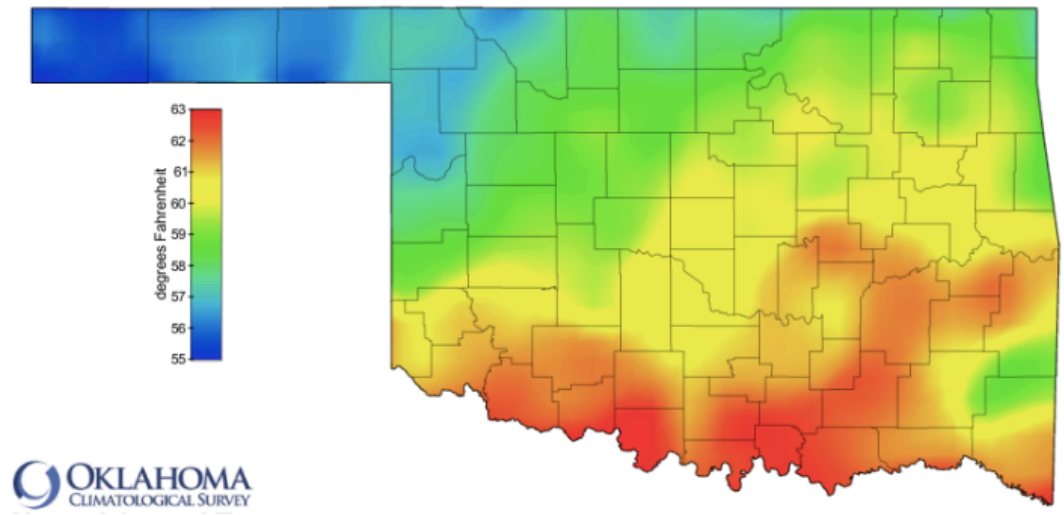


Figure 3.4: Oklahoma annual mean temperature for a 30-year period (1981 - 2010) using quality assured observations from the National Weather Service cooperative observer network, taken from http://climate.ok.gov/index.php/climate/map/mean_annual_temperature2/oklahoma_climate, accessed on 12th Aug 2019.

Table 3.1 shows the specification of just the subset considered in the assimilation experiments. To closely look at the land cover for each station, we show an aerial map and station photos of the two sites, one from the less vegetated in the north-west (Boise city) and the other, from densely vegetated in the south-east (Mt Herman). The other stations are in between these two extremes. The reason for considering sites with different properties is that that various sites will hold different soil moisture dynamics and we will be able to test the data assimilation in different scenarios.

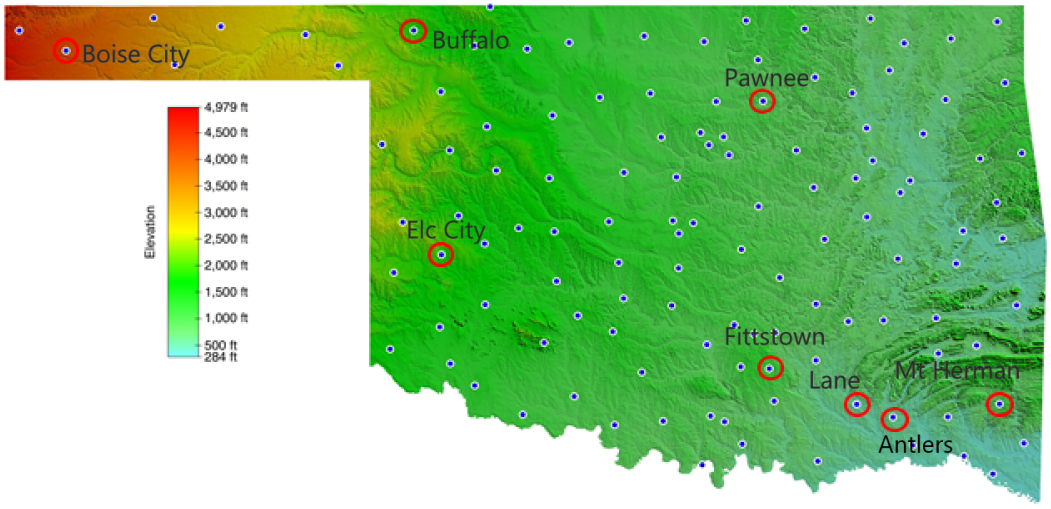


Figure 3.5: Oklahoma topographic map with Mesonet sites, taken from `https://www.mesonet.org/index.php/site/about/oklahoma_topographic_map`, accessed on 30th July 2019. The blue dots represent all sites and blue dots circled in red are sites we have used for the data assimilation experiments.

Station	Soil class	Latitude	Longitude	Elevation (m)
Antlers	sandy loam	34.25^0 N	95.67^0 W	171.9 m
Boise City	clay	36.69^0 N	102.5^0 W	1267 m
Buffalo	loam	36.8^0N	99.64^0 W	559 m
Elk City	loam	35.33^0N	99.4^0 W	584 m
Lane	silty loam	34.31^0N	96.0^0 W	181 m
Mt. Herman	loam	34.31^0N	94.82^0 W	284 m
Pawnee	loam	36.36^0N	96.77^0 W	282.9 m

Table 3.1: Oklahoma mesonet sites used for the experiments

Figure 3.6 and Figure 3.7 show that Boise city is covered by only grass and no woody trees around the site. Whereas Figure 3.8 and Figure 3.9 show that the land cover of Mt Herman station is only grasses but surrounded by big trees. As one of the criteria, all Mesonet stations do not have big trees inside the stations. The impact of having trees in the surrounding is that, satellite soil moisture observation is sensitive to vegetation cover and affects the quality of observations.

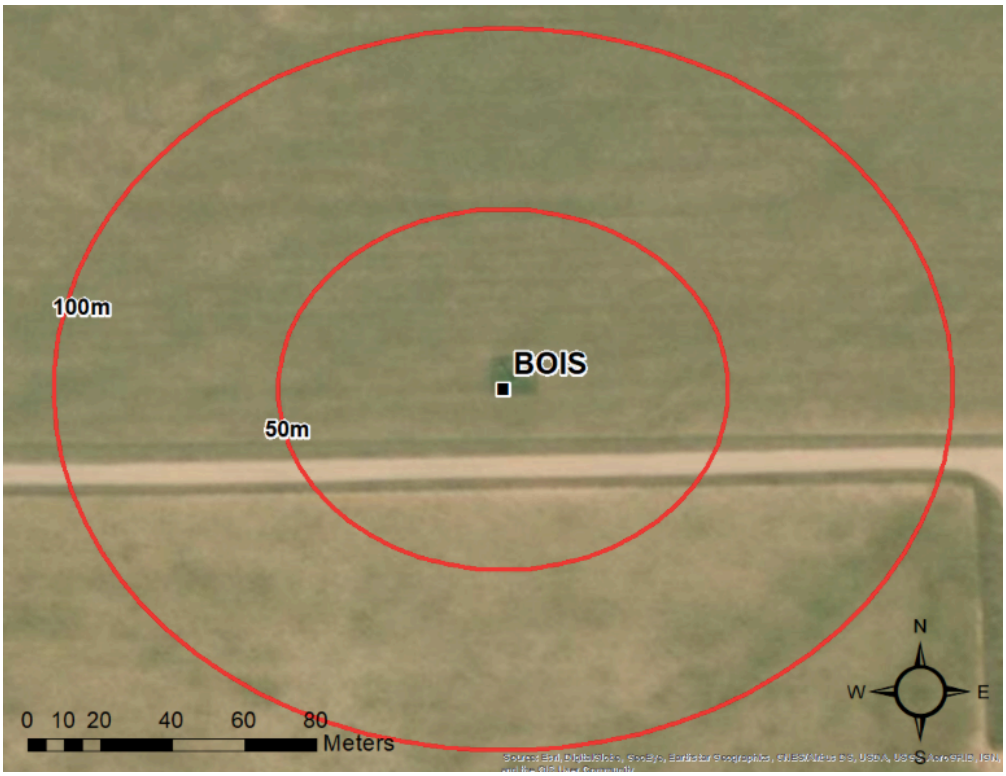


Figure 3.6: Aerial photo of Boise city station, picture taken on July 23, 2019 `http://climate.ok.gov/index.php/climate/climate_trends/temperature_history_annual_statewide/CD00/tavg/Annual`, accessed on April 7, 2020.

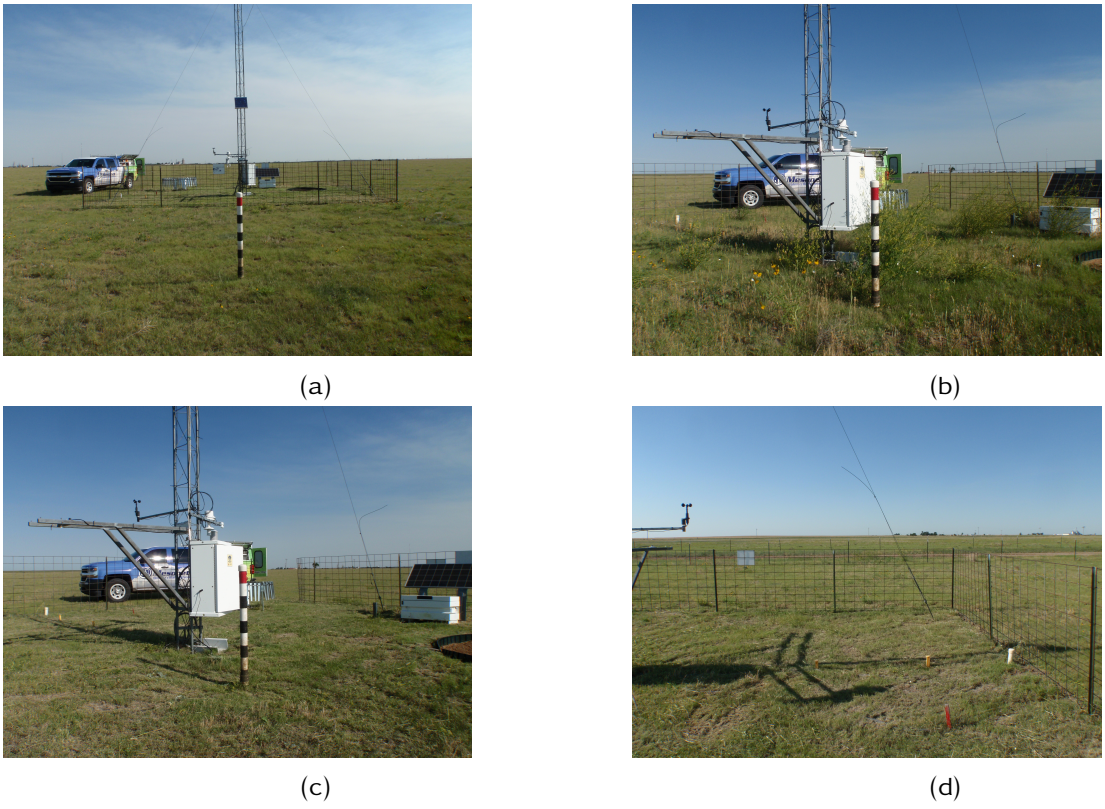
3.2.2 *In-situ* soil moisture data

In this section, a description of soil moisture data used in the data assimilation experiments is given. The three soil moisture data are the JULES predicted soil moisture, Mesonet ground measurement soil moisture data and SMAP satellite soil moisture data.

For Mesonet sites, soil moisture is measured indirectly from a reference temperature difference (ΔT_{ref}) using the heat-dissipation sensor, CSI 229-L, Illston et al. (2008). The initial and final temperature is recorded before and after introducing a 50 mA current into the resistor for 21 seconds. The temperature difference is measured at four depths of the soil, 5 cm (103 sites), 25 cm (101 sites), 60 cm(76 sites) and 75cm (53 sites) and every 30-min. Then volumetric soil moisture content θ ($\mathrm{m}^3\mathrm{m}^{-3}$) can be found after converting the reference temperature difference into matric potential MP (kPa) and then into volumetric water content.

The matric potential is given by:

$$MP = -c\exp(a\Delta T_{\text{ref}}), \tag{3.7}$$

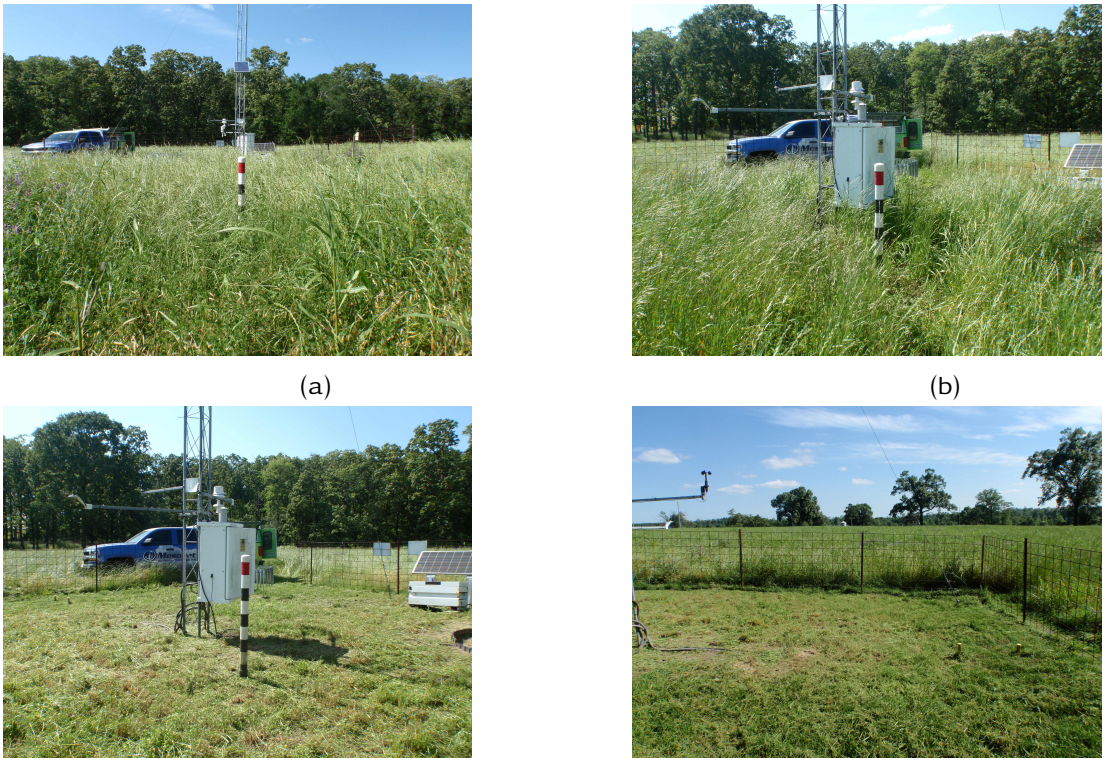


(a) (b)

(c) (d)

Figure 3.7: Boise city station as of 2017. Sub-figure a: vegetation cover outside the station, before trimming. Sub-figure b: vegetation cover inside the station before trimming. Sub-figure c: vegetation cover inside the station after trimming. Sub-figure d: as in sub-figure c but near the soil moisture measuring instruments. Pictures are taken from Oklahoma Mesonet website `https://www.mesonet.org/index.php/site/sites/station_names_map#`, access date Sept 5, 2019.

Figure 3.8: Aerial photo of Mt Herman station, picture taken on July 23, 2019 `http://climate.ok.gov/index.php/climate/climate_trends/temperature_history_annual_statewide/CD00/tavg/Annual`, accessed on April 7, 2020.



(a) (b)

(c) (d)

Figure 3.9: As in Figure 3.7 but for MtHerman station.

where $\Delta T_{\text{ref}}(^{\text{o}}\text{C})$ is the reference temperature difference, $a = 1.788^{o}C^{-1}$ and $c = 0.717kPa$ are calibration constants.

Then volumetric water content θ of the soil can be obtained using either the Van Genuchten relationship

$$\theta = \theta_r + \frac{\theta_s - \theta_r}{(1 + (-\alpha \times MP)^n)^{(1-\frac{1}{n})}}, \tag{3.8}$$

where θ_s $(\text{m}^3\text{m}^{-3})$ is soil moisture content at saturation, θ_r $(\text{m}^3\text{m}^{-3})$ is residual volumetric soil moisture. and The constants α and n are empirical constants. For Mesonet sites, soil characteristics θ_r, θ_s, α and n are given for each soil layer. Alternatively, volumetric water content can be obtained using the Brooks and Corey relationships, given in Equation 3.4 and Equation 3.5.

Illston et al. (2008) compared soil moisture data derived from the 229-L sensor and Van Genuchten equations with soil moisture products obtained from neutron probe and gravimetric methods. Based on the comparison, the maximum uncertainty from the 229-L was approximately 0.05 m^3m^{-3}.

3.2.3 SMAP soil moisture data

The Soil Moisture Active Passive (SMAP) satellite has been operational since April 2015 and provides soil moisture products at different spatial resolutions. Following the malfunctioning of the active (radar) microwave sensor in June 2015, only the passive(radiometer) microwave sensor continues to provide land process information. The passive microwave sensor detects microwave emissions from the soil surface in terms of brightness temperature, which is proportional to soil physical temperature. Different algorithms are used to convert brightness temperature into soil dielectric property and then to soil moisture. The proportionality constant between physical soil temperature and brightness temperature is soil microwave emissivity which is characterized by soil moisture variations (Karthikeyan et al., 2017).

The SMAP soil moisture resolution for the radiometer (Passive part) is 36km by 36km brightness temperature observations, and the 9km by 9km resolution data are obtained by disaggregation from the original data by Backus-Gilbert optimal interpolation method, Chaubell (2016). In the method, the surface brightness temperature in the required pixel is calculated by a weighted sum of the surface brightness temperature of the nearby pixels with the original resolution, (Poe,

1990; Long and Brodzik, 2016).

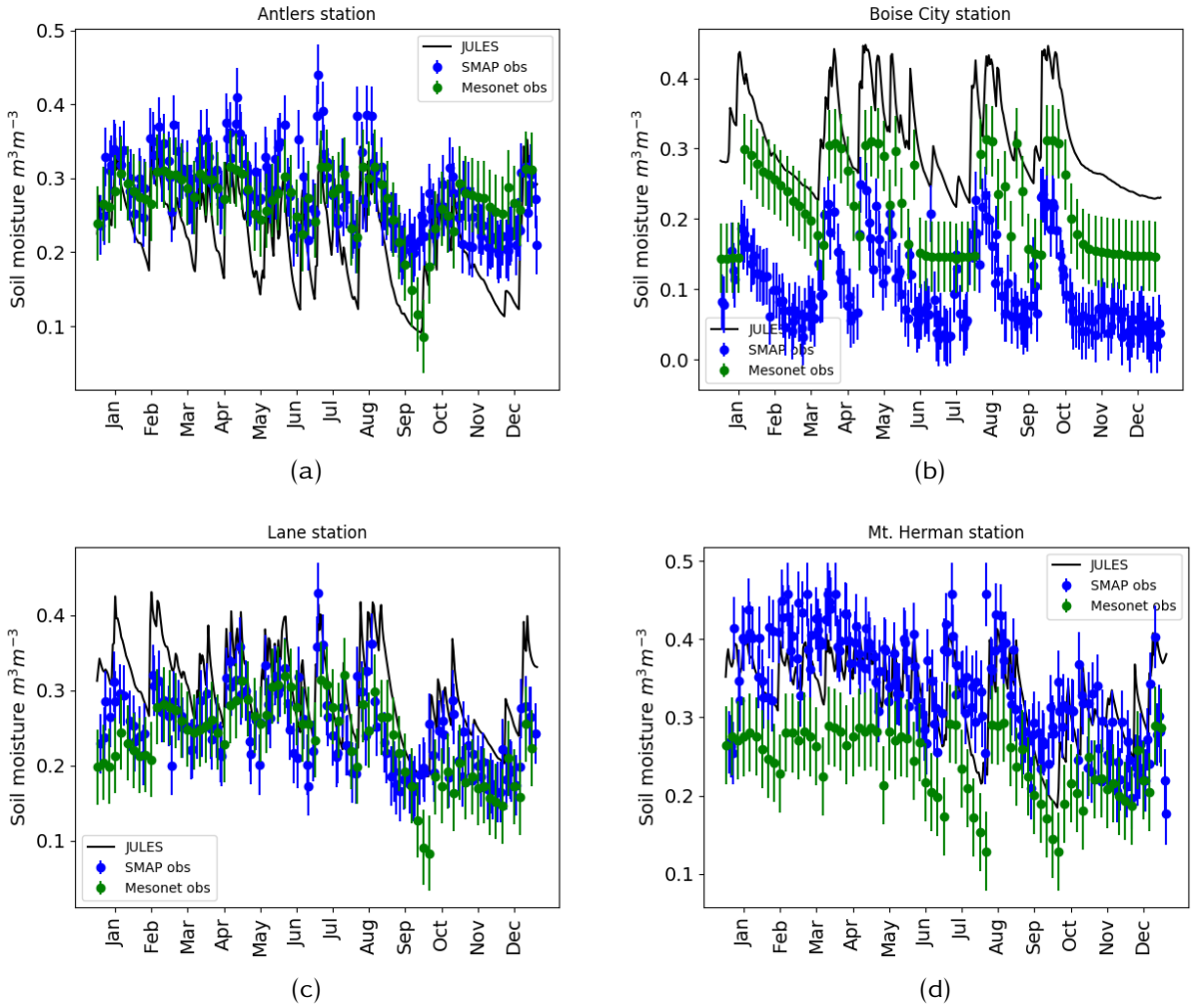


Figure 3.10: Comparison of volumetric soil moisture for the top soil layer in Mesonet sites for the year 2017. Blue dots: SMAP satellite observed data. Blue vertical lines: SMAP observation error bars. Green dots: Mesonet ground measurement data. Green vertical lines: Error bars for Mesonet observations. Black line: JULES model output.

Figure 3.10 shows a comparison of ground measurement volumetric soil moisture data from stations in Mesonet Oklahoma, satellite observed volumetric soil moisture data for the same stations and model estimates from the JULES model using the soil properties and meteorological forcing of the corresponding stations. The error bars are observation error standard deviations, $0.04 m^3 m^{-3}$ and $0.05 m^3 m^{-3}$ for SMAP and for Mesonet *in-situ* observations. The resolution for mesonet ground measurement volumetric soil moisture data is a point measurement, i.e. 10 cm diameter soil at each measuring depth (5 cm, 25 cm 60 cm and 75 cm), Illston et al. (2008) discusses the description of instrumentation for soil moisture measuring at Oklahoma Mesonet. On the other hand, SMAP soil moisture observations considered in these plots are 9 km × 9 km.

The frequency of Mesonet observations is at the same observation frequency with SMAP, two to three days apart.

Looking at Figure 3.10, soil moisture observations for Mt Herman (Figure 3.10d) appear to have a less defined pattern, compared to ground measurement observations. As photos from stations show, Mt Herman station has forests in the neighbourhood which has been included in the original resolution, a potential source of observation noise. Zhang et al. (2019a) found that SMAP soil moisture observations are less accurate for densely vegetated sites by comparing with *in-situ* measurement soil moisture data. In relation to forest coverage in the neighbourhood, the SMAP soil moisture observations are overestimated compared to both ground measurements and the JULES model run with C3 (temperate) grass cover configuration. For stations Antlers and Lane, Figure 3.10a and Figure 3.10c respectively, SMAP observation overestimated and observations are with noise but better than the case for Mt Herman station, and SMAP soil moisture observations are reasonably representative of the ground observations with some expected noises.

On the other hand, for Boise City station, Figure 3.10c, SMAP soil moisture is underestimated compared to both the ground measurement soil moisture and the JULES model run but less observation noise. Looking at Figure 3.6 and Figure 3.7, there is no forest in the neighbourhood, but some bare soil is included rather. As a result, the satellite observations looked drier as it is the disaggregated, including the bare soil with high evaporation areas.

From Figure 3.10, it is evident that there are systematic and random errors of the SMAP data and the output from JULES. Keeping this in mind, data assimilation experiments are performed in chapter 6 without any bias correction to see by how much the raw data is usable. The performance of the data assimilation is evaluated by comparing with *in-situ* soil moisture data.

Chapter 4

Stochastic forcing for ensemble spread generation in soil moisture data assimilation with ETKF

This chapter looks at uncertainty representation in the rainfall forcing data and in the DRBC model. As such rainfall is generated stochastically based on observations and also Gaussian model error is implemented, which both help to generate ensemble spread for Ensemble open-loop (EnOL) soil moisture predictions. *In-situ* soil moisture data are assimilated using an ETKF for state estimation. The RMSE and error-spread score ES are used to evaluate posterior soil moisture accuracy and uncertainty, respectively.

4.1 Introduction

The accuracy of a modelled dynamics in representing the true state and degree of uncertainty depends on several factors. Uncertainties of the initial conditions, parameter values and forcing data impact the accuracy of the model prediction in addition to the model physics. In the case of atmospheric models especially, altering the initial condition alone can result in a completely different trajectory for the same model (Magnusson et al., 2008). Choice of parameter values also plays a major role in modelling, especially for land surface models where parameter values are often not well-known (Pinnington et al., 2020). Forcing data also has a direct effect on the accuracy of the model state. Hence, to account for those uncertainties and encompass a general

behaviour of the model under different circumstances, it is essential to consider multiple runs of a model instead of a single prediction.

Ensemble methods consider multiple model integrations where each model integration, also called ensemble member, represents a distinct evolution of the model by differing initial conditions, forcing data and/or model formulation. Often, the mean is assumed to be a single forecast which represents the best available estimate of the model dynamics, (Whitaker and Loughe, 1998). Ensemble spread represents the uncertainty in the model forecast whereby the smaller the spread, the smaller the presumed uncertainty and vice versa.

Ensemble initialisation can be achieved by perturbing either initial conditions or parameters or the driving data. Given an initial condition, by assuming errors follow a known distribution, ensemble members can be initialised using perturbed background state with errors sampled from the chosen distribution (Browne and Wilson, 2015). This technique works best to provide spread among ensemble members throughout the model integration window for chaotic models like atmospheric models, as small perturbations can lead to a significant deviation in the dynamics. However, for models like land surface models, perturbations of initial conditions do not give spread among ensemble members. The dynamics of each ensemble member approaches a dynamical attractor and finally become identical after a while, which is called ensemble collapse. Figure 4.1 shows that, after about three months, the dynamical system loses memory of the initial condition perturbation and ensemble collapses.

Ensemble methods are often used for ensemble prediction systems (Magnusson et al., 2008; Wilks, 2007). In this chapter, however, ensemble methods are used to represent prior distributions of soil moisture predictions for ensemble data assimilation method, ETKF. The hydrology part of the JULES model, the DRBC model, as described in subsection 3.1.2, is used for soil moisture prediction. Ensemble initialisation and stochastic forcing mechanisms for ensemble spread enhancement are discussed in section 4.2 and section 4.3 respectively.

4.2 Initial condition perturbation

To start with, consider the DRBC model as a perfect model, perfect parameters and also perfect forcing rainfall data. Assuming the DRBC model to act as a function $\mathcal{M}$ and the forcing rainfall data appeared explicitly, the time evolution of the state variable $\mathbf{x}$ described in Equation 2.1

becomes:

$$\mathbf{x}_{t+1} = \mathcal{M}_{t+1,t}(\mathbf{x}_t, p_t) \tag{4.1}$$

where t is the time step and p observed rainfall. Equation 4.1 shows that, besides the perfect model assumption, it results in a deterministic approach where only one model prediction is considered as a source of information about the predicted state.

The data assimilation method we will use here is the ETKF, described in subsection 2.1.1, a sequential method which requires ensemble members of a model. Initial condition perturbations are often used to generate ensemble members by drawing samples from a priori error covariance matrix by assuming a Gaussian error. Hence, from a reference state $\mathbf{x}_{\text{ref}}$, initialisation of ensembles by perturbation follows as:

$$\mathbf{x}_0^{b,i} = \mathbf{x}_{\text{ref}} + \zeta_i, \text{ where } \zeta_i \sim \mathcal{N}(0, \mathbf{P}_0^b). \tag{4.2}$$

where $i = 1, 2, 3, \ldots N_e, \mathbf{P}_0^b$ is the background error covariance matrix. Hence following Equation 4.1 to get the ensemble members, time evolution of each ensemble member is given by

$$\mathbf{x}_{t+1}^{b,i} = \mathcal{M}_{t+1,t}\left(\mathbf{x}_t^{b,i}, p_t\right) \text{ for } t = 0, 1, 2, \ldots T_{\text{max}}. \tag{4.3}$$

From Equation 4.3, the difference between each ensemble member is only the initial condition; the meteorological forcing data, including rainfall and parameter values, are the same for each model integration.

Figure 4.1 shows top level soil moisture ensemble members ($N_e = 200$) predicted by the DRBC model over a year by following Equation 4.2 and Equation 4.3. The reference state, meteorological forcing data and parameter values for each site is taken from Oklahoma Mesonet observation in 2016 and the background error covariance matrix $\mathbf{P}_0^b = 0.1^2\mathbf{I}_{4\times4}\ \mathrm{m}^3\mathrm{m}^{-3}$.

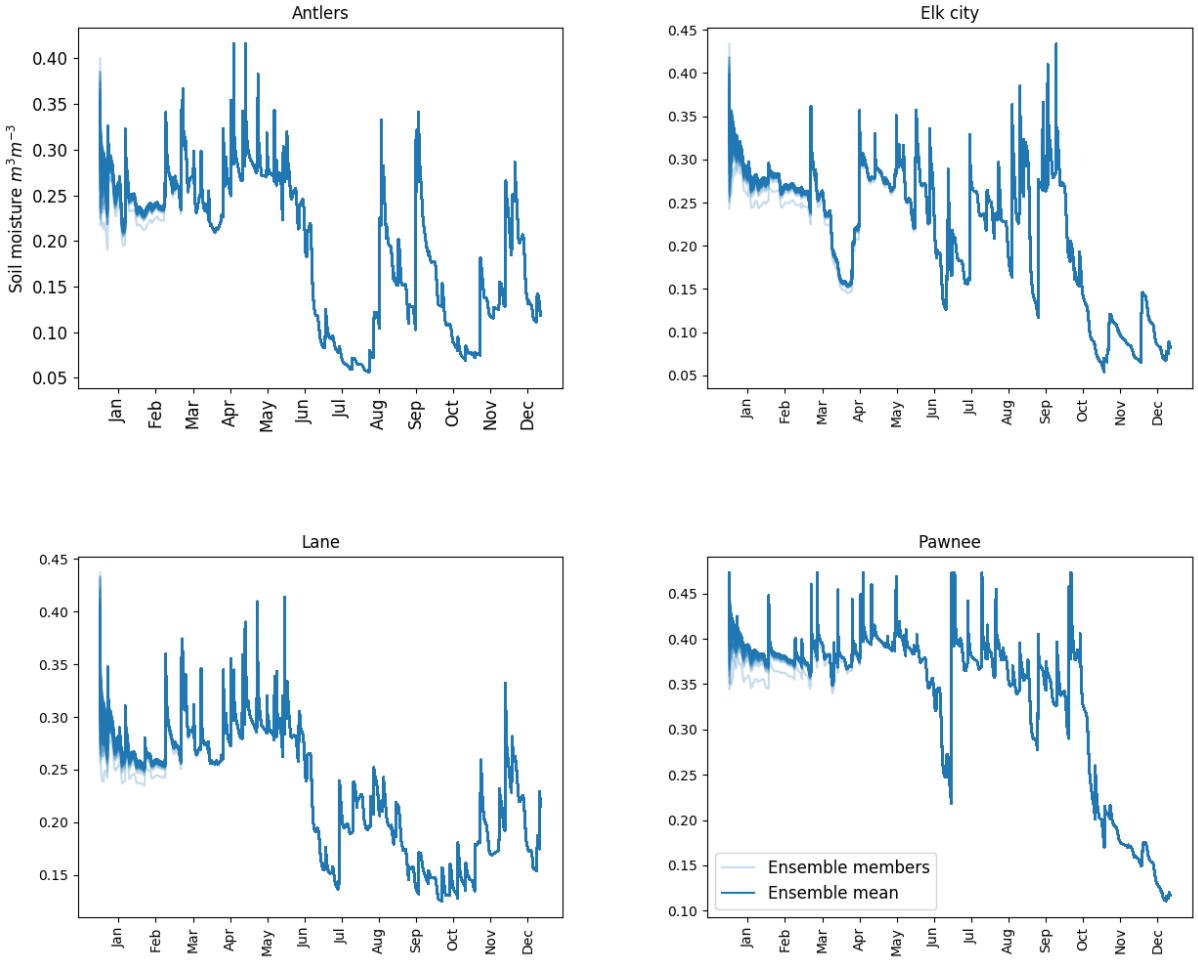


Figure 4.1: Open loop DRBC model estimate of volumetric soil moisture for Messonet sites, Oklahoma. The light blue lines are the 200 ensemble members and the blue line is ensemble mean.

Figure 4.1 shows that for all the four sites initial condition perturbation does not give spread among ensemble members except the first three months of the model integration. After that, the system loses memory of the initial condition, and each ensemble member converges to a similar prediction irrespective of their difference initialisation. Increasing the perturbation at the beginning increases spread among ensemble members and for a bit longer time, however, it does not give spread throughout the integration period. Besides, the perturbation should be reasonable (not too big) since there is information about the uncertainties in the background. Hence, initial condition perturbation does not give appropriate spread among ensemble members for the DRBC model and other ensemble generating techniques, stochastic forcing, are introduced in the following section.

4.3 Stochastic forcing

To stop the ensemble collapse in the DRBC model, observed in Figure 4.1, two stochastic forcing methods are deployed. The first method is by introducing stochastic forcing through rainfall and second by considering the model error.

Since rainfall measurements are prone to errors (Ettema and Viterbo, 2001), it is crucial to take into account the errors in the observed rainfall. Rainfall is a driving meteorological data for all land surface models, and the amount of noise is determined from the observed rainfall itself.

In this case, the model equation given in Equation 4.1 becomes,

$$\mathbf{x}_{t+1} = \mathcal{M}_{t+1,t}(\mathbf{x}_t, V_t) \tag{4.4}$$

where V_t is generated rainfall for a given observed rainfall at time t. Equation 4.4 shows that, the stochastic forcing is acting on the model state every time step the model is integrated as opposed to the initial condition perturbation. For each ensemble member of soil moisture estimates, different generated rainfall forcing generated from the same observed data are used. This assures that each ensemble member will have spread among them as a result of a different rainfall forcing.

Either due to simplifications in model parametrisation or the missing processes, the model is by no means perfect in representing the observed physics. Maggioni et al. (2012) pointed out that soil moisture estimates from a model are directly affected by errors in the model formulation, in addition to errors in the meteorological deriving data and errors in model parameters. Particularly for the DRBC model, the hydrological cycle is considered separately, and vegetation dynamics is not modelled. Hence there is a strong case for representing the effects of model error.

By assuming a Gaussian error, but perfect forcing the model becomes

$$\mathbf{x}_{t+1} = \mathcal{M}_{t+1,t}(\mathbf{x}_t, p_t) + q_t, q_t \sim \mathcal{N}(0, \mathbf{Q}), \tag{4.5}$$

where $\mathbf{Q}$ is model error covariance matrix. As in the case of generated rainfall discussed above, each ensemble member of soil moisture estimate will have different stochastic forcing and ensures ensemble spread between them is gained.

When stochastic forcing from both generated rainfall and model error is considered, the model with stochastic forcing can be generalised as

$$\mathbf{x}_{t+1} = \mathcal{M}_{t+1,t}(\mathbf{x}_t, V_t) + q_t, q_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{4.6}$$

So far, the stochastic forcing from generated rainfall and model error is described in general forms, subsection 4.3.1 and section 4.4 describe the actual stochastic forcing used in this thesis.

4.3.1 Stochastic rainfall generation

Rainfall forcing data can be generated stochastically using observed rainfall data to account for uncertainties in the observed rainfall or to obtain generated rainfall from nearby stations where there is no observed rainfall for a particular station. In either case, the Gamma distribution is a widely used distribution to generate stochastic rainfall. The probability density function of the Gamma distribution, $f(p)$, as given by Roger Stern and Richard Coe (1984) and Wilks and Wilby (1999) and many other authors is

$$f(p) = \frac{\left(\frac{p}{\beta}\right)^{\alpha-1} \exp\left(\frac{-p}{\beta}\right)}{\beta\Gamma(\alpha)} \tag{4.7}$$

where $p \geq 0, \alpha, \beta > 0$, p is a random variable, in this case the amount of rainfall. α and β are constants where α is shape parameter, β is scale parameter and Γ is the gamma function.

The shape and scale parameters for the gamma distribution are approximated using the moments method, Wilks (1990)

$$\beta = \frac{\sigma^2}{\mu}, \alpha = (\frac{\mu}{\beta}) \tag{4.8}$$

where μ is the mean rainfall for the considered period of time. However, rain is accompanied by dry conditions, and the Dirac delta distribution is used to represent the dry events.

To calculate parameters for the Gamma distribution and probability of being rainy for a par-

ticular time step (30 minutes in our case to match with a model time step), we need to consider a broad range in the data so that we encounter rainy events. For Mesonet rainfall dataset, we considered 31 days moving window (similar to Basinger et al. (2010)). The distribution we draw rainfall from is $G(p)$, similar to the one developed by Hyndman and Grunwald (2000) where

$$G(p) = rf(p) + (1 - r)\delta_0(p), \tag{4.9}$$

$f(p)$ is the gamma distribution, $\delta_0(p)$ is the Dirac delta function, r is the probability of occurrence of rainfall.

Basinger et al. (2010) generate rainfall for rainwater harvesting with a moving window by considering 30 years of rainfall data as a domain to draw samples. Here, a moving window is used as in the case of Basinger et al. (2010), but we draw from a gamma distribution instead of from a rainfall data.

From the observed rainfall data, the probability of a rainy event is given by the ratio of the number of rainy periods divided by the total number of observations considered in that window. Mathematically the probability of rain is given by:

$$r = \frac{N_R}{N_T} \tag{4.10}$$

where N_R is a number of rainy (wet) periods, and N_T is the total number of observations in the rainfall data. If the probability of being wet is r, then the remaining probability (the probability of being dry) will be $1 - r$. That means, we draw from a Gamma distribution with probability r and draw from a Dirac delta distribution with probability $1 - r$. Then the generated rainfall ensemble is used to force the DRBC model to get different state estimates for each rainfall ensemble so that the mean will represent the best possible state estimate.

4.4 Experimental design

For numerical data assimilation experiments, we used ETKF as implemented in Employing Message Passing Interface for Researching Ensemble (EMPIRE) at the University of Reading (Browne and Wilson, 2015) by coupling with the DRBC model. Soil moisture observations are from Mesonet sites, converted from temperature difference into volumetric soil moisture using the Brooks and

Corey equations as given in subsection 3.2.2. The reason for choosing the Brooks and Corey over the Van Genuchten is that to be consistent with the DRBC model as the Brooks and Corey are considered in the model.

Since the available observation of volumetric soil moisture are at three depths in the soil, 5 cm, 25 cm and 60 cm, the observation operator becomes

$$\mathbf{H} = \mathbf{I}_{3\times 4}. \tag{4.11}$$

The observation error covariance matrix $\mathbf{R}$, as given by Illston et al. (2008), is a diagonal matrix (since measurements in each point are independent to each other) with entries 0.05^2. Hence

$$\mathbf{R} = 0.05^2\mathbf{I}_{3\times 3}. \tag{4.12}$$

Each ensemble member is initialised with a Gaussian perturbation from a reference state. The initial perturbation is chosen in such a way that uncertainty in the background is more than the uncertainty in the observations, to put more trust in the observations than the model. The background error covariance matrix we choose is

$$\mathbf{P}_0^b = 0.1^2\mathbf{I}_{4\times 4}. \tag{4.13}$$

To test different scenarios for the data assimilation, we considered the DRBC model to be perfect in the first instance, followed by considering a model error with Gaussian distributed and diagonal covariance matrix, $\mathbf{Q}$, with different magnitudes. The choice of the size of $\mathbf{Q}$ is in such a way such that the incorporation of the model error does not distort the model dynamics. The three covariance matrices we considered for the model error are $\mathbf{Q_0} = \mathbf{0}, \mathbf{Q_1} = 0.002^2\mathbf{I}_{4\times 4}$ and , $\mathbf{Q_2} = 0.01^2\mathbf{I}_{4\times 4}$ where $\mathbf{Q}_0$ represents a perfect model.

On the other hand, we considered the forcing rainfall data as perfect and also with errors. To account for the observation error in the rainfall, we used the stochastic rainfall generator discussed above. Generated rainfall events vary in standard deviation, σ, but the mean is maintained. We considered three generated rainfall patters with varying standard deviations: $\sigma_1 = \sigma_0, \sigma_2 = 2\sigma_0$ and $\sigma_3 = 5\sigma_0$ where σ_0 is the standard deviation of observed rainfall. The reason for considering generated rainfall with different standard deviations is to represent rainfall patterns in the nearby places and also at different times.

The background ensemble is initialized as in Equation 4.2 and the time evolution follows Equation 4.6 for different combination of $\mathbf{Q}$ and σ given above.

The model time step is set to 30 minutes, observation frequency is 72 hours, and the number of ensemble members is 200. The background state of soil moisture $\mathbf{x}_0$ is obtained from the four soil layers from Mesonet sites.

4.5 Diagnostic tools

The data assimilation performance is often evaluated using the RMSE and time series plots of the posterior state compared to the prior, which is the mean of ensemble open-loop model run EnOL. In this chapter, in addition to the traditional metrics, appropriateness of ensemble spread of the analysis is examined based on the error-spread score proposed by Christensen et al. (2015). The method is motivated by Leutbecher and Palmer (2008) and compares ensemble RMSE with root mean square deviation (RMSD), also known as ensemble standard deviation of the mean, for forecast verification. Leutbecher and Palmer (2008) discussed that for a perfect ensemble of the forecast, ensemble mean RMSE should be proportional to the RMSD. In an ideal situation, the proportionality constant is close to one, and the scatter plot of ensemble-mean RMSE, and RMSD will lie at 45 degrees. The use of ES in this work is different from the use in Christensen et al. (2015) such that, series of analysis over the assimilation window is evaluated instead of the forecast. Christensen et al. (2015) argued that the first two moments of a forecast distribution are not enough to check if the forecast is proper or not. For that reason, the third-moment, skewness, is included in the new metric they proposed. The error-spread score given in Christensen et al. (2015) is

$$ES_i = ({s_i}^2 - {e_i}^2 - s_i e_i g_i)^2 \tag{4.14}$$

where s_i is the standard deviation of distribution of ensemble members, $e_i = z_i - m_i$ is error in the ensemble mean m_i, z_i is verification (observation) and g_i is skewness of distribution of ensemble members. Then ES is calculated by taking the average of all forecast-verification pairs at different grid box and at different time. In our case however, there is only one grid box and the averaging will be over time.

The way to interpret the ES score is that the distance between the ensemble mean and obser-

vations should be proportional to the uncertainty of ensemble mean. This is because if the mean is far from the observations (with larger RMSE) and with small uncertainty (smaller standard deviation), it gives false confidence on the mean. On the other hand, if the mean has smaller RMSE but larger variance, the mean is accurate but has a small believability. That means the accuracy and the associated uncertainty need to be proportional, i.e. the smaller the value of ES is the better (certain) the forecast.

However, smaller ES can be achieved as long as the RMSE and RMSD are proportional. Considering the case where RMSE of the mean is larger, the forecast is far from the observed reality. In this case, the forecast should not be considered as appropriate since it is not accurate. Forecast appropriateness needs accuracy and precision, which means smaller RMSE and smaller RMSD. Hence in our case, the forecast is considered to be appropriate when both RMSE and ES are smaller, as close as possible to zero.

4.6 Results and discussions

In this section, posterior soil moisture predictions of the DRBC model obtained from assimilating *in-situ* soil moisture data with data assimilation set-ups given in section 4.4 are presented and discussed. Posterior soil moisture time series, RMSE and ES from different set-ups are compared. The comparison is made for five sites and three soil layers (Appendix B) where data were assimilated. Though the fourth layer soil moisture estimates are available, the discussion does not include the fourth layer as there are no observations to compare with. For ease of representation, each ES is scaled (divided) by the ES from posterior soil moisture for data assimilation with a deterministic model run, (with σ_0 and Q_0). Compared to the ES with EnOL, data assimilation with σ_0 and Q_0 did not show a significant improvement due to ensemble collapse. Hence stochastic forcing is introduced. In general, data assimilation has helped to reduce ES, though very small improvement without stochastic forcing.

Figure 4.2 shows time series plots of top layer soil moisture ensemble mean and shades in respective colours within ±1 standard deviation from the mean together with respective RMSE and ES score plots for Antlers station, described Table 3.1. Each time series plot represents soil moisture ensemble mean for a single rainfall forcing but varying model error. In Figure 4.2a, for example, observed rainfall forcing, with standard deviation σ_0 (standard deviation of observed

rainfall) and model error with varying covariance matrix is used. The blue, purple and black lines are for $\mathbf{Q_0}, \mathbf{Q_1}$ and $\mathbf{Q_2}$ respectively. The associated uncertainties, ensemble spread with $\pm$ 1 standard deviation from the means are displayed with blue, purple and grey (instead of black) shades. Figure 4.2e and Figure 4.2f are RMSE and ES plots for the corresponding time series plots. The colours used for the bar plots of RMSE and ES correspond to the colours of the shades in the respective time series plots. The green dots and vertical lines are soil moisture observations with observation error standard deviation of $\pm$ 0.05.

In Figure 4.2a, the blue line shows that posterior soil moisture estimates with perfect model set-up are far from observations. This has been shown with large RMSE and large ES in Figure 4.2e and Figure 4.2f respectively. As discussed in Figure 4.1, EnOL soil moisture of the DRBC model have shown a very narrow ensemble spread for a perfect model run with initial condition perturbation. Having small spread is associated with high certainty; hence observations are given minimal weight in the data assimilation system. In this case, the posterior is almost the same as the EnOL and far from the observations. Comparing the blue line with observations, the difference is larger during the drying up periods, between June and August, for example. The most likely reason is, the DRBC model does not include the vegetation dynamics, and as a result, evaporation from bare soil is higher. Hence, the drier trend for the DRBC on the top layer. Occasionally, the DRBC overestimate soil moisture. But for deeper layers, as there is no transpiration from plants, the DRBC estimate tends to be wetter than the observations.

When stochastic forcing is introduced via model error with covariance matrix $\mathbf{Q_1}$ and $\mathbf{Q_2}$, Figure 4.2a purple and black lines, the posterior soil moisture ensemble means are getting closer to the observations. Hence corresponding RMSE and ES have reduced, the first column of Figure 4.2e and Figure 4.2f respectively. The value of RMSE and ES decreases as model error increases. By applying model error only, error-spread score reduction ranges from 88% – 99% across the five stations and three soil layers. With this method, the best reduction in the error-spread score (99%) is obtained for Pawnee station first layer, Figure B.4f and Lane station third layer, Figure B.12f. The smallest reduction (88%) is for Elk City station second layer, Figure B.7f and Pawnee station second layer, Figure B.9f. There is no dependence between the soil layers and the amount of reduction in the error-spread score when the model error is used. However, when generated rainfall is used with a different value of $\mathbf{Q}$, the values of RMSE and ES are similar to the value with $\mathbf{Q_2}$ and observed rainfall. For example, for Antlers station, implementing $\mathbf{Q_2}$ is sufficient to get smaller RMSE and smaller ES.

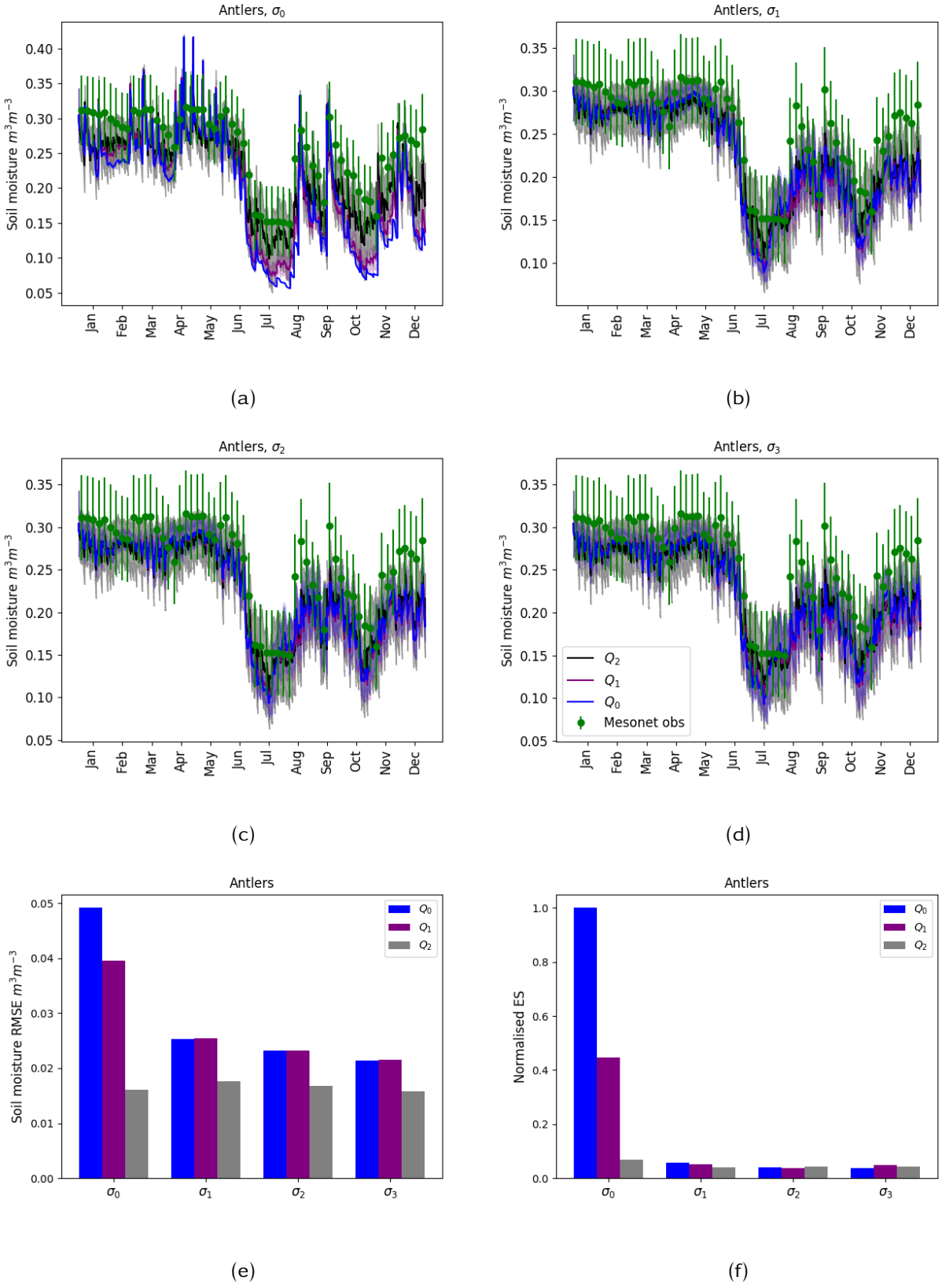


Figure 4.2: Top layer posterior soil moisture ensemble mean for different values of **Q** and σ, as described in section 4.4. The shades for each mean is ± 1 std from the mean. The bar plots at the bottom are corresponding RMSE and ES score for each *sigma* and varying **Q**. The forcing data and parameter values are from Antlers station, Oklahoma Mesonet for the year 2016.

Similar patterns are observed for results given in Figures B.1 - B.4, once $\mathbf{Q}_2$ is implemented, the variation in σ does not have a significant impact on the magnitude of the RMSE and ES. However, with $\mathbf{Q}_1$, the combination of model error with $\mathbf{Q}_1$ and generated rainfall has improved ES score, for example, Elk city top layer presented in Figure B.2f.

When generated rainfall forcing is used separately ($\mathbf{Q} = \mathbf{0}$), a substantial reduction in the ES score is observed for the top layer though out the five stations we considered. The percentage improvement ranges from 92% (Lane station) - 98% (Pawnee station). For the top layer, reduction in the ES score as a result of generated rainfall is slightly smaller percentage compared to applying model error. For second and third layers, however, generated rainfall brought small improvement or even increased RMSE and ES score. This is because only a small fraction of rainfall can reach the deeper layers, and as a result, the improvements in error-spread reduction we get is much smaller than the top layer. In this set-up for the top layer, minimum RMSE and ES values are obtained with σ_3 except for Lane station, Figure B.3e, where σ_2 gives the minimum RMSE and ES. The only exception is Lane station where generated rainfall reduces the error-spread score by 92% and model error by 89%. Whereas for the second and third layers, model error performs much better (minimum 88%) than generated rainfall in all the five stations (maximum 62%).

Figures B.5 - B.13 are times series plots of posterior soil moisture ensemble mean together with the corresponding RMSE and ES for deeper layers, layers two and three. For the deeper layers, there is additional stochastic forcing from the top layer due to the propagation of information downwards by the model besides the forcing considered in each layer.

Similar to what is observed for the top layer, Q_2 resulted in the smallest RMSE and corresponding ES, most of the time. In some cases, however, increasing from Q_1 to Q_2 has increased RMSE or/and corresponding ES. Figures 4.3e and 4.3f show that Q_2 gives smallest RMSE but ES is larger than the result obtained with Q_1. Looking at the time series plots in Figure 4.3, the posterior soil moisture ensemble mean with Q_2 is closer to the observations compared to the mean with Q_1 but the spread with Q_2 is larger than the spread with Q_1. This shows that the predicted error (spread) is much larger than the observed error. In this case, the decision is open as to whether the result with Q_1 or with Q_2 should be considered appropriate. Here we argue that the result with Q_1 is better to be considered as the magnitude of the RMSE (0.015) is smaller compared to the observation error and with high certainty, than RMSE of 0.005 with very high uncertainty.

Comparing the performance with the top layer, deeper layers shown very smaller or even

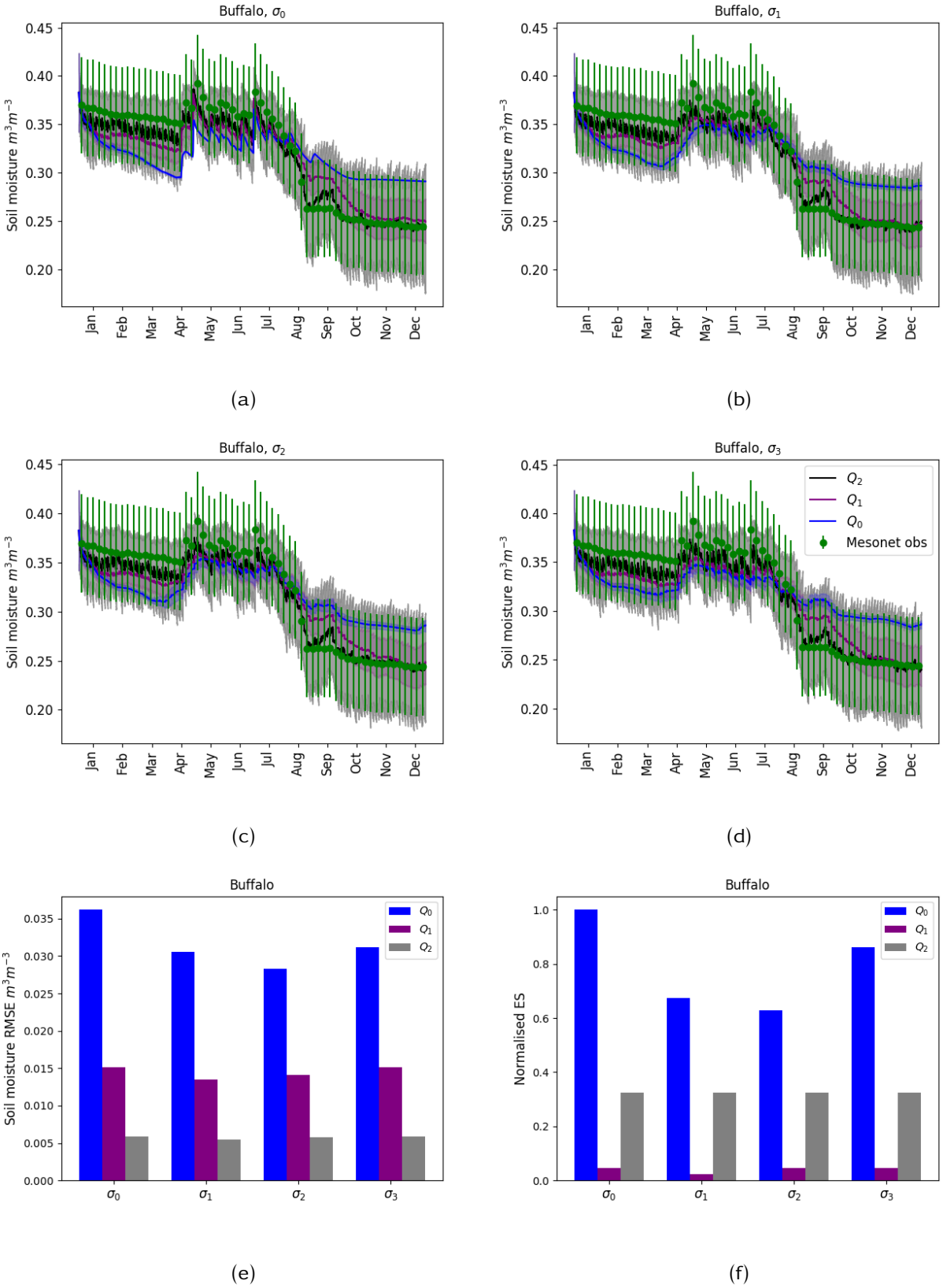


Figure 4.3: As for Figure 4.2 but for Buffalo station third layer.

negative impact as variability (spread) in stochastic rainfall increases. For deeper layers, σ_1 has shown the best performance, i.e. smaller RMSE and ES, in most cases where σ_3 was the best for the top layer. It was observed that when stochastic forcing increases, especially using generated rainfall, RMSE and/or ES increase instead of decrease. For example, in Figure B.8f and Figure B.10f, as σ increases, the value of ES also increases. This is because the contribution from the upper layer and also the soil moisture dynamics is smoother and adding more stochastic forcing pushes the dynamics away from observations.

In general, we have seen that top layer soil moisture observations are more dynamic than deeper layers, as a response of rainfall. As we go deeper, the dynamics smooths out and becomes less variable. Compared to using model error, high-frequency soil moisture observations are not well captured when generated rainfall is used. This is due to the dampening of the dynamics when the ensemble mean is calculated using different patterns of generated rainfall. This phenomenon is true whether generated rainfall is applied alone or with model error and for all rainfall patterns we considered. The effect of generated rainfall is more pronounced on the top layer as rainfall has direct contact into the top layer. The incorporation of model error enhances well the DA in reducing the gap between the observations and the model prediction irrespective of the frequency of the model dynamics and soil layers.

4.7 Summary

In general, both methods of stochastic forcing, generated rainfall and model error, help reduce the RMSE and ES score. This shows that the methods help to generate appropriate ensemble spread. From the DA experiments results (Figure 4.2 - Figure B.13), it is shown that generated rainfall reduces RMSE and ES score for the top layers better than for the deeper layers. When each one of the methods is implemented separately, model error performs better than generated rainfall in reducing the RMSE and ES score. Most of the time, applying Q_2 has given the best combination of RMSE and ES except Figure 4.3.

Generated rainfall dampens the model dynamics, make less variable, for soil moisture ensemble mean plots. Even though it helps to capture the dynamics when the water is drying out quickly (when observed rainfall is used), most of the observations are not well captured when the dynamics are peaking quickly. The performance of the generated rainfall depends on the soil

layer and also the observed soil moisture dynamics the model is trying to capture. However, the inclusion of model error works for all soil layers and for different dynamics considered whether high frequency or drying out quickly.

When both model error and generated rainfall are applied, the effect of model error is more influential mainly for the error-spread metric. Considering soil moisture ensemble mean time series plots, the effect of generated rainfall is reflected in dampening the dynamics. It is also observed that the larger the ES with the perfect model, the largest percentage reduction when stochastic forcing is applied and vice-versa.

Chapter 5

Parameter estimation using the Dirichlet distribution to initialise model ensemble for 4DEnVar: twin experiments

This chapter aims to suggest a novel technique of parameter estimation for soil moisture data assimilation using 4DEnVar with the JULES model for parameter estimation.

5.1 Introduction

Numerical results in chapter 4 showed that initial condition perturbation for the DRBC (hydrology part of the JULES) model did not provide ensemble spread except for the first few time steps. The ETKF has been implemented for soil moisture estimation with the DRBC model by assimilating ground measurements of soil moisture data. Other ensemble spread generating techniques have been deployed to gain ensemble spread: introducing stochastic forcing via the forcing data (rainfall) and allowing for a model error. Error spread (Christensen et al., 2015) is used to evaluate the performance.

In this chapter, we use a twin experiment to explore the impact of assimilating top layer soil moisture data using the 4DEnVar with the JULES model to estimate soil parameters: percentage

sand, silt and clay. The reason for using 4DEnVar in this chapter is that parameter estimation with the ETKF (used in chapter chapter 4) is likely to result in a time-variant estimation of parameters for each assimilation step (Pinnington et al., 2018). Having time-variant parameters contradicts the fact that parameters are constant through the year for a particular area. Even if parameters change due to changes in organic matter of the soil naturally, it takes longer time than the assimilation window, which is one year in our experiment. If the soil parameters are changing with human intervention, sequential data assimilation could be a way to estimate soil parameters. On the other hand, 4DEnVar - as an example of variational data assimilation, will yield time-invariant posterior parameters throughout the assimilation window. Another difference in this chapter from chapter 4 is, here the full JULES model is used to consider more realistic assumption on the model while the DRBC model in chapter 4. As discussed in chapter 4, the DRBC model did not consider processes like vegetation and bare soil evaporation was considered.

Parameter perturbation is better than initial condition perturbation (seen in chapter 4) at keeping ensemble members spread for land surface models like JULES. Parameters act upon the model state every time step when the model is integrated, as opposed to the initial condition perturbation where the effect is only at the beginning of the model integration. Soil texture parameters can be sampled from different distributions. This technique is considered as parameter perturbation since random errors are being added to the mean parameter where errors are from the chosen distribution. In doing so, the actual error is not known but represented by the error statistics when parameters are sampled. To start with, considering a diagonal covariance matrix is the easiest to choose as correlations are not known. However, the correlation is not always zero. For example, in our experiment, percentage sand, silt and clay are highly correlated as their sum should be a hundred and the increase/decrease of one parameter will result in reduction/increase of another. As a result, errors are also correlated. Hence, considering a diagonal covariance matrix is not realistic. However, even if we know that errors are correlated, determining the value of the correlation is a difficult task.

The Dirichlet distribution section 5.2 is a potential distribution to sample soil texture. Soil texture parameters are needed to be positive, and their sum is bounded above by a hundred. The Dirichlet distribution also shares this property of boundedness and positivity of individual parameters distribution. Hence sampling soil parameters from the Dirichlet distribution is meaningful, and samples from the Dirichlet distribution will have non-zero correlations. Ensemble initialisation of the state (soil moisture) is then attained by running the JULES model so that the initial

condition soil moisture is consistent with the soil texture parameters.

Considering the link between the properties of the Dirichlet distribution and soil texture parameters, sampling from a Dirichlet distribution is a promising method to get parameter ensembles. On the other hand, in the data assimilation system, background errors (and observation errors) are assumed to follow a Gaussian distribution. To this end, it is unclear whether or not soil texture parameters sampled from a Dirichlet distribution can be used for soil moisture data assimilation. Hence for comparison, we sampled soil texture parameters from the Dirichlet and Gaussian distributions with the same correlations. Here the Gaussian distribution is truncated so that samples are positive. To examine the benefits of having correlations in the Gaussian distribution (which comes from the Dirichlet distribution), we have also sampled from a Gaussian distribution without correlations. In section 5.2 and section 5.3 , basic descriptions of the Dirichlet and Gaussian distributions are given.

5.2 The Dirichlet distribution

The Dirichlet distribution is a multivariate distribution characterised by a vector with dimension D, say $\alpha_1, \alpha_1, ..., \alpha_D$, is given by (Lin, 2016)

$$p(x_1, ..., x_D|\alpha_1, ..., \alpha_D) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1}, \tag{5.1}$$

where

$$\sum_{i=1}^{D} x_i = 1, x_i > 0, \tag{5.2}$$

and Γ is the Gamma function. Note that the condition for the vector α_i is, $\alpha_i > 0$. The two key parameters are the concentration parameter

$$\Omega = \sum_{i=1}^{D} \alpha_i \text{ and the base measure } \lambda_i = \frac{\alpha_i}{\Omega}. \tag{5.3}$$

The first and second moments are defined as:

$$\mathbb{E}(x_i) = \frac{\alpha_i}{\Omega} (0 < \mathbb{E}(x_i) < 1), \tag{5.4}$$

$$\mathrm{Var}(x_i) = \frac{\alpha_i(\Omega - \alpha_i)}{\Omega^2(\Omega + 1)}, \tag{5.5}$$

$$\mathrm{Cov}(x_i, x_j) = \frac{-\alpha_i \alpha_j}{\Omega^2(\Omega + 1)} \text{ for } i \neq j. \tag{5.6}$$

From Equation 5.6 we can see that samples from a Dirichlet distribution inherit correlations among themselves.

If α_i in Equations 5.4 - 5.6 is divided by a scaling factor k (positive constant), then the new moments become

$$\mathbb{E}_2(x_i) = \frac{\alpha_i}{\Omega} = \mathbb{E}(x_i). \tag{5.7}$$

$$\mathrm{Var}_2(x_i) = \frac{\alpha_i(\Omega - \alpha_i)}{\Omega^2(\frac{\Omega}{k} + 1)} \neq \mathrm{Var}(x_i), \tag{5.8}$$

$$\mathrm{Cov}_2(x_i, x_j) = \frac{-\alpha_i \alpha_j}{\Omega^2(\frac{\Omega}{k} + 1)} \neq \mathrm{Cov}(x_i, x_j) \text{ for } i \neq j. \tag{5.9}$$

From Equations 5.7 - 5.9 we can see that dividing each vector α_i by a positive constant k maintains the means but not the variances and covariances. This property of the Dirichlet distribution makes it possible to adjust the variance and covariance of the draws as required without changing the mean. The parameter k is a scale factor which controls the variances and covariances of the distribution. As the value of k increases (decreases), the variance and covariance also increase (decrease) proportional to k, ($k \neq 0$) while maintaining the mean. The problem can be reduced into two dimensions as clay fractions can be calculated as a residual:

$$\text{clay} = 100 - (\text{sand} + \text{silt}).$$

On the other hand, Equation 5.2 shows that a random draw from a Dirichlet distribution results in parameters sum up to unity, and each parameter is positive. This property makes the Dirichlet distribution a potential candidate to represent positive quantities with a bounded sum.

5.3 The Gaussian distribution

The Gaussian or normal distribution, a widely used distribution in many disciplines, for a single random variable, x can be written as

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-1}{2} \frac{(x-\mu)^2}{\sigma^2}, \tag{5.10}$$

where μ is the sample mean and σ is sample standard deviation.

For an n-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution is given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5.11}$$

where $\boldsymbol{\mu}$ is an n-dimensional mean vector, $\boldsymbol{\Sigma}$ is an $n \times n$ covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Gaussian distribution is defined over $\mathbb{R}$.

5.4 Ensemble initialisation

From 5.2 we have seen that to sample from a Dirichlet distribution, we only need to provide percentage sand, silt and clay, α_i in Equation 5.1. Variance and covariances will be automatically assigned based on the α_i. To control the variance and covariances, scale factor k is introduced, as in Equation 5.4 - Equation 5.9. Whereas to sample from a Gaussian distribution, we need to provide a covariance matrix in addition to the means. As the Gaussian distribution is not bounded, there is a chance of getting negative samples, which will not represent soil texture proportions. Hence, we add constraints in Equation 5.2 for the Gaussian distribution.

If a sample does not satisfy Equation 5.2, then it will be rejected, and re-sampling continues until the criteria are met. Ensemble initialisation techniques with the two distributions are compared by setting the first two moments the same. The covariance matrix for the Gaussian distribution is provided from the samples obtained from the Dirichlet distribution. The Gaussian distribution with a diagonal covariance matrix (called Gaussian-diag) is also considered to examine the advantages (or disadvantages) of having correlations. Hence, parameter samples are drawn from the three distributions, Dirichlet, Gaussian and diagonal Gaussian. The methodology

of ensemble initialisation from those distributions is detailed in 5.4.1.

5.4.1 Methods

The following steps are for drawing soil parameters samples from the three distributions, i.e. the Dirichlet, Gaussian and diagonal Gaussian distributions.

1. Sampling soil texture percentages from the Dirichlet distribution: Dirichlet
   (a) Draw sand, silt and clay samples (fractions sum up to unity) from a Dirichlet distribution by providing the mean sand, silt and clay percentage α_i together with the number of samples, N_e, need to be drawn and multiply the resulting samples by a hundred.
2. Sampling from a Gaussian distribution with correlations: Gaussian
   (a) Calculate the covariance matrix, Σ, between sand and silt for the samples drawn from the Dirichlet distribution.
   (b) Draw sand and silt percentage samples from a multinomial Gaussian distribution, Equation 5.11, by providing the covariance matrix Σ and the same mean sand and silt percentage used for the Dirichlet distribution. Samples are accepted if sand, silt $\in [0, 100]$, if not redraw continues. Then calculate the corresponding clay percentage samples as a residual, clay = 100– (sand+silt).
3. Sampling from a Gaussian distribution without correlation: Gaussian-diag
   (a) Set the off-diagonal elements of the covariance matrix Σ to zero, i.e the only non-zero entries of the new covariance matrix will be variances. Let the new covariance matrix be Σ_d .
   (b) Draw sand and silt samples as in step 2b but with Σ_d.

In subsection 5.4.2, numerical illustration of the methods is given.

5.4.2 Numerical illustration of the methods

To illustrate the methods discussed in subsection 5.4.1, soil parameter samples drawn from the three sampling methods are plotted in figures 5.2, 5.3 and 5.4.

Depending on the proportions of percentage sand, silt and clay, a given soil sample can fall in one of the many soil classes. Figure 5.1 shows possible soil texture percentage combinations and resulting soil classes. For example, the area shaded yellow represents a soil class classified as clay. For this class, the sand percentage should be between 0 and 45; silt percentage should be between 0 and 40 and the clay percentage between 40 and 100. At any point in the soil triangle, percentage sand, silt and clay add up to a hundred. For example, the black dot in the yellow shaded area represents a soil sample with sand, silt and clay percentage combinations of $30, 20, 50$. The three sides of a triangle represent sand, silt and clay. Percentage sand at a given point is read by following a line parallel to silt axis, percentage silt is read by following a line parallel to clay axis and percentage clay read by following a line parallel to the sand axis (red arrows).

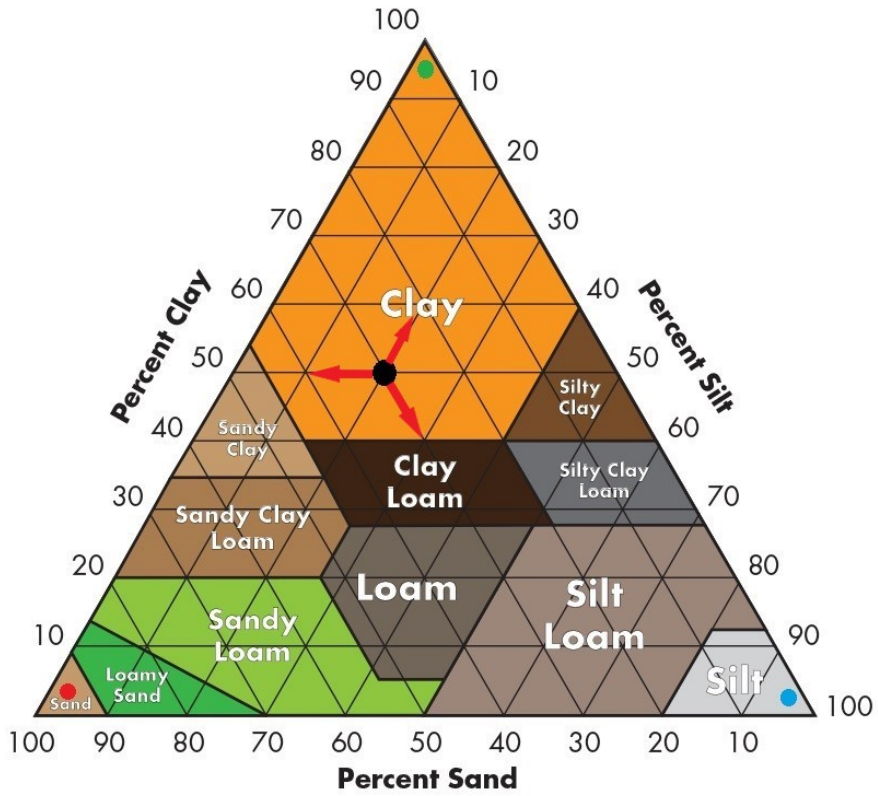


Figure 5.1: Soil class types and possible combinations of percentage sand, silt and clay for each soil class.

To address a range of scenarios, mean soil parameters are taken from a variety of soil classes as given in 5.1. From Figure 5.1, soil class clay (yellow), silt loam (light gray), sandy loam (light green) and loam (gray) are considered. The mean sand, silt and clay (α_i) are taken

from Oklahoma Mesonet stations for each soil class, Table 5.1. In Oklahoma Mesonet, for a particular station with known soil texture class, the mean sand, silt and clay percentages are calculated by taking the sample mean over Oklahoma Mesonet stations with the same soil class (Scott et al., 2013). For example, for Antlers station, the soil type is classified as sandy loam. Based on 58 sites with sandy loam soil class in Oklahoma Mesonet, the average sand percentage is 66.5 with a standard deviation of 9.5 and the average clay percentage is 12.8 with a standard deviation of 3.9. Note that the calculation of silt percentage is omitted in the paper as silt = 100% - (sand +clay). Table 5.1 contains a list of mean percentage parameters for the sites considered in these experiments. To encompass cases with smaller and larger variances from the original Dirichlet distribution, k = 0.5, 1. and 2. are considered respectively. As a result of the change of covariances and variances proportional to k, the corresponding Gaussian and Gaussian-diag distributions also considered with the three values of k.

Station	Soil class	Truth soil parameter		
		% Sand	% Silt	% Clay
Antlers	sand loam	66.5 (9.5)	20.7	12.8 (3.9)
Boise City	clay	17.4 (9.4)	29.8	52.8 (8.6)
Lane	silty loam	21.1 (7.6)	60.2	18.7 (4.6)
Mt Herman	loam	41.1 (6)	38.2	20.7 (4.2)

Table 5.1: Mean percentage sand, silt and clay (Truth parameters) for Mesonet sites (Scott et al., 2013) used in the numerical experiments.

By performing steps 1a - 3b with the scale factor k = 0.5, 1. and 2., Figure 5.2 - Figure 5.4 are obtained. The number of samples, N_e, is 200. In Figure 5.2 - Figure 5.4, "Truth" refers to the mean percentage parameter used to sample from distributions.

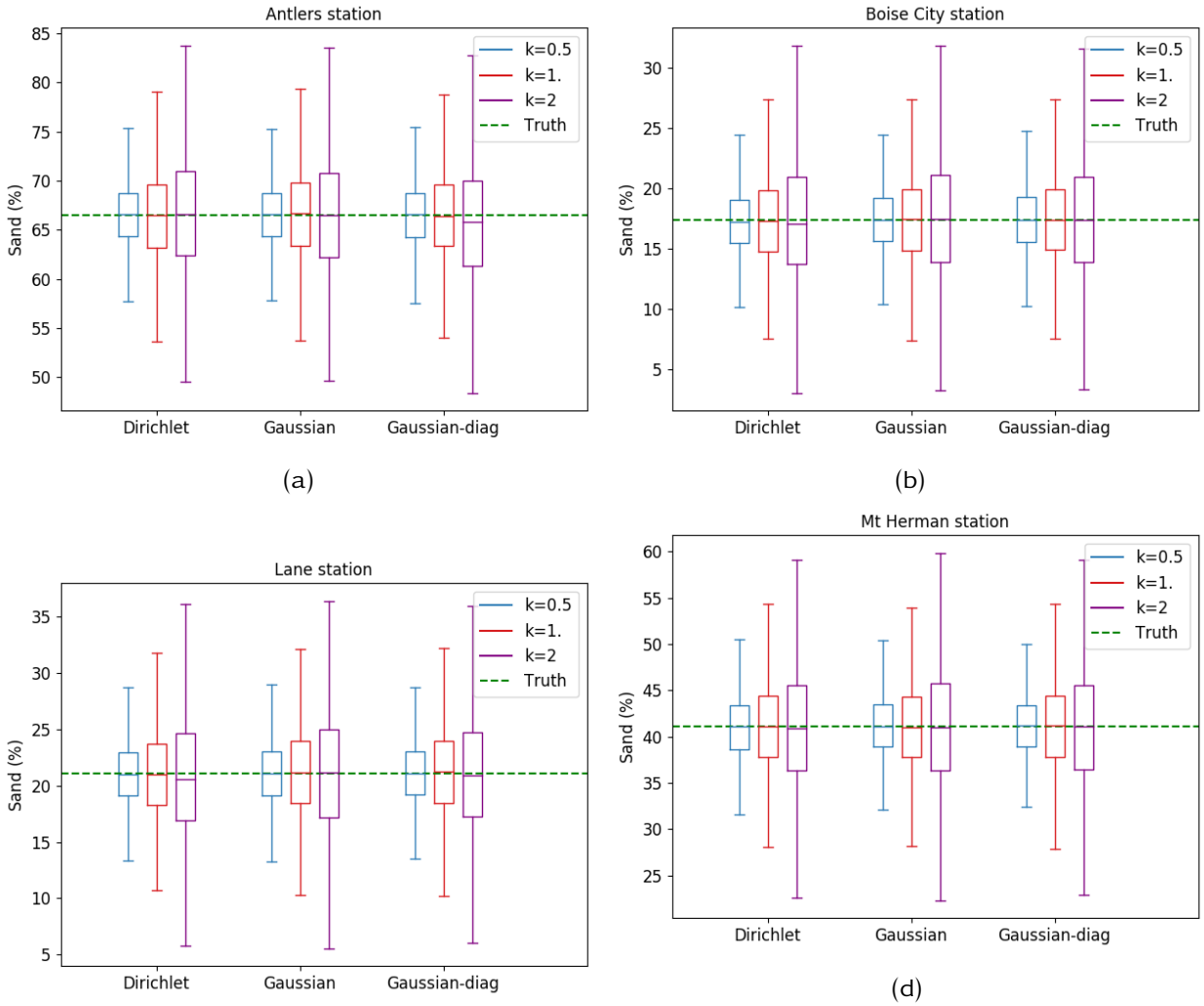


Figure 5.2: Sand percentage samples drawn from Dirichlet, Gaussian and diagonal Gaussian distributions for Mesonet sites Antlers, Boise City, Lane and Mt Herman.

Figure 5.2 - Figure 5.4 are box plots of sand, silt and clay parameter samples. The boxes are samples within the interquartile range (IQR - 50% of samples) where the solid horizontal lines (with respective colours) in the box representing the median of each sample. The minimum and maximum values are $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$ respectively where Q_1 and Q_3 are lower quantile and upper quantiles respectively, 49.3% of samples are found outside the IQR and between the whiskers. Samples outside the minimum and maximum values account for 0.7%.

Figure 5.2 shows boxplots of sand percentage samples for the four Mesonet sites, the three ensemble initialisation techniques and for three values of k. As k is a scaling factor for the standard deviation of the samples, the standard deviation increases as k increases from 0.5 to 2., Equation 5.8 and Equation 5.9. The green dashed lines are truth sand percentages for each site.

Since the four stations are from different soil classes, the truth sand percentage is different for each site. Soil class of each of the stations is given in Table 5.1. For Antlers station, the Gaussian-diag has shown a slight bias. On the other hand, samples from Gaussian-diag (especially for k) showed a slightly shorter IQR and also shorter whiskers, as a result of rejections of samples in the truncation. Even if truncation is considered for the Gaussian distribution, the probability of samples being outside the lower and upper bound is higher for the Gaussian-diag distribution. As a result, more re-sampling results in shorter IQR and shorter whiskers for the Gaussian-diag distribution.

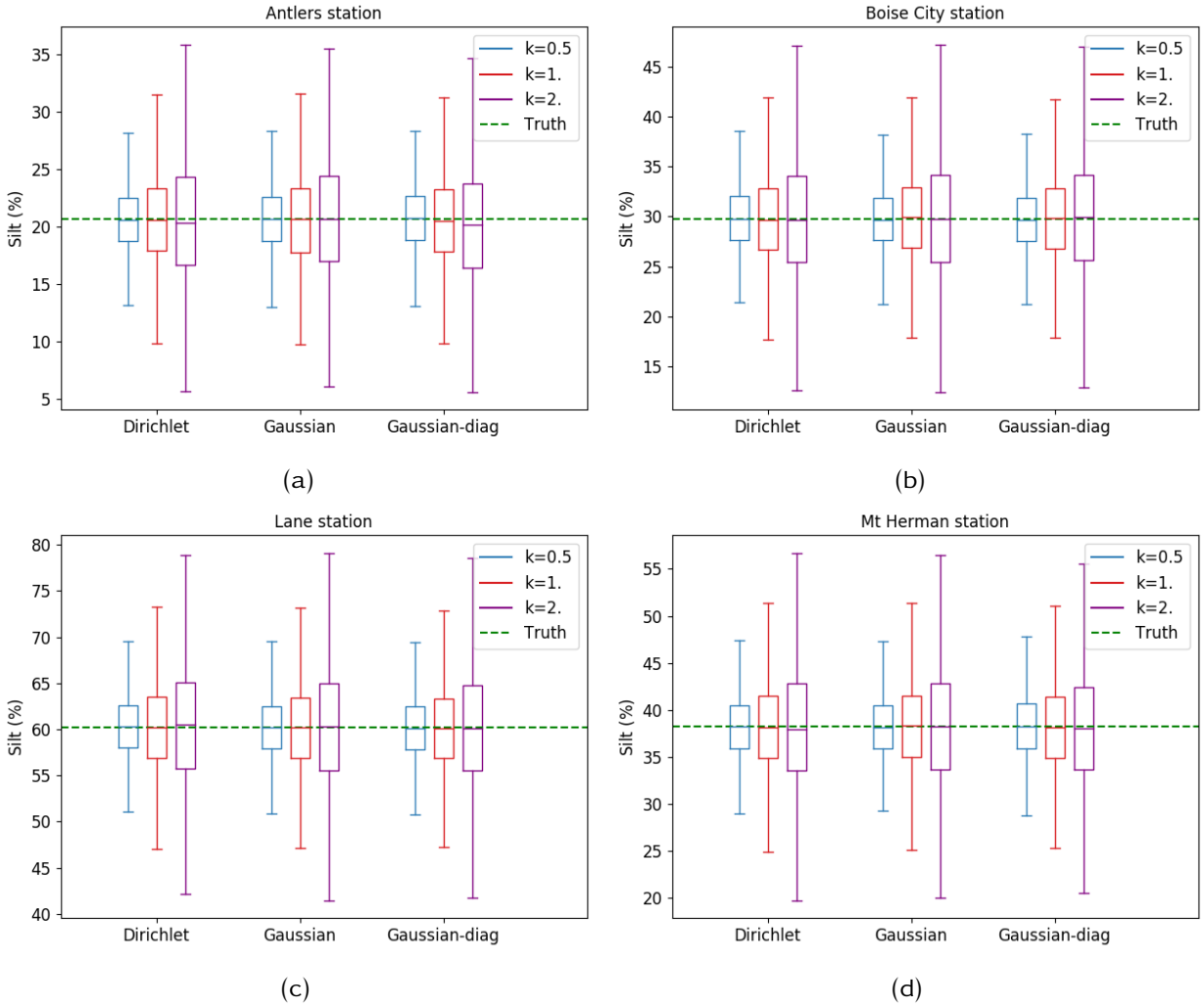


Figure 5.3: As in Figure 5.2, but for prior silt percentages.

Figure 5.3 is as in Figure 5.2 except it is for silt percentage samples. The effect of k is the same as in the case of sand percentages for all the methods and all the stations.

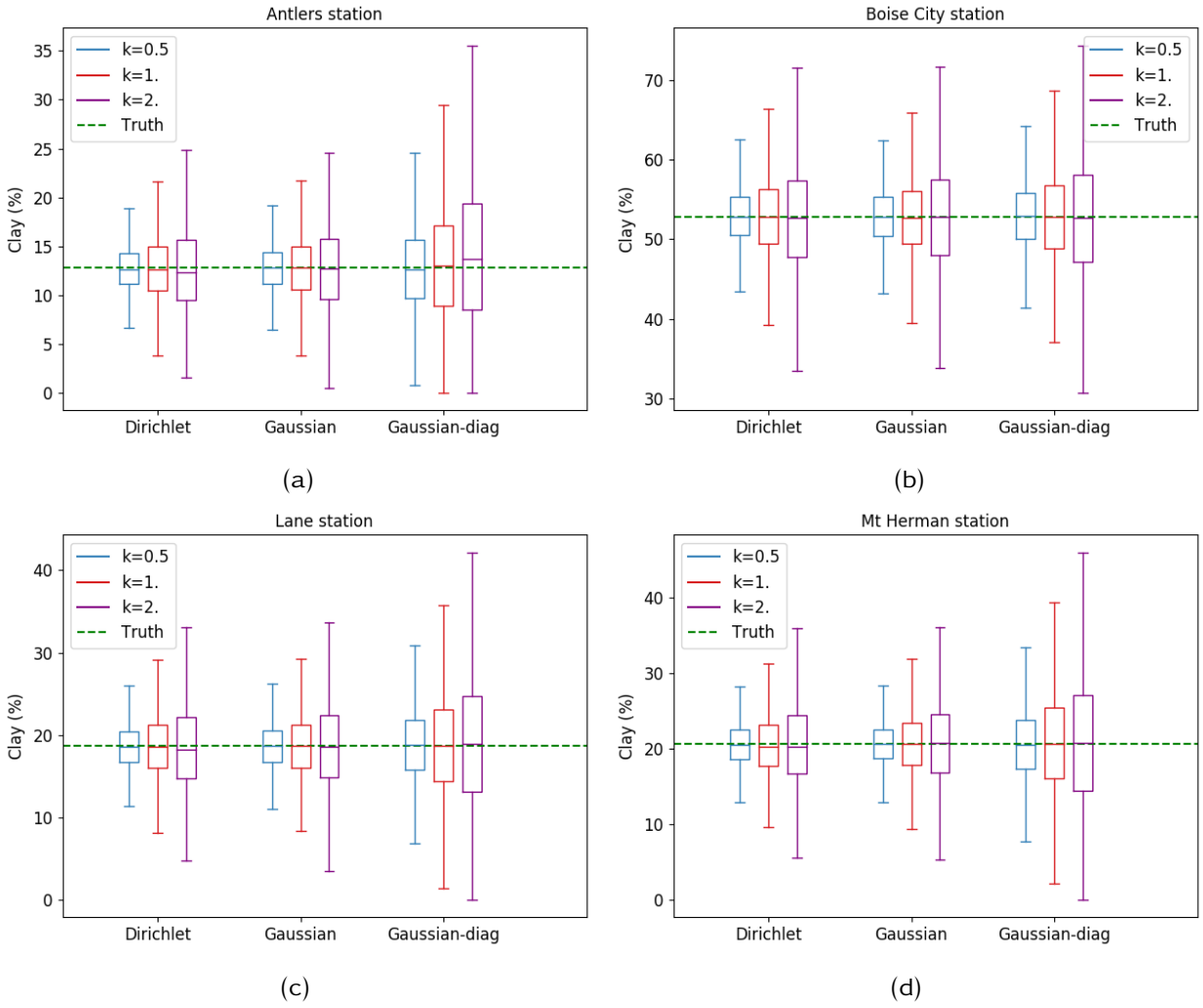


Figure 5.4: As in Figure 5.2, but for prior clay percentages.

Figure 5.4 is similar to figures Figure 5.2 but for clay percentage samples. The Dirichlet and Gaussian distributions have similar responses for varying k, but the Gaussian-diag has shown a different pattern for clay percentage samples than sand and silt. Note that clay percentage is calculated as a residual parameter for the Gaussian and Gaussian-diag distributions. The Gaussian-diag has shown a relatively shorter IQR and shorter whiskers for sand and silt percentages. Corresponding to sand and silt percentage, Figure 5.4 shows that the Gaussian-diag has longer IQR and longer whiskers as the sum of sand, silt, and clay is a hundred for each sample. In general, the three distributions resulted in similar soil texture samples except for sand and silt samples from the Gaussian-diag have smaller standard deviations, and clay percentage samples have larger standard deviations compared to the corresponding samples from the Dirichlet and Gaussian distributions.

So far, we have sampled from the three distributions where each true percentage of the soil

texture is far from their lower and upper bounds, 0 and 100. Another aspect of comparison of the methods is handling of extreme cases, where one or more of the parameters are closer to zero or a hundred. Section 5.4.2 illustrates the comparison with numerical examples where the soil class is dominated by only one of the soil texture percentages. In reality, there are chances where one of the soil texture parameters is dominant over the other two. For example, near the beaches, there is almost entirely sand. On the contrary, there are cases where all the three soil texture parameters are almost equally shared. Here we are investigating the performance of the above three distributions using the soil parameter values in Table 5.2 to demonstrate such cases.

	Soil class	Soil parameter		
		% Sand	% Silt	% Clay
	Sandy	94	3	3
	Silty	3	94	3
	Clay	3	3	94
	Clay loam	33.3	33.3	33.3

Table 5.2: Example proportions of sand, silt and clay percentages for sandy, silty, clay and clay loam soil classes.

Figures Figure 5.5 - Figure 5.7 are soil texture samples obtained using soil texture mean as in Table 5.2 and sampling from the three distributions. Results showed that there is more bias when the soil is dominated by one soil parameter. In such cases, samples have a higher probability of hitting the bounds and bias is introduced to avoid samples outside the bounds. The bias in all the three distributions. When the three soil textures are equal, the three distributions behave similarly except in case of the clay percentage, and the Gaussian-diag distribution has shown larger variance, as in the case of the previous section. The Dirichlet distributions look correctly skewed in each case (i.e. skewed negative when truth close to the upper bound, skewed positive when the truth is close to the lower bound and un-skewed when far from the bounds).

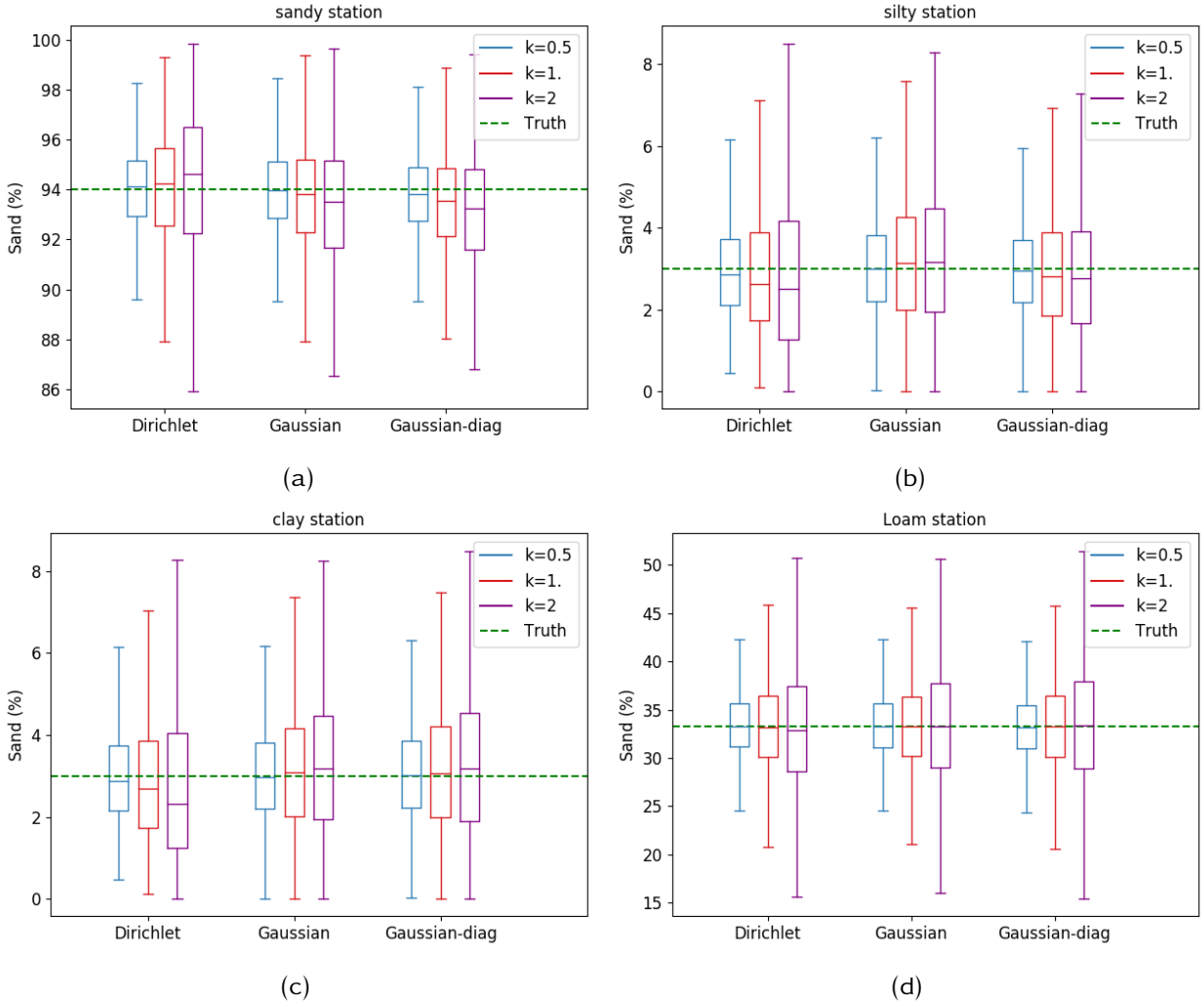


Figure 5.5: Sand percentage samples drawn from Dirichlet, Gaussian and diagonal Gaussian distributions for sandy, silty, clay and loam soil classes with extreme values.

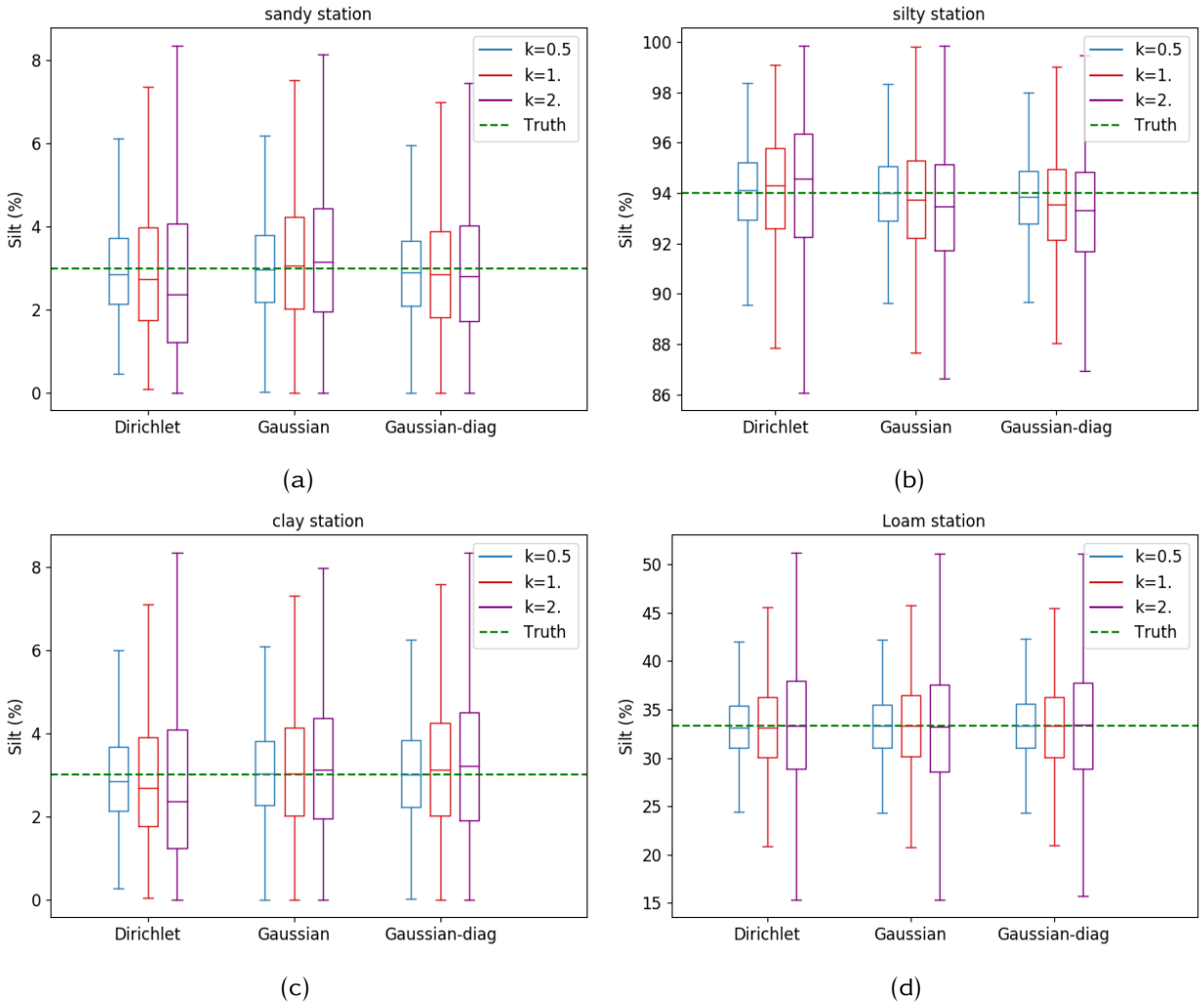


Figure 5.6: As in Figure 5.5 but for silt.

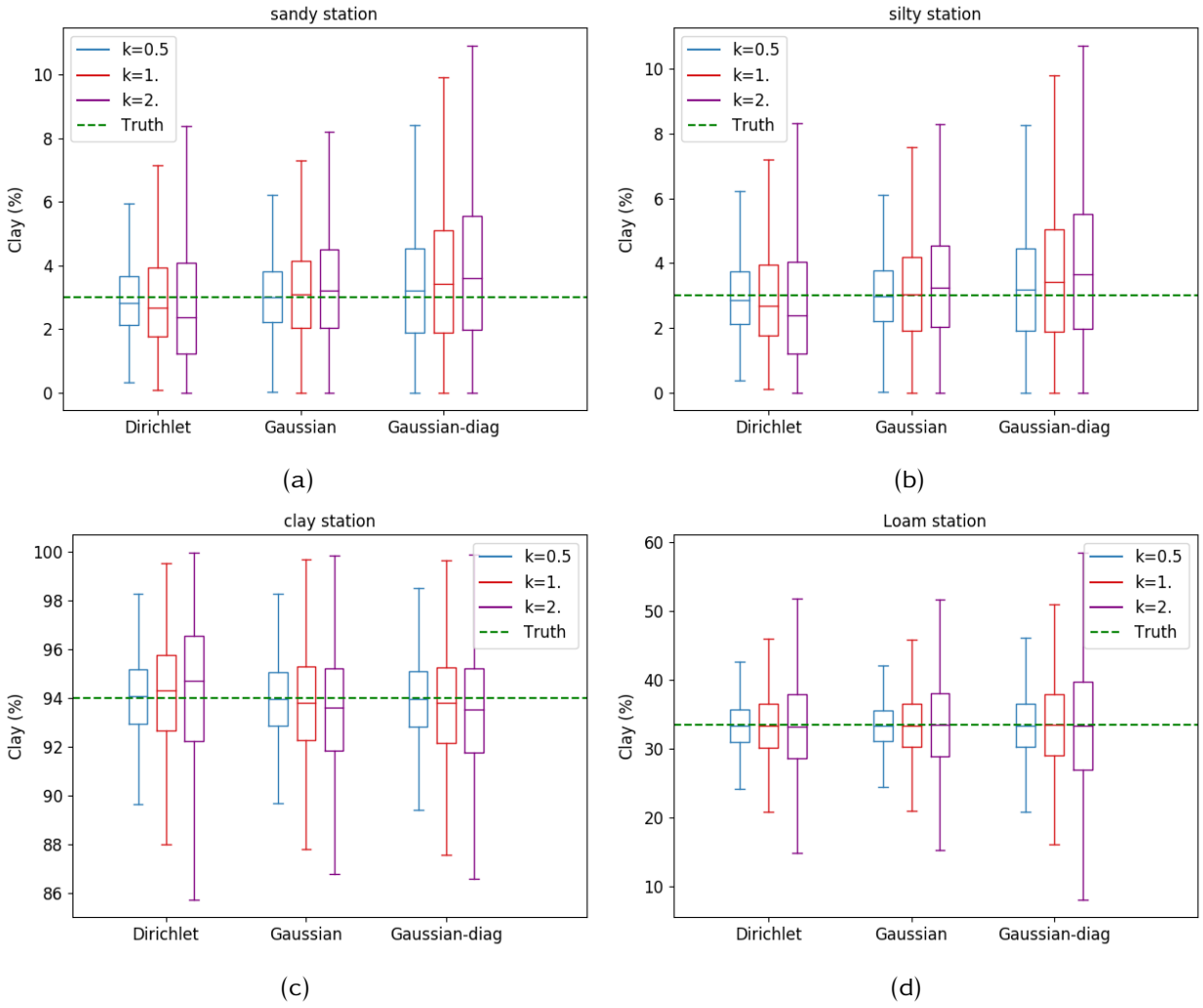


Figure 5.7: As in Figure 5.5 but for clay

In this section, we have shown that soil texture parameters can be sampled from the three distributions. Soil texture parameters determine the moisture content in the soil. In section 5.5 we have shown how soil moisture is sensitive to soil texture parameters using the JULES model to estimate soil moisture.

5.5 Sensitivity analysis of soil moisture for soil parameters

Here, the sensitivity of soil moisture estimates from the JULES model due to changes in soil texture is investigated. The soil texture percentage sand, silt and clay varies from the minimum value (zero) to a maximum value (100) and increasing by ten. Soil moisture ensembles from the JULES model were generated for a time length of one year, 2016. A column of soil with four layers is considered for the JULES model. Figure 5.8 is a ternary diagram where the colour map is for

top layer annual mean soil moisture estimate for Antlers station. The forcing data is 0.5-degree resolution from WATCH Forcing Data methodology applied to ERA-Interim (WFDEI), Weedon et al. (2014). Soil texture values inside a triangle are similar to discussed in the case of figure 5.1 but the silt and clay axis are exchanged and the arrows accordingly.

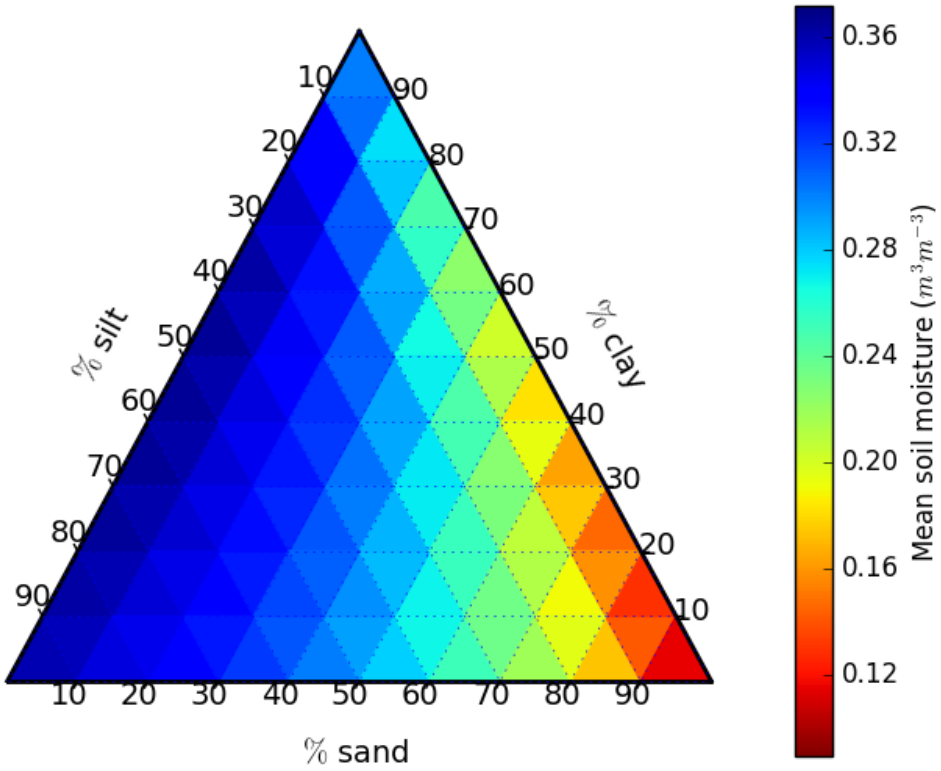


Figure 5.8: Annual mean soil moisture predicted by the JULES model for the year 2016 with varying soil texture parameter combinations with the forcing data from Antlers station, Oklahoma Mesonet.

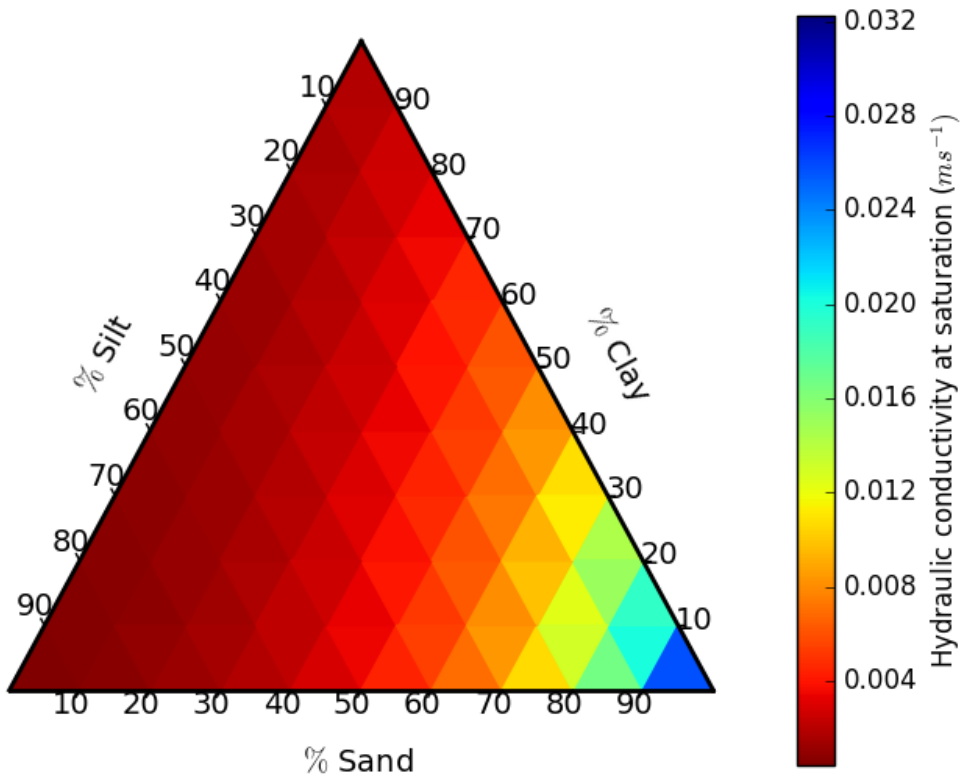


Figure 5.9: Hydraulic conductivity at saturation obtained from a pedo-transfer function for varying soil texture parameter combinations.

From Figure 5.8, we can see that the minimum soil moisture estimate is obtained for sandy soil. On the contrary, the maximum soil moisture estimate is attained with a small percentage of sand and a high percentage of silt and/or clay. What we have observed is intuitive in that water passes quicker in larger soil particles than fine soil particles and vice-versa. This phenomenon is supported by Figure 5.9; high water infiltration is when all soil is sand and infiltration decreases with sand percentage.

In the region where percentage silt > 20 and percentage sand < 40 (wet region), the soil moisture value is wet ($\geq 0.3 \mathrm{m}^3\mathrm{m}^{-3}$) irrespective of the corresponding clay percentage. In this region for a given percentage of clay, annual mean soil moisture estimate increases when the percentage silt increases and as the percentage sand decreases. For percentage sand < 10, annual mean soil moisture estimate reaches its maximum, for any combination of silt and clay. For the region where percentage silt < 20 and % sand > 40, the soil remains drier no matter how the percentage clay varies. When the silt percentage varies (any line parallel to the silt axis), the

variation in annual mean soil moisture is marginal for any combination of sand and clay. This shows that sand and clay percentage are the more important proportions and silt percentage is the least important proportion to get the soil moisture estimate right.

Another interesting fact is that different combinations of soil percentages can give similar annual mean soil moisture estimates (equifinality). As a result, conclusions from the comparison of ensemble generating methods for parameter space (soil texture percentages) does not imply for soil moisture estimates. i.e. one method being better in analysis soil texture proportions does not necessarily mean that it is better for the soil moisture estimates and vice versa. Similar experiments were conducted for all the five stations described in Figure 5.1 and similar sensitivity (of annual mean soil moisture as soil texture parameters vary) was observed except the magnitude of the annual mean soil moisture estimate for each particular station.

5.6 4DEnVar twin experiments with the JULES model for parameter estimation

In data assimilation (DA), a twin experiment is a technique to test whether assimilating a model-predicted set of observations can retrieve a chosen variable in the model using a DA method. In such a set-up, observations are perfectly represented by the model within a range of observation error. In reality, observations do not always obey the underlying physics of the model.

Here, twin experiments are conducted to estimate soil texture parameters, sand, silt and clay (%). These parameters have a direct effect on the amount of moisture in the soil layer (see section 5.5). Perturbed soil moisture predictions from the JULES model with Gaussian perturbation are assimilated into the JULES model using the 4DEnVar DA technique. The numerical implementation of the 4DEnVar for the JULES model was provided by Ewan Pinnington, described in Pinnington et al. (2020). In their work, model-predicted and observed data are assimilated with the JULES model and successfully estimated parameters which determine the harvestable material and selected variables in the JULES-Crop model. Details of the experimental set-up for this chapter is given in subsection 5.6.1.

5.6.1 Experimental design

Data assimilation experiments are performed to test the performance of ensemble initialisation techniques described in subsection 5.4.1 for soil moisture data assimilation. Soil moisture ensembles are generated by running the JULES model (version 4.8) using parameters sampled from the three ensemble initialisation techniques, the number of ensemble members is $N_e = 200$. These ensemble initialisation techniques can be used for any model which uses soil texture parameters. The meteorological data used to force the JULES model are from WFDEI data with 0.5-degree resolution, (Weedon et al., 2014).

A model truth was obtained by running the JULES model using the truth soil parameters given in Table 5.1. Synthetic observations were sampled from the model truth with 4% Gaussian noise, which matches the SMAP soil moisture observation error. Then observation error with a standard deviation of 0.04 is considered. The square root of the background error covariance matrix is calculated from the ensembles of parameters, see Equation 2.9 but used for parameters instead of state space. The observation frequency is every five days with assimilation window of one year, for 2016. After assimilating synthetic observations, diagnostic tools explained in subsection 5.6.2 are used to evaluate the performance of each data assimilation experiment.

5.6.2 Diagnostic tools for experimental results

To evaluate the performance of assimilating soil moisture from the top layer, posterior in parameter space (analysis sand, silt and clay percentages) and posterior in state space (analysis soil moisture) are compared with truth parameters and truth state space respectively.

For the parameters, boxplots of prior and posterior values are presented, (Figure 5.11, Figure 5.12 and Figure 5.13). For a given boxplot, the boxplot being longer shows that samples have larger variance and less precise in locating the variable in the target. On the other hand, a shorter boxplot refers to samples being precise and with a smaller uncertainty. If a boxplot for posterior is shorter than the boxplot of the corresponding prior boxplot for the same method, then the analysis is better than the prior in predicting the truth, more accurate analysis than the prior. Hence the analysis is skilful as a result of top layer soil moisture data assimilation. Otherwise, the posterior parameter is considered as not skilful in parameter space.

To investigate the performance of soil moisture data assimilation from a top layer in state

space, RMSE is calculated for each data assimilation experiment for prior and posterior soil moisture compared to the truth soil moisture. The analysis RMSE in the state space for each ensemble is calculated as

$$\text{RMSE}_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t^{a,i} - x_t)^2}, \tag{5.12}$$

where x is truth volumetric soil moisture, x^a is posterior volumetric soil moisture, i represents ensemble members, N is the total length of analysis and truth, 365 in this case as we consider daily time step. The prior RMSE in state space is calculated as in Equation 5.12 but for prior volumetric soil moisture x^b instead of x^a.

Prior and posterior RMSE metrics for state-space are calculated for the top layer where soil moisture observations are assimilated, Figure 5.14. For the second layer, the corresponding RMSE is calculated to investigate the effect of assimilating top layer soil moisture for root-zone soil moisture estimates, Figure 5.15. For both soil layers, if analysis RMSE is less than the corresponding prior RMSE, then the posterior soil moisture is considered to be skilful. However, as a twin experiment, this shows that the data assimilation set-up is working. In a twin experiment, one would expect that assimilating soil moisture estimates to reduce the RMSE in soil moisture prediction. Among the three ensemble initialisation methods, the one with the smallest posterior RMSE will be considered as the best. Based on these metrics for parameter and state space, performances of each data assimilation experiment is discussed in subsection 5.6.4.

5.6.3 Effect of ensemble size

The presumption is that the more samples included, the better to represent the distribution under consideration. However, the accuracy of representing the distribution is not proportional to the number of samples, and the accuracy will saturate at some point in time. Besides, there are factors which restrict the number of samples which can be considered. Computational time and storage are among the challenges. Having a smaller sample size results in sampling error where the distribution is misrepresented. Hence, there should be a reasonable compromise to include a smaller number of samples with minimal sampling error. Here, we investigated the minimum number of samples needed to be included to represent the PDF; the higher sample sizes the better the PDF is represented. In the ideal scenario is, the RMSE being zero is a target. However, in this

case, the benchmark is for RMSE to be less than the observation error, 0.04.

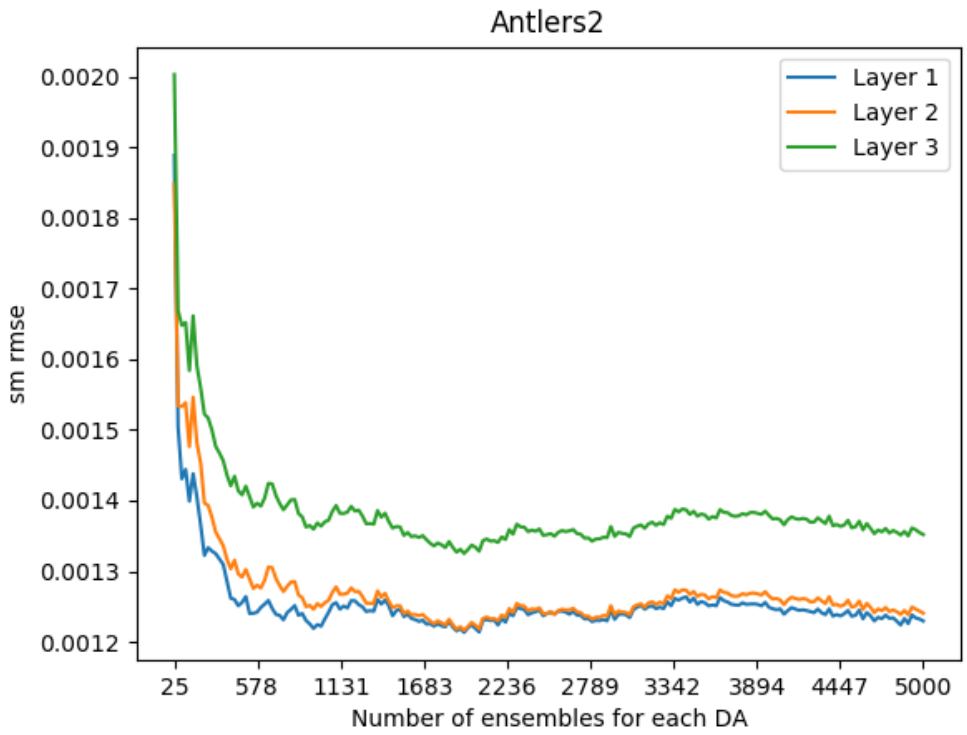


Figure 5.10: Posterior soil moisture RMSE for the three soil layers, with different ensemble size for data assimilation experiments. The forcing data is from Antlers station and ensemble initialisation technique is using the Dirichlet distribution with scale factor k = 1.5.

Figure 5.10 shows RMSE for posterior soil moisture for varying ensemble size. In total, 200 data assimilation, experiments are performed, and corresponding analysis soil moisture is obtained. The ensemble size for each experiment varies from 25 to 5000 with an increment of 25. Then the RMSE is calculated for each soil moisture analysis using Equation 5.12.

The magnitude of RMSE is less than 0.04, observation error, even with the smallest ensemble size, N_e = 25. In this case, one can use 25 ensemble members if computing time and storage are unbeatable issues, especially for global models which could take too long to run a single ensemble member. The ECMWF ensemble forecasting system uses 50 ensemble members . In our case, the JULES model takes minutes to integrate for a single ensemble member as it is only for a single grid. Besides, the parallel computing facilities at the University of Reading makes it possible to consider a larger ensemble size. Therefore, 200 ensemble numbers will be used for the data experiments throughout the chapter.

5.6.4 Results and discussions

As discussed in section 5.5, soil texture parameters are an important part of soil moisture estimation in land surface models, JULES in this case. Getting the right set of parameters is necessary for soil moisture prediction. Hence, here soil texture parameters are updated in the data assimilation, and soil moisture analysis is obtained by running the JULES model using the updated parameters. As part of assessing the success of data assimilation experiments, prior and posterior soil texture parameters are compared. Besides, posterior soil texture parameters obtained from the three ensemble initialisation techniques are compared.

Figure 5.11 - Figure 5.13 are boxplots displaying prior and posterior soil texture parameters from Dirichlet, Gaussian and Gaussian-diag ensemble initialisation techniques discussed in subsection 5.4.1. The green dotted line is the population mean; we called it truth instead of mean so that it is not confused with the sample mean. For all the sites, a single data assimilation experiment is conducted, and analysis ensemble members are updated by using the background perturbation and analysis error covariance, as implemented in Pinnington et al. (2020).

Figure 5.11 shows that as a result of assimilating top layer soil moisture data, percentage sand values are closer to the actual value of the corresponding prior parameter. The length of the IQR for prior is much larger compared to the posterior in all the cases. This is observed for the four stations with different soil class and all the three data assimilation experiments corresponding to the three ensemble initialisation techniques. We can see that the posterior parameters are more precise than the prior, which makes the posterior sand percentage skilful.

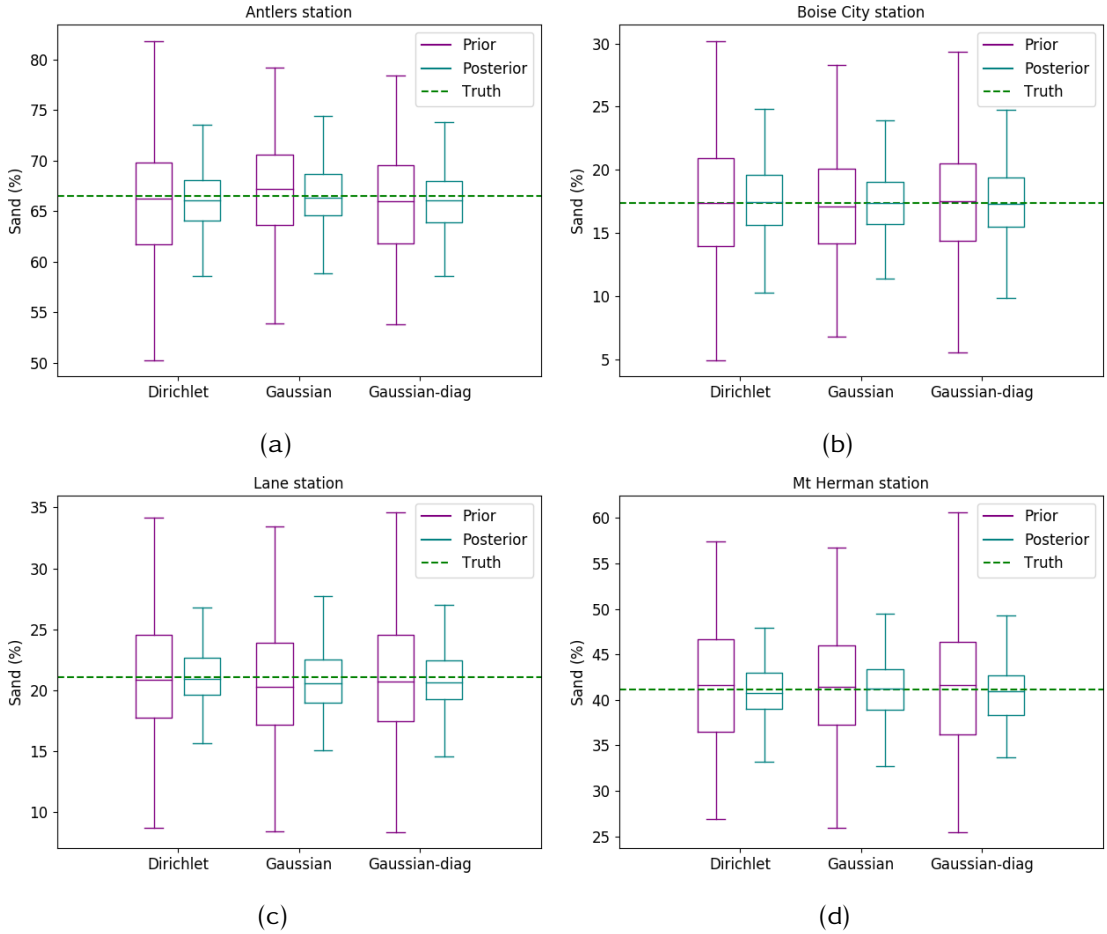


Figure 5.11: Prior and posterior percentage sand with 200 ensemble members. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques with $k = 2$. used for sampling prior parameters.

Figure 5.12 is similar to Figure 5.11 but for silt. Compared to posterior sand percentages, posterior silt percentages are less skilful. This is because soil moisture prediction in the JULES model is less sensitive to the silt percentage compared to sand and clay percentage (section 5.5). Predicted soil moisture values do not change that much as silt percentage varies. The objective of the assimilation is targeting to reduce the difference between observed and predicted soil moisture values. From the ternary diagram (5.8), we have seen that a broader range of silt percentage parameter resulted in similar soil moisture values. Hence, if the soil moisture estimate is attained from a particular combination of parameters, there is no need for the DA system to go closer to the truth parameter value.

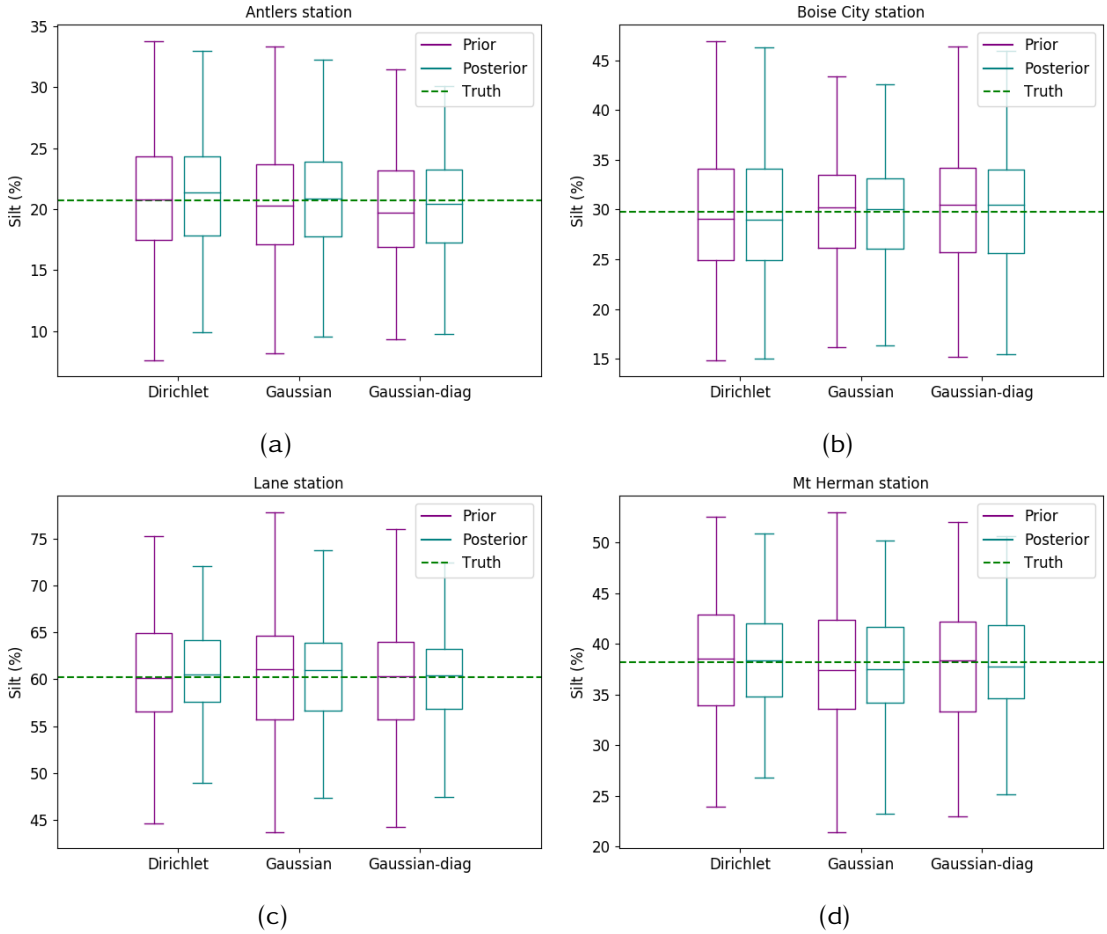


Figure 5.12: As in Figure 5.11 but for percentage silt.

Figure 5.13 is as in Figure 5.11 but for a percentage of clay. For all stations and all ensemble initialisation techniques, posterior clay percentage has a smaller variance than the corresponding prior. This result shows that assimilating top layer soil moisture has helped to recover truth clay parameters. The result is consistent with the discussion in section 5.5 that soil moisture predicted with the JULES model is highly sensitive to changes in clay percentage.

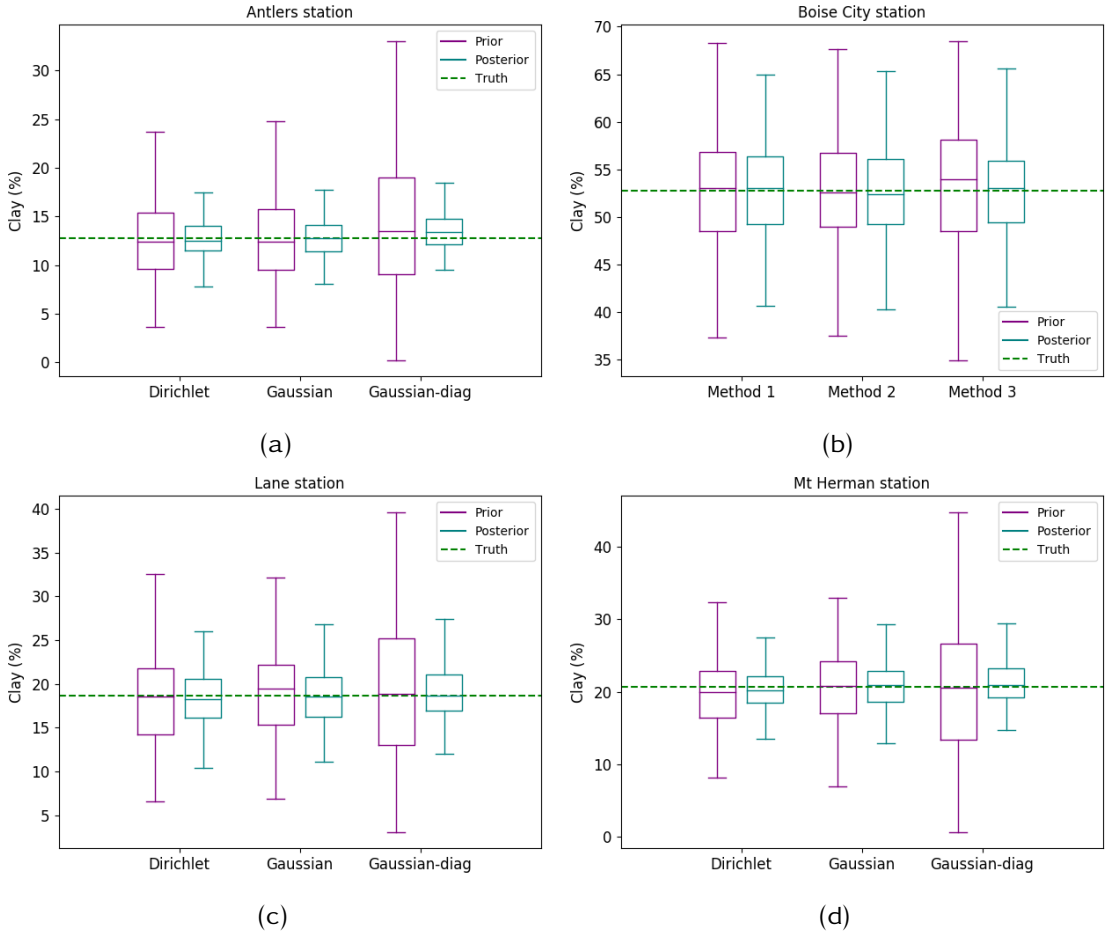


Figure 5.13: As in Figure 5.11 but for percentage clay.

In most cases, we have seen that posterior parameters are more skilful than prior parameters. The exception in silt percentages is that soil moisture is less sensitive to changes in silt percentage. As we have seen in section 5.5, soil texture parameters have a direct impact on the soil moisture estimate. Having skilful posterior parameters is presumed to obtain skilful posterior soil moisture estimate.

Comparing the performances amongst the three distributions, the Dirichlet and Gaussian distributions have shown similar accuracies, both for prior and posterior parameters. The similarity is consistent for all the soil texture parameters and all the sites. The Gaussian-diag distribution has also resulted in similar accuracy for posterior parameters. The prior sand and silt samples from the Gaussian-diag distribution have smaller variance, and clay parameter samples have larger variance compared to sand, silt and clay samples from the Dirichlet and Gaussian distributions. This difference is because of rejections and re-sampling of sand and silt parameters to avoid negative samples, and clay is calculated as a residual to complement the sum to a

hundred.

The hypothesis was the Dirichlet distribution to outperform the other distributions as a result of the physical properties shared with the soil texture parameters. However, it turns out that all the three distributions can be used interchangeably to sample soil texture parameters, by imposing positivity, boundedness and sum to a hundred to the Gaussian and Gaussian-diag distributions and with more input, covariance matrix, for the Gaussian distributions. The differences in the prior parameters did not have much effect on the data assimilation, and the posterior parameters from the three distributions are similar.

It is not surprising that the Dirichlet and Gaussian distributions performing similarly as both are using the same first and second moments during sampling. However, it is surprising to see that Gauss-diag is also performing similarly in the data assimilation.

The posterior soil moisture corresponding to the posterior soil texture parameters is obtained by running the JULES model with the posterior soil texture parameters. To evaluate the performance of different data assimilation experiments for soil moisture, analysis soil moisture RMSE is plotted for all the four stations, Figure 5.14. In addition to the top layer where observations are assimilated, second layer soil moisture estimates are also impacted as a result of data assimilated in the top layer. To examine the impact, RMSE is calculated and plotted for the second layer as well, Figure 5.15.

Figure 5.14 shows that posterior soil moisture RMSE is less than the prior soil moisture RMSE in all the three methods and all sites. This shows that assimilating top layer soil moisture to estimate soil parameters is a useful technique to constrain the JULES model for soil moisture prediction. This result is consistent with the results in parameter space: posterior soil parameters obtained by assimilating top layer soil moisture data are more accurate and precise compared to the prior soil parameters.

The impact of assimilating top layer soil moisture is not restricted to the top layer soil moisture estimate; it also greatly affected the second layer. Figure 5.15 shows that, as a result of assimilating top layer soil moisture, the RMSE of posterior soil moisture for the second layer is much smaller than the corresponding RMSE of prior soil moisture estimates. The characteristics for each site and ensemble initialisation technique are consistent with what is observed in the top layer. This is as expected since soil parameters are assumed to be constant across a soil column at different depth.

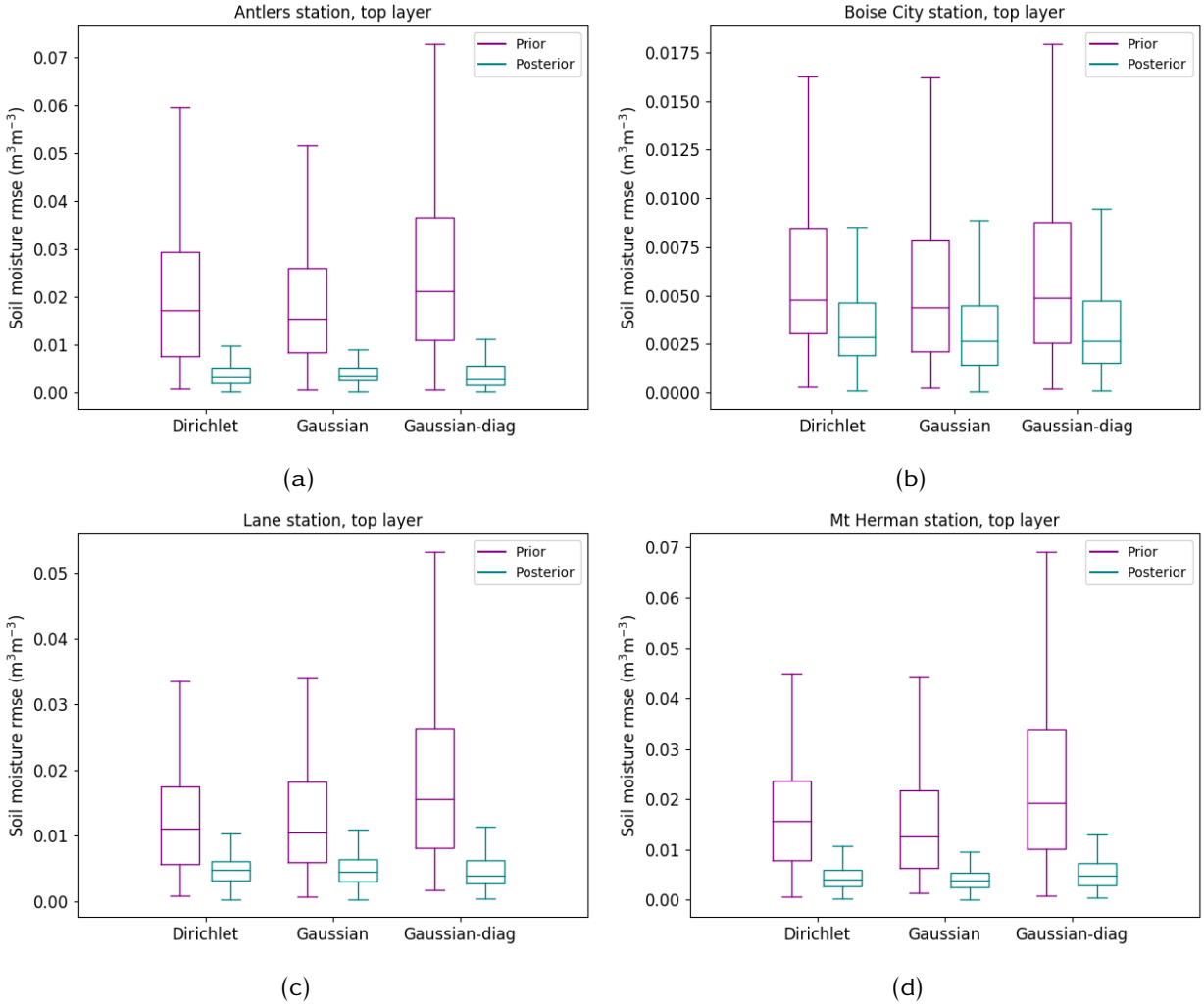


Figure 5.14: Top soil layer prior and posterior soil moisture RMSE for 200 ensemble members. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques which resulted in prior soil moisture ensemble members and corresponding posterior soil moisture ensemble members after data assimilation.

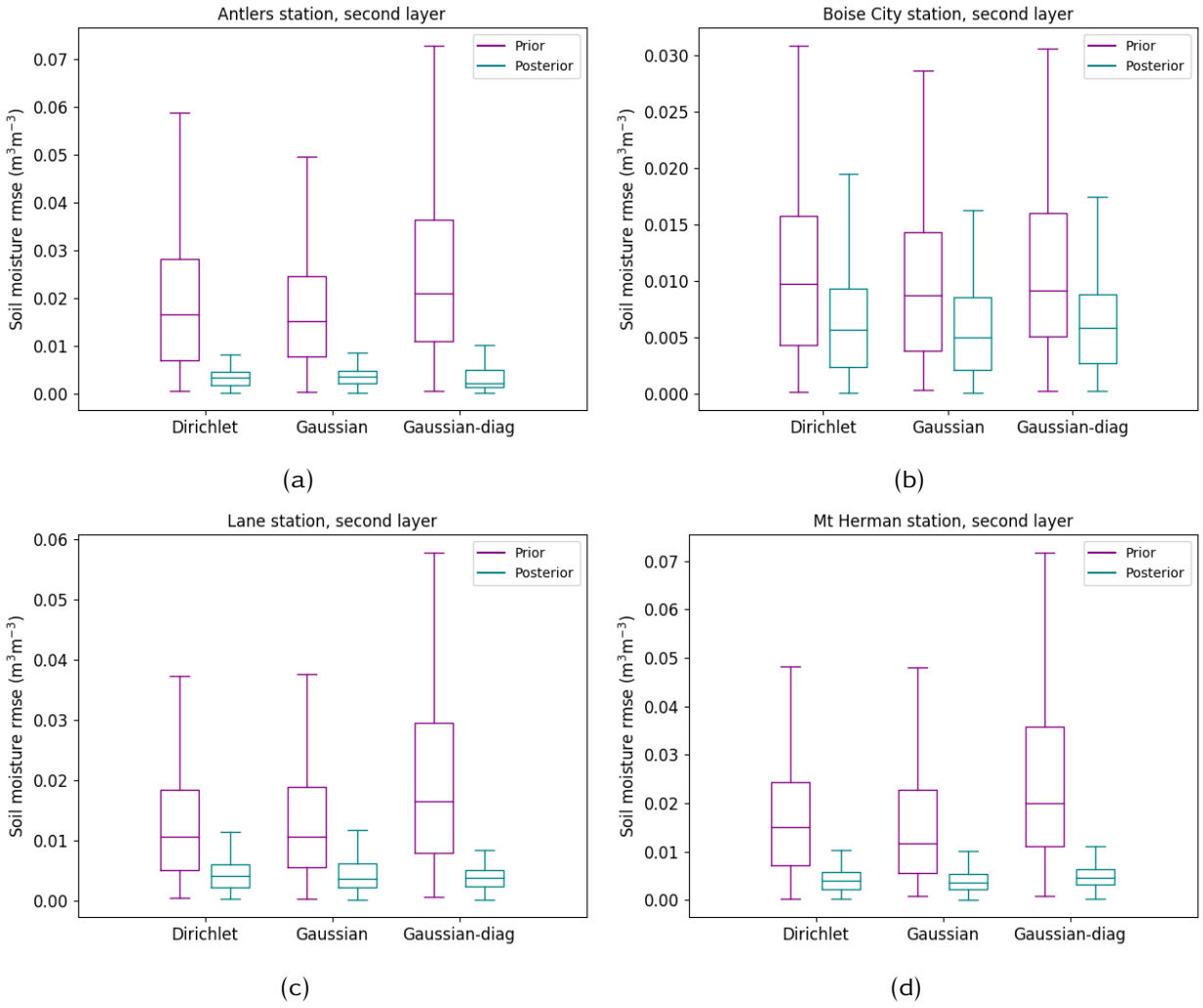


Figure 5.15: Second layer prior and posterior soil moisture RMSE for 200 ensemble members. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques which resulted in prior soil moisture ensemble members and corresponding data assimilation experiments which resulted in the posterior soil moisture ensemble members.

Figure 5.14b shows that RMSE for Boise City is different from RMSE for other sites in two aspects: the RMSE both for prior and posterior soil moisture prediction is smaller, and the RMSE reduction of posterior soil moisture from the RMSE of prior soil moisture for Boise City is smaller compared to any other site considered in this experiment. This is because for this station, the uncertainty in the JULES model is smaller compared to other sites and the influence (contribution) from observations is smaller - the system is overconfident on the model.

Comparing the three ensemble initialisation methods, prior soil moisture RMSE from the Gaussian-diag is larger than the corresponding RMSE from Dirichlet and Gaussian distributions. However, for the posterior soil moisture RMSE, the difference is negligible and all data assimilation results using parameters from the three distributions showed a great performance in re-

ducing the posterior soil moisture RMSE. As we have seen for parameter space, the results for state-space also shows that the performance of the three ensemble initialisation techniques is similar, and any of the distributions can be used for soil moisture data assimilation.

5.7 4DEnVar twin experiments with the JULES model for parameter estimation with a wrong background

In section 5.6 twin experiments have shown that a model predicted soil moisture assimilation could retrieve truth soil texture parameters where the background ensemble members encompass the truth: background ensemble members were sampled around the truth parameters. Here, we are investigating whether the truth parameters will be retrieved if the background soil texture parameter ensemble members are far from the truth soil parameter. As such, the truth soil parameters are set to be 33.3% for sand, silt and clay while the background ensemble members are as in section 5.6. Then the experiment is repeated with the remaining experimental set-ups similar to section 5.6.

5.7.1 Results and discussions

Here, the skill of the posterior parameters is investigated by how close it is from the truth parameter and by its uncertainty reduction. The investigation is similar to the case where a known background, but here the position of the truth is different. We have considered the case when the truth is outside the background ensemble members. i.e. observations are far from the background as we are considering synthetic observations by perturbing the truth. Such a case happens where there is no enough information about the background. Figure 5.16 - Figure 5.18 are comparisons of prior and posterior soil texture parameters from the three methods of ensemble initialisation for this set up.

Figure 5.16 shows prior and posterior sand percentage for three ensemble initialisation techniques and for the four sites. In all the cases, the posterior sand percentage has moved towards the truth sand percentage and with smaller uncertainty. The extent of being close to the truth varies across the sites due to the differences between the prior parameter and truth for each site and also due to the differences in the meteorological forcing for each site. Similar to the case we have seen in section 5.6, the Dirichlet and Gaussian distributions behave very similarly,

and truth parameters are estimated in a right direction though could not exact the truth in some cases, Figures 5.16a and 5.16c. The Gaussian-diag has shown slightly worse performance than the Dirichlet and Gaussian distributions.

Figure 5.17 is as in Figure 5.16 but for silt. The main difference in Figure 5.17 is that in Figure 5.17a, the posterior silt percentage for Gaussian-diag is worse than the prior silt percentage (and also worse than the Dirichlet and Gaussian distributions) in predicting the truth silt. In section 5.5 we have seen that soil moisture is less sensitive to silt percentage, hence having worse silt posterior is not necessarily a bad thing for soil moisture prediction.

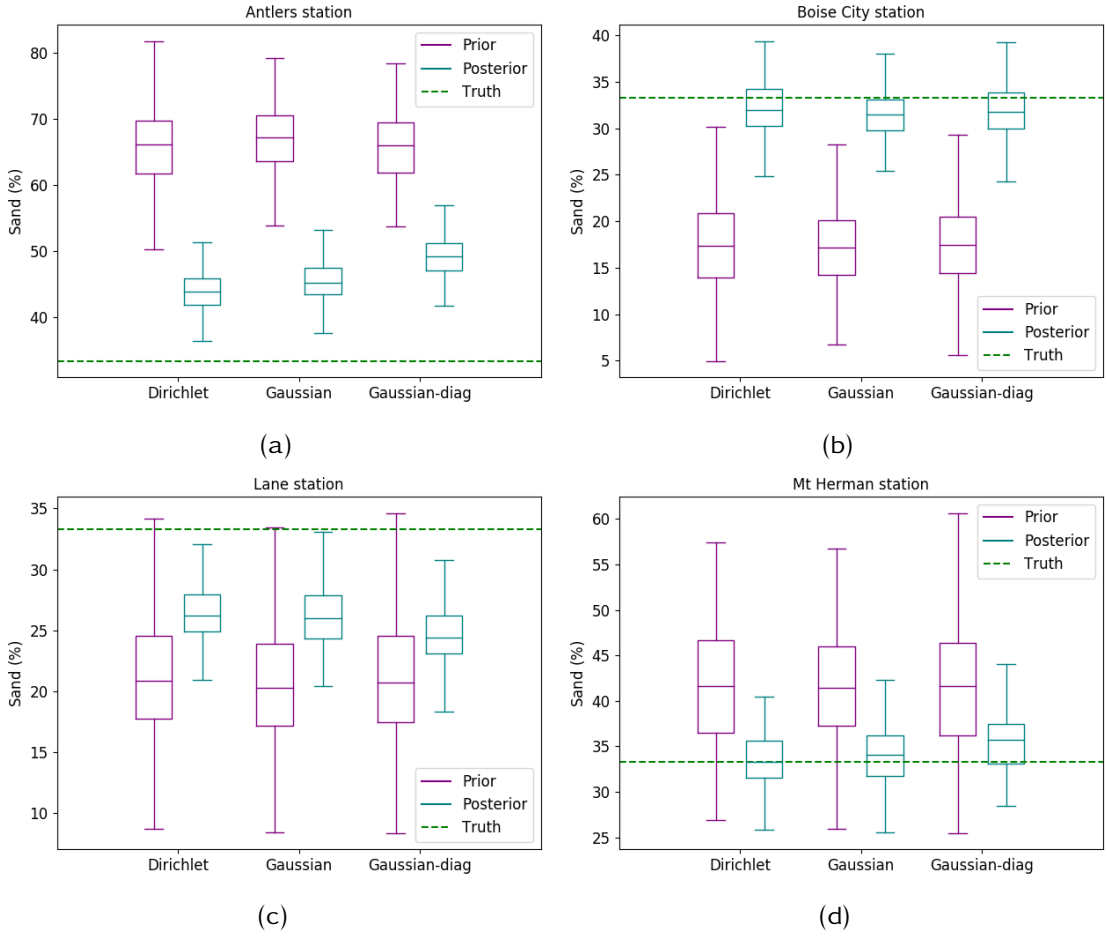


Figure 5.16: Prior and posterior percentage sand with 200 ensemble members. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques with $k = 2$. used for sampling prior parameters. Note that the truth is $(33, 33.3, 33.3)$ for percentage sand, silt and clay.

Figure 5.18 shows the prior and posterior clay percentage where the truth is outside the distribution of the background ensemble members. In most cases, the posterior clay percentage is close to the truth clay percentage than the prior clay. In Figure 5.18a, the Gaussian-diag posterior has passed the truth and goes further. As the clay is calculated as a residual, it is complementing

the other two parameters we have seen above.

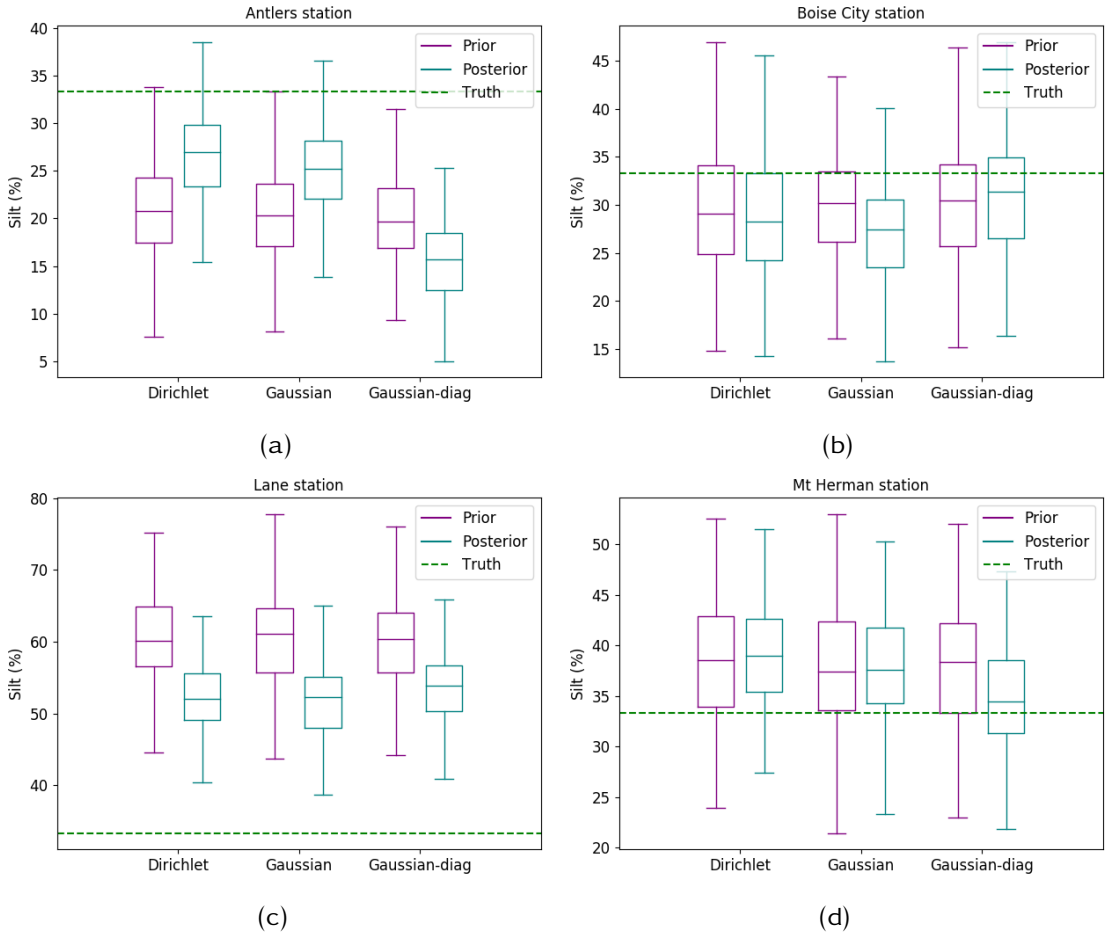


Figure 5.17: As in Figure 5.16 but for silt.

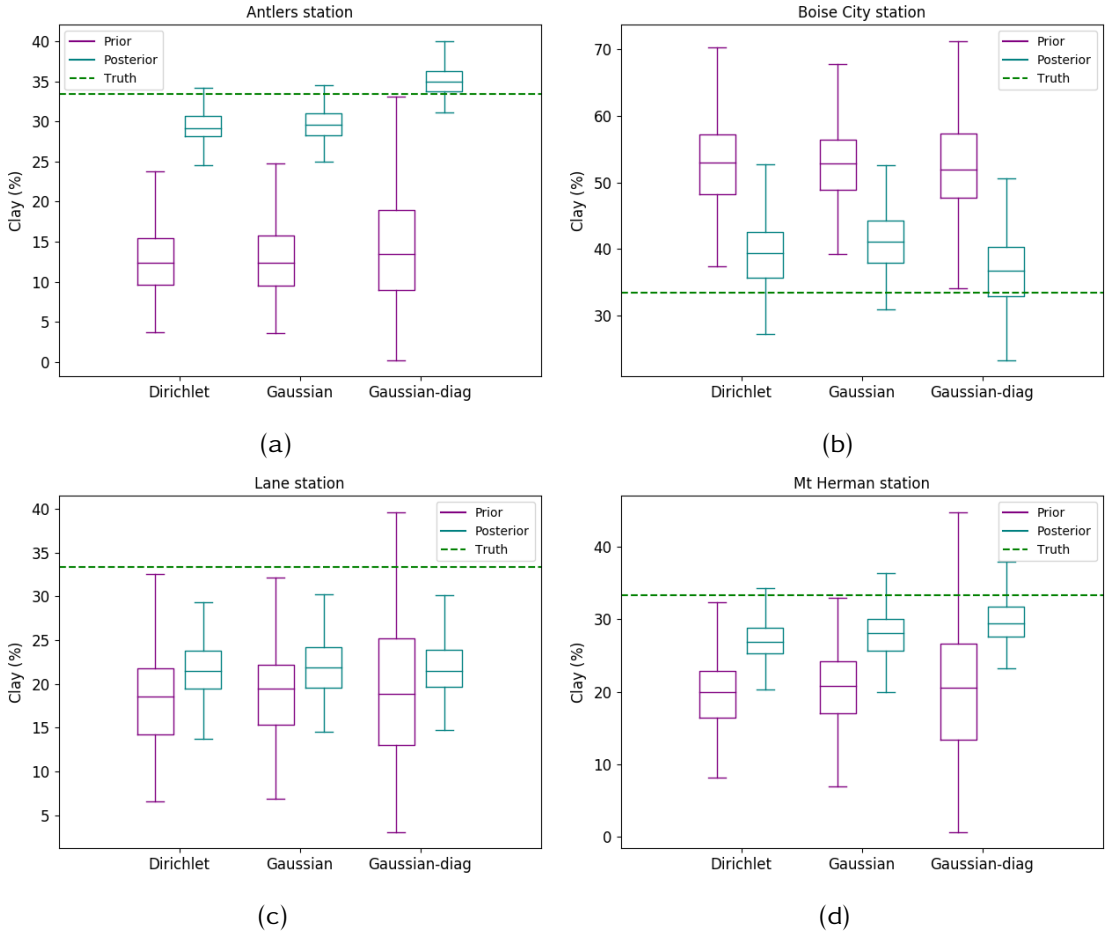


Figure 5.18: As in Figure 5.16 but for clay.

In general, from Figures 5.16 - 5.18 we have seen that even if the background parameters are specified wrongly, far from the truth, the data assimilation is able to push the parameter predictions towards the truth parameters. An exception has been observed for silt percentages resulted from the Gaussian-diag distribution where the prior is better than the posterior parameters. Figure 5.19 and Figure 5.20 are corresponding plots for soil moisture RMSE using the prior and posterior soil parameters given in Figures 5.16 - 5.18.

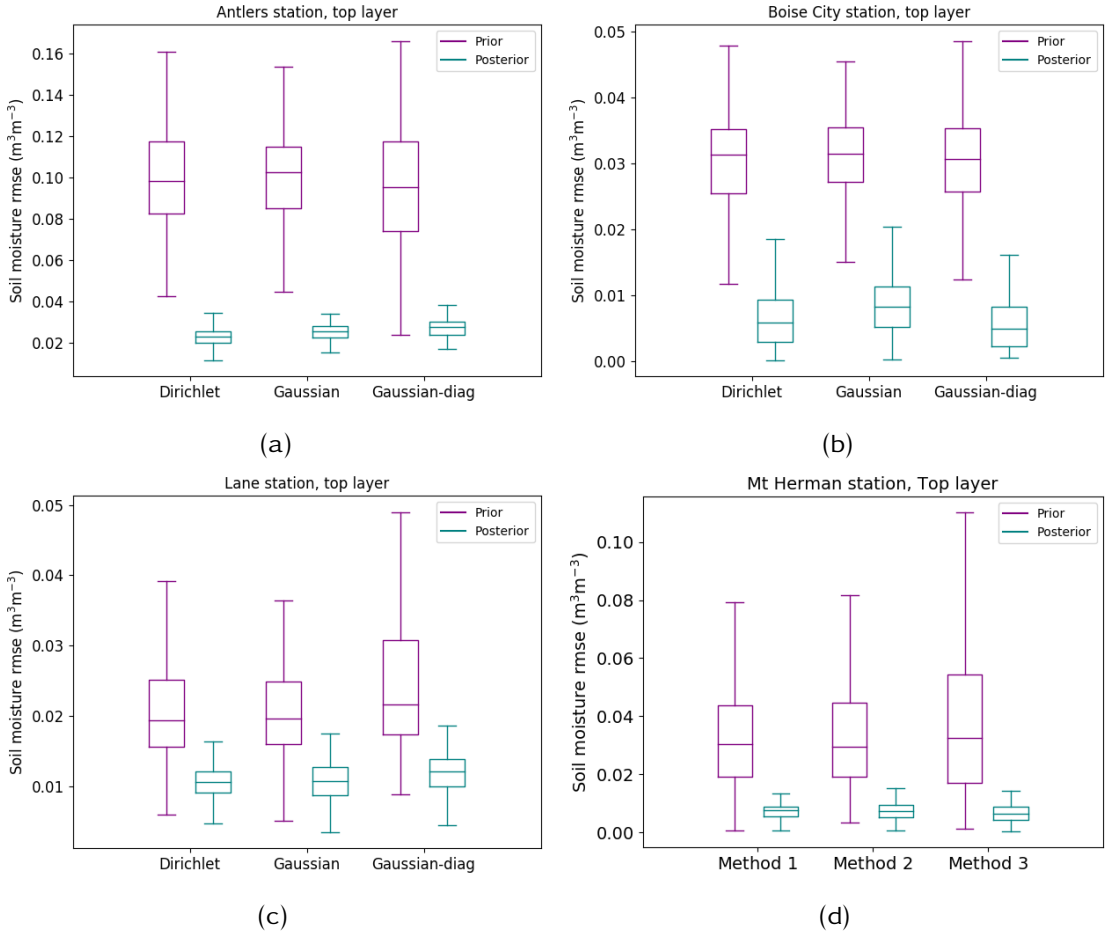


Figure 5.19: Prior and posterior soil moisture RMSE for 200 ensemble members, for the top layer. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques which resulted in prior soil moisture ensemble members and corresponding data assimilation experiments which resulted in the posterior soil moisture ensemble members.

Figure 5.19 and Figure 5.20 show that posterior soil moisture after assimilating observations have a better skill even if the background ensemble members were far from the observations. The general pattern we have seen from the above experiments is that the performance of the Dirichlet and Gaussian is very similar whereas the performance of Gaussian-diag distribution is slightly better in some cases and worse in other cases. The comparison was made for informed and uninformed background ensembles and different meteorological forcing.

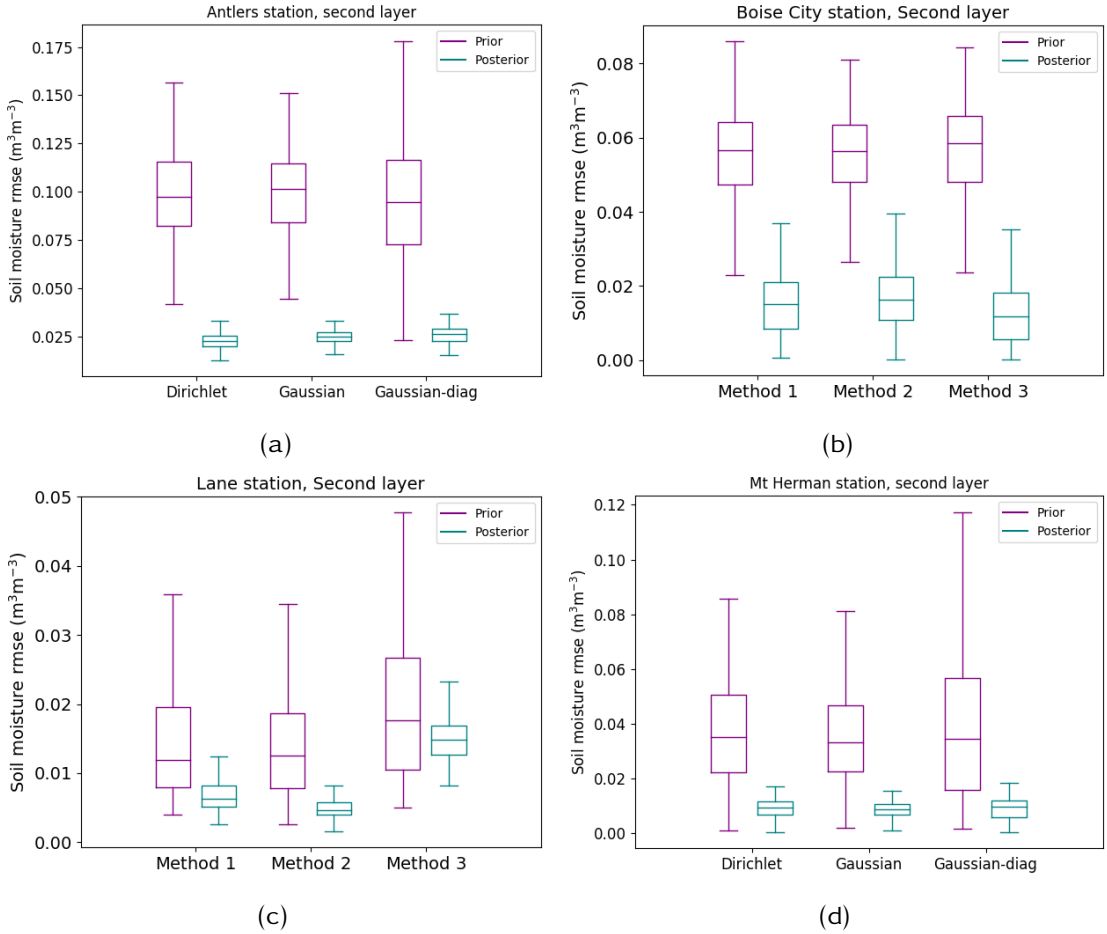


Figure 5.20: As in Figure 5.19 but for the second soil layer.

5.8 Comparison of methods for handling extreme soil texture

In section 5.6 and section 5.7 we have seen that assimilating top layer soil moisture into the JULES model can improve model predicted soil moisture by parameter estimation. For both cases, soil texture parameters were taken from the Mesonet sites. For all the four sites we considered, soil texture percentages were somewhere in the middle of the soil triangle, i.e. none of the three soil texture parameters was entirely dominant in contribution. However, in reality, there are cases where only one of the soil textures is entirely dominant than the other two. Contrary to having one soil texture being dominant, we also considered where all the three soil texture parameters have exactly equal contribution to the soil class texture. Hence in this section, we are comparing the performance of the three distributions by considering the case where the background soil texture parameters are near the corners of the soil triangle as well as at the centre. Similar meteorological forcing data are used for all the cases, to make the comparison solely on the soil

type. The truth soil texture parameters considered in this experiment are as in Table 5.2 and the background ensemble members are sampled around the truth soil parameters. Then the twin experiment is repeated with the rest of the experimental set-ups as in section 5.6.

5.8.1 Results and Discussion

Figure 5.21 shows prior and posterior sand percentage where the prior percentage is sampled using the respective truth as a mean from the Dirichlet, Gaussian and Gaussian-diag. In all the cases, the posterior sand percentage is closer to the truth and has smaller uncertainty compared to the corresponding prior. Both for prior and posterior parameters, Figure 5.21d shows less bias and the truth parameters are retrieved. On the other hand, for Figures 5.21c, 5.21a and 5.21b, both prior and posterior parameters are biased since the truth (population mean) parameter is closer to the boundaries (0 or 100).

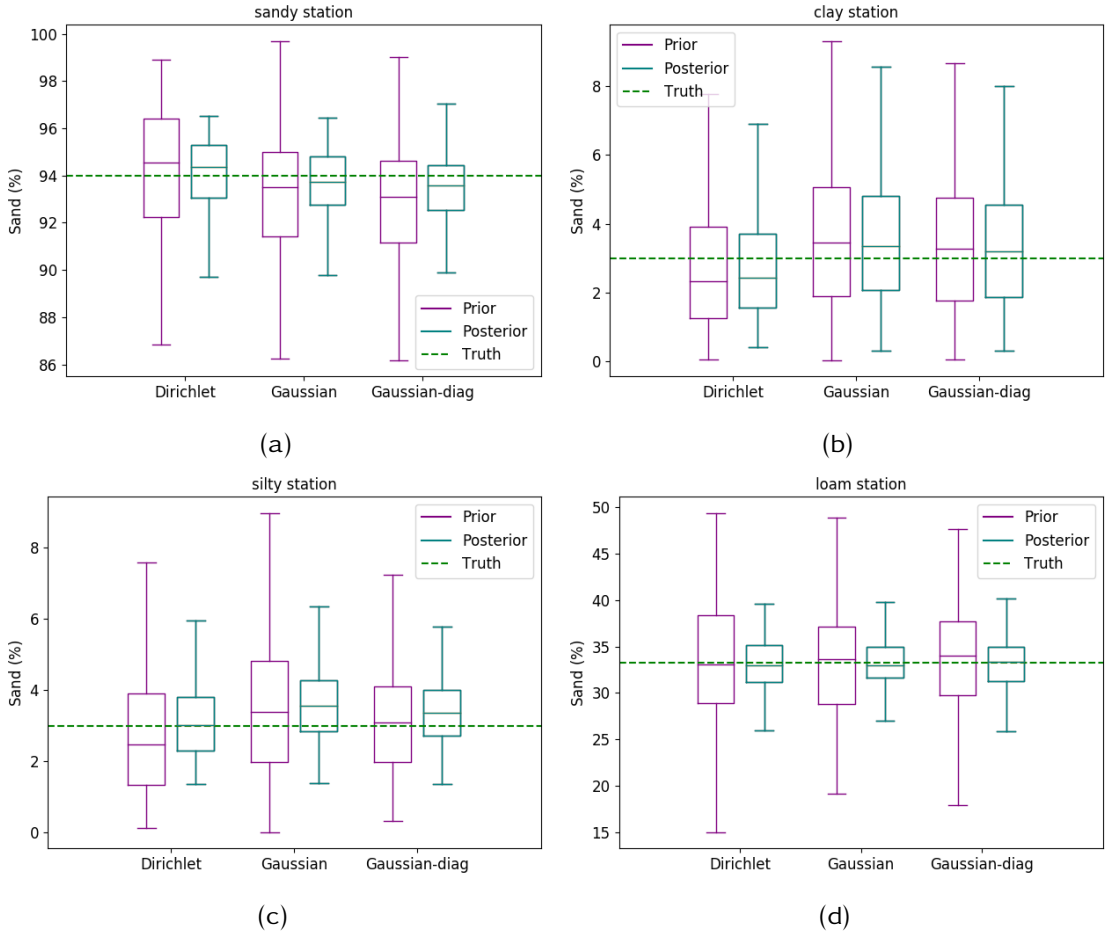


Figure 5.21: Prior and corresponding posterior percentage sand for k = 2 and 200 ensemble members. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques used for sampling prior parameters. The Truth was used as a mean for sampling from all the three distributions.

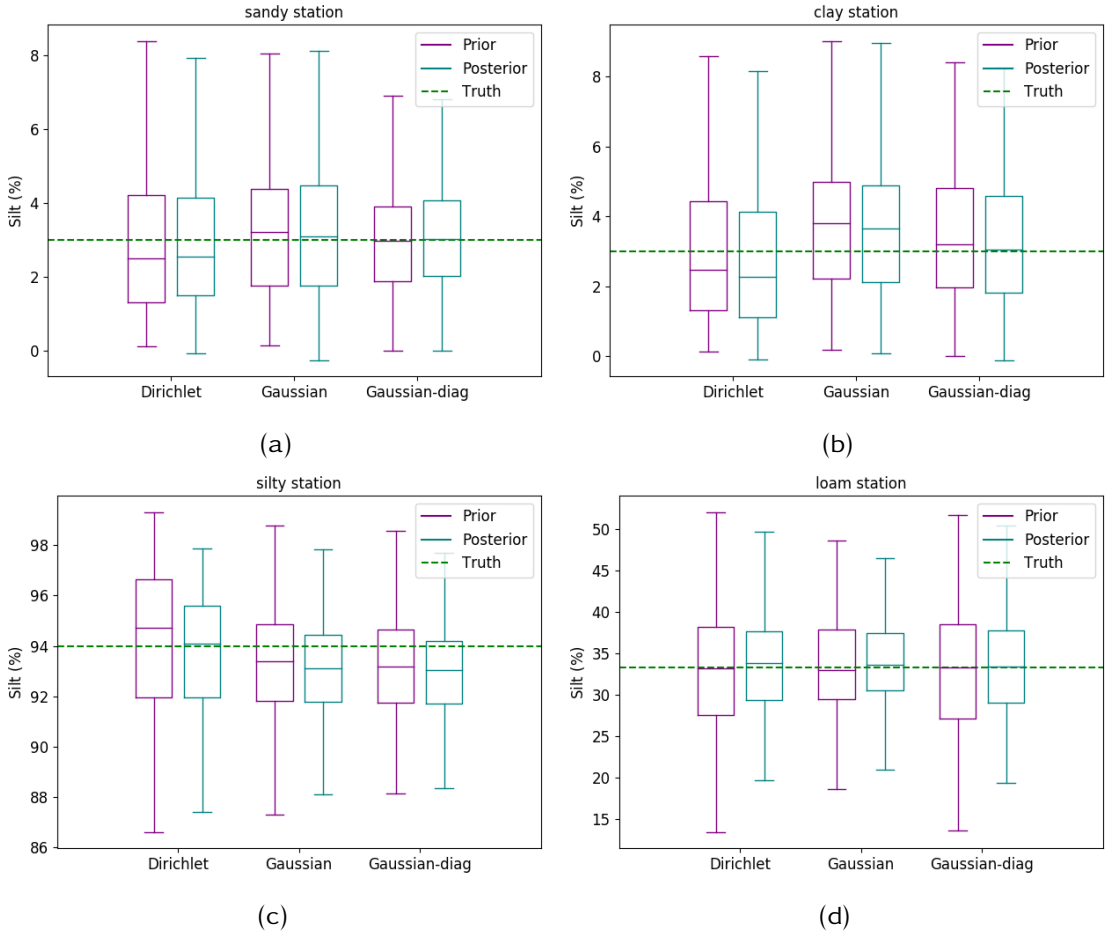


Figure 5.22: As in Figure 5.21 but for silt.

Figure 5.22 is as in Figure 5.21 but for silt. Posterior sand percentages are less skilful than the corresponding sand posterior as soil moisture prediction is less sensitive to silt percentage changes.

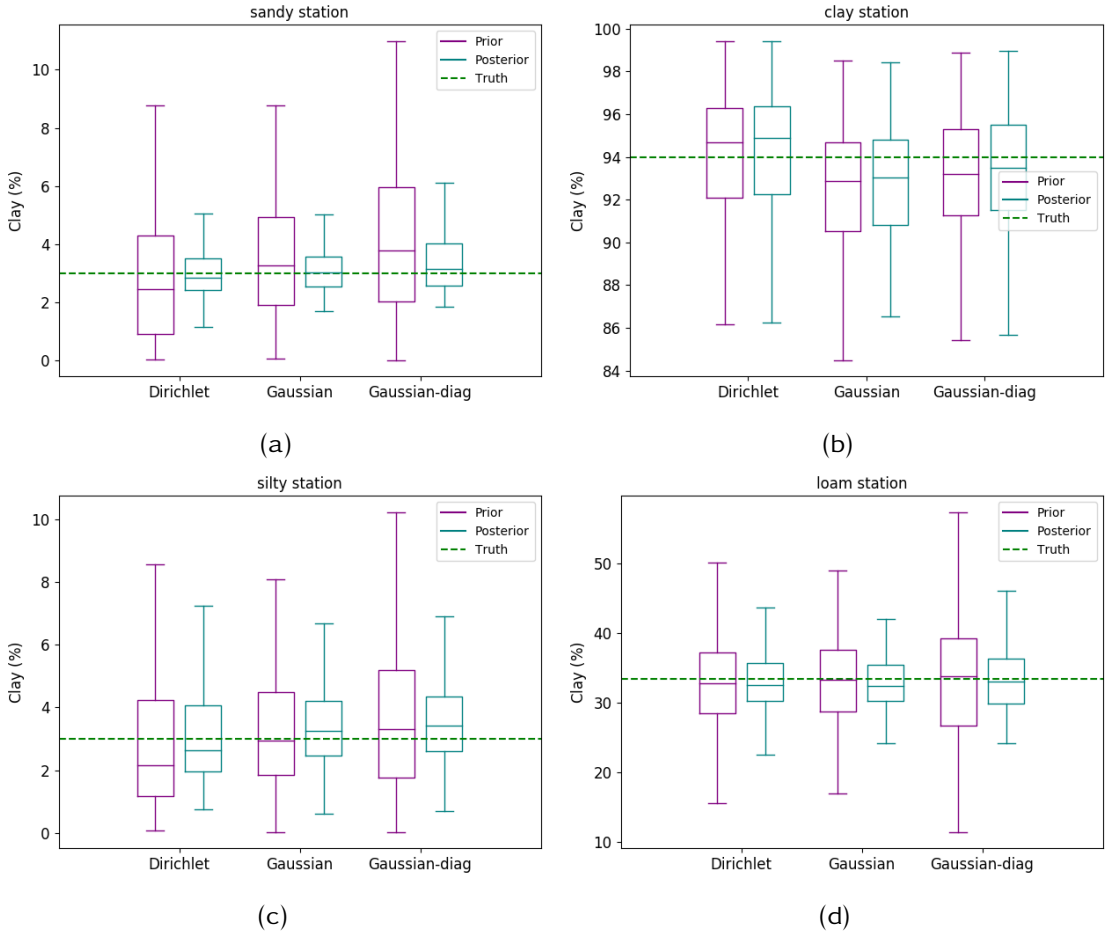


Figure 5.23: As in Figure 5.21 but for clay.

From Figure 5.21 - Figure 5.23 we have seen that, when the background soil texture percentages are equally shared among the three parameters (Figure 5.21d, Figure 5.22d and Figure 5.23d), all the three distributions have shown similar performance and less bias, for prior and posterior parameters. Whereas, if one of the soil textures is dominant, prior parameters are biased and also posterior parameters as a result. Comparing the three distributions, there is no substantial difference between them, all have shown better posterior than the prior. Figure 5.24 and Figure 5.25 are prior and posterior soil moisture RMSE for top soil layer and second layer corresponding to prior and posterior parameters plotted in Figure 5.21 - Figure 5.23.

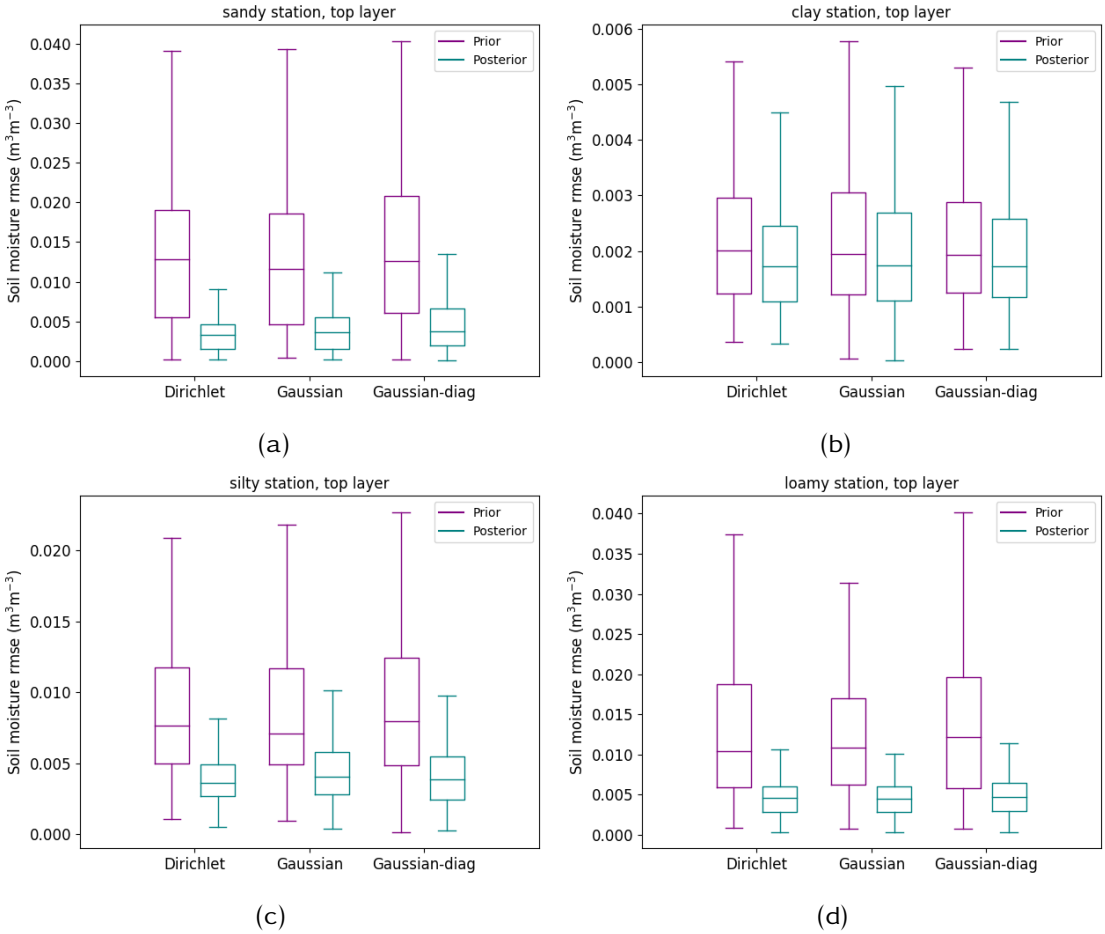


Figure 5.24: Prior and posterior soil moisture RMSE for 200 ensemble members, for the top layer. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques which resulted in prior soil moisture ensemble members and corresponding data assimilation experiments which resulted in the posterior soil moisture ensemble members.

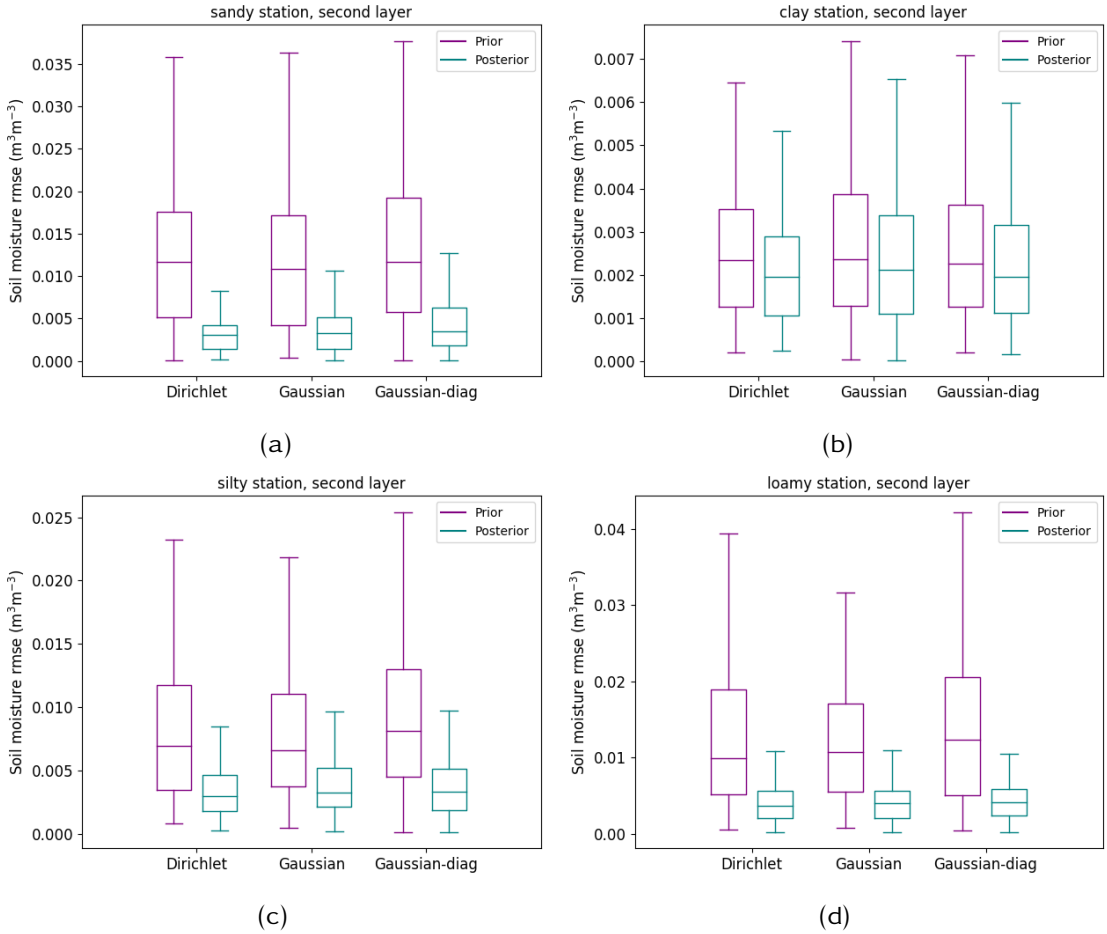


Figure 5.25: Prior and posterior soil moisture RMSE for 200 ensemble members, for the second layer. Dirichlet, Gaussian and Gaussian-diag represent the three ensemble initialisation techniques which resulted in prior soil moisture ensemble members and corresponding data assimilation experiments which resulted in the posterior soil moisture ensemble members.

Figure 5.24 and Figure 5.25 show that at any corner of the soil triangle and at the centre, the three ensemble initialisation methods can be used for sampling soil texture parameters. The posterior soil moisture is more skilful than the prior, for all the three ensemble initialisation techniques. When the mean parameters are closer to the boundaries, parameter ensembles become closer to the truth. As a result, the RMSE is very small, even for the prior soil moisture. In comparison, when the mean parameter is at the centre, parameter ensembles have a larger variance, and prior soil moisture RMSE is larger, Figure 5.24d and Figure 5.25d.

5.9 Summary

In this chapter, a novel ensemble initialisation technique using the Dirichlet distribution is introduced for synthetic soil moisture data assimilation with the JULES model using 4DEnVar for

parameter estimation. Cases where known and unknown background ensembles are considered to investigate the ability of the data assimilation set-up. In both cases, numerical results showed that the JULES model is successfully constrained by assimilating top layer synthetic soil moisture data to improve soil texture parameters as well as soil moisture estimates. Performance of the Dirichlet distribution is compared to the Gaussian distribution. When the first two moments are the same, prior and posterior soil parameters from the Dirichlet and Gaussian distribution are similar. This might seem obvious; however, the new information here is that the Dirichlet distribution can be used in the data assimilation where background errors are assumed to be normally distributed. For the Gaussian distribution without correlations (Gauss-diag), numerical results showed that posterior parameters and posterior soil moisture estimates are similar, but the prior soil moisture has larger RMSE than the corresponding prior from Gaussian and Dirichlet distributions.

Prior soil texture parameters from different soil classes were sampled to compare the distributions with various circumstances. The same meteorological forcing data were considered and used in the data assimilation. Even though the Dirichlet distribution shares the physical properties of soil texture parameters, the effect on sampling soil texture parameters is not significant compared to the Gaussian distribution with and without correlations. In conclusion, based on numerical results for posterior soil moisture estimates, any of the three distributions can be used to sample soil texture parameters.

In a twin experiment, the data assimilation is presumed to retrieve the truth soil parameters. However, for real soil moisture data, this might not be the case. As we are considering a perfect model, any mismatch between the observation and model estimate soil moisture will be compensated by adjusting the soil parameters, even if it is due to model error. As a result, the posterior soil texture parameter may not necessarily be a valid parameter value.

As soil parameters determine hydraulic parameters which in turn determines the amount of moisture in the soil, improving the prediction of soil parameters is presumed to improve soil moisture estimates of a numerical model. By assimilating top layer soil moisture data into the JULES land surface model, the truth soil parameters were restored, and as a result, posterior soil moisture estimates have smaller RMSE compared to the prior soil moisture. This result is in agreement with the results discussed in (Scott et al., 2013), where improving soil property database lead into improved soil moisture estimates in Oklahoma Mesonet sites. To verify these

experiments with real observations, in chapter 6 we have assimilated ground measurement, and satellite observed soil moisture data for Mesonet site into the JULES model using the 4DEnVar for parameters estimation.

Chapter 6

Parameter estimation using the Dirichlet distribution to initialise model ensemble for 4DEnVar: with observed data

In this chapter, ground measurements and satellite observed soil moisture data are assimilated into the JULES model with 4DEnVar to improve soil moisture estimates from the JULES model by parameter estimation. Performance of the prediction system is verified by hindcasting for the following year where data is not assimilated.

6.1 Introduction

In chapter 5, we have successfully performed twin experiments with the 4DEnVar for parameter estimation where the prior parameters are drawn using the Dirichlet and Gaussian distributions. Results show that, for parameters and state, the posterior is closer to the truth compared to the prior. While twin experiments have advantages because of the availability of truth parameter and state to compare the data assimilation results with, it is vital to validate the method with real data for actual use of the system. For example, in the twin experiment, the model is a perfect representation of the observations, which is not attainable in reality. As such, in this chapter, ground

measurement and satellite observed soil moisture data from Mesonet sites (see subsection 3.2.2 and subsection 3.2.3) are assimilated into the JULES model. To strengthen the validation, soil moisture hindcast for the following year after data assimilation is obtained and compared with observations in that year. This step is important as soil moisture can be obtained for the future forecast and can be used for early warning systems like drought monitoring and flood warning systems if the forcing data forecast are available from climate models. From the twin experiment, improvements for the top layer soil moisture estimates have improved soil moisture estimates for the deeper layer too. Here we also investigate if this holds for assimilation using real data or not.

From the comparison of twin experiment results in chapter 5, we have seen that posterior parameters obtained by using the Gaussian and Dirichlet distributions show similar accuracy when both distributions are set to have the same correlations for the prior. In this chapter, the Dirichlet distribution will be used to draw prior soil texture parameters and initialise model ensembles, since the Dirichlet distribution shares positivity and boundedness properties with soil texture parameters (section 5.2). In section 6.2, the experimental set-up and results from assimilating ground measured soil moisture data from Oklahoma Mesonet sites into the JULES model are presented. With a similar experimental set-up and forcing data as in section 6.2, high-resolution satellite observed soil moisture data (SMAP 9 km by 9 km resolution data) assimilation results are given in section 6.3. Then we investigate whether or not *in-situ* observations can be used to validate the satellite data assimilation results.

6.2 *In-situ* soil moisture data assimilation

In this experiment, the JULES model is configured in such a way that the Oklahoma Mesonet site description is met. To match Mesonet soil moisture measuring points 5 cm, 25 cm and 60 cm deep from the surface (Illston et al., 2008), the standard JULES discretisation of each soil layer thickness is changed into 10 cm, 30 cm, 40 cm and 2.2 m, from top to bottom. This corresponds to measurement points are at the centre of each soil layer. The fourth layer thickness is chosen such that the total depth of the column is 3 m as the JULES standard setting, (Best et al., 2011).

6.2.1 Experimental design

In chapter 5, all experiments are based on synthetic data and the model is considered to be perfect, observation errors are Gaussian and the maximum value of k, which determines the magnitude of uncertainty in the background, used is 2. However, in real data assimilation applications, the model does not represent reality perfectly, and the related error is not perfectly Gaussian. In this experiment, the value of k is expected to be larger than the one used for the perfect model assumption. Otherwise, the model ensemble predictions will have small spread while they are away from the observations, which we have seen in chapter 4 with a perfect model assumption. This will make the system too confident on the model prior, and observations will be disregarded. Hence, the value of $k = 10$ is selected so that all the four stations with different soil type have reasonable spread for soil moisture ensembles.

The forcing data to drive the JULES model is from Mesonet meteorological data except for the longwave radiation (LW). LW is taken from the Watch Forcing Data methodology applied to ERA-Interim data (WFDEI) (Weedon et al., 2014) 36 km resolution data, disaggregated from 3 hrs to 30 min by cubic interpolation. Years 2016 - 2018 inclusive are considered for the forcing data where 2016 is for spin-up the JULES model, 2017 for DA experiments and 2018 for hindcast based on the assimilation results.

The background soil texture parameters, used as the mean value to draw from a Dirichlet distribution, are taken from the measured values for each station, Scott et al. (2013). Compared to the SMAP footprint, stations represent only a fraction of a pixel. 200 ensemble members of prior soil texture parameters and the corresponding soil moisture ensemble are used for the experiments. The JULES model is configured with 100% temperate ($C3$) grasses, a leaf area index of 6 to generate soil moisture ensemble members corresponding to each ensemble member of soil texture parameters. As we have seen aerial photos in chapter 3, there is a difference in vegetation coverage between stations. However, here we considered 100% temperate ($C3$) grasses for all the stations. This tests the potential of the data assimilation experiment while little information on the land cover is provided. The Van Genuchten equation was used to calculate the hydraulic characteristics parameters. Gaussian observation error with error covariance matrix $\mathbf{R} = 0.05^2\mathbf{I}m^3m^{-3}$ is used, based on the validation experiment done by comparing Mesonet ground measurement soil moisture observations with the gravimetric method, Illston et al. (2008) and observation frequency is every five days. Volumetric soil moisture data was obtained by using the

Van Genuchten equation, to be consistent with the JULES model. Using the above experimental set-up, results of assimilating *in-situ* surface soil moisture data for the four sites in Oklahoma Mesonet is given in the following subsection.

6.2.2 Results and discussion

Here *in-situ* soil moisture data described in subsection 3.2.3 is assimilated into the JULES model and soil texture parameters are updated. State updates are then obtained by running the JULES model using the updated parameters.

Figure 6.1 shows data assimilation results using the 4DEnVar data assimilation method. The posterior soil moisture is a single JULES run corresponding to the posterior soil texture parameter obtained from the data assimilation as opposed to posterior soil texture parameter ensembles as in chapter 5. Hence, the measure of data assimilation performance is by comparing how close the posterior soil moisture estimate x_a is to the observations Obs compared to the prior soil moisture estimate x_b.

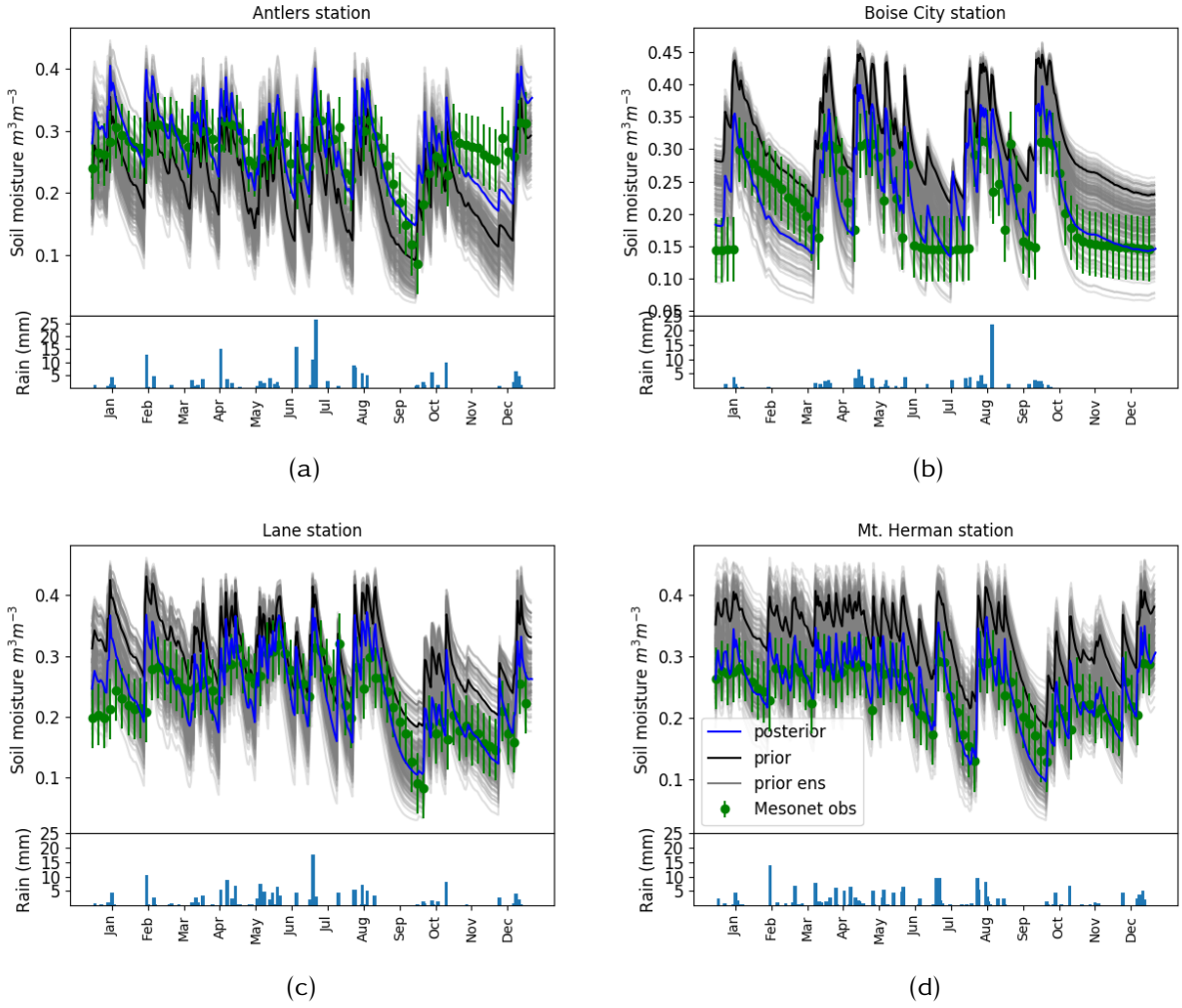


Figure 6.1: Top layer soil moisture data assimilation results for stations in Oklahoma Mesonet and daily rainfall. Black line: JULES prior trajectory. Blue line: JULES posterior trajectory. Light gray lines: JULES prior ensemble members trajectory. Green dots: ground measurement soil moisture data from Oklahoma, Mesonet, 2017. Green vertical lines: error bar for observations. Blue bars: Daily rainfall data.

Figure 6.1 shows that, the JULES posterior trajectory is closer to respective observations than the JULES prior soil moisture trajectory. The improvement is as a result of posterior parameters obtained from assimilating soil moisture observations. In all the stations and throughout the year, we can see that the JULES model responded to the forcing rainfall data and mimics the trajectory of ground observations even if the magnitude is different.

The primary objective of data assimilation is to minimise the distance between the model estimate and observed state. As such, figure 6.2 is a figure plotted to assess the performance of data assimilation experiment results presented in figure 6.1 in terms of the distance of prior and posterior soil moisture estimates from soil moisture observations in each month.

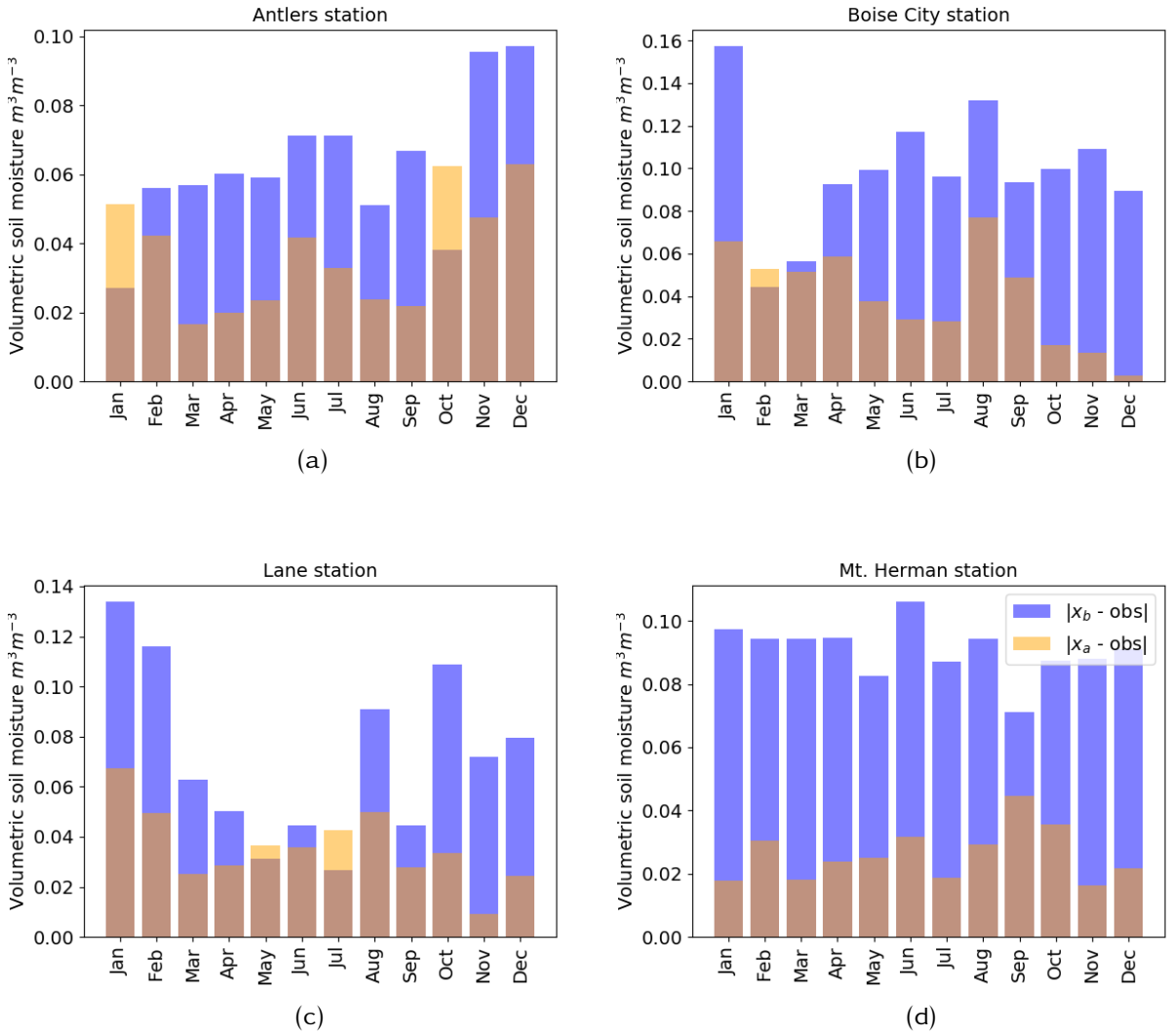


Figure 6.2: Corresponding plot for Figure 6.1. Blue bars: distance of background soil moisture from observations. Orange bars: distance of posterior soil moisture from observations. Brown bars: where the blue and orange bars overlap. Observations are top layer ground measurement soil moisture from Oklahoma Mesonet in 2017.

From figure 6.2 we can generalise that, posterior soil moisture estimates are closer to the observations compared to the prior, we see more blue color than the orange. However, looking at particular times, Figure 6.2a representing Antlers station for example, prior soil moisture is closer to observations compared to the posterior in January and October. This is due to the fact that 4DEnVar, as one of the variational data assimilation methods, considers all the observations to come up with an optimal trajectory to fit into the observations when the new initial condition (analysis) is used and the model is integrated. Hence, it is not surprising to see the prior being closer to the observations than the analysis sometimes. Note that as we are estimating parameters instead of the state, the new initial condition (analysis) is obtained by using the analysis parameters to run the JULES model. Looking at the prior soil moisture for Antlers station, (Fig-

ure 6.2a), it matches well the observations in Jan and Oct. In addition, the prior error is smaller than the observation error in Jan and Oct. Hence the posterior which is obtained by considering the whole year is likely to be worst than the prior for Jan and Oct. The same scenario is observed for Boise City station in Feb (Figure 6.2b) and for Lane station in July (Figure 6.2c). Whereas for Mt Herman station, Figure 6.2d, the prior soil moisture is far from the observations throughout the year and as a result the posterior is closer to the observations throughout the year.

Considering the overall performance across the whole year (the whole assimilation window), the posterior soil moisture estimates are closer to the observations than the prior soil moisture model estimate, as shown in Table 6.1. Up to a 60% reduction is obtained by assimilating Mesonet *in-situ* observations. Looking at specific times; however, posterior soil moisture with better improvement is observed where the prior soil moisture is far from the observations and vice versa.

6.2.3 Hindcasting

One way of verifying the prediction system is by hindcasting, i.e. forecasting retrospectively, and compare with the observations.Comparing posterior with observations in the same time frame of assimilation is a useful metric, however, comparing against the same observations which are assimilated is somehow straightforward and does not necessarily show that the result is reliable. It is worth checking with other sets of observations other than the ones which are assimilated. Here, in addition to calculating the distance reduction by the posterior soil moisture estimates to the observations, soil moisture is estimated for the following year data is assimilated is performed and the hindcast soil moisture is compared with *in-situ* observations, Figure 6.3.

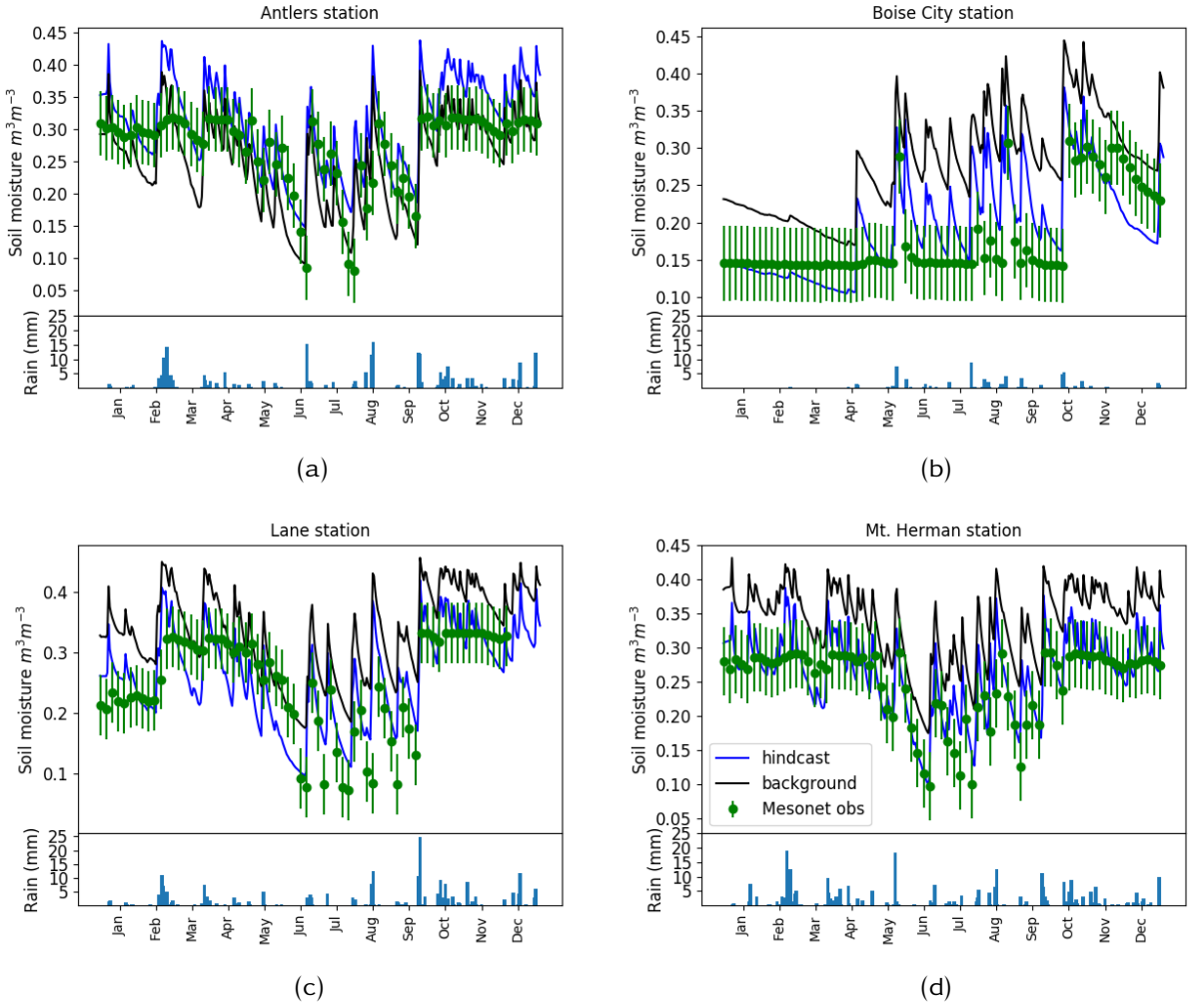


Figure 6.3: Hindcast soil moisture for 2018 based on the posterior soil texture parameters corresponding to the result obtained from assimilated Mesonet soil moisture data for 2017, depicted in figure 6.1. Black line: JULES open-loop trajectory. Blue line: JULES hindcast trajectory. Green dots: ground measurement soil moisture data from Oklahoma, Mesonet, 2018. Green vertical lines: error bar for observations. Blue bars: Daily rainfall data.

Figure 6.3 depicts open-loop, hindcast and observed volumetric soil moisture for the year 2018. Open-loop runs x_o and hindcast x_f are soil moisture estimates for the year beyond the assimilation window using prior and posterior soil texture parameters respectively. Similar to the case in Figure 6.1 in the previous year, the soil moisture variables responded well for the forcing rainfall data. For Boise City station, for example, Figure 6.3b, for the duration from January to March there was almost no rainfall, and as a result soil moisture observation and model estimates are drier than any other time.

Figure 6.4 shows the distance of the open-loop and hindcast soil moisture from the observations for each month. Similar to the assimilation period, the forecast soil moisture is skilful compared to the open-loop most of the time. However, when the open-loop is closer to the ob-

servations, the hindcast is less skilful, for example, Figure 6.4a, Sept-Dec. Note that Mesonet observations are missing for Lane station in December 2018, Figure 6.3c and Figure 6.4c. Hence the RMSE given in Table 6.1 is calculated over the available observations only.

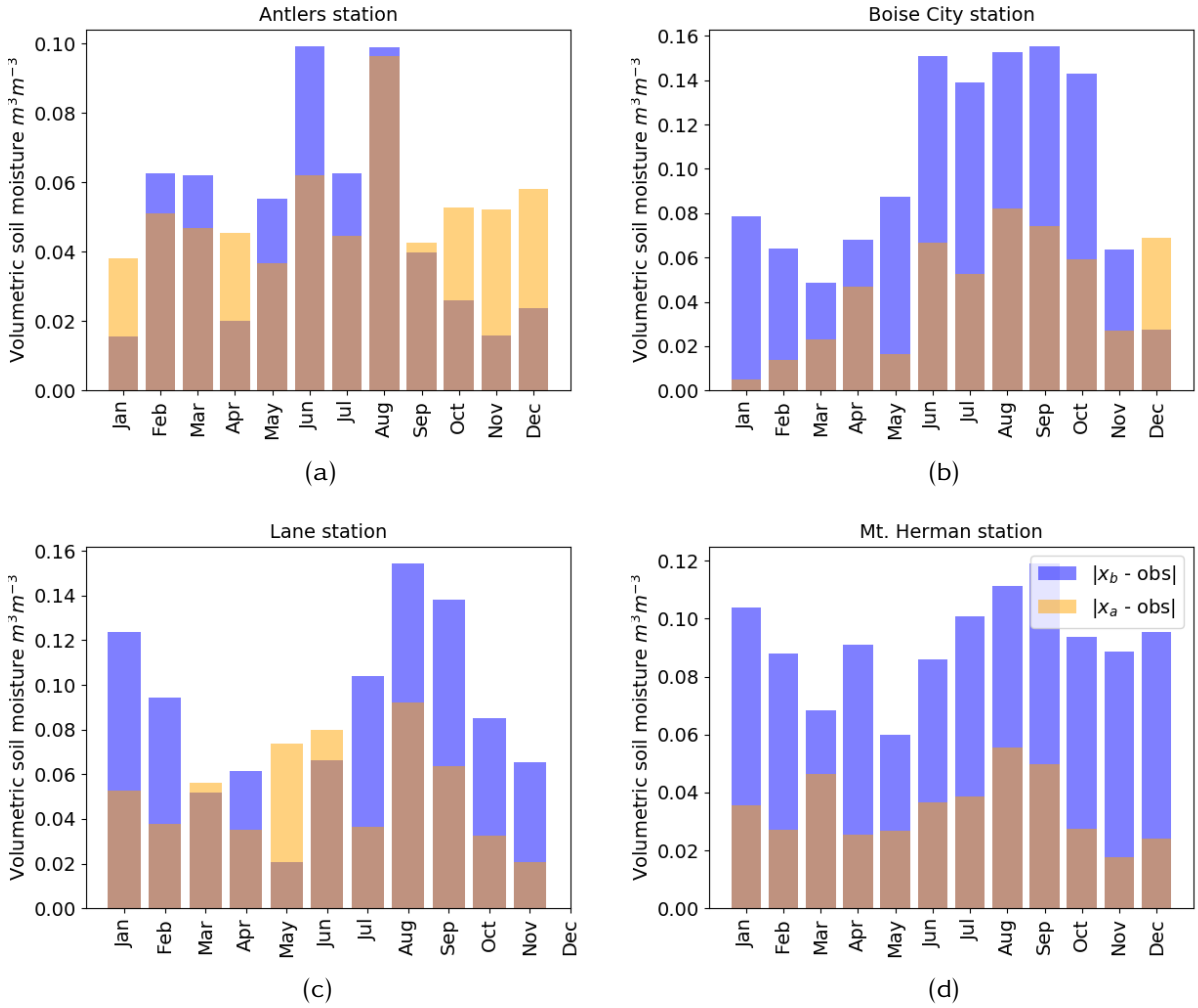


Figure 6.4: Corresponding plot for Figure 6.3. Blue bars: distance of background soil moisture from observations. Orange bars: distance of posterior soil moisture from observations. Brown bars: where the blue and orange bars overlap. Observations are ground measurement top layer soil moisture from Oklahoma Mesonet for 2018.

Comparing the hindcast with open-loop model runs, most of the time, the hindcast is closer to the observations. However, there is a variation of performance across stations. For example, for Antlers station, the overall improvement of the hindcast (RMSE = 0.07) compared to the open-loop model run (0.063) is slightly worse. For Lane station, the hindcast has RMSE = 0.071 and open-loop run RMSE = .105. Comparing among the hindcast, Antlers station is slightly better than Lane station. In general, soil moisture hindcast is skilful, and it reassures the performance of the data assimilation results we see above, subsection 6.2.2.

Another aspect we looked in the twin experiments is the impact of assimilating surface soil moisture for deeper layers. In chapter 5 we have seen that deeper layers benefited from the improvement on the top layer. The following subsection investigates whether or not the same result holds with *in-situ* soil moisture data assimilation.

6.2.4 Investigating root-zone soil moisture content

Here we assess how the data assimilation on the top layer influences the layer below it. The assessment is important because the soil moisture content in the second layer is vital for several land surface processes. However, soil moisture observations for deeper layers are absent (satellite observations) or scarce in general. On the other hand, data assimilation makes it possible for the deeper layers to gain information from observations on the top layer Reichle et al. (2001). For Mesonet sites, soil moisture is observed for the second layer as well, at 25 cm from the surface, and observations are compared with the posterior soil moisture as a verification. In chapter 5 we have seen that in a twin experiment, assimilating soil moisture on the top layer resulted in improved posterior soil moisture. As in chapter 5, in this experiment, the JULES model is configured with constant soil texture parameters for all the soil layers where which is not true in reality. For mesonet sites, soil texture varies significantly with depth Scott et al. (2013).

Any alteration in the first layer will alter the soil moisture estimate for deeper layers in two ways. First, the posterior parameter (if different from the prior) will alter soil moisture estimation in the deeper layers. Second, the difference in posterior soil moisture for the top layer will affect posterior soil moisture in the deeper layers as a result of infiltration. Hence, if the prior soil moisture in the deeper layers is already close to the observations, the posterior soil moisture in the deeper layer is most likely to be worse as a result of alterations from the top layer. Note that there is no data assimilated for the deeper layer, we are assessing the impact of assimilating on the top layer.

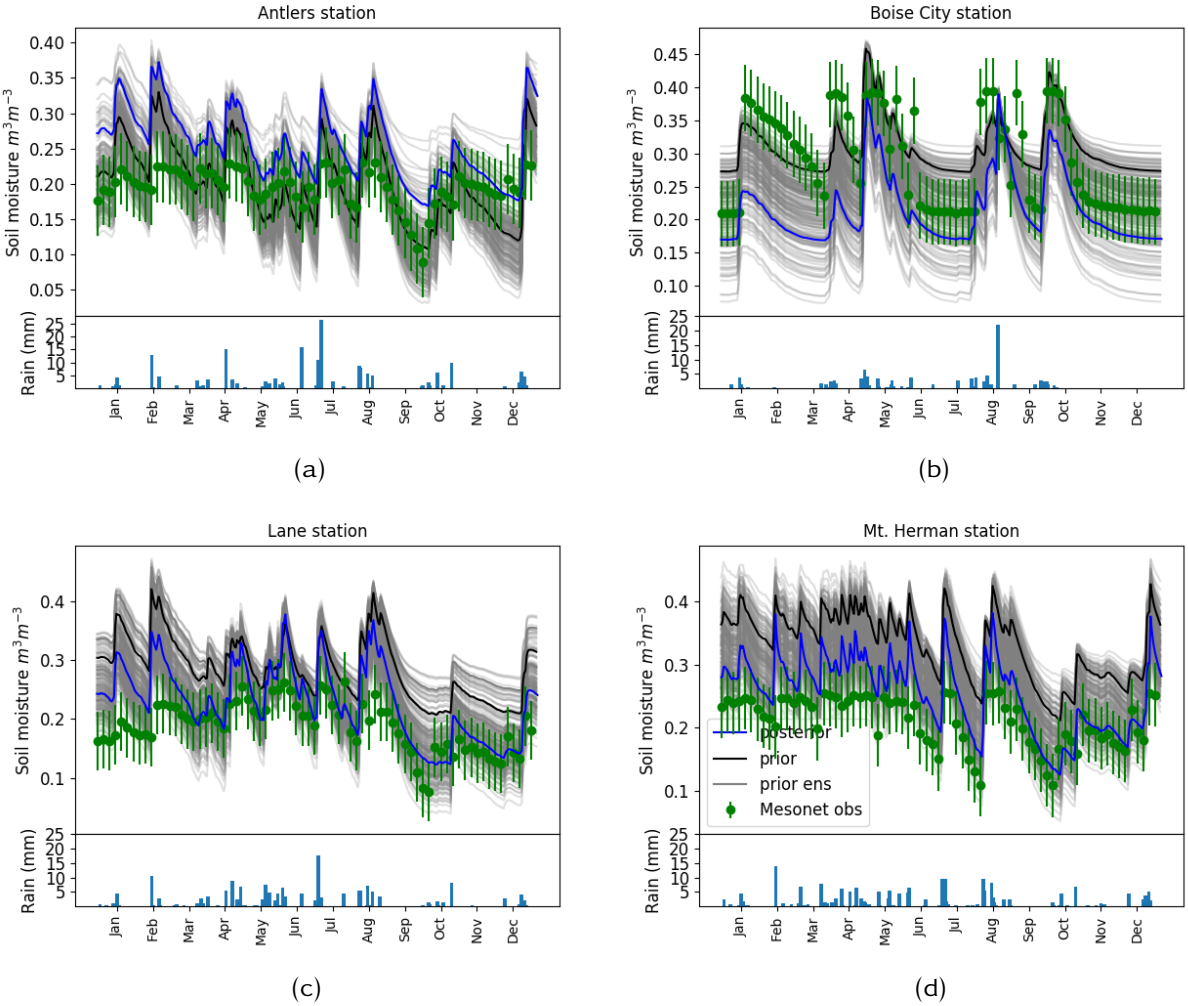


Figure 6.5: Similar to figure 6.1 but for the second layer.

Figure 6.5 shows observed, background and posterior soil moisture for the second layer. Figure 6.5a and Figure 6.5b shows that posterior soil moisture which resulted from adjustments on the top layer is less skilful than the background. However, Figure 6.5c and Figure 6.5d shows that posterior soil moisture is more skilful than the corresponding background soil moisture.

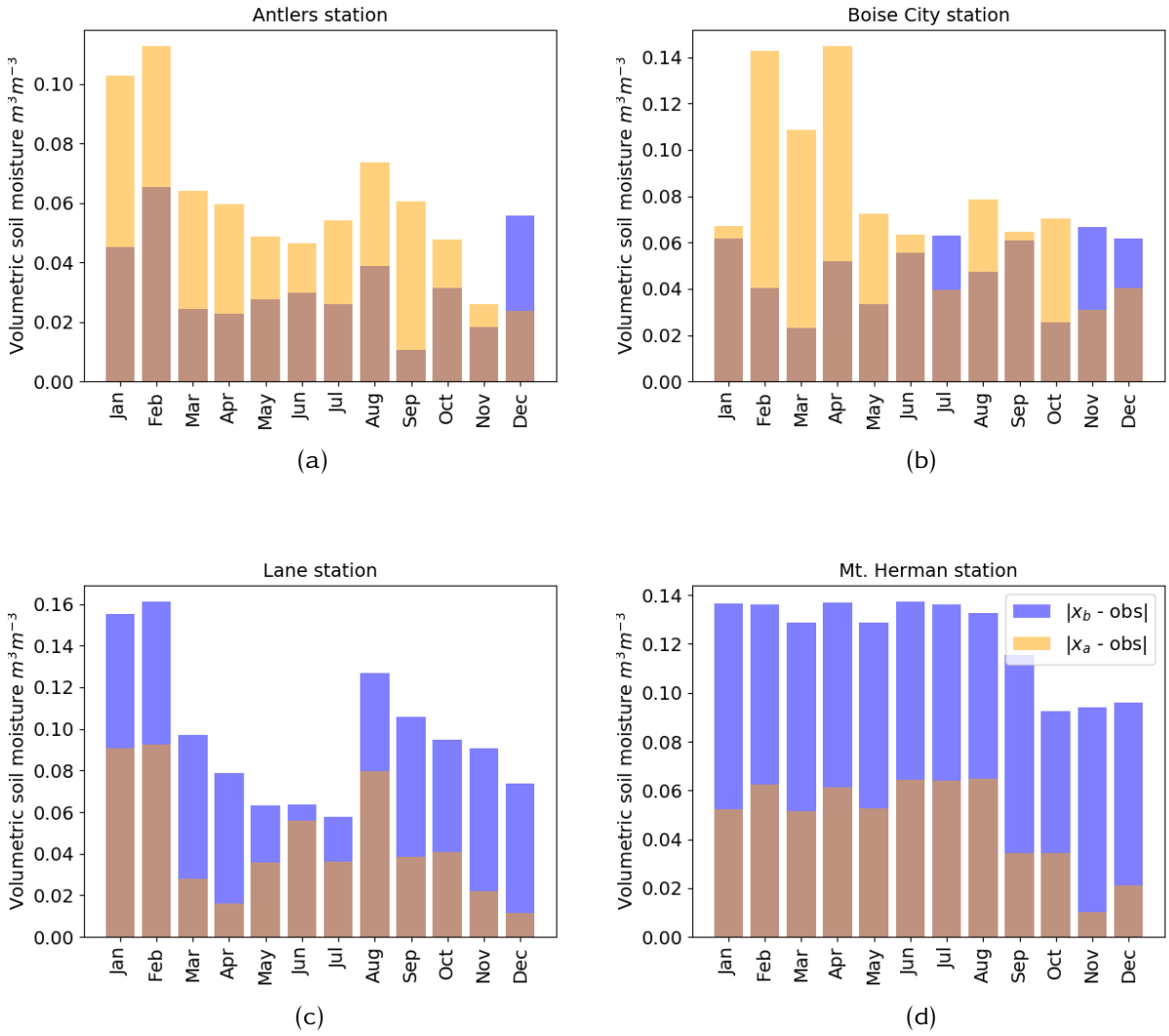


Figure 6.6: Similar to Figure 6.2 but for second soil layer.

Figure 6.6 is a corresponding plot to Figure 6.5, the distance of the background and posterior soil moisture from the observation at each observation time. From the two figures we have learned that, with constant soil texture parameter assumption, posterior soil moisture for the second layer improves the background soil moisture estimates for the two sites and made worse for the other two sites.

6.3 Satellite soil moisture (SMAP) data assimilation

In subsection 6.2.1, *in-situ* soil moisture observations from Oklahoma Mesonet were assimilated into the JULES model where the forcing data is also observed meteorological data from Mesonet. In this case, soil moisture observations and model output soil moisture estimates are represen-

tative in spatial resolution. However, as we have discussed in section 1.2, *in-situ* soil moisture observations have sparse spatial coverage, especially for the developing world. Hence, using satellite soil moisture observations is a feasible alternative due to spatial coverage. The caveat of using satellite observations where *in-situ* data is absent is that there is no way to validate whether assimilating satellite observations improve the model forecast skill or not. So, here we are assimilating satellite soil moisture observations from SMAP for Mesonet stations where *in-situ* observations are available to validate the data assimilation results. It is important to note that data assimilation results with SMAP observations for Mesonet sites being skilful does not imply the same for other sites, but it gives insight that there is a chance.

The SMAP soil moisture resolution for the radiometer (Passive part) is 36km by 36km and the 9km by 9km resolution data are obtained by disaggregation from the original data by Backus-Gilbert optimal interpolation method, Chaubell (2016). In the method, the surface brightness temperature in the required pixel is calculated by a weighted sum of the surface brightness temperature of the nearby pixels with the original resolution, Poe (1990) and Long and Brodzik (2016).

Here the assimilated soil moisture data is the L3, 9 km by 9 km SMAP satellite observed soil moisture data, as discussed in subsection 3.2.3. The experimental set-up in this section is as in subsection 6.2.1 except that soil moisture observations are 9 km by 9 km SMAP satellite observed soil moisture, with observation frequency of two to three days and observation error of 0.04, (Zhang et al., 2019a). Posterior soil moisture obtained from assimilating the SMAP soil moisture data is verified against *in-situ* soil moisture data. In addition, soil moisture is hindcast based on the assimilation results and compared with ground measurement observations.

6.3.1 Results and discussion

Here, data assimilation results obtained by assimilating SMAP high-resolution data is presented. Figure 6.7 is a time series plot for prior ensemble members, posterior, background and SMAP high-resolution soil moisture. The bottom box in each plot is observed rainfall for each station. As discussed in subsection 3.2.3, observations for Antlers and Mt Herman are noisy and the JULES model prior overestimated soil moisture for Boise city.

Figure 6.7 shows that for Boise city and Lane stations posterior soil moisture estimates (blue lines) have moved towards the observations from the background (black lines). For Antlers station, most of the time, the posterior is closer to the observations compared to the prior. However,

for Mt Herman station, the background is already close to the observations, and as a result the posterior did not move that much closer to the observations. Besides, since observations are noisy, it is hard to tell whether it is the posterior or the background is closer to the observations. Figure 6.8 is a distance between the background and posterior from the observations for each month and clearly shows the performance of the data assimilation.

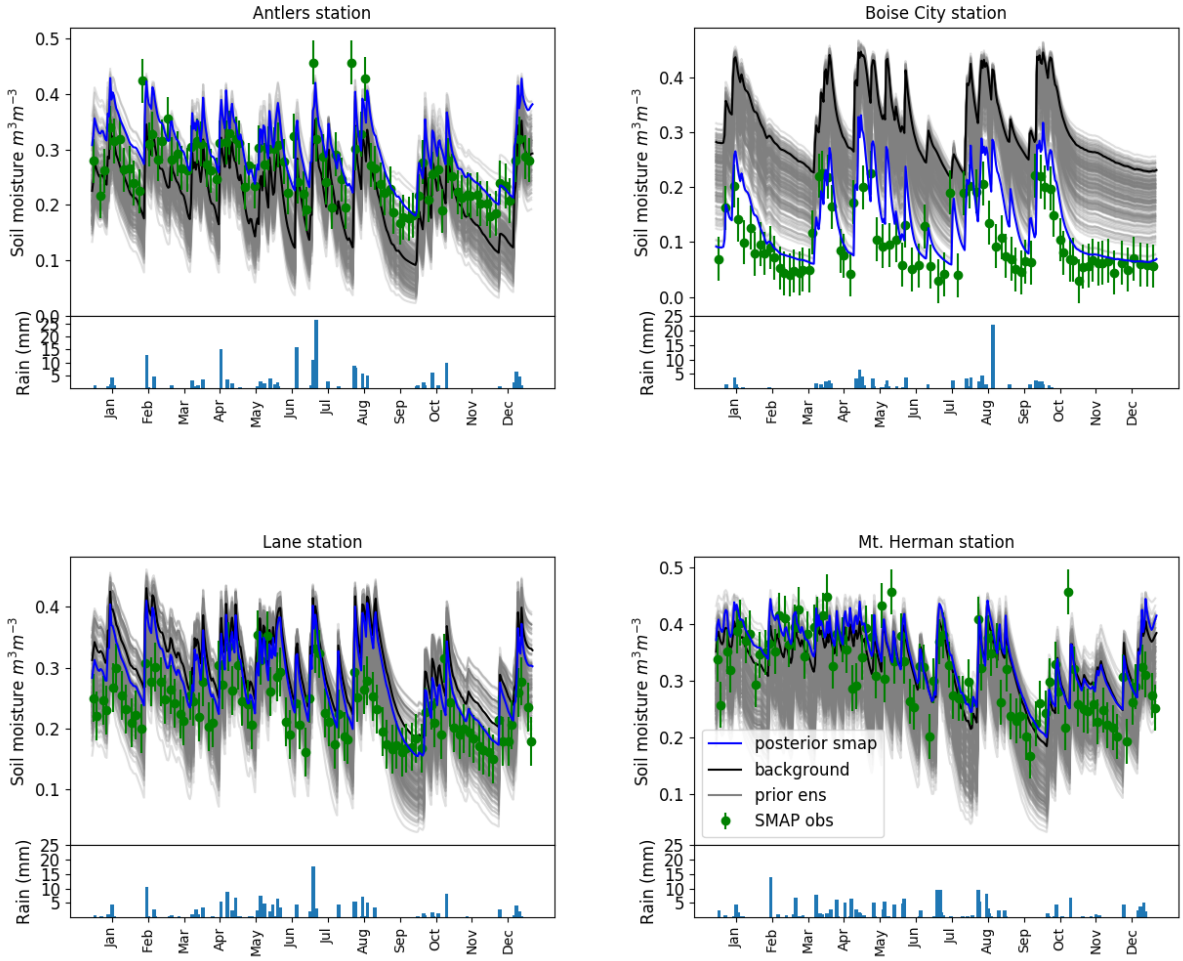


Figure 6.7: Top layer prior, background, posterior and SMAP satellite observed volumetric soil moisture for stations in Oklahoma Mesonet, for the year 2017. The prior parameters are drawn from a Dirichlet distribution with $k = 10$.

Figure 6.8 shows the distance from the observation for background and posterior soil moisture estimates for each month. The result shows that maximum improvement is attained for Boise City station and little or no improvement for Mt Herman station. Compared to similar plot from *in-situ* soil moisture data assimilation, Figure 6.4, posterior estimates are closer to the observations than the background. The exception is for Mt Herman station with improvements only a few times, Figure 6.8d, where observations are noisy. Mt Herman station is near big trees and Zhang

et al. (2019b) showed that SMAP observations are less accurate where the vegetation cover is dense.

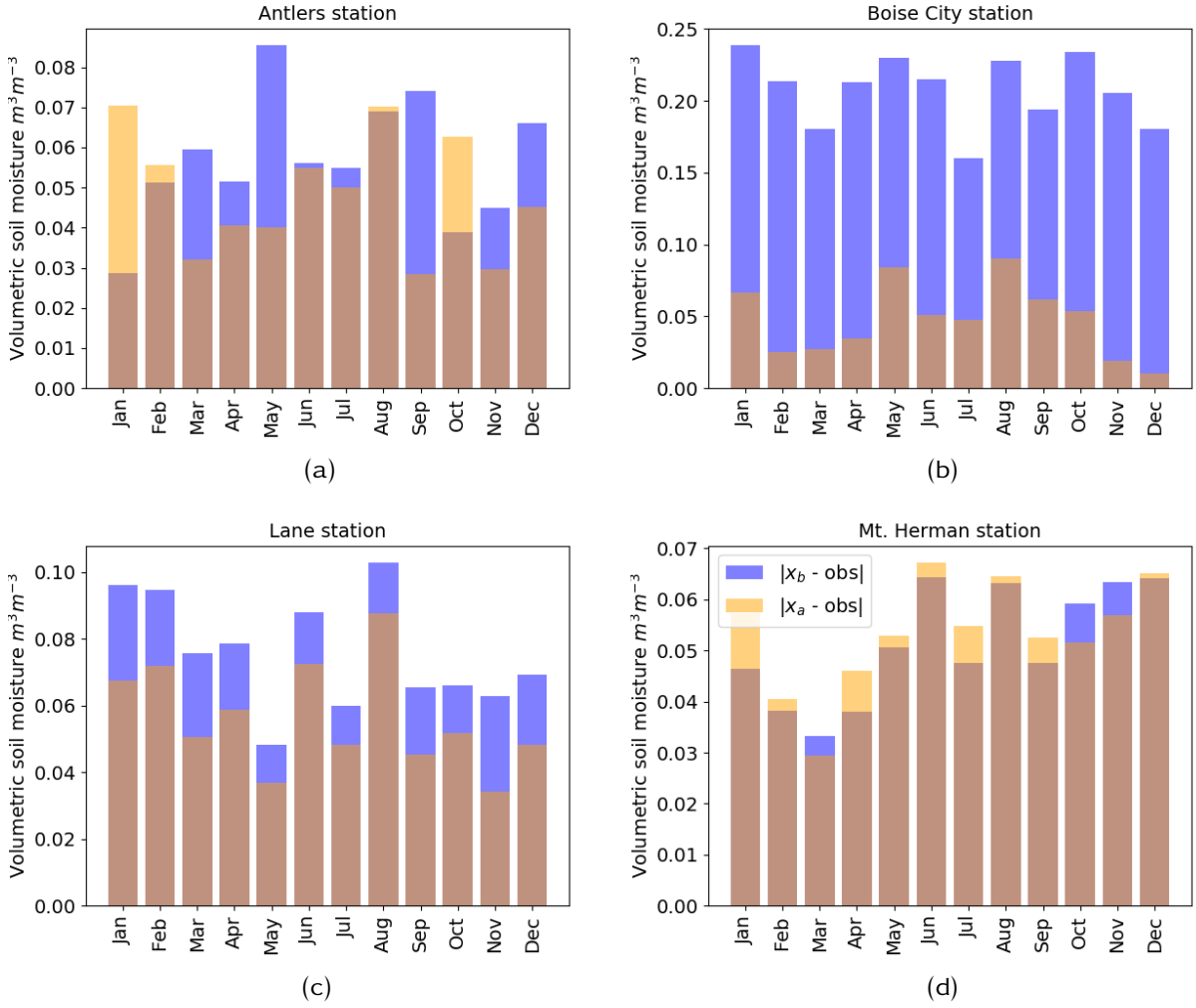


Figure 6.8: Blue bars: distance of background soil moisture from the concurrent soil moisture observations. Orange bars: distance of analysis soil moisture from the concurrent soil moisture observations. Brown bars: where the blue and orange bars overlap. Observations are high resolution soil moisture from SMAP. Both observation and model estimates are for top layer.

6.3.2 Hindcasting

Figure 6.9 is a soil moisture hindcast based on the high-resolution SMAP soil moisture and Figure 6.10 is the corresponding plot which shows the distance of the hindcast from the observations. Similar to the year data assimilation is performed, hindcast soil moisture is closer to the observations compared to the open-loop model run in the following year.

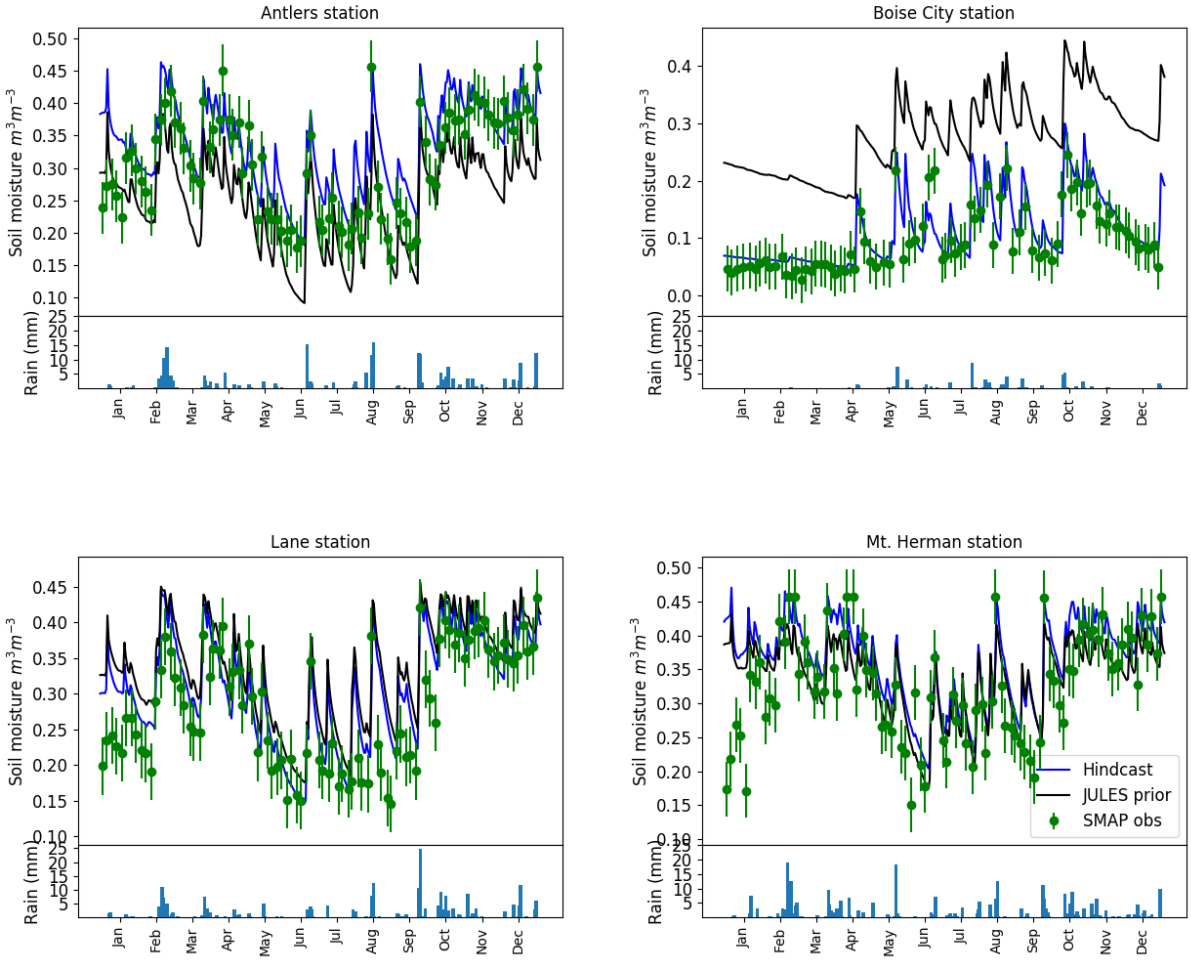


Figure 6.9: Hindcast of soil moisture from assimilating SMAP 9km resolution satellite observed soil moisture.

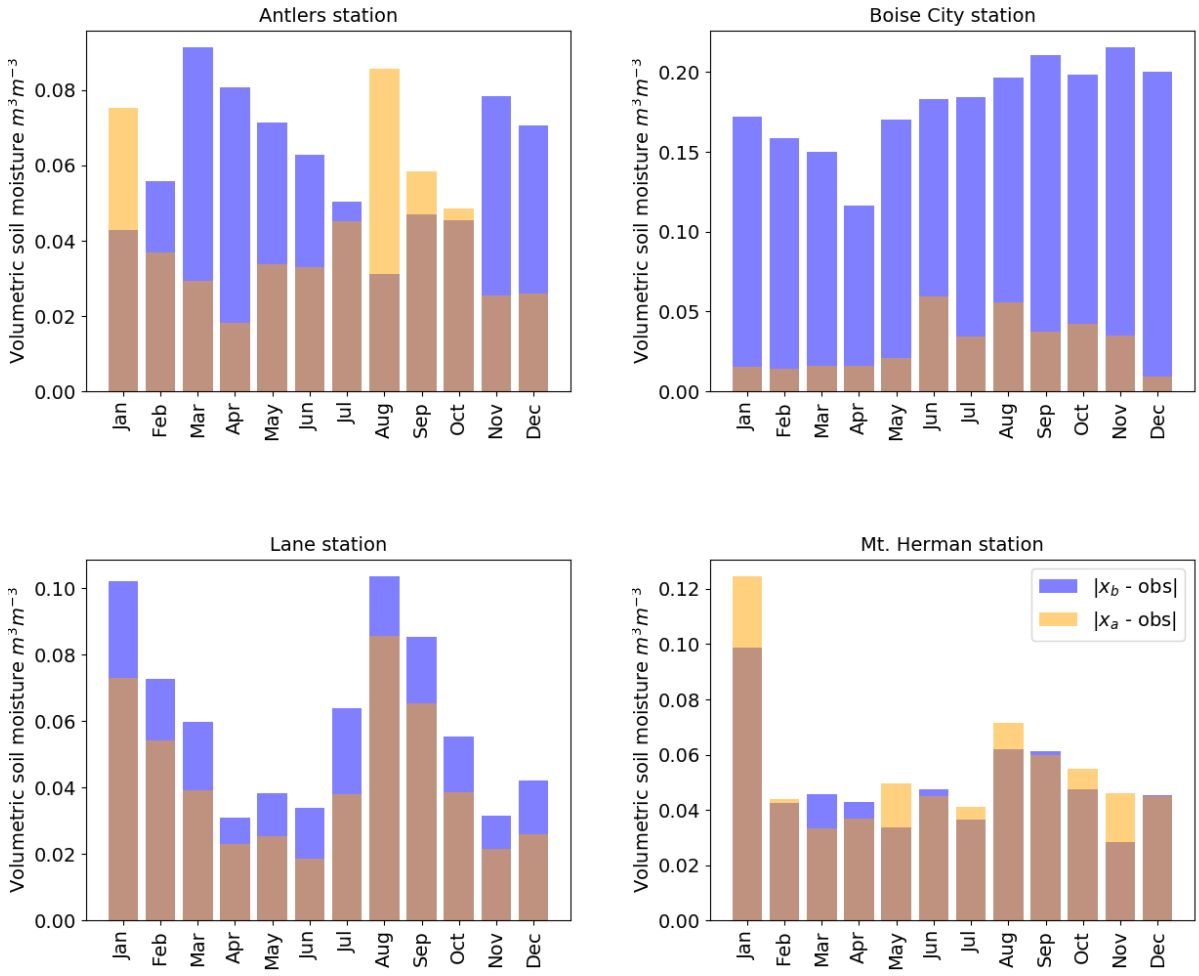


Figure 6.10: Hindcast of soil moisture from assimilating SMAP 9km resolution satellite observed soil moisture.

Based on the data assimilation experiments with *in-situ* and satellite observations, we can conclude that it is possible to constrain the JULES model soil moisture estimates via parameter estimation. Table 6.1 is a summary of the data assimilation results. Up to a 70% reduction in posterior RMSE is observed compared to the prior as a result of assimilating the SMAP satellite observed soil moisture data, in the case of Boise City station.

Table 6.1: RMSE of posterior x_a, prior x_b, hindcast x_f and open-loop model run for the following year x_o with respect to the corresponding observations.

	RMSE for Mesonet DA				RMSE for SMAP DA			
	x_a	x_b	x_f	x_o	x_a	x_b	x_f	x_o
Antlers	0.049	.071	.07	.063	.059	.073	.054	.072
Boise City	.054	.11	.063	.115	.059	.213	.039	.185
Lane	.048	.086	.071	.105	.065	.083	.054	.070
Mt Herman	.037	.095	.047	.101	.068	.065	.069	.062

Table 6.1 shows the RMSE for the prior and posterior soil moisture during the assimilation window. In addition, the RMSE of the hindcast and open-loop model run soil moisture for the following year after assimilation is presented. Generally RMSE for x_a is smaller than the RMSE of x_b, except for Mt Herman station in satellite data assimilation. Comparing prior and open-loop run soil moisture estimates, the JULES model predicts *in-situ* observations better for 2017 than 2018 but vice versa for satellite observations. As a result, for satellite data assimilation, x_f is smaller than x_a.

6.3.3 Verification of analysis and forecast soil moisture

From Figure 6.7, we have discussed results from assimilation of satellite soil moisture data. In Figure 6.11, the results are compared with ground measurements from the respective sites.

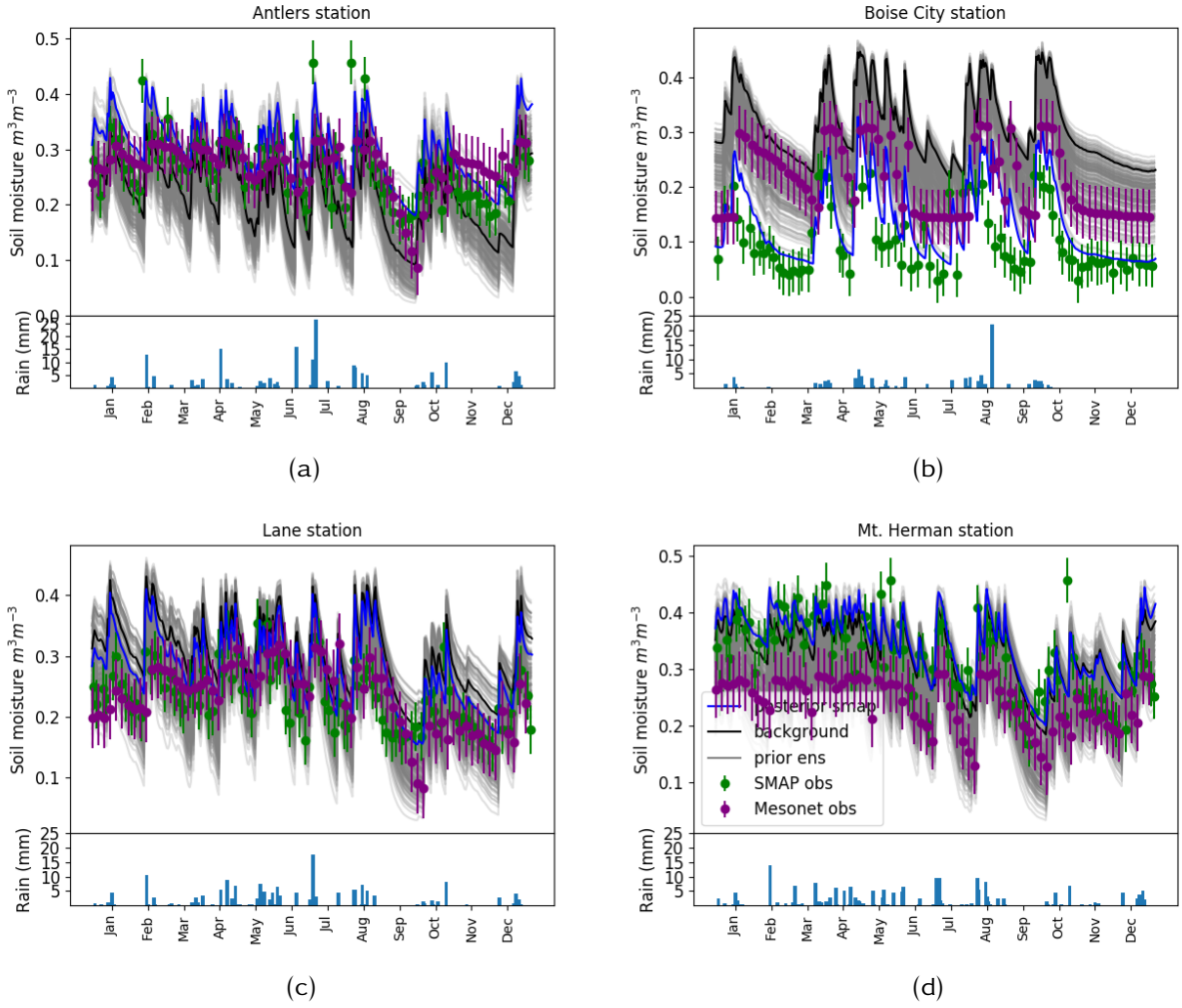


Figure 6.11: Verification of analysis soil moisture from assimilating SMAP 9km resolution satellite observed soil moisture with ground measurements from Mesonet, top layer.

Figure 6.11 shows that the posterior soil moisture estimate, obtained after assimilating the SMAP 9 km soil moisture data, are closer to the *in-situ* soil moisture data than the prior, except for Mt Herman station. For Mt Herman station, the SMAP observations were closer to the prior soil moisture estimates at the beginning, and also observations are noisier, Figure 6.11d shows that. As a result, the posterior soil moisture has not moved much towards the *in-situ* soil moisture.

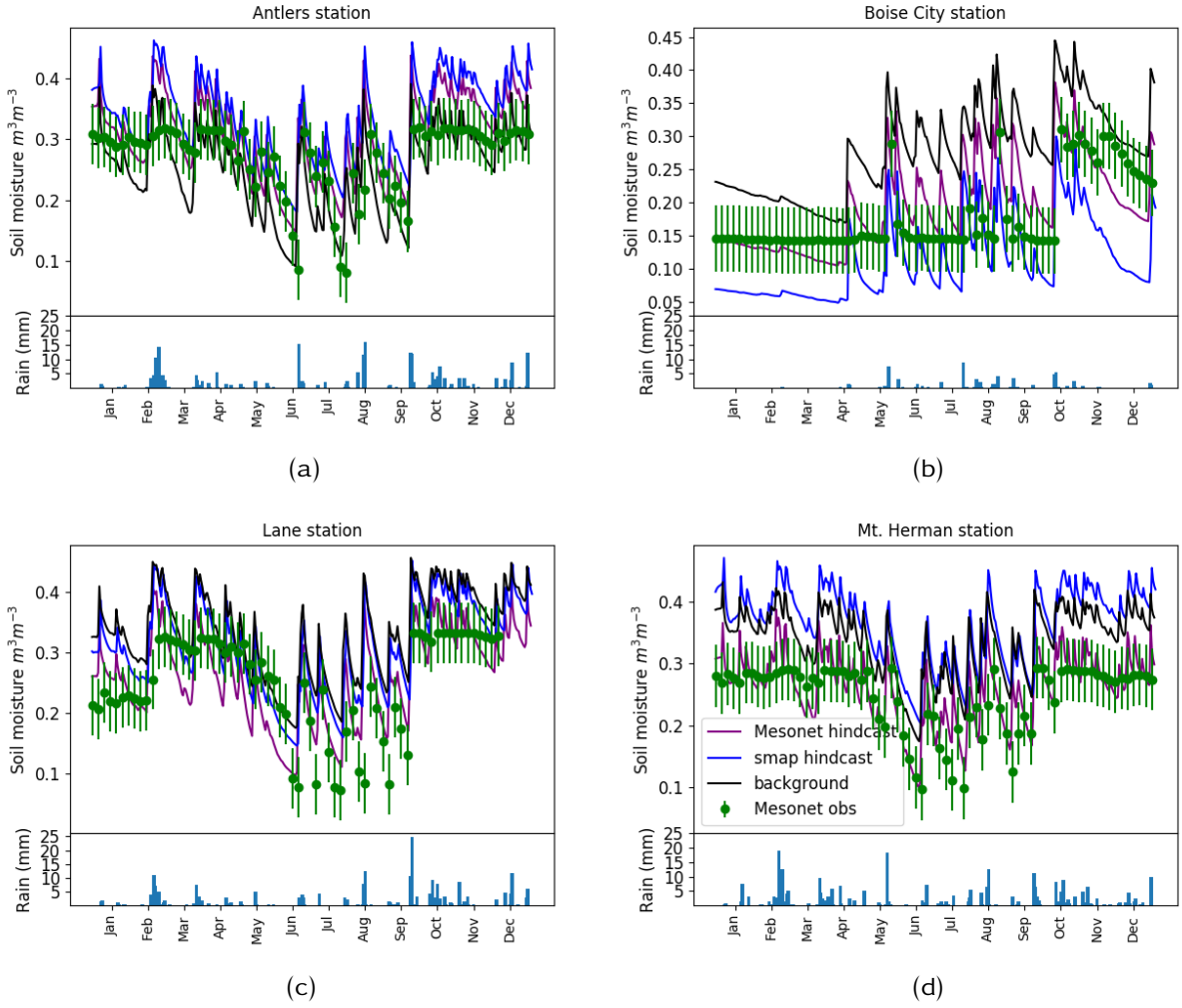


Figure 6.12: Hindcast soil moisture from assimilating SMAP 9km resolution satellite observed soil moisture with ground measurements from Mesonet, top layer.

Figure 6.12 is similar to Figure 6.11 but for the following year after the assimilation period. Figure 6.12 shows that hindcast obtained from *in-situ* soil moisture data assimilation in the previous year is closer to the *in-situ* observations of the current year except Antlers station, Figure 6.12a. Looking at the open-loop run, it is already closer to the observations. Hence, the hindcast which was obtained based on the previous year's observations, is worse than the open-loop-run.

For satellite observations, Boise city station (Figure 6.12b) and Lane station (Figure 6.12c) show improved skill hindcast. However, Antlers (Figure 6.12a) and Mt Herman (Figure 6.12d) stations show less skilful hindcast compared to the open-loop run. The case of Antlers is more of due to the fact that the open-loop itself is closer to the observations and for Mt Herman, observation noise is also a factor, as we have seen in the previous results.

6.4 Summary

In this chapter, ground measurement and satellite observation soil moisture data are assimilated in the JULES model. Soil texture parameters are updated. Based on the updated parameters, analysis soil moisture state is obtained for the assimilation window as well as forecast beyond the assimilation window. For all the experiments, the analysis soil moisture has smaller RMSE than the prior soil moisture as a result of improved posterior soil texture parameters, except Mt Herman for satellite observations. Based on posterior soil texture parameters from *in-situ* soil moisture data assimilation, retrospective soil moisture forecasts which are more skilful compared to open-loop soil moisture estimates are obtained (Table 6.1). For satellite observations, skilful soil moisture hindcast are obtained for all the stations we considered except Mt Herman station. Observations for Mt Herman station are noisy, and also the prior soil moisture estimates were closer to the observations. As a result, the forecast which incorporated information from the observations are less skilful compared to the prior soil moisture estimates.

For Mesonet sites, we have an *in-situ* soil moisture to compare with. Hence we investigate if assimilating the SMAP soil moisture data has increased the models forecast skill. Hence we compare the analysis and hindcast with *in-situ* observations. Figure 6.12 shows that the forecast obtained from assimilating satellite soil moisture data are close to ground measurements compared to the background, with Mt Herman station as an exception. This shows that the SMAP 9km resolution data represents the state of surface soil moisture with some degree of uncertainty. In the case of Mt Herman, assimilating the SMAP soil moisture resulted in worse hindcast than the open-loop JULES model run. This is due to the fact that Mt Herman station has trees in the surrounding area, which can be included for the SMAP 9k observations, which is responsible for observation noise, Figure 3.9. Based on the data assimilation experiments, we can expect the SMAP soil moisture data to be representative of the state of the soil moisture when the site is away from the woody vegetations and vice-versa, which agrees with the finding by Zhang et al. (2019a)

Chapter 7

Conclusion

As both models and data are not perfect representations of the soil moisture state separately, data assimilation has been a way of improving state estimates. As such, data assimilation experiments require determining of the model, parameters and observations uncertainties beforehand. The success of the data assimilation experiments highly depends on how those uncertainties are represented. The use of data assimilation is invaluable for sites where *in-situ* observations are sparse or absent.

In Africa, and many parts of the developing world, *in-situ* soil moisture observations are not available. The few available observations, if any, are patchy and investigation of spatial and long term variability of the state of soil moisture is impossible. On the other hand, an accurate and reliable estimate of the state of soil moisture is a crucial factor in mitigating the recurring drought lead food insecurity and related socio-economic impacts caused by extreme weather events.

This research has looked at different techniques to constrain a numerical model with ground measurement and satellite soil moisture data for a better prediction of soil moisture. Here the study sites are chosen from data-rich automated networks in Oklahoma, Mesonet so that we can verify the results. However, the methods can be used for data-sparse sites as well. In chapter 1, the following objectives were set to be addressed in this thesis.

1. **Apply stochastic forcing to generate ensemble spread for ETKF**
2. **Determine whether or not non-Gaussian distributions can be used to initialise model ensembles for 4DEnVar**

3. **Investigate the improvement of soil moisture forecast skill as a result of posterior parameters**

The following sections address each of the objectives in turn and give summary for the chapters 4, 5 and 6.

7.1 Apply stochastic forcing to generate ensemble spread for ETKF

In chapter 4, Mesonet soil moisture data is assimilated into the DRBC model using the ETKF data assimilation method. As the DRBC model is not chaotic, initial condition perturbation did not give enough spread among ensemble members, leading to ensemble collapse. This makes the data assimilation unsuccessful as observations will not have an impact, and the posterior will be determined solely by the prior. Hence this chapter implemented stochastic forcing on the basis of errors in rainfall observations and errors in the numerical model for appropriate posterior soil moisture forecast. The forecast was verified based on the RMSE and ES score, which measures the accuracy of the model estimate and the associated uncertainty. The following conclusions were drawn from chapter 4.

- Both methods of stochastic forcing, generated rainfall and model error, help to gain ensemble spread for the prior model estimates. As a result, reduction in the RMSE and ES is observed for posterior soil moisture. Comparing the two stochastic forcing methods, accounting for model error outperforms generated rainfall.
- Stochastic forcing via generated rainfall helped the DRBC model ensemble members to gain spread throughout the integration window. As a result, filter degeneracy is alleviated, and the assumed certainty of the model forecast within the DA is reduced. Hence, the contribution from observations has increased compared to the case where observed rainfall forcing was used.
- Accounting for model error makes it possible for the DRBC model to compensate for the misrepresentation of soil moisture estimates due to the missing processes, errors in parameter values or/and other sources of uncertainty. As a result, the posterior soil moisture with the imperfect model assumption is a better representation of what is observed in reality compared to a perfect model assumption.

- The appropriateness of ensemble spread and the amount of stochastic forcing, via generated rainfall and model error, is measured by the RMSE and ES. ES indicates the representativeness of ensemble spread for the ensemble mean error, where RMSE is the accuracy of the ensemble mean. The smaller the RMSE and the smaller the ES are the best measures to determine the appropriateness of ensemble spread and associated stochastic forcing.

7.2 Determine whether or not non-Gaussian distributions can be used to initialise model ensembles for 4DEnVar

In chapter 5, synthetic soil moisture observations are assimilated into the JULES model using the 4DEnVar data assimilation method. Soil texture parameters being positive and bounded, there is a need for sampling techniques where the error distribution is not Gaussian. A new method of parameter sampling using the Dirichlet distribution is implemented. Parameters are then used to initialise soil moisture ensembles and also influence the model run at each time step. Because of this, they maintain the spread on the ensemble without any additional stochastic forcing, unlike what is observed in chapter 4. To investigate the robustness of the method, different parameter backgrounds are considered in the experiments. Data assimilation results are compared with the existing methods of using the Gaussian distribution. The following points are concluded based on the experiments in chapter 5.

- The Dirichlet distribution is a viable distribution to sample soil texture parameters. For example, the parameter samples from the Dirichlet distributions are implicitly correlated; it respects the reality that the increase/decrease of one of the soil texture parameters results in increase/decrease on the other.

- Sampling from different combinations of sand silt and clay proportions, the data assimilation results show that posterior parameters moved towards the truth, in all the cases.

- Data assimilation experiments showed that the Gaussian distribution with correlations, from the Dirichlet distribution, is more consistent compared to not having correlations, especially when the background parameters are wrongly determined. The results from using the Dirichlet distribution are similar to the Gaussian distribution when both have the same means and correlations.

- Assimilating top layer soil moisture has a potential of improving soil moisture estimates for the deeper layers. In this twin experiment, constant parameter values are considered for all layers and improvements on the top layer is expected to favour the deeper layers as well. However, in practice, this set-up is not expected to bring improvements for deeper layers where soil texture parameters vary with the soil depth.

7.3 Investigate the improvement of soil moisture forecast skill as a result of posterior parameters

In Chapter 6, ground measurement and satellite observed soil moisture data are assimilated into the JULES model for a year-long assimilation window. As in chapter 5, soil texture parameters are estimated, and corresponding soil moisture estimates are obtained using the JULES model. The posterior soil moisture estimates obtained by assimilating satellite observations are compared with ground measurement soil moisture data for the respective sites. In addition to the year of assimilation, soil moisture forecast for the following year is performed based on the posterior parameters. The following conclusions were drawn:

- When *in-situ* observations are assimilated, posterior soil moisture for the assimilation window as well as a retrospective soil moisture forecasts have shown increased skill, as evidenced by a reduction in RMSE.
- When sites are less vegetated, the SMAP 9 km soil moisture data assimilation is also able to reduce the RMSE for the posterior and hindcast. However, when the site is nearby a denser vegetation, this is not the case.
- Compared to prior estimates, posterior soil moisture estimates have shown improved agreement with *in-situ* observations, at least in the pattern of the dynamics on the time series plots.
- Posterior soil moisture for the deeper layer has improved in some cases but got worse in others. This is because a uniform parameter value is considered for all the layers, which is not true in reality. We recall that we have seen improvement in all the cases in chapter 5 where synthetic observations were assimilated.

7.4 Key findings

The key findings of this thesis are the following:

- We show that both generated rainfall and model error can enhance ensemble spread and improved posterior surface soil moisture estimates. However, for the deeper layers generated rainfall alone did not give substantial improvement and considering model error is necessary. The advantage of using generated rainfall is soil moisture estimates are bounded as governed by the model physics.
- Performance of stochastic forcing using the model error covariance matrix Q is not limited to the top layer like generated rainfall, however, characterising the magnitude of Q is difficult, and soil moisture estimates could be non-physical for larger perturbations.
- We show that the Dirichlet distribution, a non-Gaussian distribution, can be used to initialise model ensembles for 4DEnVar. Automatic assignment of correlations in the covariance matrix makes the Dirichlet distribution preferable over the Gaussian distribution, apart from that both distributions resulted improved the accuracy of posterior parameters.
- Posterior parameters obtained from assimilating in-situ and satellite observations showed improvement in soil moisture forecast skills beyond the assimilation window.

7.5 Future work

Uncertainty representation is one of the key procedures for a successful data assimilation, to obtain the optimal result by combining observations with the prior knowledge represented by the model (Maggioni et al., 2012). In this thesis, we have explored techniques of error representation in the model, rainfall forcing and parameters in different data assimilation experiments. Hence we have identified the following points for further investigation.

- **Parametrising the uncertainty in the model via Q.** In chapter 4 we have considered the different magnitude of model error and compare the effect on the analysis soil moisture based on the RMSE and ES score. For this study, we managed to control the amount of added noise by ES but did not characterise the model error covariance matrix Q. For a bet-

ter representation of model error, it would be better to parametrise the covariance matrix, based on the difference between the observations and model evolution, for example.

- **Considering the uncertainty of parameters, the model and the forcing all together.** As we discussed above, neither models nor parameters are perfect. However to make to things manageable and understand the effect of each separately, in chapter 4 we assume that parameters are perfect and in chapter 5 and chapter 6 we assume a perfect model. Considering all sources of uncertainty together would be useful to make a more realistic representation of the reality and a further improvement in soil moisture prediction.

- **Implementing ensemble initialisation for each layer separately.** In chapter 5 and chapter 6 we assumed constant parameters for all the layers. The assumption did not hamper the deeper layer soil moisture prediction in chapter 5 since the observations are model output with the same assumption. However, in chapter 6 where *in-situ* and satellite observations are assimilated, we have seen cases where deeper layer soil moisture predictions got worse while the top layer improved. Hence, considering different parameters for each layer would be important to improve deeper layer soil moisture predictions.

- **Considering different observation error for different sites.** In section 6.3 we considered constant observation error for each site irrespective of the vegetation cover near the stations. However, SMAP soil moisture observations are affected by the vegetation cover (Zhang et al., 2019a), and we have seen that for one of the stations, Mt Herman station. Hence, taking the vegetation cover into account in the observation error would be helpful for better use of satellite observations.

Appendix A

Bare soil evaporation

Bare soil evaporation used in chapter 4, as given by Essery et al. (2009) is given as follows. Evaporation from the top surface E_1 is given by a product of fraction of bare soil and total surface evaporation,

$$
\begin{aligned}
E_1 &= e_1 E_s, \\
e_1 &= \frac{(1-fr)g_{soil}}{g_s} \\
fr &= \left(1 - \frac{\exp(-LAI)}{2}\right) \\
g_{soil} &= \frac{1}{100}\left(\frac{\theta_1}{\theta_c}\right)^2 \\
E_s &= (1-fa)\Psi_s E_0, \\
g_s &= g_c + (1-fr)g_{soil} \\
\Psi_s &= \frac{g_s}{g_s + ChU_1}
\end{aligned}
$$

where E_0 is potential evaporation given by

$$
E_0 = \frac{1.2}{ra}(q_{sat} - q_1), \tag{A.1}
$$

ra is aerodynamic resistance, Ch is a surface exchange coefficient between the surface and lower level atmosphere for sensible and latent heat fluxes, page 5 and 6 on Essery et al. (2009), U_1 is atmospheric wind speed, q_{sat} is saturated humidity and q_1 is specific humidity.

Appendix B

Numerical results from chapter 4

Figures B.1 - Figure B.13 are numerical results from chapter 4 which are not included in the main document are given. The general trend of the results are: generated rainfall improved posterior soil moisture for the top layer. For the deeper layers, slight improvement or negative impact was observed and considering model error was needed.

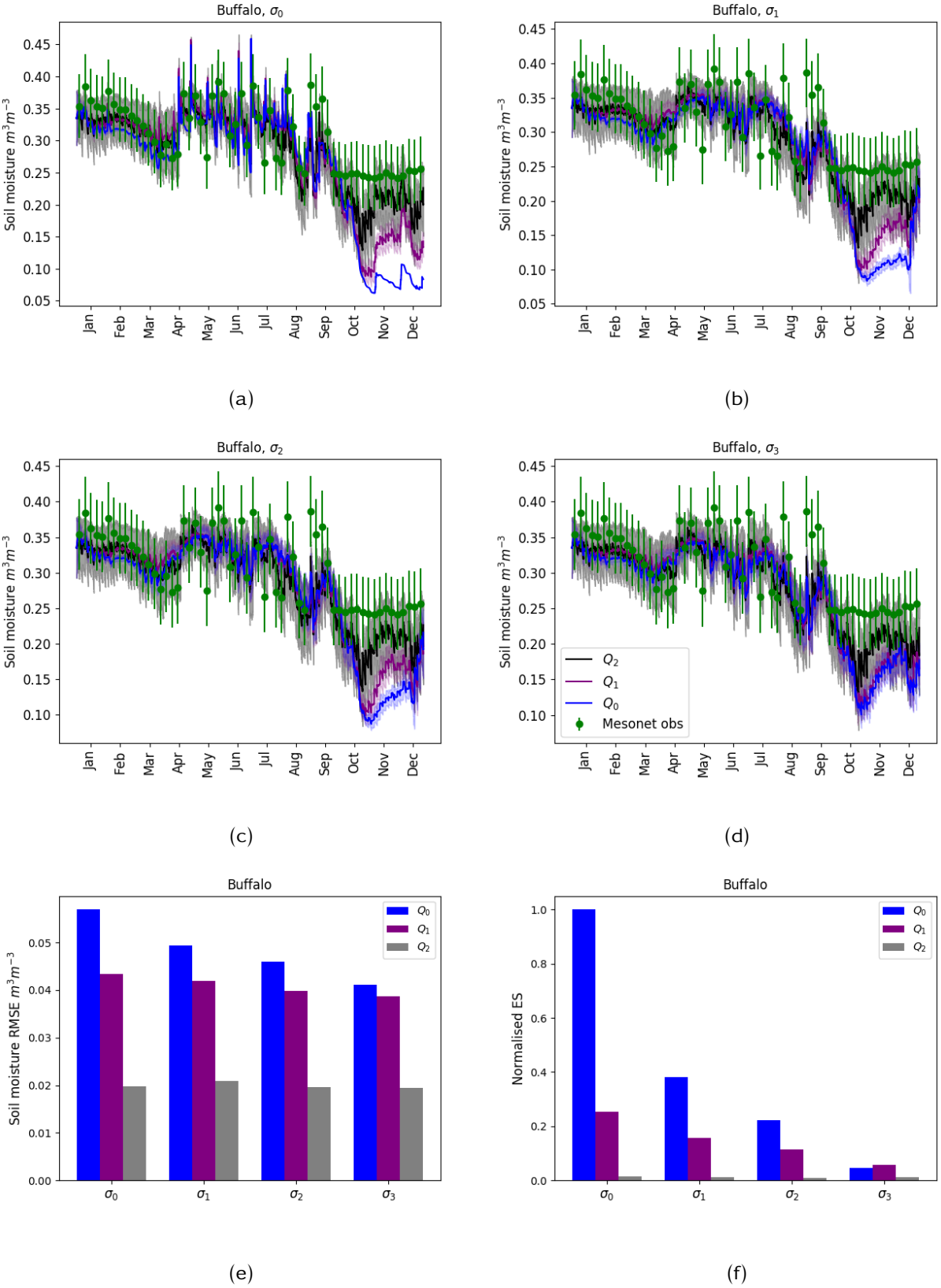


Figure B.1: Top layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Buffalo station, Oklahoma Mesonet for the year 2016.

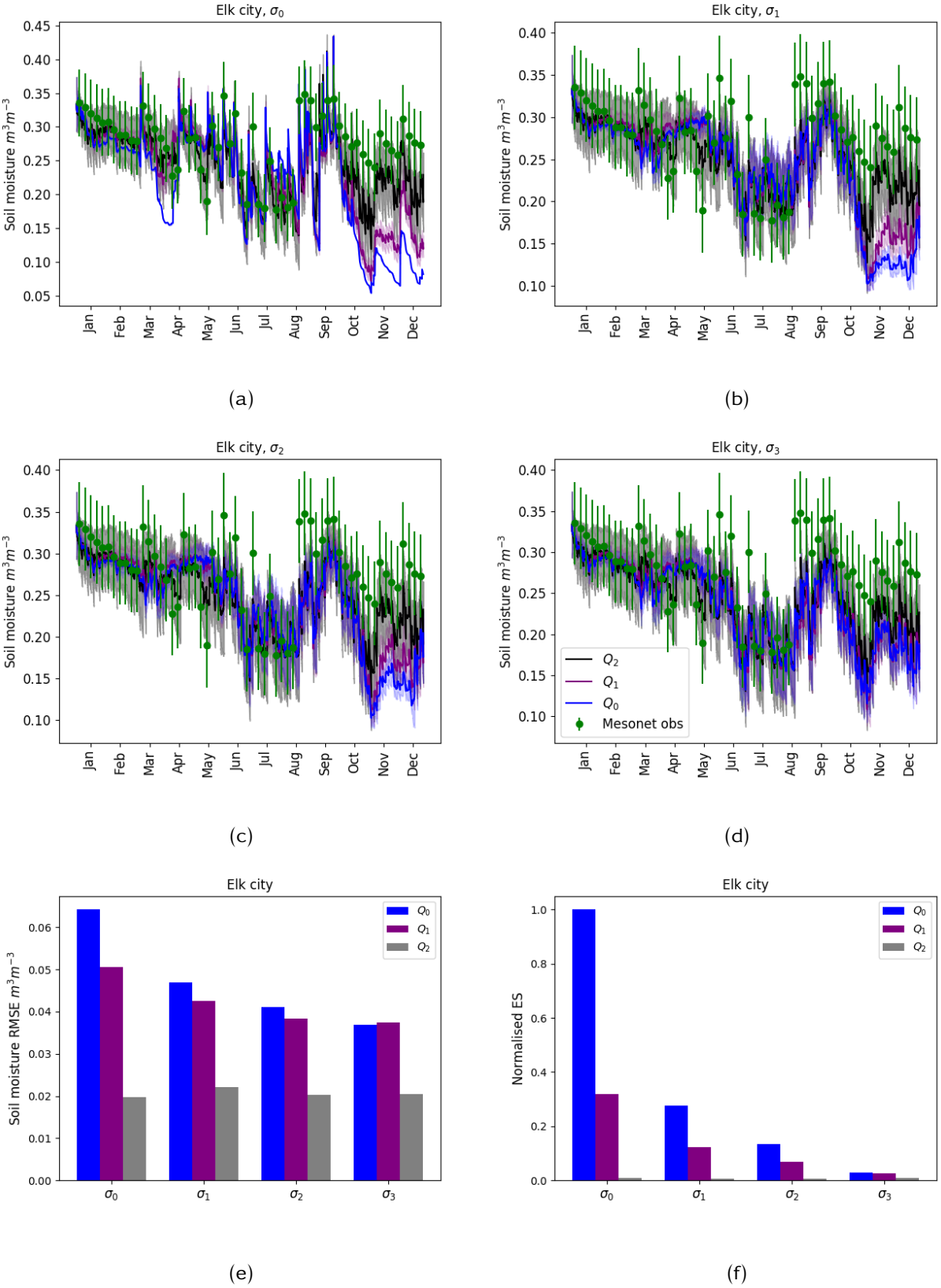


Figure B.2: Top layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Elk city station, Oklahoma Mesonet for the year 2016.

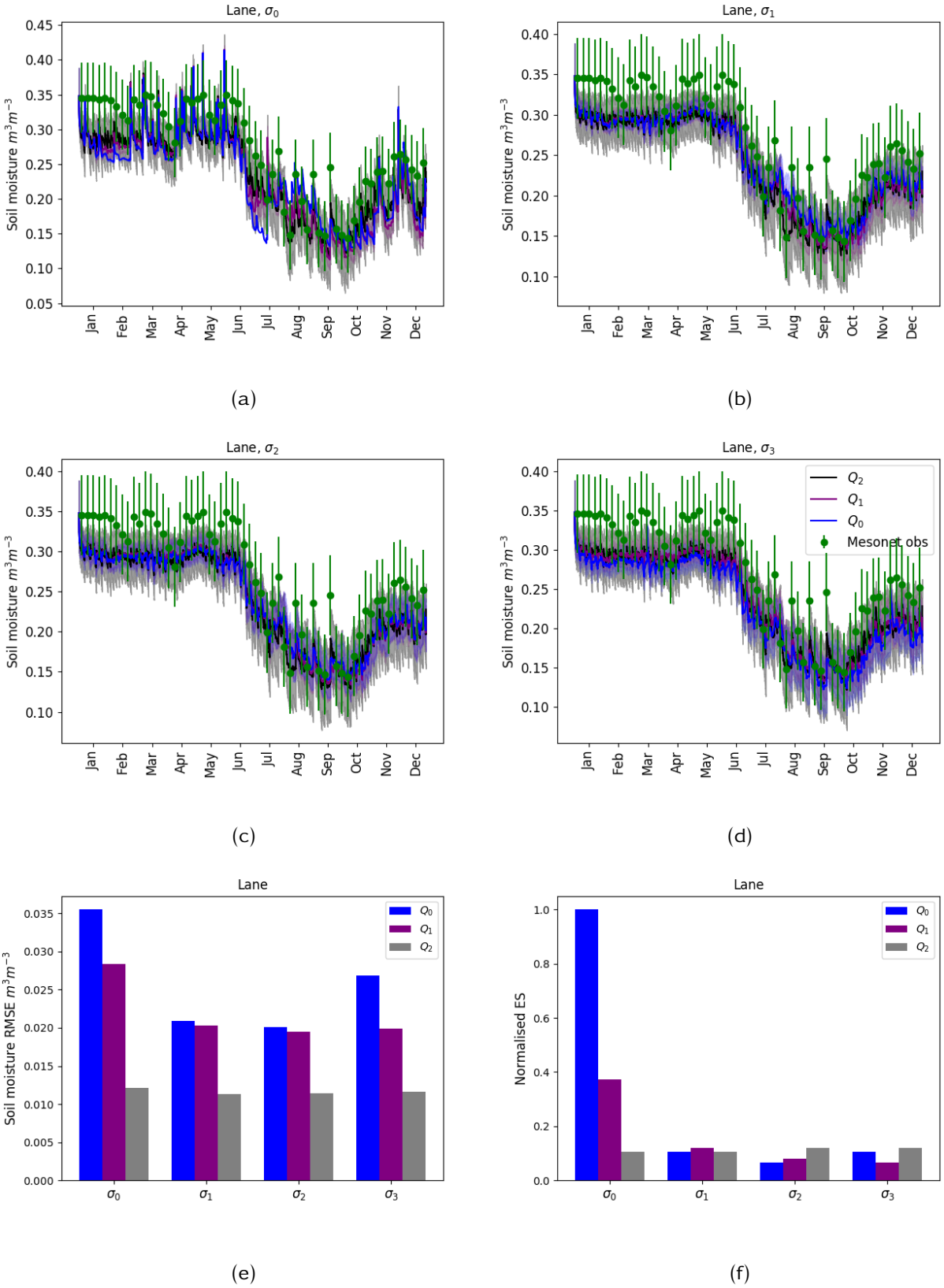


Figure B.3: Top layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Lane station, Oklahoma Mesonet for the year 2016.

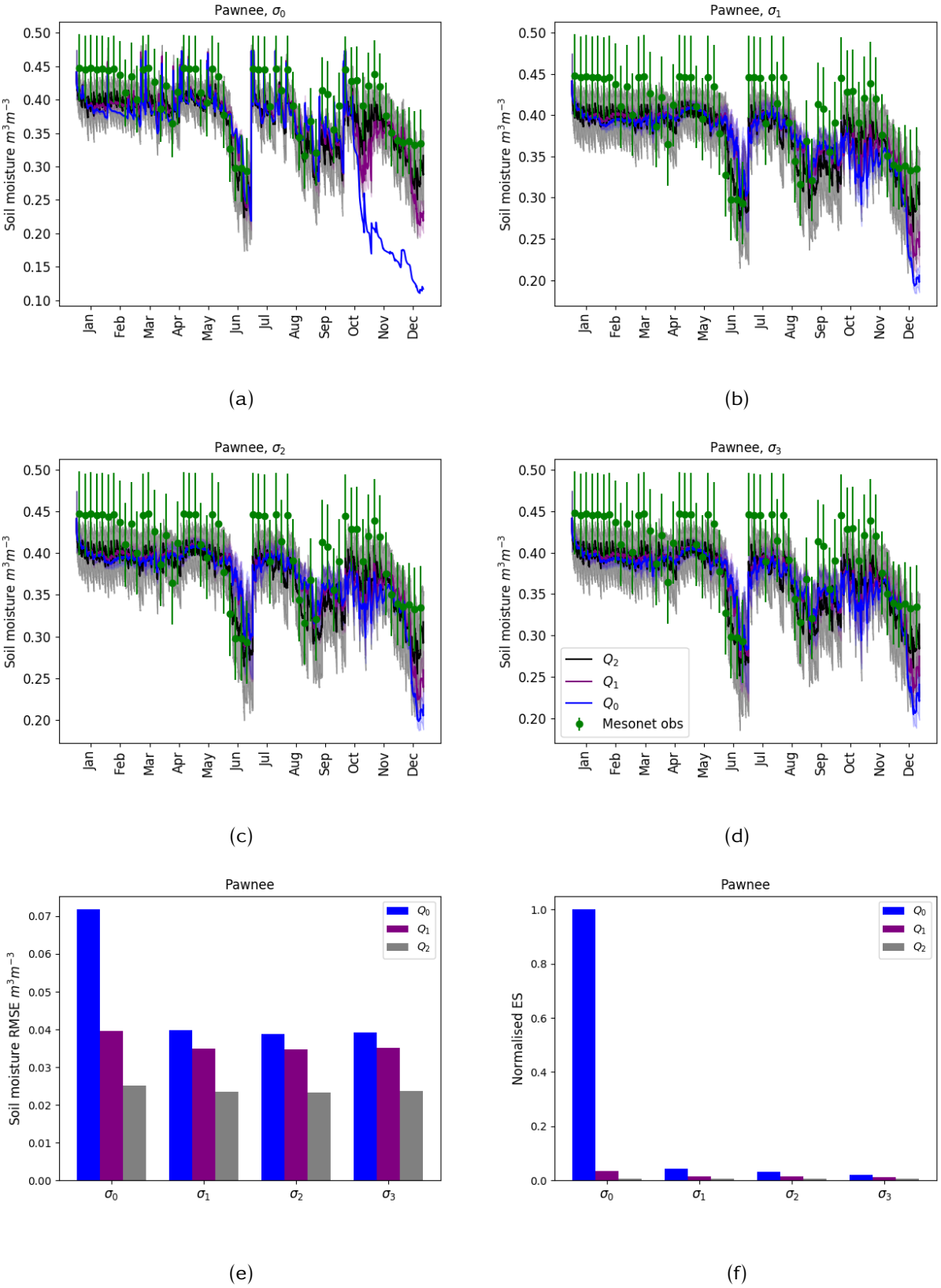


Figure B.4: Top layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Pawnee station, Oklahoma Mesonet for the year 2016.

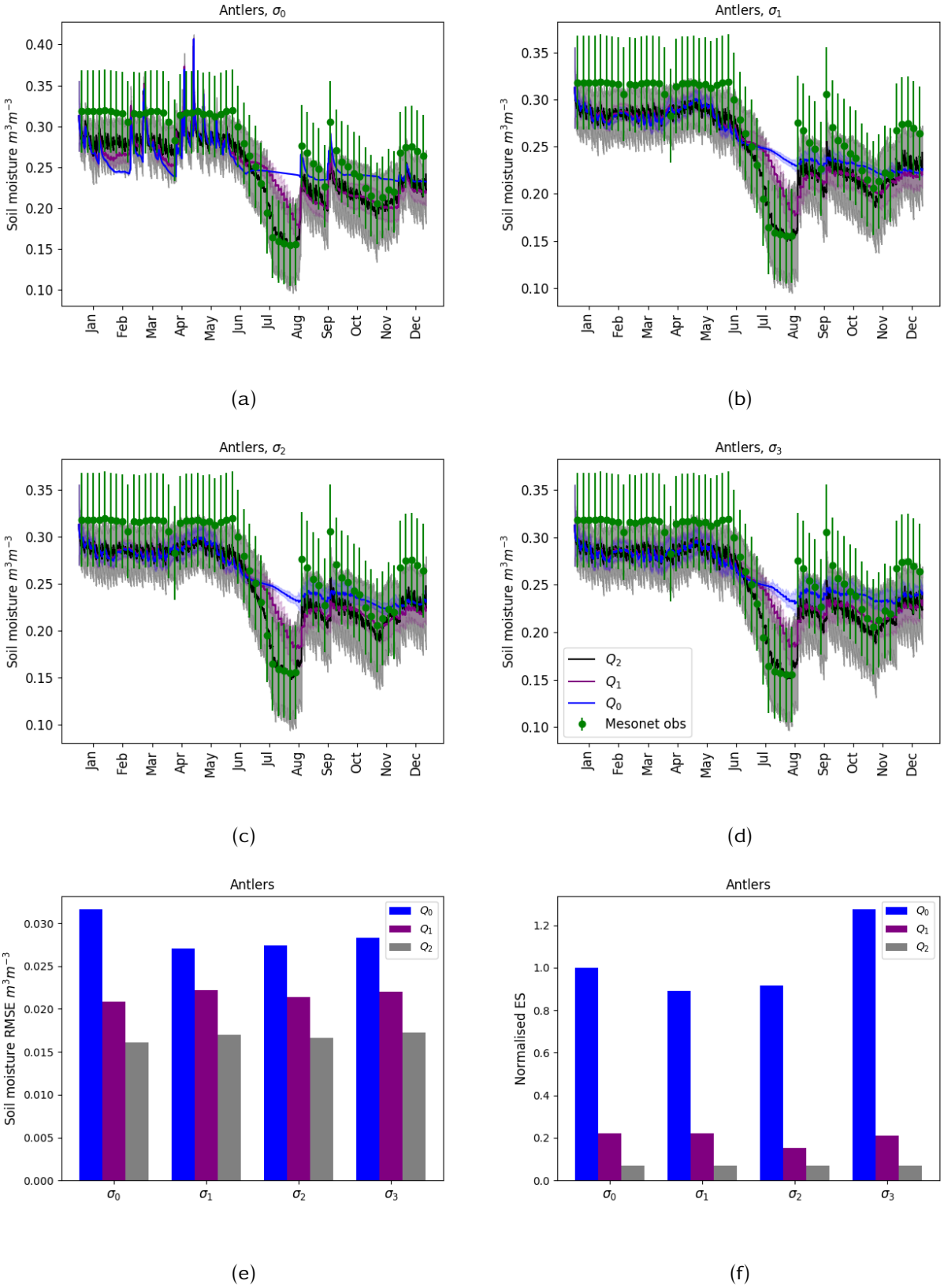


Figure B.5: Second soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Antlers station, Oklahoma Mesonet for the year 2016.

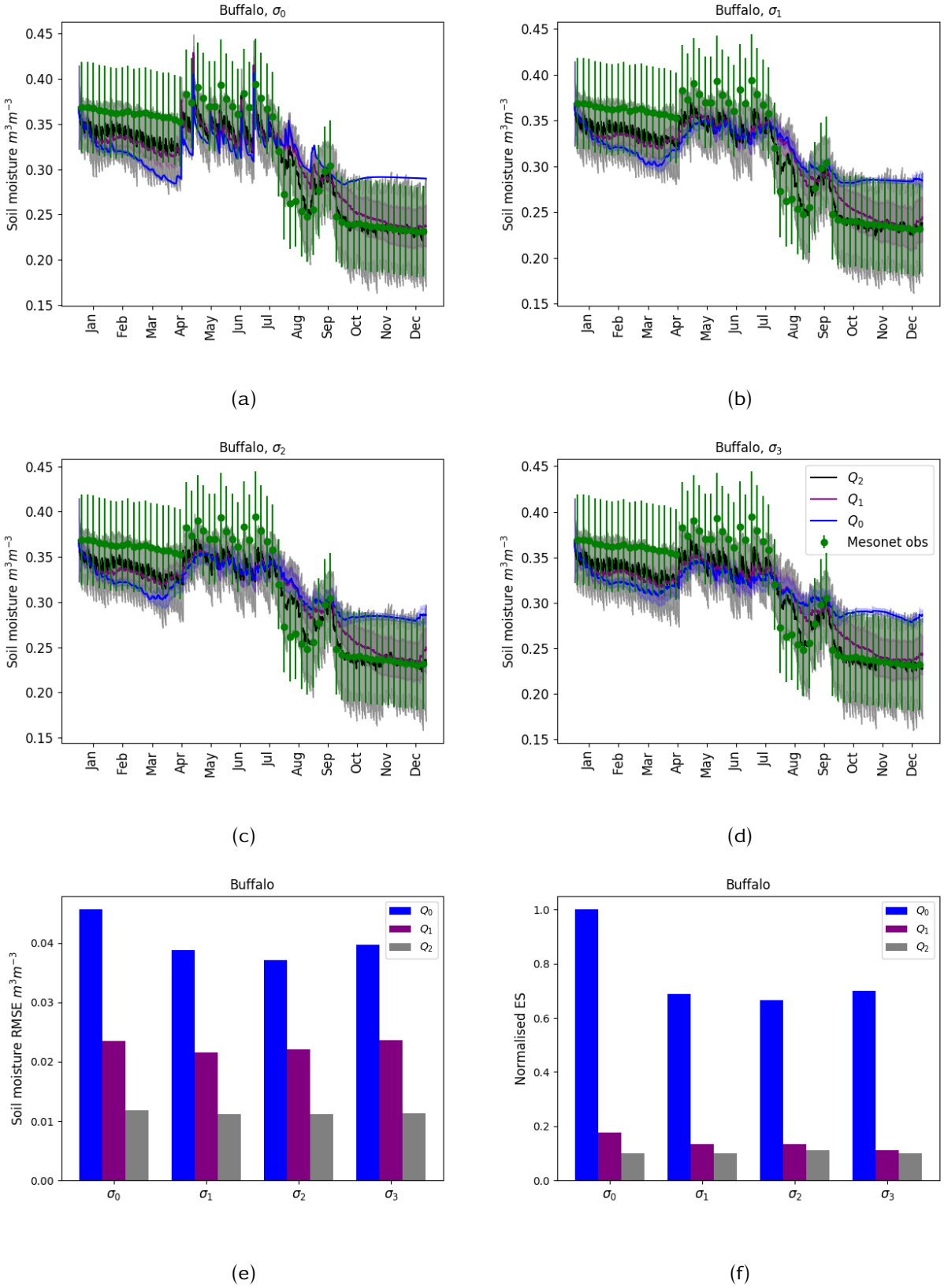


Figure B.6: Second soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Buffalo station, Oklahoma Mesonet for the year 2016.

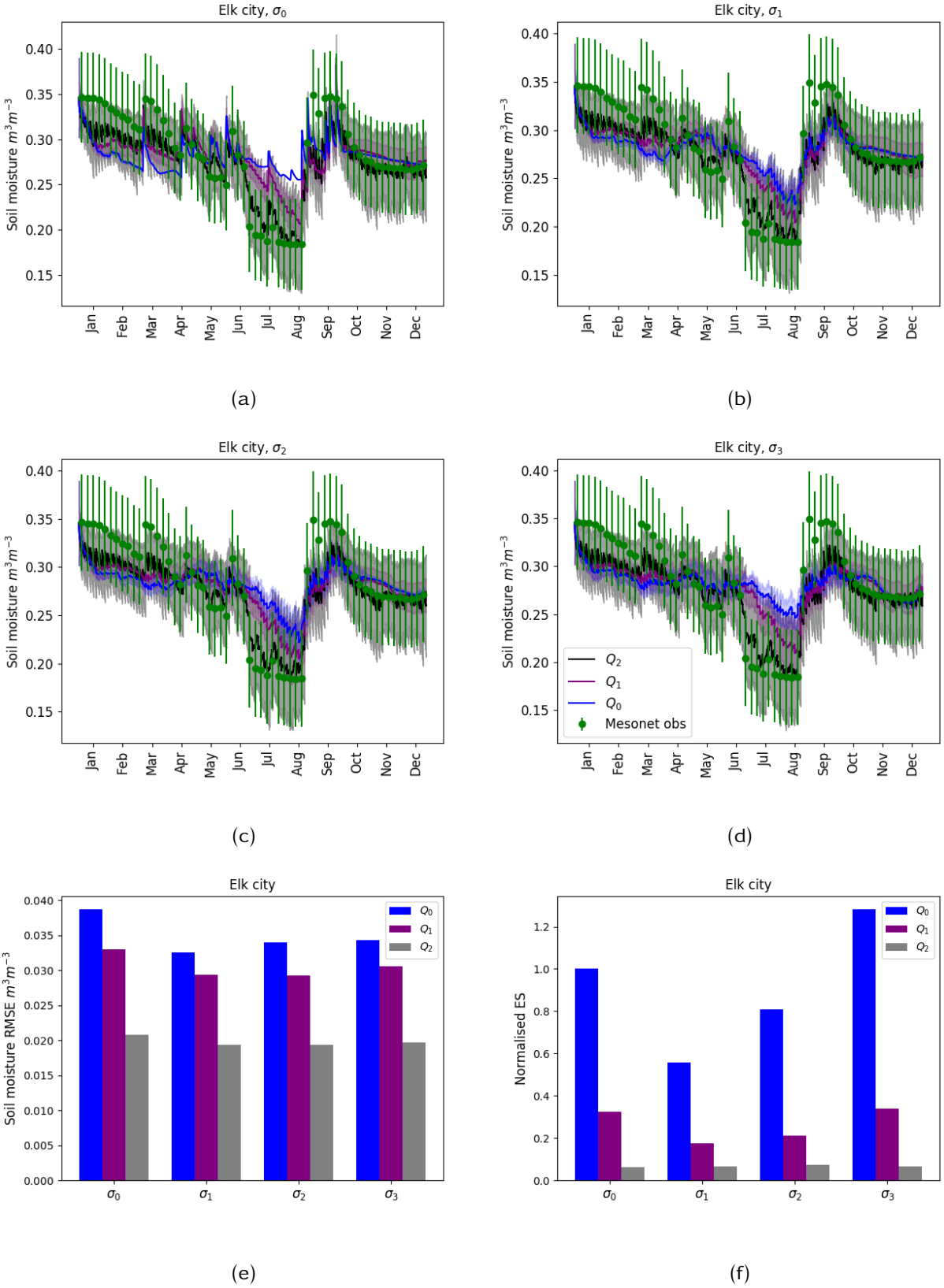


Figure B.7: Second soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Elk city station, Oklahoma Mesonet for the year 2016.

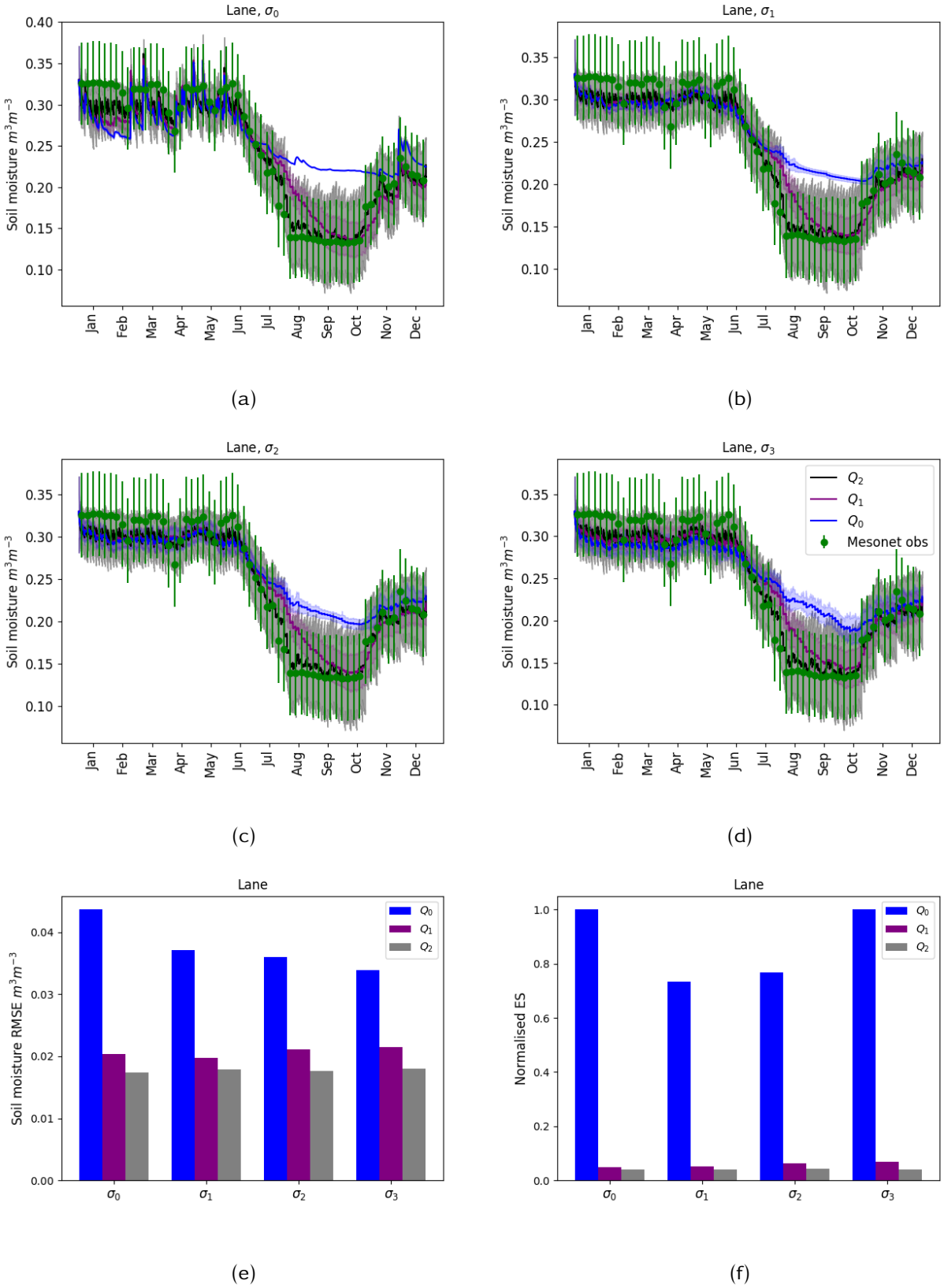


Figure B.8: Second soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Lane station, Oklahoma Mesonet for the year 2016.

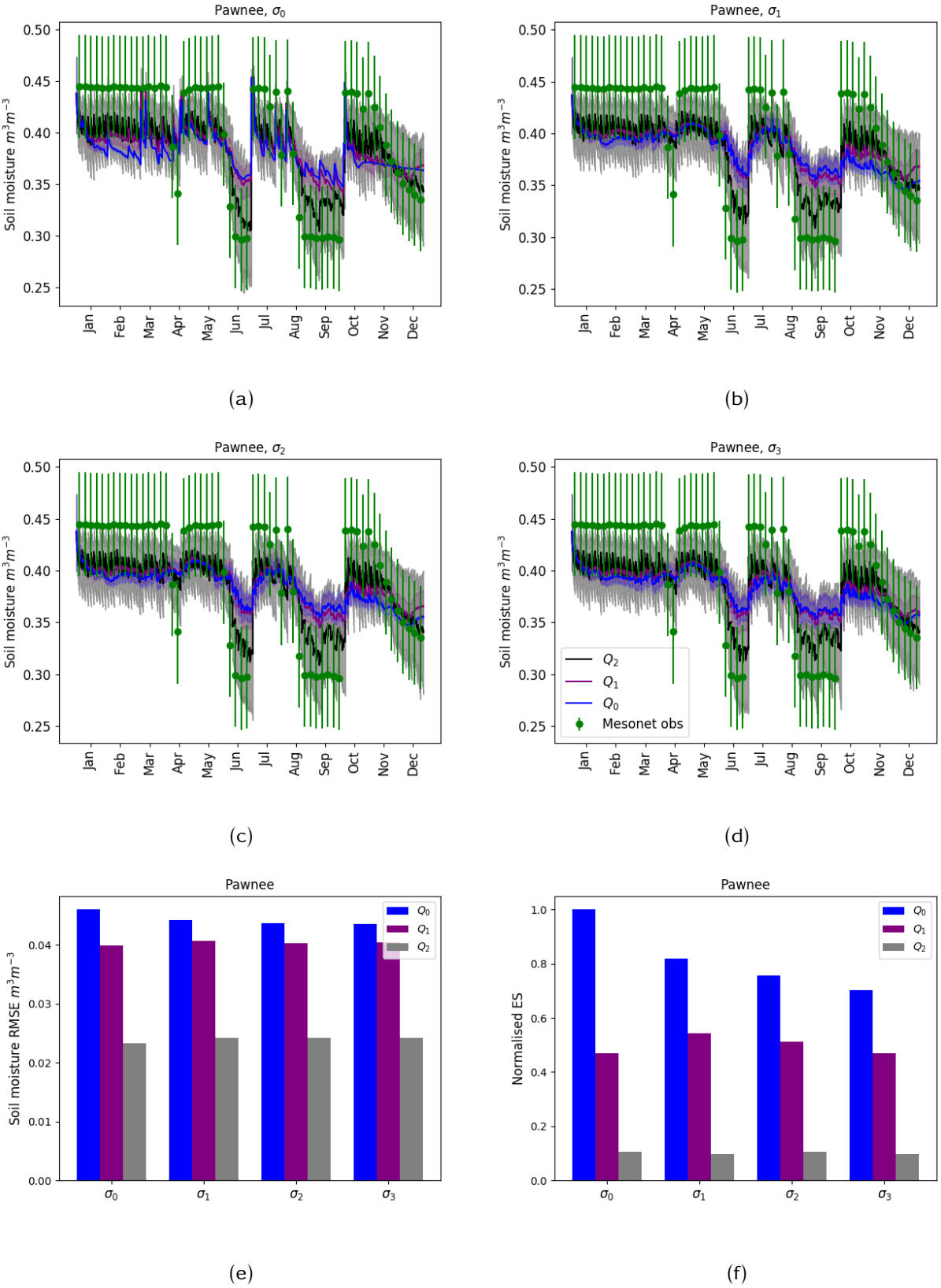


Figure B.9: Second soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Pawnee station, Oklahoma Mesonet for the year 2016.

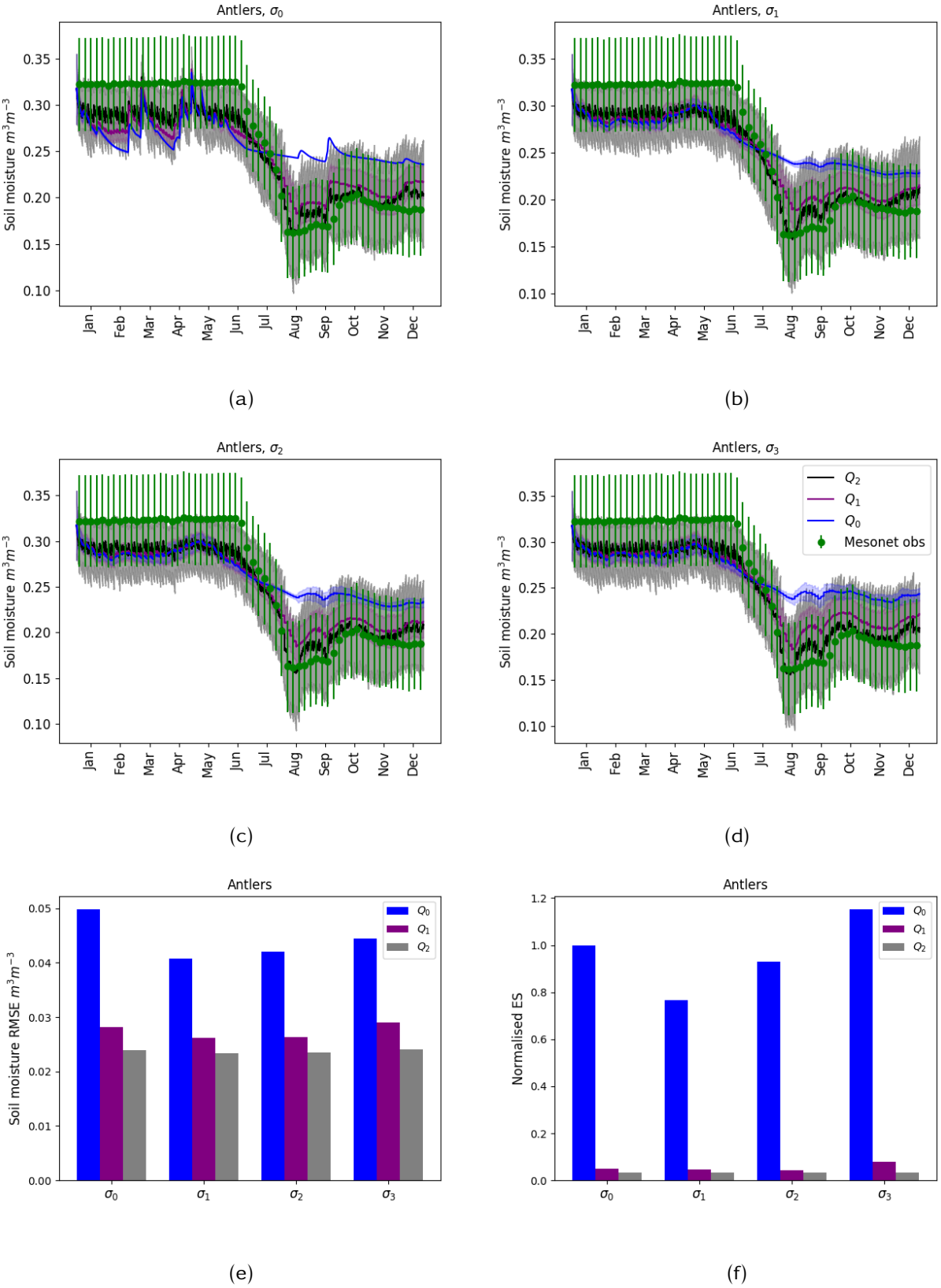


Figure B.10: Third soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Antlers station, Oklahoma Mesonet for the year 2016.

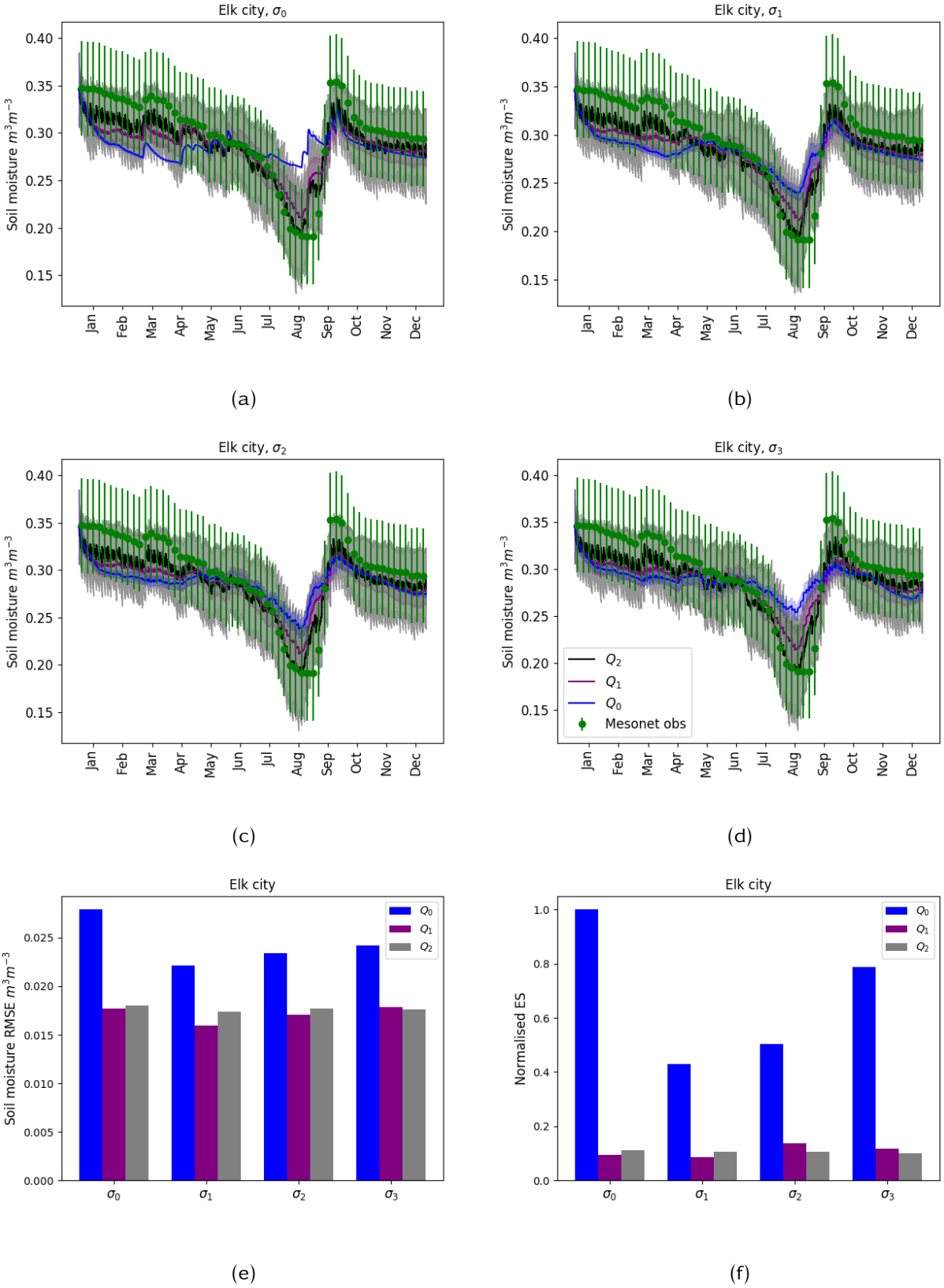


Figure B.11: Third soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is $\pm$ 1 std from the mean. The forcing data and parameter values are from Elk city station, Oklahoma Mesonet for the year 2016.

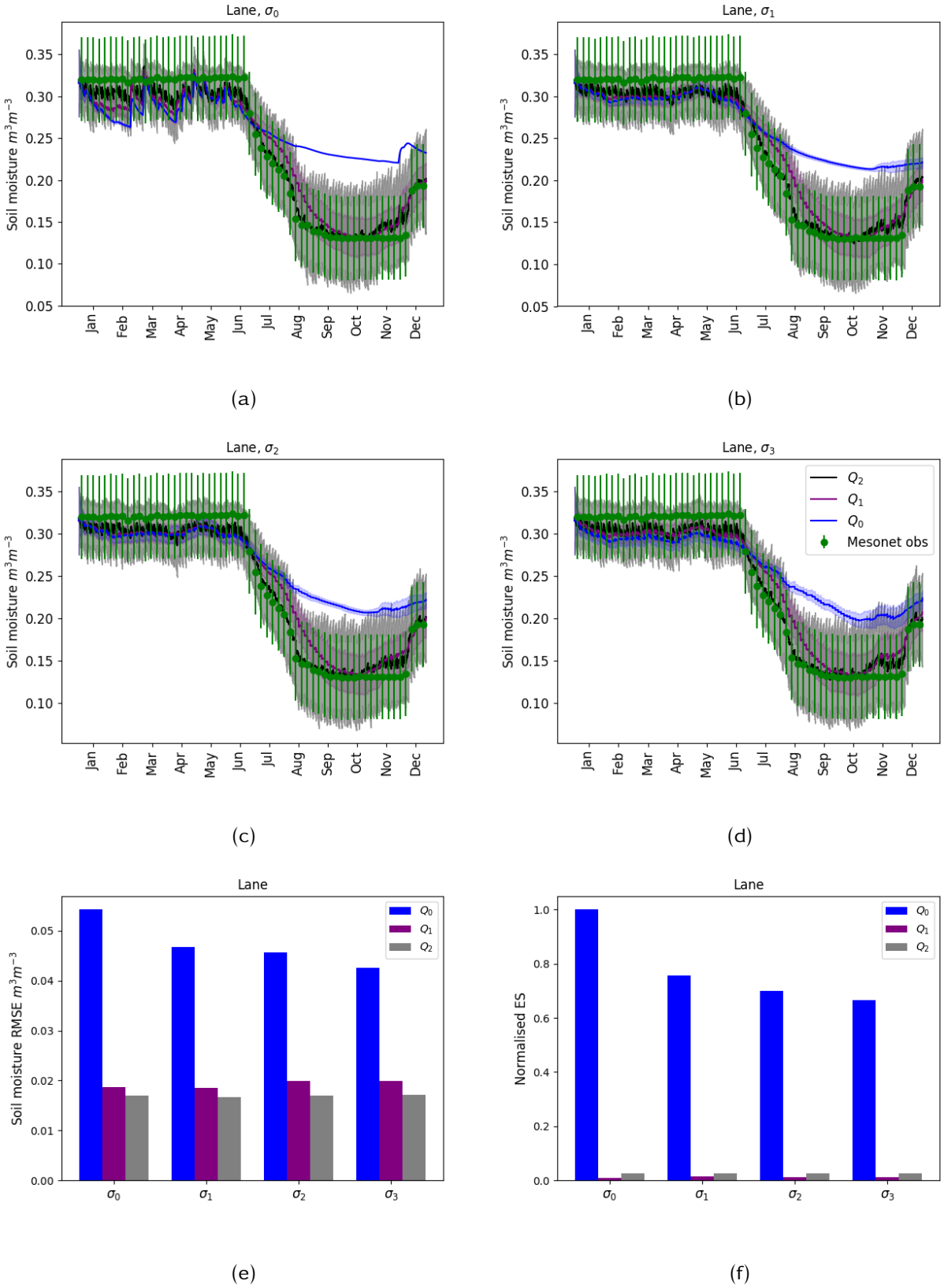


Figure B.12: Third soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Lane station, Oklahoma Mesonet for the year 2016.

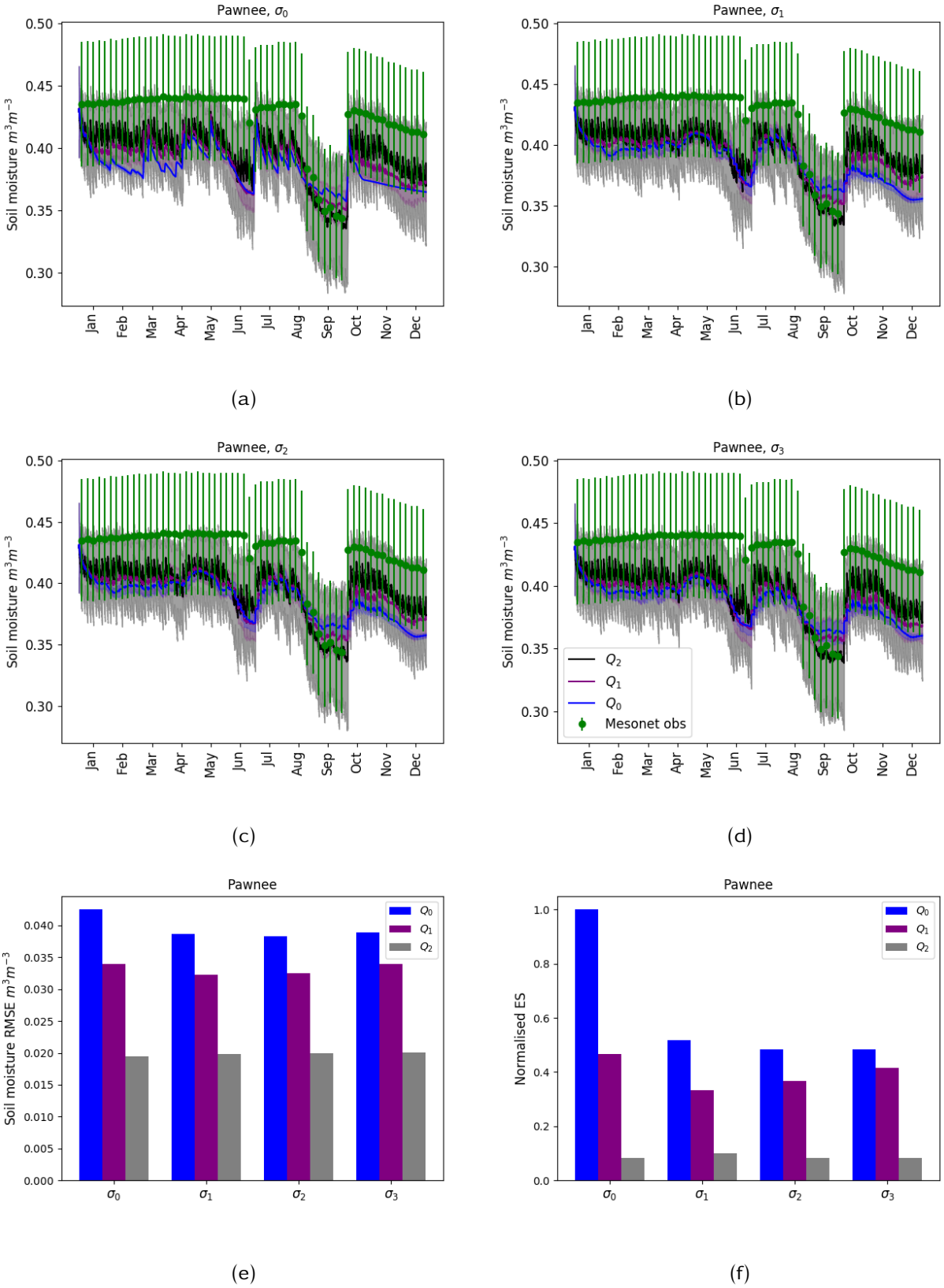


Figure B.13: Third soil layer posterior soil moisture ensemble mean for different values of Q and σ. The shades for each mean is ± 1 std from the mean. The forcing data and parameter values are from Pawnee station, Oklahoma Mesonet for the year 2016.

Bibliography

Bannister, R. N., 2017: A review of operational methods of variational and ensemble-variational data assimilation. *Q. J. R. Meteorol. Soc.*, **143 (703)**, 607–633, doi:10.1002/qj.2982, 1601.03446.

Basinger, M., F. Montalto, and U. Lall, 2010: A rainwater harvesting system reliability model based on nonparametric stochastic rainfall generator. *J. Hydrol.*

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55, doi:10.1038/nature14956.

Baugh, C., P. de Rosnay, H. Lawrence, T. Jurlina, M. Drusch, E. Zsoter, and C. Prudhomme, 2020: The impact of smos soil moisture data assimilation within the operational global flood awareness system (GloFAS). *Remote Sens.*, **12 (9)**, doi:10.3390/RS12091490.

Best, M., et al., 2011: The Joint UK Land Environment Simulator (JULES), model description. Part 1: Energy and water fluxes. *Geosci. Model Dev.*, **4**, 677–699.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. *Mon. Weather Rev.*, **129 (3)**, 420–436.

Black, E., E. Tarnavsky, R. Maidment, H. Greatrex, A. Mookerjee, T. Quaife, and M. Brown, 2016: The use of remotely sensed rainfall for managing drought risk: A case study of weather index insurance in Zambia. *Remote Sens.*, **8 (4)**, doi:10.3390/rs8040342.

Boyd, E., R. J. Cornforth, P. J. Lamb, A. Tarhule, M. Issa Lélé, and A. Brouder, 2013: Building resilience to face recurring environmental crisis in African Sahel. *Nat. Clim. Chang.*, **3 (7)**, 631–637, doi:10.1038/nclimate1856.

Browne, P. A. and S. Wilson, 2015: A simple method for integrating a complex model into an ensemble data assimilation system using MPI. *Environ. Model. Softw.*, **68**, 122–128.

Buehner, M., P. L. Houtekamer, C. Charette, H. L. Mitchell, and B. He, 2010: Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic NWP. Part I: Description and Single-Observation Experiments. *Mon. Weather Rev.*, **138 (5)**, 1550–1566.

Chaubell, J., 2016: Soil Moisture Active Passive (SMAP) Project Algorithm Theoretical Basis Document SMAP L1B Enhancement Radiometer Brightness Temperature Data Product. Tech. rep., Jet Propulsion Laboratory, California, California Institute of Technology, 1–24 pp.

Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2015: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Q. J. R. Meteorol. Soc.*, **141 (687)**, 538–549.

Clark, D. B., et al., 2011: The Joint UK Land Environment Simulator (JULES), model description Part 2: Carbon fluxes and vegetation dynamics. *Geosci. Model Dev.*, **4 (3)**, 701–722, doi:10.5194/gmd-4-701-2011.

Cooper, P. J. M., J. Dimes, K. P. C. Rao, B. Shapiro, B. Shiferaw, and S. Twomlow, 2008: Coping better with current climatic variability in the rain-fed farming systems of sub-Saharan Africa: An essential first step in adapting to future climate change? *Agric. Ecosyst. Environ.*, **126 (1-2)**, 24–35.

Darvishi, M. and G. Ahmadi, 2014: Data assimilation techniques and modelling uncertainty in geosciences. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, **40 (2W3)**, 85–90.

Dee, D. P., et al., 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, **137 (656)**, 553–597, doi:10.1002/qj.828.

Dorigo, W., A. Gruber, P. Van Oevelen, W. Wagner, M. Drusch, S. Mecklenburg, A. Robock, and T. Jackson, 2011a: The international soil moisture network - An observational network for soil moisture product validations. *34th Int. Symp. Remote Sens. Environ. - GEOSS Era Towar. Oper. Environ. Monit.*, 2–5.

Dorigo, W. A., et al., 2011b: The International Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements. *Hydrol. Earth Syst. Sci.*, **15 (5)**, 1675–1698, doi:10.5194/hess-15-1675-2011.

Dunne, S. and D. Entekhabi, 2005: An ensemble-based reanalysis approach to land data assimilation. *Water Resour. Res.*, **41 (2)**, 1–18, doi:10.1029/2004WR003449.

Engda, T. A. and T. J. Kelleners, 2016: Soil moisture-based drought monitoring at different time scales: A case study for the U.S. Great Plains. *J. Am. Water Resour. Assoc.*, **52 (1)**, 77–88.

Essery, R., M. Best, and P. Cox, 2009: JULES Technical Documentation MOSES 2 . 2 Technical Documentation. Tech. rep., Met Office.

Ettema, J. and P. Viterbo, 2001: ECMWF soil moisture data assimilation The simplified Extended Kalman Filter analysis. Tech. rep., ECMWF, 97–104 pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99 (C5)**, 10 143–10 162, doi:10.1029/94JC00572.

Fairbairn, D., 2009: Comparison of the Ensemble Transform Kalman Filter with the Ensemble Transform Kalman Smoother. **(August)**, URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.6139{&}rep=rep1{&}type=pdf.

Fairbairn, D., S. R. Pring, A. C. Lorenc, and I. Roulstone, 2014: A comparison of 4DVar with ensemble data assimilation methods. *Q. J. R. Meteorol. Soc.*, **140 (678)**, 281–294.

Fandel, C. A., D. D. Breshears, and E. E. McMahon, 2018: Implicit assumptions of conceptual diagrams in environmental science and best practices for their illustration. *Ecosphere*, **9 (1)**, doi:10.1002/ecs2.2072.

Fuhlendorf, S. D. and D. M. Engle, 2004: Application of the fire-grazing interaction to restore a shifting mosaic on tallgrass prairie. *J. Appl. Ecol.*, **41 (4)**, 604–614, doi:10.1111/j.0021-8901.2004.00937.x.

Green, J. K., S. I. Seneviratne, A. M. Berg, K. L. Findell, S. Hagemann, D. M. Lawrence, and P. Gentine, 2019: Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature*, **565 (7740)**, 476–479, doi:10.1038/s41586-018-0848-x, URL http://dx.doi.org/10.1038/s41586-018-0848-x.

Hamill, T. M. and J. S. Whitaker, 2011: What constrains spread growth in forecasts initialized from ensemble Kalman filters? *Mon. Weather Rev.*, **139 (1)**, 117–131, doi:10.1175/2010MWR3246.1.

Han, E., W. T. Crow, T. Holmes, and J. Bolten, 2014: Benchmarking a Soil Moisture Data Assimilation System for Agricultural Drought Monitoring. *J. Hydrometeorol.*, **15 (3)**, 1117–1134.

Han, E., V. Merwade, and G. C. Heathman, 2012: Implementation of surface soil moisture data assimilation with watershed scale distributed hydrological model. *J. Hydrol.*, **416-417**, 98–117, doi:10.1016/j.jhydrol.2011.11.039, URL http://dx.doi.org/10.1016/j.jhydrol.2011.11.039.

Heathman, G. C., P. J. Starks, L. R. Ahuja, and T. J. Jackson, 2003: Assimilation of surface soil moisture to estimate profile soil water content. *J. Hydrol.*, **279 (1-4)**, 1–17, doi:10.1016/S0022-1694(03)00088-X.

Houborg, R., M. Rodell, B. Li, R. Reichle, and B. F. Zaitchik, 2012: Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations. *Water Resour. Res.*, **48 (7)**.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Phys. D Nonlinear Phenom.*, **230 (1-2)**, 112–126.

Hyndman, R. J. and G. K. Grunwald, 2000: Applications: Generalized Additive Modelling of Mixed Distribution Markov Models with Application to Melbourne's Rainfall. *Aust. N. Z. J. Stat.*, **42 (2)**, 145–158.

Illston, B. G., J. Basara, D. K. Fischer, R. L. Elliott, C. Fiebrich, K. C. Crawford, K. S. Humes, and E. Hunt, 2008: Mesoscale monitoring of soil mositure across a statewide network. *J. Atmos. Ocean. Technol.*, **25 (2)**, 167–182.

Kalman, R. E., 1960: A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.*, **82 (1)**, 35, doi:10.1115/1.3662552, URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:A+New+Approach+to+Linear+Filtering+and+Prediction+Problems{#}0{%}5Cnhttp://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430402.

Karthikeyan, L., M. Pan, N. Wanders, D. N. Kumar, and E. F. Wood, 2017: Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms. *Adv. Water Resour.*, **109**, 106–120, doi:10.1016/j.advwatres.2017.09.006, URL http://dx.doi.org/10.1016/j.advwatres.2017.09.006.

Köpken, C., G. Kelly, and J. N. Thépaut, 2004: Assimilation of Meteosat radiance data within the 4D-Var system at ECMWF: Assimilation experiments and forecasts impact. *Q. J. R. Meteorol. Soc.*, **130 (601 PART B)**, 2277–2292, doi:10.1256/qj.02.230.

Koster, R. D., Z. Guo, G. Bonan, E. Chan, and P. Cox, 2014: Regions of Strong Coupling Between Soil Moisture and Precipitation. *Science (80-.).*, **1138 (2004)**, 10–13.

Lawless, A. S., 2013: Variational data assimilation for very large environmental problems. *Large Scale Inverse Probl.*, De Gruyter, 1–37, doi:10.1515/9783110282269.55.

Legates, D. R., R. Mahmood, D. F. Levia, T. L. DeLiberty, S. M. Quiring, C. Houser, and F. E. Nelson, 2011: Soil moisture: A central and unifying theme in physical geography. *Prog. Phys. Geogr.*, **35 (1)**, 65–86, doi:10.1177/0309133310386514.

Leutbecher, M. and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227 (7)**, 3515–3539.

Lievens, H., et al., 2015: SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia. *Remote Sens. Environ.*, **168**, 146–162, doi:10.1016/j.rse.2015.06.025, URL http://dx.doi.org/10.1016/j.rse.2015.06.025.

Lin, J., 2016: by. Ph.D. thesis, Queen's University.

Liu, C., Q. Xiao, and B. Wang, 2008: An Ensemble-Based Four-Dimensional Variational Data Assimilation Scheme. Part I: Technical Formulation and Preliminary Test. *Mon. Weather Rev.*, **136 (9)**, 3363–3373.

Liu, C., Q. Xiao, and B. Wang, 2009: An Ensemble-Based Four-Dimensional Variational Data Assimilation Scheme. Part II: Observing System Simulation Experiments with Advanced Research WRF (ARW). *Mon. Weather Rev.*, **137 (5)**, 1687–1704.

Liu, P. W., T. Bongiovanni, A. Monsivais-Huertero, J. Judge, S. Steele-Dunne, R. Bindlish, and T. J. Jackson, 2016: Assimilation of Active and Passive Microwave Observations for Improved Estimates of Soil Moisture and Crop Growth. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **9 (4)**, 1357–1369, doi:10.1109/JSTARS.2015.2506504.

Liu, Z., M. Notaro, and R. Gallimore, 2010: Indirect vegetation-soil moisture feedback with application to Holocene North Africa climate1. *Glob. Chang. Biol.*, **16 (6)**, 1733–1743, doi:10.1111/j.1365-2486.2009.02087.x.

Long, D. G. and M. J. Brodzik, 2016: Optimum Image Formation for Spaceborne Microwave Radiometer Products. *IEEE Trans. Geosci. Remote Sens.*, **54 (5)**, 2763–2779, doi:10.1109/TGRS.2015.2505677.

Lu, H., T. Koike, and P. Gong, 2011: Monitoring soil moisture change in Africa over past 20 years with using passive microwave remote sensing. *Proc. - 2011 19th Int. Conf. Geoinformatics, Geoinformatics 2011*, 1–5, doi:10.1109/GeoInformatics.2011.5980961.

Luo, L., J. Sheffield, and E. F. Wood, 2008: Towards a Global Drought Monitoring and Forecasting Capability. *Environ. Eng.*, 1–9.

Luo, X. and I. Hoteit, 2013: Covariance inflation in the ensemble kalman filter: A residual nudging perspective and some implications. *Mon. Weather Rev.*, **141 (10)**, 3360–3368, doi:10.1175/MWR-D-13-00067.1.

Maggioni, V., R. H. Reichle, and E. N. Anagnostou, 2012: The Impact of Rainfall Error Characterization on the Estimation of Soil Moisture Fields in a Land Data Assimilation System. *J. Hydrometeorol.*, **13 (i)**, 1107–1118, doi:10.1175/JHM-D-11-0115.1.

Magnusson, L., E. Källén, and J. Nycander, 2008: Nonlinear Processes in Geophysics Initial state perturbations in ensemble forecasting. *Nonlin. Process. Geophys*, **15**, 751–759, URL www.nonlin-processes-geophys.net/15/751/2008/.

Maidment, R. I., D. Grimes, E. Allan, Richard P. Tarnavsky, M. Stringer, T. Hewison, R. Roebeling, and E. Black, 2014: Journal of Geophysical Research: Atmospheres And Time series (TARCAT) data set. *J. Geophys. Res. Atmos.*, **119**, 10 619–10 644, doi:10.1002/2014JD021927.Received, URL FromVrieling,Meronietal.2016.

Margulis, S. A., D. McLaughlin, D. Entekhabi, and S. Dunne, 2002: Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment. *Water Resour. Res.*, **38 (12)**, 35–1–35–18, doi:10.1029/2001wr001114.

Massari, C., S. Camici, L. Ciabatta, and L. Brocca, 2018: Exploiting satellite-based surface soil moisture for flood forecasting in the Mediterranean area: State update versus rainfall correction. *Remote Sens.*, **10 (2)**, doi:10.3390/rs10020292.

McColl, K. A., S. H. Alemohammad, R. Akbar, A. G. Konings, S. H. Yueh, and D. Entekhabi, 2017: The global distribution and dynamics of surface soil moisture. *Nat. Geosci.*, **10 (February)**, in press.

Milan, M., et al., 2019: Hourly 4D-Var in the Met Office UKV operational forecast model. *Q. J. R. Meteorol. Soc.*, **(December 2019)**, 1281–1301, doi:10.1002/qj.3737.

Mohanty, B. P., M. H. Cosh, V. Lakshmi, and C. Montzka, 2017: Soil Moisture Remote Sensing: State-of-the-Science. *Vadose Zo. J.*, **16 (1)**, vzj2016.10.0105, doi:10.2136/vzj2016.10.0105.

Morzfeld, M., D. Hodyss, and C. Snyder, 2016: The practical irrelevance of the collapse of the ensemble Kalman filter and other particle filters. Tech. rep., University of Arizona, 1–20 pp. doi:10.1080/16000870.2017.1283809.

Narasimhan, B. and R. Srinivasan, 2005: Development and evaluation of Soil Moisture Deficit Index (SMDI) and Evapotranspiration Deficit Index (ETDI) for agricultural drought monitoring. *Agric. For. Meteorol.*, **133 (1-4)**, 69–88.

Notaro, M., 2008: Statistical identification of global hot spots in soil moisture feedbacks among IPCC AR4 models. *J. Geophys. Res. Atmos.*, **113 (9)**, 1–8, doi:10.1029/2007JD009199.

O'Neill, P., et al., 2017: Assessment of version 4 of the SMAP passive soil moisture standard product. *Int. Geosci. Remote Sens. Symp.*, **2017-July (8)**, 3941–3944, doi:10.1109/IGARSS.2017.8127862.

Pappenberger, F., J. Bartholmes, J. Thielen, H. L. Cloke, R. Buizza, and A. de Roo, 2008: New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.*, **35 (10)**, 1–7, doi:10.1029/2008GL033837.

Pinnington, E., T. Quaife, and E. Black, 2018: Impact of remotely sensed soil moisture and precipitation on soil moisture prediction in a data assimilation system with the JULES land surface model. *Hydrol. Earth Syst. Sci.*, **22 (4)**, 2575–2588, doi:10.5194/hess-22-2575-2018.

Pinnington, E., T. Quaife, A. Lawless, K. Williams, T. Arkebauer, and D. Scoby, 2020: The Land Variational Ensemble Data Assimilation Framework: LAVENDAR v1.0.0. *Geosci. Model Dev.*, **13 (1)**, 55–69, doi:10.5194/gmd-13-55-2020.

Pitman, A. J., 2003: The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol.*, **23 (5)**, 479–510, doi:10.1002/joc.893.

Poe, A. G., 1990: Optimum Interolation of Imaging Microwave Radiometer Data. *IEEE Trans. Geosci. Remote Sens.*, **28 (5)**, 800–810, doi:10.1109/36.58966.

Reichle, R. H., 2000: Variational Assimilation of Remote Sensing Data for Land Surface Hydrologic Applications. Ph.D. thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 193 pp.

Reichle, R. H., D. Entekhabi, D. B. Mclaughlin, and R. M. Parsons, 2001: Downscaling of radio brightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach. *Water Resour.*, **37 (9)**, 2353–2364.

Reichle, R. H., J. P. Walker, R. D. Koster, and P. R. Houser, 2002: Extended versus ensemble Kalman filtering for land data assimilation. *J. Hydrometeorol.*, **3 (6)**, 728–740, doi:10.1175/1525-7541(2002)003<0728:EVEKFF>2.0.CO;2.

Roger Stern and Richard Coe, 1984: A Model Fitting Analysis of Daily Rainfall Data. *R. Stat. Soc.*, **147 (1)**, 1–34.

Scott, B. L., T. E. Ochsner, B. G. Illston, J. B. Basara, and A. J. Sutherland, 2013: New soil property database improves oklahoma mesonet soil moisture estimates. *J. Atmos. Ocean. Technol.*, **30 (11)**, 2585–2595, doi:10.1175/JTECH-D-13-00084.1.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture-climate interactions in a changing climate: A review. *Earth-Science Rev.*, **99 (3-4)**, 125–161, WebofScience.

Sheffield, J. and E. F. Wood, 2008: Global trends and variability in soil moisture and drought characteristics, 1950-2000, from observation-driven simulations of the terrestrial hydrologic cycle. *J. Clim.*, **21 (3)**, 432–458.

Silvestro, F. and N. Rebora, 2014: Impact of precipitation forecast uncertainties and initial soil moisture conditions on a probabilistic flood forecasting chain. *J. Hydrol.*, **519 (PA)**, 1052–1067, doi:10.1016/j.jhydrol.2014.07.042, URL http://dx.doi.org/10.1016/j.jhydrol.2014.07.042.

Silvestro, F., L. Rossi, L. Campo, A. Parodi, E. Fiori, R. Rudari, and L. Ferraris, 2019: Impact-based flash-flood forecasting system: Sensitivity to high resolution numerical weather prediction systems and soil moisture. *J. Hydrol.*, **572 (July 2018)**, 388–402, doi:10.1016/j.jhydrol.2019.02.055, URL https://doi.org/10.1016/j.jhydrol.2019.02.055.

Steinemann, A. C. and L. F. N. Cavalcanti, 2006: Developing Multiple Indicators and Triggers for Drought Plans. *J. Water Resour. Plan. Manag.*, **132 (3)**, 164–174.

Tadele, Z., 2017: Raising Crop Productivity in Africa through Intensification. *Agronomy*, **7 (1)**, 22.

Tian, X., Z. Xie, and A. Dai, 2008: An ensemble-based explicit four-dimensional variational assimilation method. *J. Geophys. Res.*, **113 (D21)**, 1–13.

Velpuri, N. M., G. B. Senay, and J. T. Morisette, 2015: Evaluating New SMAP Soil Moisture for Drought Monitoring in the Rangelands of the US High Plains. *Rangelands*, **38 (4)**, 183–190.

Wagner, W., et al., 2013: The ASCAT soil moisture product: A review of its specifications, validation results, and emerging applications. *Meteorol. Zeitschrift*, **22 (1)**, 5–33, doi:10.1127/0941-2948/2013/0399.

Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo, 2014: Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.*, **50**, 7505–7514, doi:10.1002/2014WR015638.Received.

Whitaker, J. S. and A. F. Loughe, 1998: The Relationship between Ensemble Spread and Ensemble Mean Skill. *Mon. Weather Rev.*, **126 (12)**, 3292–3302.

Wilks, D. S., 1990: Maximum Likelihood Estimation for the Gamma Distribution Using Data Containing Zeros. *Am. Meteorol. Soc.*, **3 (12)**, 1495–1501.

Wilks, D. S., 2007: Ensemble Forecasting. Tech. Rep. 514, ECMWF, 313–367 pp. doi:10.1016/b978-0-12-815823-4.00008-0.

Wilks, D. S. and R. L. Wilby, 1999: The weather generation game: a review of stochastic weather models. *Prog. Phys. Geogr.*, **23 (3)**, 329–357, doi:10.1177/030913339902300302.

Wu, G., , B. Dan, and X. Zheng, 2016: Soil Moisture Assimilation Using a Modified Ensemble Transform Kalman Filter Based on Station Observations in the Hai River Basin. *Adv. Meteorol.*, 1–12.

Wu, G. and X. Zheng, 2018: *The Error Covariance Matrix Inflation in Ensemble Kalman Filter*. Open science, 33–54 pp., doi:10.5772/intechopen.71960.

Wu, G., X. Zheng, L. Wang, S. Zhang, X. Liang, and Y. Li, 2013: A new structure for error covariance matrices and their adaptive estimation in EnKF assimilation. *Q. J. R. Meteorol. Soc.*, **139 (672)**, 795–804, doi:10.1002/qj.2000.

Zeng, N., 1999: Enhancement of Interdecadal Climate Variability in the Sahel by Vegetation Interaction. *Science (80-.).*, **286 (5444)**, 1537–1540.

Zhang, J., W.-C. Wang, and J. Wei, 2008: Assessing land-atmosphere coupling using soil moisture from the Global Land Data Assimilation System and observational precipitation. *J. Geophys. Res.*, **113 (D17)**, D17 119.

Zhang, R., S. Kim, and A. Sharma, 2019a: A comprehensive validation of the SMAP Enhanced Level-3 Soil Moisture product using ground measurements over varied climates and landscapes. *Remote Sens. Environ.*, **223 (January)**, 82–94, doi:10.1016/j.rse.2019.01.015.

Zhang, R., S. Kim, and A. Sharma, 2019b: A comprehensive validation of the SMAP Enhanced Level-3 Soil Moisture product using ground measurements over varied climates and landscapes. *Remote Sens. Environ.*, **223 (January)**, 82–94, doi:10.1016/j.rse.2019.01.015, URL https://doi.org/10.1016/j.rse.2019.01.015.

Zhao, L., Z. L. Yang, and T. J. Hoar, 2016: Global soil moisture estimation by assimilating AMSR-E brightness temperatures in a coupled CLM4-RTM-DART system. *J. Hydrometeorol.*, **17 (9)**, 2431–2454, doi:10.1175/JHM-D-15-0218.1.

Zhao, W. and A. Li, 2015: A Review on Land Surface Processes Modelling over Complex Terrain. *Adv. Meteorol.*, **2015**, doi:10.1155/2015/607181.

Zheng, W., X. Zhan, J. Liu, and M. Ek, 2018: A Preliminary Assessment of the Impact of Assimilating Satellite Soil Moisture Data Products on NCEP Global Forecast System. *Adv. Meteorol.*, **2018**, doi:10.1155/2018/7363194.

Zheng, X. and E. A. Eltahir, 1998: A soil moisture-rainfall feedback mechanism 2. Numerical experiments. *Water Resour. Res.*, **34 (4)**, 777–785, doi:10.1029/97WR03497.