Global mapping of root zone soil moisture at subkilometric resolution

Roiya Souissi

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITE DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par Roiya SOUISSI
Le 14 décembre 2022
Cartographie de l'humidité de sol en zone racinaire à l'échelle globale et avec une résolution kilométrique

---

Jury

Mme Catherine OTTLE, Rapporteure
M. Niko VERHOEST, Rapporteur
M. Ahmad AL BITAR, Examinateur
M. Marwan FAHS, Examinateur
M. Lionel JARLAN, Examinateur
Mme Chiara CORBARI, Examinatrice
Mehrez ZRIBI, Directeur de thèse

Ecole doctorale: **SDU2E – Sciences de l'Univers, de l'Environnement et de l'Espace**
Spécialité: **Surfaces et interfaces continentales, Hydrologie**

Unité de Recherche : **CESBIO – Centre d'Etudes Spatiales de la Biosphère**

Thèse dirigée par:
Mehrez ZRIBI et Ahmad AL BITAR

Abstract

In a global context of water crisis, soil moisture is considered as a crucial variable for agriculture since it heavily relies on water resources. Given its importance in land–atmosphere feedbacks, it is recognized as an Essential Climate Variable (ECV) with both its surface and root-zone components. Root-zone soil moisture (RZSM) is particularly interesting since plants draw water and nutrients from the root zone. Soil moisture can be directly or indirectly measured. Over the three past decades, remote sensing techniques have been providing surface soil moisture (SSM) retrievals. However, RZSM is currently inaccessible by satellite sensors since their penetration depth is limited to few centimeters. Besides in-situ measurements, RZSM can be derived using physically-based methods, data assimilation techniques or data-driven methods. Data-driven techniques, like Artificial Neural Networks (ANN), are especially promising since they do not require explicit relationships between the inputs and the target which is an advantage compared to the first two options that are very prone to inaccuracies.

This PhD aims at predicting RZSM at large scales and kilometric resolutions using ANNs. The work is structured in three main parts. The first was centered on the prediction of RZSM at different locations around the world using an ANN model that relies on only SSM in-situ data provided by the International soil Moisture Network (ISMN). A transferability analysis demonstrated that no soil moisture network trained alone was able to well reproduce. However, the model was able to capture the variabilities of the RZSM when trained on stations from different networks. Overall, the predictions were good in areas of alternate wet and dry cycles but less good for instance in regions with high evaporation rates. These limitations motivated us to complexify the method such as it becomes a physics-aware data-driven approach. To do so, physical process-related variables were added to the ANN model. More precisely, soil water index (SWI) which is computed by a recursive exponential filter, was considered to depict the infiltration process. An evaporation efficiency, whose formulation relies on a remote sensing-retrieved potential evapotranspiration (PET), was also considered to represent the evaporation process. A normalized difference vegetation index (NDVI) was used to infer vegetation dynamics. Several ANN models were built such that the features include SSM and process-related variables. The models were trained on good-quality ISMN stations and tested on the rest of the previously considered ISMN stations. Additional tests were conducted on stations external to the ISMN database in order to assess the robustness of the method namely over Tunisia, Italy and India. Results showed that the ANN model made up of SSM and all process-related features was the best performing in most cases. The individual impact of each process-related variable on the prediction quality was also highlighted through an analysis across climate classes. For instance, evaporation efficiency was relevant in regions of high evaporation rates and NDVI was most beneficial over agricultural fields. The robustness of the approach was validated in the case of Tunisia but no significant improvement was recorded in Italy and India.

The last part of the work consisted in studying the feasibility of mapping RZSM over continental Europe at 1km resolution using the previously developed ANN model. It was not calibrated again but only applied on unseen test datasets which consisted of remotely-sensed variables. Maps of RZSM at 1 km resolution were produced using SSM data from the Copernicus SSM1km product which is based on Sentinel-1 measurements. For validation purposes, RZSM maps at 9km and 36km were also generated using the ERA5-Land reanalysis SSM and the Soil Moisture Active

Passive (SMAP) level-3 SSM products, respectively. Validations against ERA5-Land reanalysis RZSM product and against in-situ RZSM data were conducted. The C-band SSM information was proved unreliable in complex sceneries and highly impacting the quality of the RZSM predictions. The L-band SSM information has been shown more reliable but it was hampered by the coarse spatial resolution. The ERA5-Land SSM product injected in the ANN model allowed better predictions than the previous products and outperformed the RZSM reanalysis product in some areas when compared against in-situ data.

The results obtained in this work highlight the feasibility of global mapping of RZSM at high resolution using an ANN model. The use of more qualitative SSM data provided by future missions would allow better quantification of RZSM.

Keywords: root-zone soil moisture, artificial neural networks, process-related variables, sub-kilometric resolution, sentinel-1, SMAP, ERA5-land, ISMN.

Résumé

Dans un contexte mondial de crise de l'eau, l'humidité du sol est une variable cruciale pour l'agriculture qui dépend fortement des ressources en eau. Vu son rôle dans les interactions terre-atmosphère, elle est reconnue comme une variable climatique essentielle avec sa composante de surface et de zone racinaire. L'humidité du sol de la zone racinaire (RZSM) est particulièrement intéressante car les plantes puisent l'eau dans la zone où se développent les racines. L'humidité du sol peut être mesurée directement ou indirectement. Au cours des trois dernières décennies, les techniques de télédétection ont permis d'observer l'humidité du sol de surface (SSM). Cependant, la RZSM est actuellement inaccessible par les capteurs satellitaires dont la profondeur de pénétration est limitée. Outre les mesures terrain, la RZSM peut être obtenue via des méthodes basées sur la modélisation physique, techniques d'assimilation de données ou des méthodes basées sur les données. Ces dernières sont prometteuse et moins sujettes aux erreurs car ne nécessitent pas l'explicitation des relations gouvernant les données en entrée et la cible. Les réseaux de neurones artificiels (ANN) en sont un exemple.

Cette thèse est structurée en trois axes et se focalise sur la prédiction de la RZSM à large échelle et à résolution subkilométrique en utilisant les ANNs. La première partie concerne la prédiction de la RZSM à différents endroits du monde à l'aide d'un modèle ANN qui repose uniquement sur les données in-situ de la SSM fournies par le réseau international de l'humidité du sol (ISMN). Une analyse de transférabilité a démontré qu'aucun réseau d'humidité du sol entraîné seul n'était capable de bien reproduire la RZSM. Le modèle entraîné sur des stations de différents réseaux était capable de suivre les variabilités de la RZSM. Dans l'ensemble, le modèle s'est avéré fiable dans les zones où alternent des cycles humides et secs mais moins performant sur les zones à forts taux d'évaporation par exemple. Ces limitations nous ont menés à évaluer l'impact de l'ajout de variables d'entrée dans le modèle ANN. L'indice d'eau du sol (SWI), basé sur un filtre exponentiel récursif, a été considéré pour décrire le processus d'infiltration. Une efficacité d'évaporation, dont la formulation repose sur une évapotranspiration potentielle (PET) issue de la télédétection, a également été considérée pour représenter le processus d'évaporation. L'indice de végétation par différence normalisée (NDVI) a été utilisé pour déduire la dynamique de la végétation. Plusieurs modèles ANN ont été construits de manière à ce que les données d'entrée comprennent la SSM et une ou plusieurs variables reliées à des processus. Les modèles ont été entraînés sur des stations ISMN avec mesures de bonne qualité et testés sur le reste des stations ISMN également considérées dans la partie précédente. Des tests supplémentaires ont été effectués sur des stations non incluses dans ISMN à savoir sur la Tunisie, l'Italie et l'Inde. Les résultats ont montré que le modèle ANN composé des entrées SSM et de toutes les variables liées aux processus, était le plus fiable dans la plupart des cas. L'impact individuel de chaque variable sur la qualité des prédictions a été également mis en évidence via une analyse climatique.

Le dernier axe concerne la cartographie de la RZSM à l'échelle de l'Europe continentale et à résolution subkilométrique en utilisant le modèle ANN précédemment développé qui n'a pas été recalibré mais uniquement appliqué sur de nouvelles données de test issues de la télédétection. Des cartes de RZSM à 1km ont été produites en utilisant le modèle ANN telles que SSM était fournie par le produit Copernicus SSM1km, basé sur les données Sentinel-1. Pour les valider, des cartes RZSM à 9 km et 36 km ont été générées à partir du produit SSM de réanalyse ERA5-Land et du produit SSM de niveau 3 de la mission Soil Moisture Active Passive (SMAP), respectivement. Des

comparaisons avec le produit RZSM de réanalyse ERA5-Land et avec des données in-situ ont été effectuées. Les données SSM en bande C se sont avérées peu fiables dans les scènes complexes et avaient un impact négatif sur la qualité des prédictions RZSM. Les données SSM en bande L ont permis de s'affranchir de certaines limitations du produit précédent mais elles sont entravées par une résolution spatiale grossière. Le produit de réanalyse SSM ERA5-Land, injecté dans le modèle ANN, a généré de meilleures prédictions que les deux produits précédents et aussi par rapport au produit de réanalyse RZSM.

Les résultats obtenus dans ce travail soulignent la faisabilité de la cartographie globale de RZSM à haute résolution en utilisant un modèle ANN. L'utilisation de données plus qualitatives de la SSM qui seront fournies par de futures missions permettrait une meilleure quantification de la RZSM.

Mots clés: humidité du sol en zone racinaire, réseaux de neurones artificiels, variables reliées aux processus, résolution subkilométrique, Sentinel-1, SMAP, ERA5-land, ISMN.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my PhD supervisors Mehrez Zribi and Ahmad Al Bitar for their continuous assistance at every stage of this PhD, their insightful ideas and their patience and support. I would like to thank them for the precious brainstormings we had during our weekly meetings and their valuable suggestions to improve the quality of the work. Their emotional support and understanding during the COVID period was also of great help in such uncertain time. I could have surrendered to distress if it wasn't for their follow up and encouragement.

I would like also to thank the thesis committee members for their valuable comments and suggestions that inspired me at different steps of the work. I also thank the jury members for the time and attention they devoted to my work. Moreover, I am grateful to the project partners namely Chiara Corbari and Marco Mancini for the interest they have been showing to my work and for the discussions we had through ZOOM meetings. My gratitude extends to the diligent and kind staff at CESBIO to name but a few Laurence Keppel, Dominique Tarrisse, Delphine Maria and Laura Léal. Special thanks also to Eric Brune for his kindness and cheerfulness.

I sincerely thank all my colleagues at CESBIO for their welcoming and helpful attitude. I am delighted and proud of the relationships that I have developed in the lab. I am tremendously grateful for my co-workers who turned into real friends namely Emna Ayari, Nadia Ouaadi and Nitu Ojha. They have always been there for me through all my joys and distress. I will be also very nostalgic of the small moments I spent while sharing the office with David Morin and formerly Hervé Thevenon, who I consider real friends. I thank both of them for their guidance, kindness and emotional and technical support.

My biggest gratitude goes to my family, my number one supporter. I will be forever thankful to my mother, late father and brother for their unconditional love and support. I have no words to acknowledge the sacrifices they made to make my life better. I would like to wholeheartedly dedicate this thesis to the memory of my beloved father who had always believed in me and encouraged me to pursue this PhD. I am sorry he cannot see me defend my work but at the same time glad to know that he witnessed part of the process. Last but not least, I thank my husband Amine for his priceless love, encouragement and patience. He has been accompanying me at every step of the way while recognising the emotional demands that research was placing on me. I still find it difficult to express all my gratitude to him because it is so immeasurable.

TABLE OF CONTENTS

GENERAL INTRODUCTION 12

Societal, Political, Scientific Context 13

Research axes 15

INTRODUCTION GENERALE (FRANÇAIS) 18

Contexte sociétal, politique et scientifique 19

Axes de recherche 22

CHAPTER 1: STATE-OF-THE-ART 24

1.1. Soil moisture definition 25

1.2 Processess related to soil moisture 26
1.2.1 Infiltration process 27
1.2.2 Rainfall-Runoff 27
1.2.3 Evapotranspiration 28

1.3 Measuring Soil moisture 29
1.3.1 Direct measurements 29
1.3.2 Indirect measurements 29
1.3.2.1 In-situ measurements 30
1.3.2.2 Remote sensing observations 31

1.4 Root-Zone Soil moisture estimation 38
1.4.1 Physically-based models 39
1.4.2 Data assimilation 40
1.4.3 Data-driven methods 42

CHAPTER 2: PREDICTION OF RZSM BASED ON SSM USING ANN 50

2.1. Introduction 51

2.2 Conclusion 51

2.3 Article 52

CHAPTER 3: PREDICTION OF RZSM BASED ON SSM AND PROCESS-RELATED INPUTS USING ANN 74

3.1 Introduction 75

3.2 Conclusion 75

3.3 Article 76

CHAPTER 4: RZSM SPATIAL MAPS AT LARGE SCALE AND 1KM RESOLUTION ... 112

4.1 Introduction **113**

4.2 Conclusion **113**

4.3 Article **114**

GENERAL CONCLUSION AND PERSPECTIVES **146**

CONCLUSION GENERALE ET PERSPECTIVES (FRANÇAIS) **150**

ACRONYMS **155**

REFERENCES **159**

LIST OF FIGURES

General introduction

Figure 1: The 17 SDGs from the United Nations (source: https://www.un.org/sustainabledevelopment/news/communications-material/) 13
Figure 2: Map of water stress due to agriculture, by basin, in 2018 (Source: FAO and UN-Water, 2021, modified to be consistent with UN, 2021.) 14

Chapter 1: State-of-the-art

Figure 3: Soil water content at saturation, field capacity and permanent wilting point (Datta et al., 2017). .. 25
Figure 4: Diagram of unsaturated soil zone holding the soil moisture components and saturated soil zone. . 26
Figure 5: Hydrologic processes interacting with soil moisture (source: KGS Pub. Inf. Circ. 22, last access: 5 September 2022) 27
Figure 6: Diagram of the land water balance for a given surface soil layer; dS/dt represents the change in water content. 29
Figure 7: Some past and current microwave missions for SM retrieval 35
Figure 8: Venn diagram of the artificial intelligence and some of its integrated technologies. 42
Figure 9: Machine learning workflow 44
Figure 10: Structure of a neuron 45
Figure 11: Examples of different architectures of artificial neural networks: (a) Single-layer feedforward network (b) Multilayer feedforward network (c) Recurrent neural network. 46

LIST OF TABLES

Chapter 1: State-of-the-art

Table 1. Overview of remote sensing techniques for soil moisture estimation (after Wang and Qu, 2009; Babaeian et al., 2019). 32

Foreword

This PhD was conducted in the Centre des Etudes Spatiales de la BIOsphère (CESBIO) under the supervision of Mehrez ZRIBI (Research Director at the Centre National de la Recherche Scientifique (CNRS)) and Ahmad AL BITAR (Research engineer at CNRS).

This PhD was cofunded by the Agence Nationale de Recherche (ANR) and Centre National d'Etudes Spatiales (CNES) under the grants *RET-SIF-ERANETMED–ANR-17-NMED-0004-01* and *SMARTIES-PRIMA–ANR-NMED*.

During this PhD, two thesis committees took place on October 16, 2020 and January 21, 2022 respectively. Olivier MERLIN (Research Director at CNRS), Philippe MAISONGRANDE (Land and Hydrology Program Manager at Centre National d'Etudes Spatiales (CNES)) and Thierry PELLARIN (Research Director at CNRS) are the three members of the committee.

Préface

Cette thèse a été réalisée au Centre des Etudes Spatiales de la Biosphère (CESBIO) sous la direction de Mehrez ZRIBI (directeur de recherche au Centre National de la Recherche Scientifique (CNRS)) et Ahmad AL BITAR (Ingénieur de recherche au CNRS).

Cette thèse a été cofinancée par l'Agence Nationale de Recherche (ANR) et le Centre National d'Etudes Spatiales (CNES) sous les bourses RET-SIF-ERANETMED-ANR-17-NMED-0004-01 et SMARTIES-PRIMA-ANR-NMED.

Au cours de cette thèse, deux comités de thèse ont eu lieu respectivement le 16 octobre 2020 et le 21 janvier 2022. Olivier MERLIN (directeur de recherche au CNRS), Philippe MAISONGRANDE (Responsable de la thématique Surfaces, biosphère continentales, hydrologie au Centre National d'Etudes Spatiales (CNES)) et Thierry PELLARIN (directeur de recherche au CNRS) sont les trois membres du comité.

General introduction

This chapter presents the broad context and background to this PhD. The three main research axes are also briefly introduced.

Table of contents

Societal, Political, Scientific Context 13

Research axes 15

Societal, Political, Scientific Context

Nowadays, the term "water" is closely associated with the war and peace terminologies. The ninth edition of the World Water Forum, which was held between March 22 and 27, 2022 in Senegal, is the best illustration since it was entitled "Water security for peace and sustainable development". In a context of climate change, growing world population and increasing urbanization, water is becoming scarcer. This shortage can threaten peaceful coexistence between countries sharing cross-border river basins, as the President of the Republic of Senegal recalled in his welcome message at the last World Water Forum, hence the urgency of making this resource a global political priority.

Water appears in the Sustainable Development Goals (SDGs) presented in the UN 2030 agenda, namely through SDG 2 “zero hunger” and SDG 6 “guarantee water for all” (Figure 1).

Figure 1: The 17 SDGs from the United Nations (source: https://www.un.org/sustainabledevelopment/news/communications-material/)

The aggregated SDG indicator 6.4.2 describes water stress. In 2018, the average value of this indicator was equal to 18% at the global scale. At the regional scale, the indicator was below 25% (equal to 8.3%) for Europe which makes it at that time unexposed to water stress; in Eastern and Western Asia water stress was low to medium (indicator between 45% and 70%), However, it was high in Central Asia and South Asia with levels above 70% and critical in North Africa (indicator above 100%).

In this context of water crisis, 3.2 billion people living in rural areas are threatened, according to the Food and Agriculture Organization (FAO). The latest report of FAO, entitled “The State of the World’s Land and Water Resources for Food and Agriculture”, highlights the threats to key agrifood systems around the world. In Africa, several countries suffer from the lack of water resources and are unable to cultivate and to subsidize the food needs of their people, says FAO Director-General Mr. Qu Dongyu during the World Water Forum in Dakar (2022).

Moreover, different businesses are threatened in several agricultural areas. These include smallholders and farmers whose crops are deteriorating due to drought events, as well as pastoral breeders whose animals are dying of thirst and fishermen whose fishing grounds are gradually shrinking.

According to (FAO, 2021), 10% of inland renewable water resources from rivers and aquifers have been used. 70% of freshwater withdrawals are for agriculture and constitute a large share of withdrawals in North Africa, Central Asia and the Middle East-West Asia region.

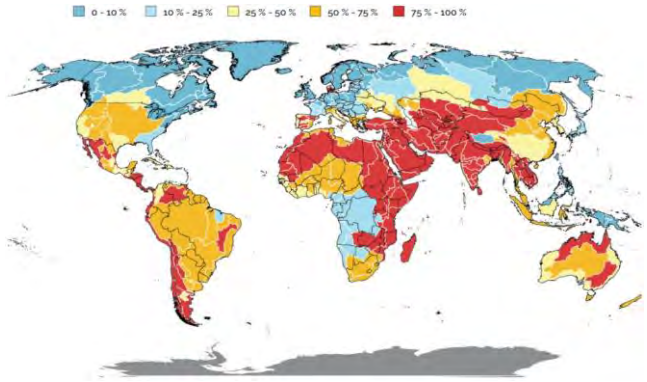


Figure 2: Map of water stress due to agriculture, by basin, in 2018 (Source: FAO and UN-Water, 2021, modified to be consistent with UN, 2021.)

Another major factor that increases the threats to water security consists in climate change. According to (IPCC, 2022), heavy precipitation events are more frequently occurring in many regions compared to the 1950s. This led to more substantial annual maximum one-day precipitation in many regions against longer dry spells in other regions, all compared to the 1950s. Besides, the global glacier mass loss rate went beyond 0.5 m water equivalent per year during the last two decades, according to the same report, which highly affected the cultural uses of water among vulnerable high mountain and polar communities. Overall, unprecedented and severe floods and droughts are more likely to occur due to human-induced climate change. The report also highlighted that 7% of global disasters between 1970 and 2019 were drought-related whereas 31% of all economic losses in the same time range were flood-related due to the increasing warming. These extreme events affect both the quantity and quality of water. Actually, the availability of water is highly impacted by droughts mainly over arid areas like large parts of Africa, the USA, China and India. On the other side, floods and heavy rainfall events are likely to affect the quality of water and may make it undrinkable. Moreover, the salination of groundwater resources can increase due to harsher storms and higher sea levels in small islands and coastal regions for example.

In order to achieve the objectives inherent to the sustainable management of water (SDG 6), the FAO is participating in several projects focusing on water scarcity such as AQUASTAT database, the Water Productivity Open-access portal (WaPOR) and The Global Framework on Water Scarcity in Agriculture (WASAG). The characterization of surface conditions and more specifically the soil hydraulic properties is a major challenge to accomplish the water related sustainable goals. Actually, soil moisture is one of the major elements of the surface water budget that characterize the level of stress of land surface ecosystems. Although only 1 mg for each kilogram of water on Earth is stored as soil moisture (which makes around 0.0001% of earth's water), this tiny amount of water governs different hydrological, agricultural, and meteorological processes ranging from boundary layer dynamics to the global water cycle (Islam and Engman, 1996). Since the superficial reservoir has a small capacity, the monitoring and the accurate quantification of the spatial and temporal variability of soil moisture in deep layers is more important than the surface soil moisture. Hence, the prediction of root-zone soil moisture (RZSM) is essential for crop management,

irrigation planning, flooding and drought mapping, weather prediction and quantification of carbon fluxes within soils.

Although in-situ measurements provide accurate estimations of RZSM, they are lacking over large spatial scales (Dorigo et al., 2011). Alternatively, remote sensing technology provides seamless and large-scale soil moisture retrievals but with a limited sensing depth of few centimeters. A common approach to produce continuous estimates of this variable consists in observing surface meteorological data in physically-based models like land surface models (LSMs) in an uncoupled manner (Koster et al., 2009). Nevertheless, the errors in forcing parameters and the deficiencies in the representation of land surface processes lead to uncertainties in these estimates. Shallow remote sensing soil moisture accounted for the integration of in situ surface data or satellite sensors data like passive microwave brightness temperature or radar backscattering coefficients into hydrologic models to predict RZSM through data assimilation techniques (Kolassa et al., 2017; Lievens et al., 2017; De Lannoy et al., 2019). Data-driven techniques, such as Artificial Neural Networks (ANNs), have proven efficient in RZSM prediction (Kornelsen and Coulibaly, 2014, Pan et al., 2017; Souissi et al, 2020; Souissi et al., 2022).

All the aforementioned techniques have been applied in several studies in order to estimate RZSM at local, regional and continental scales. However, no attempt has been made yet to predict this component at large scales and high spatial resolutions concurrently. The interest of the different spatial resolutions varies with the field of application. For instance, meteorological applications need coarse resolution soil moisture data, hydrological applications are usually centered on the watershed scale whereas agricultural applications require high-resolution data (Stefan et al., 2021). The resolution at which RZSM is currently being predicted at large scales and the resolution at which land processes occur within the soil profile are disparate. This observation motivated this work to focus on the prediction of root zone soil moisture at large scale and subkilometric resolution. In the literature, ANNs have been used as surrogate models, calibration tools or as physics-aware methods. The classic way of tackling soil moisture estimation problems with ANNs as a physics-aware method is to start from the analysis of the radiometric observations i.e. remote sensing observations. Then, physics which is depicted by different physically-based variables or models is injected into the ANN to finally obtain the target, namely soil moisture. In this PhD, we used ANNs as a physics-aware method but in a different processing order. Actually, we started from multi-location in-situ soil moisture measurements to train and test an ANN model. Process-related variables were then added as model inputs. After this hybridization of the approach, the locally-trained model is tested using remote sensing data in order to produce spatial maps of RZSM at large scale and 1 km resolution.

Research axes

This work is centered on different research axes namely:

1/ Assess the robustness, accuracy and transferability with which RZSM can be estimated using only surface soil moisture (SSM) in a data-driven method. 2/ Explore how best to include process-related information along with SSM in ANN models intended to predict RZSM. 3/ Study the feasibility of mapping RZSM at large scale and subkilometric resolution using ANN and remote sensing-based inputs.

- **Prediction of RZSM based on SSM using ANN**

In this part, an ANN model was developed to estimate RZSM based on only in-situ SSM information. Different experiments were conducted on the model with regard to the temporal sampling of the SSM features, the scaling technique and of the split of the training, validation and test sets. In-situ SSM measurements were provided by the International Soil Moisture Network (ISMN) at different locations around the world such that different climates and soil types were considered. This particularity is important to assess the transferability of the approach. The accuracy of the method was investigated across the climate classes and soil textures. The contribution of each soil moisture network was also assessed and led to a data filtering. This approach is detailed in the scientific paper entitled «Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe» and published in *Water* journal, and will be presented in chapter 2.

- **Prediction RZSM based on SSM and process-related inputs using ANN**

The findings of the first axis mainly with regard to the performance limitations of the first approach over regions where the link between surface and root zone is weak, have led us to further complexify the method. Given that different hydrological processes like diffusion processes connect RZSM to SSM, we decided to investigate the impact of adding process-related inputs in addition to SSM in ANN models and seek the best combination that ensures the more accurate RZSM predictions. The soil water index (SWI) was computed with a recursive exponential filter and used to account for the infiltration process. The evaporation process was illustrated through an evaporation efficiency computed based on a Moderate Resolution Imaging Spectroradiometer (MODIS) remote-sensing potential evapotranspiration (PET) dataset and a simplified analytical model. Vegetation growth was interpreted through the normalized difference vegetation index (NDVI) time series. Several ANN models with different combination of features were developed.

This approach and different results are presented in the scientific paper «Integrating process-related information into an artificial neural network for root-zone soil moisture prediction», published in *Hydrology and Earth System Sciences (HESS)* journal, and will be detailed in chapter 3.

- **RZSM spatial maps at large scale and 1 km resolution**

Our starting point in this axis is the most complex ANN model which was developed in the previous part and which allowed agreement between RZSM predictions and in-situ information. Actually, the assessment of the reliability of the ANN model to yield consistent RZSM predictions over a continental scale namely Europe is a decisive step in this study. The feasibility of predicting RZSM at large scale using a locally-trained is a demonstration of its generalizability to the global scale. Different SSM products specifically radar, passive microwaves and reanalysis datasets were employed to compute the three SSM features. A comparison between the different RZSM maps produced at different resolutions (1 km, 9 km and 36km) was performed as well as a validation against in-situ RZSM collected over four European in-situ soil moisture networks.

This approach is detailed in a scientific paper entitled «Root-Zone soil moisture over Continental Europe using machine learning» and submitted in *International Journal of Applied Earth Observation and Geoinformation* journal and will be detailed in chapter 4.

Introduction générale (Français)

Ce chapitre présente le contexte général de cette thèse. Les trois principaux axes de recherche sont également brièvement introduits.

Table de matières

Contexte sociétal, politique et scientifique 19

Axes de recherche 22

Contexte sociétal, politique et scientifique

De nos jours, le terme "eau" est étroitement lié aux terminologies de guerre et de paix. La neuvième édition du Forum mondial de l'eau, qui s'est tenue du 22 au 27 mars 2022 au Sénégal, en est la meilleure illustration puisqu'elle était intitulée "La sécurité de l'eau pour la paix et le développement durable". Dans un contexte de changement climatique, de croissance de la population mondiale et d'urbanisation croissante, l'eau devient de plus en plus rare. Cette pénurie peut menacer la coexistence pacifique entre les pays partageant des bassins hydrographiques transfrontaliers, comme l'a rappelé le Président de la République du Sénégal dans son message de bienvenue au dernier Forum mondial de l'eau, d'où l'urgence de faire de cette ressource une priorité politique mondiale.

L'eau apparaît dans les Objectifs de développement durable (SGD) présentés dans l'agenda 2030 de l'ONU, notamment à travers l'SGD 2 « Éliminer la faim, assurer la sécurité alimentaire, améliorer la nutrition et promouvoir une agriculture durable » et l'SGD 6 « Garantir l'accès de tous à l'eau et à l'assainissement et assurer une gestion durable des ressources en eau » (Figure 1).

Figure 1: Les 17 objectifs de développement durable des Nations Unies (source: https://www.un.org/sustainabledevelopment/news/communications-material/)

L'indicateur SDG agrégé 6.4.2 décrit le stress hydrique. En 2018, la valeur moyenne de cet indicateur était égale à 18% à l'échelle mondiale. A l'échelle régionale, l'indicateur était inférieur à 25% (égal à 8,3%) pour l'Europe ce qui la rend non exposée au stress hydrique à l'époque; en Asie de l'Est et de l'Ouest le stress hydrique était faible à moyen (indicateur entre 45% et 70%), En revanche, il était élevé en Asie centrale et en Asie du Sud avec des niveaux supérieurs à 70% et critique en Afrique du Nord (indicateur supérieur à 100%).

Dans ce contexte de crise de l'eau, 3,2 milliards de personnes vivant en milieu rural sont menacées, selon l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO). Le dernier rapport de la FAO, intitulé « L'État des ressources en terres et en eau pour l'alimentation et l'agriculture dans le monde », met en évidence les menaces qui pèsent sur les principaux systèmes agroalimentaires dans le monde. En Afrique, plusieurs pays souffrent de la pénurie des ressources en eau et sont donc incapables de cultiver et de subvenir aux besoins alimentaires de leur population

soit environ 10% de la population mondiale, indique le Directeur général de la FAO, M. Qu Dongyu, lors du Forum mondial de l'eau à Dakar (2022). En outre, différents projets sont menacés dans plusieurs zones agricoles. Il s'agit notamment des petits exploitants et des agriculteurs dont les cultures se dégradent en raison des épisodes de sécheresse, ainsi que des éleveurs pastoraux dont les animaux meurent de soif, mais aussi des pêcheurs dont les zones de pêche se réduisent progressivement.

D'après (FAO, 2021), 10 % des ressources en eau renouvelables continentales provenant des rivières et des aquifères ont été utilisées. 70 % des prélèvements d'eau douce sont destinés à l'agriculture et constituent une part importante des prélèvements en Afrique du Nord, en Asie centrale et dans la région Moyen-Orient-Asie occidentale.

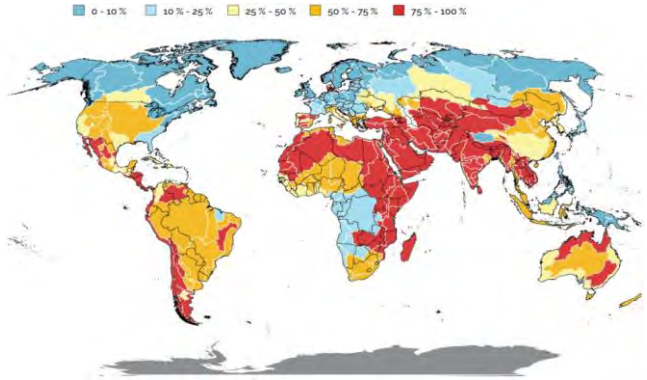


Figure 2: Carte du niveau de stress hydrique dû à l'usage agricole, par bassin, en 2018 (Source: FAO and UN-Water, 2021, modified to be consistent with UN, 2021.)

Le changement climatique est un autre facteur majeur qui prolifère les menaces sur la sécurité de l'eau. Selon (IPCC, 2022), les événements de fortes précipitations sont plus fréquents dans de nombreuses régions par rapport aux années 1950. Cela a conduit à des précipitations maximales journalières annuelles plus importantes dans de nombreuses régions et à des périodes de sécheresse plus longues dans d'autres régions, le tout par rapport aux années 1950. En outre, le taux de perte de masse des glaciers à l'échelle mondiale a dépassé 0,5 m d'équivalent eau par an au cours des deux dernières décennies, selon le même rapport, ce qui a fortement affecté l'utilisation en eau pour les cultures parmi les communautés vulnérables de hautes montagnes et polaires, par exemple. Par ailleurs, des inondations et des sécheresses sévères sont plus susceptibles de se produire en raison du changement climatique induit par l'homme. Le rapport souligne également que 7 % des catastrophes survenues à l'échelle mondiale entre 1970 et 2019 étaient liées à la sécheresse, tandis que 31 % de toutes les pertes économiques sur la même période étaient liées aux inondations, en raison du réchauffement croissant. Ces événements extrêmes affectent à la fois la quantité et la qualité de l'eau. En effet, la disponibilité de l'eau est fortement affectée par les sécheresses, principalement dans les zones arides comme en Afrique, les États-Unis, la Chine et l'Inde. D'autre part, les inondations et les fortes précipitations sont susceptibles d'affecter la qualité de l'eau et de la rendre impotable. En outre, la salinisation des ressources en eau souterraine peut augmenter en raison de tempêtes plus violentes et de l'élévation du niveau de la mer dans les petites îles et les régions côtières, par exemple. Afin d'atteindre les objectifs liés à la gestion durable de l'eau (SDG 6), la FAO intervient dans plusieurs projets axés sur la pénurie de l'eau, tels que la base de données AQUASTAT, le portail de suivi de la productivité de l'eau (WaPOR) et Le Cadre mondial contre la pénurie d'eau dans l'agriculture (WASAG). Le suivi des conditions de surface et plus

particulièrement des propriétés hydrauliques du sol est un défi majeur pour atteindre les objectifs de durabilité liés à l'eau. En effet, l'humidité du sol est l'un des principaux éléments du bilan hydrique qui caractérise le niveau de stress des écosystèmes de surface. Bien que seulement 1 mg par kilogramme d'eau sur terre soit stocké sous forme d'humidité du sol (ce qui représente environ 0,0001% de l'eau de la terre), cette petite quantité d'eau régit différents processus hydrologiques, agricoles et météorologiques allant de la dynamique de la couche limite jusqu'au cycle global de l'eau (Islam et Engman, 1996). Étant donné que le réservoir superficiel a une faible capacité, la quantification précise de la variabilité spatiale et temporelle de l'humidité du sol dans les couches profondes est essentielle. La prédiction de l'humidité du sol dans la zone racinaire (RZSM) est donc importante pour la gestion des cultures, la planification de l'irrigation, la cartographie des inondations et des évènements de sécheresse, les prévisions météorologiques et la quantification des flux de carbone dans les sols.

Bien que les mesures terrain fournissent des estimations précises de la RZSM, elles ne sont pas capables de fournir des mesures à large échelle (Dorigo et al., 2011). Par ailleurs, les techniques de télédétection permettent de fournir des données d'humidité du sol à grande échelle, mais avec une profondeur de pénétration dans le sol limitée à quelques centimètres. L'une des approches pour générer des estimations continues de cette variable consiste à utiliser les données météorologiques de surface observées dans des modèles basés sur la physique, comme les modèles de surface terrestre (LSM) (Koster et al., 2009). Cependant, les erreurs des paramètres de forçage et les déficiences dans la représentation des processus de surface conduisent à des incertitudes dans ces estimations. Les techniques d'assimilation de données peuvent reposer sur l'intégration de données de capteurs aéroportés comme les coefficients de rétrodiffusion radar et la température de brillance micro-ondes dans des modèles hydrologiques pour prédire les RZSM (Kolassa et al., 2017; Lievens et al., 2017; De Lannoy et al., 2019). Les méthodes basées sur les données, telles que les réseaux de neurones artificiels (ANN), ont été aussi démontrées efficaces pour la prédiction de la RZSM (Kornelsen and Coulibaly, 2014, Pan et al., 2017; Souissi et al, 2020; Souissi et al., 2022)..

Toutes les techniques susmentionnées ont été appliquées dans plusieurs études afin d'estimer la RZSM à l'échelle locale, régionale et continentale. Cependant, aucune tentative n'a encore été faite pour prédire cette composante à grande échelle et à haute résolution spatiale simultanément. L'intérêt des différentes échelles spatiales varie en fonction de l'application. Par exemple, les applications météorologiques nécessitent des données d'humidité du sol à résolution grossière, les applications hydrologiques sont généralement centrées sur l'échelle du bassin versant alors que les applications agricoles nécessitent des données à haute résolution (Stefan et al., 2021). La résolution à laquelle le RZSM est actuellement prédite à grande échelle et la résolution à laquelle les processus terrestres se produisent dans le profil du sol sont très différentes. Cette observation nous ramène à l'objectif de cette thèse à savoir la prédiction de l'humidité du sol de la zone racinaire à grande échelle avec une résolution kilométrique. Dans la littérature, les ANN ont été utilisés en tant que « surrogate models », outils de calibration ou en tant que méthodes tenant compte de la physique. La manière classique d'aborder les problèmes d'estimation de l'humidité du sol avec les ANN en tant que méthode tenant compte de la physique, consiste à commencer par l'analyse des observations issues de l'observation de la Terre. Ensuite, le réseau ANN est alimenté par des variables ou modèles liés aux processus physiques pour enfin, obtenir la cible, à savoir l'humidité du sol. Dans cette thèse, nous avons utilisé les ANNs en tant que méthode basée sur la physique mais dans un ordre de traitement différent. En fait, nous avons commencé par des mesures in-situ

de l'humidité du sol de surface pour entraîner et tester un modèle ANN. Les variables liées aux processus ont ensuite été ajoutées comme entrées du modèle. Après cette hybridation de l'approche, le modèle localement entraîné est testé en utilisant des données de télédétection pour produire des cartes spatiales de la RZSM à grande échelle et à une résolution de 1 km.

Axes de recherche

Ce travail est centré sur différents axes de recherche à savoir :

1/ Évaluation de la robustesse, la précision et la transférabilité avec lesquelles la RZSM peut être estimée en utilisant uniquement l'humidité du sol de surface (SSM) à travers une méthode basée sur les données. 2/ Investigation de la meilleure approche pour inclure des informations liées aux processus physiques avec la SSM dans des modèles ANN conçus pour prédire la RZSM. 3/ Etude de la faisabilité de la cartographie de la RZSM à grande échelle et à une résolution subkilométrique en utilisant des modèles ANN et des données de télédétection.

- **Prédiction de la RZSM dans un ANN basé uniquement sur la SSM**

Dans cette partie, un modèle ANN a été développé pour estimer la RZSM uniquement à partir de données in-situ de SSM. Différentes paramétrisations ont été appliquées sur le modèle en termes de paramétrage temporel des entrées SSM, la méthode de scaling et la division des sets d'apprentissage et de test. Les données in-situ de SSM ont été fournies par le réseau international d'humidité de sol (ISMN) à différents endroits dans le monde, de sorte que différents climats et types de sol sont couverts par les stations considérées. Cette particularité est importante pour évaluer la transférabilité de l'approche. La précision de la méthode a été étudiée en se basant sur une analyse climatique et de texture de sol.

Cette approche est détaillée dans l'article scientifique intitulé "Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe" et publié dans la revue internationale Water. Elle sera présentée dans le chapitre 2.

- **Prévision de la RZSM dans un ANN en se basant sur la SSM et des variables liées aux processus physiques**

Différents processus hydrologiques, comme les processus de diffusion, relient la RZSM à la SSM. Par exemple, la RZSM peut être dérivée de l'évaporation de surface via l'extraction des racines ou les remontées capillaires. Ceci nous a conduit à évaluer l'impact de l'ajout d'entrées liées aux processus physiques, en plus de la SSM, dans des modèles ANN et à trouver la meilleure combinaison qui assure les meilleures prédictions de la RZSM. L'indice d'eau du sol (SWI) a été calculé avec un filtre exponentiel récursif et utilisé pour représenter le processus d'infiltration. Le processus d'évaporation a été illustré par une efficacité d'évaporation calculée sur la base d'un modèle analytique simplifié et de données d'évapotranspiration potentielle (PET) issues du spectroradiomètre imageur à résolution modérée (MODIS). La dynamique de végétation a été déduite des séries temporelles de l'indice de végétation par différence normalisée (NDVI) fournies par le satellite MODIS. Plusieurs modèles ANN avec différentes combinaisons de variables d'entrée ont été développés.

Cette approche et les différents résultats sont présentés dans l'article scientifique "Integrating process-related information into an artificial neural network for root-zone soil moisture prediction", publié dans le journal Hydrology and Earth System Sciences (HESS). Elle sera détaillée dans le chapitre 3.

- **Cartes spatiales de RZSM à grande échelle et à résolution kilométrique**

L'évaluation de la fiabilité du modèle ANN précédemment développé pour produire des prédictions RZSM cohérentes à l'échelle continentale, à savoir l'Europe, est une étape décisive de cette étude. La faisabilité de l'estimation de la RZSM à grande échelle avec une méthode initialement calibrée et testée à l'échelle locale démontre sa généralisabilité à l'échelle globale. Le point de départ de cette étape est le modèle ANN présenté dans l'axe précédent. Différentes sources de SSM, notamment des données satellite radar, micro-ondes passives et de réanalyse, ont été utilisées pour calculer les features de la SSM. Une comparaison entre les différentes cartes de RZSM produites à différentes résolutions (1 km, 9 km et 36 km) a été effectuée ainsi qu'une validation avec des données RZSM in-situ collectées au niveau de quatre réseaux européens d'humidité du sol.

Cette approche est détaillée dans un article scientifique intitulé "Root-Zone soil moisture over Continental Europe using machine learning" soumis au journal *International Journal of Applied Earth Observation and Geoinformation*. Elle sera détaillée dans le chapitre 4.

Chapter 1: State-of-the-art

The following chapter encompasses elementary definitions inherent to soil moisture, the state-of-the-art of direct and indirect soil moisture retrieval techniques at various spatial and temporal scales and a review of the different methods for RZSM estimation. It identifies the limitations of the existing solutions and the needs for a better quantification of this component, hence the interest of this work.

Table of contents

1.1. Soil moisture definition 25

1.2 Processess related to soil moisture 26
1.2.1 Infiltration process 27
1.2.2 Rainfall-Runoff 27
1.2.3 Evapotranspiration 28

1.3 Measuring Soil moisture 29
1.3.1 Direct measurements 29
1.3.2 Indirect measurements 29
1.3.2.1 In-situ measurements 30
1.3.2.2 Remote sensing observations 31

1.4 Root-Zone Soil moisture estimation 38
1.4.1 Physically-based models 39
1.4.2 Data assimilation 40
1.4.3 Data-driven methods 42

1.1. Soil moisture definition

The soil moisture also called soil water content is the amount of water held in the soil at a given matric potential (Tuller and Or, 2005). Soil moisture can be expressed either as a gravimetric quantity θ_g (g/g) or as volumetric quantity θ_v (m^3/m^3). Gravimetric soil moisture represents the ratio of the mass of water present in a soil sample and the mass of the dry soil sample. This quantity can be obtained by weighing a moist soil sample (m_{wet}), oven drying it at 105°C and then reweighing it (m_{dry}).

$$\theta_g = \frac{m_{wet} - m_{dry}}{m_{dry}}$$

Volumetric soil moisture (θ_v) is defined as the volume of water held in a given soil volume may be expressed as a function of gravimetric soil moisture (θ_m) as follows:

$$\theta_v = \theta_m \frac{\rho_b}{\rho_w}$$

Where ρ_b is the dry bulk density (g/cm^3) of the soil, which is the weight of dry soil per the total soil sample volume, and ρ_w is the density of water (g/cm^3).

θ_v is replaced by θ_s which represents soil moisture at saturation and is attained when soil pores are totally filled with water and no air is present. However, a completely saturated soil is not achievable in practice (can be achievable however in peatlands for instance) due to air entrapment in the soil pores under wet conditions. Similarly, a completely dry soil is not realistic due to the presence of a residual moisture content θ_r. In agricultural contexts, other soil moisture parameters can be derived to describe the relations governing water in a soil sample, such as the plant available soil water content. This quantity is termed θ_{PAW} and is defined as the difference between the water content at field capacity θ_{FC}, and the water content at the permanent wilting point, θ_{PWP}. Field capacity is attained once free drainage (gravity forces) is over and is considered ideal for crop growth.

The permanent wilting point θ_{WP} (figure 3) indicates the stage below where plants irreversibly wilt and die because water is so firmly retained in the soil matrix and the roots water uptake is not enough to cover their need. The water content at field capacity (figure 3) depends on soil texture and the permanent wilting point depends on the plant type.

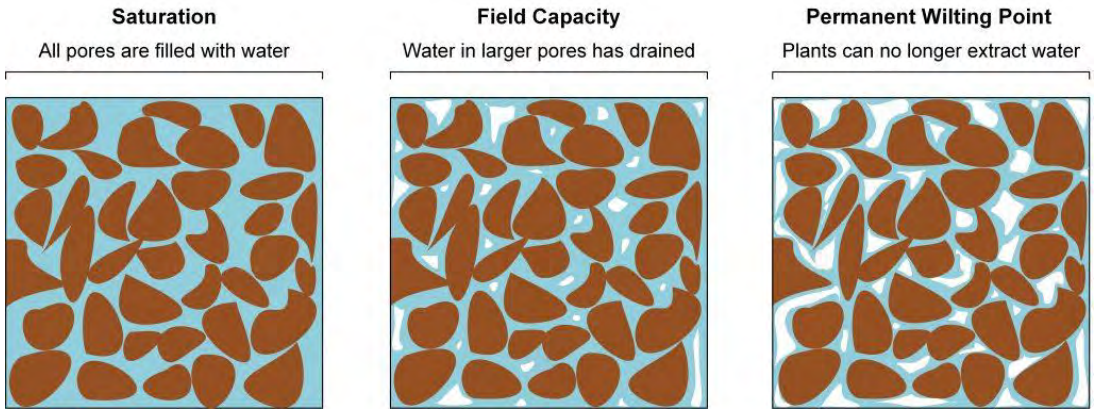


Figure 3: Soil water content at saturation, field capacity and permanent wilting point (Datta et al., 2017).

Soil moisture is often separated into two components namely surface soil moisture that corresponds to water in the upper soil (generally the top 5 centimeters) and the root-zone soil moisture that is available to plants (Figure 4) (Seneviratne et al., 2010). Soil moisture content may be termed, where applicable, surface soil moisture, near-surface soil moisture, root-zone soil moisture or vadose zone soil moisture. In the context of optical and thermal remote sensing, surface soil moisture also called skin soil moisture represents the water content held in the uppermost soil layer which thickness doesn't exceed 1mm. Near-surface soil moisture commonly denotes the average water content within the top few centimeters of the soil and is generally used in the context of microwave remote sensing. In this study, we don't make the distinction between the two terms. Only surface soil moisture will be used to refer to soil moisture in the first five centimeters of the soil. Root-Zone soil moisture refers to the water content available in the plant root zone which is available for transpiration and photosynthesis. Root water uptake by plants is one of the key components of the terrestrial water balance and a critical process controlling energy exchange between the land surface and the atmosphere and plant growth (Jarvis, 2011).

The accurate quantification of the depth of soil that roots can access, which contributes significantly to soil productivity, is quite challenging because of many factors namely the reduced pore volume, the abruptness of textural change over depth, the depth of soil to bedrock, the extremely acidic and or alkaline pH, and many other physical and chemical properties (Leenaars et al., 2018). The root profiles are also vegetation-dependant. Albers et al. (2022) consider a depth of 150 cm ideal to approximate the root zone depth outside all root-restricting zones. Many models and hypotheses have been proposed to estimate the rooting depth (Jackson et al., 1996; Musters and Bouten, 1999; Schenk and Jackson, 2002; Zeng, 2001; Leenaars et al., 2018; Rivieccio et al., 2020). This study is centered on the study of the soil moisture in the zone where roots develop and not where roots really are. The shallowest RZSM observation point we will consider in this PhD based on the soil moisture data we will be using, is equal to 30 cm and the deepest one is fixed at 55 cm.

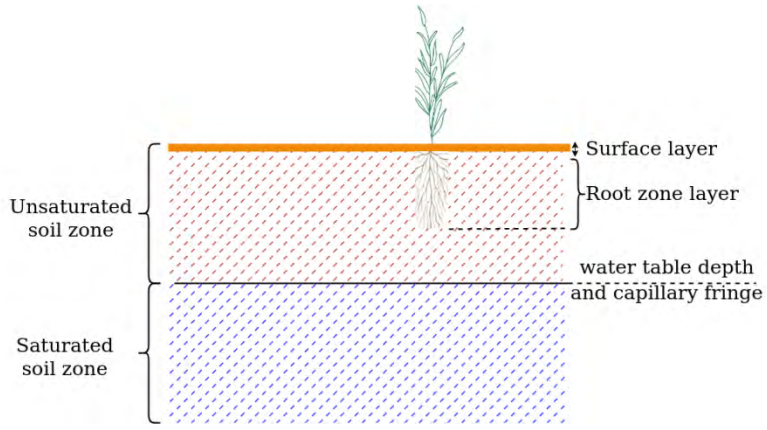


Figure 4: Diagram of unsaturated soil zone holding the soil moisture components and saturated soil zone.

1.2 Processess related to soil moisture

Soil moisture variations interact with different processes. For instance, the near-surface soil moisture variations interact with precipitation and evapotranspiration. Runoff, percolation and infiltration are also linked to the soil moisture variability (figure 3). Hereafter, some processes are detailed with respect to their interaction with soil moisture.

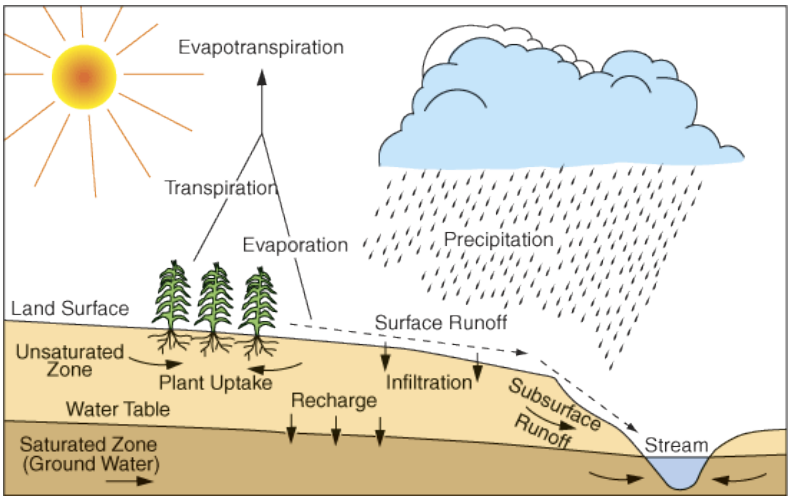


Figure 5: Hydrologic processes interacting with soil moisture (source: KGS Pub. Inf. Circ. 22, last access: 5 September 2022)

1.2.1 Infiltration process

Infiltration is one of the most important hydrological processes that are related to soil moisture. It describes the water movement from the surface into deeper soil. Several factors such as soil texture, irrigation or rainfall control this process.

The modeling of the infiltration process has gained much interest in the literature through several infiltration models which yield different levels of accuracy (Feki et al., 2018). These models are usually based on Richards' equation (Richards, 1931). Actually, the Richards' equation describes the flow of water in an unsaturated porous medium which is due to gravity and capillarity rise. The flow of the non-wetting phase usually air, is neglected (Farthing and Ogden, 2017). Due to its high nonlinearity, the numerical solutions which are proposed to resolve the Richards' equation are time-consuming especially in the case of large study areas and lead to stability issues under some conditions such as the wetting of an initially dry medium (Tinet al., 2015). Different simplifications were suggested and implemented by empirical, semi-empirical and physically-based models. Mishra et al. (2003) compared the performance of fourteen different infiltration models based on the Nash-Sutcliffe efficiency coefficient and 243 sets of in-situ infiltration data collected over India and USA such that different soil textures are covered. Feki et al. (2018) assessed the impact of modeling the infiltration process on soil moisture simulations accuracy. They included different infiltration models within a distributed hydrological model and evaluated their ability to simulate soil water content through a comparison against in-situ observations acquired in a maize field in northern Italy. Besides, they demonstrated that the Ross solution (Ross, 2003), which is a fast non-iterative solution for the non-linear 1D Richards' equation, was able to follow the soil moisture variability within the soil profile. In addition to the Ross solution, simpler analytical infiltration models allowed a good agreement between in-situ and simulated soil moisture if well calibrated.

1.2.2 Rainfall-Runoff

The quantification of stream flow that occurs in a river following a rainfall event is very important for different hydrological applications. The study of the rainfall–runoff processes encompasses

looking at where water goes when it rains, how long does water reside in a watershed, and what pathway does water take to the stream channel (Tarboton, 2003).

Woods et al. (2001) presented preliminary analysis related to rainfall-runoff response and soil moisture behavior over the Mahurangi River catchment in New Zeland and suggested that accurate measurement of soil moisture should be useful for runoff prediction. Actually, significant runoff was observed to be generated only for moisture contents above about 42%. Woods et al. (2001) suggested that at these high average moisture contents, the spatial distribution of soil water was critical in the prediction of runoff behavior and therefore should be well predicted.

Runoff is generated by different mechanisms and processes, which depend on the soil moisture status of the soil, referred to as the antecedent conditions. Actually, processes of evaporation, transpiration, percolation and drainage allow the definition of the soil moisture antecedent conditions (Tarboton, 2003).

These observations led to the development of different continuous simulation models, such as the National Weather Service (NWS) Sacramento soil moisture accounting model which is based on well-structured representation of the catchment's soil moisture storage system. This model is based on simple approximations of many soil moisture processes. Given that many of the catchment characteristics are related to the soil moisture capabilities of the catchment, a good application of the model starts with a good understanding of the three basic types of soil moisture which are hygroscopic, gravitational and capillary water and which can potentially influence catchment runoff conditions as reported in (Burnash, 1995). Hygroscopic water can be defined as soil water that is present not only in the pores but also on the surface of the soil particle and which is not available for plants. Gravitational water is the water that drains after moving through the soil by the force of gravity. Capillary water is the water which is held inside soil pores against gravity.

1.2.3 Evapotranspiration

The quantification of evapotranspiration (ET) is crucial for water resources management applications. Actually, ET establishes a link between the water, energy and carbon cycles and thus is very important for climate and hydrological applications. ET is a physically-based process that describes the water transfer from the soil layers and vegetation layer to the atmosphere. Over land, ET is made up of evaporation and plant transpiration. Evaporation is a physically-based process that describes the water transfer from different sources such as the soil, the surface of canopies, stems, or branches to the atmosphere. Transpiration is a bio-physical process that represents the evaporation of water in the vascular system of plants through leaf stomata (Verstraeten et al., 2008). Evapotranspiration is directly connected to soil moisture content which is subject to evaporation. Soil moisture has an impact on evapotranspiration since it has an influence on the partitioning of the available energy at the Earth surface into sensible and latent heat fluxes (Hirschi et al., 2020). Both land water balance and water balance for a given surface soil layer, without considering lateral exchange between adjacent soil volumes can be expressed using this same equation (figure 6):

$$\frac{dS}{dt} = P - E - R_s - R_g$$

where dS/dt is either the change of terrestrial water storage or water content within a given soil layer, P is precipitation, E is evapotranspiration, R_s is surface runoff, and R_g is groundwater flow in the case of land water balance or drainage in the case of a given soil layer water balance. S includes moisture stored in the soil, surface water, snow, ice cover and water stored in biomass.

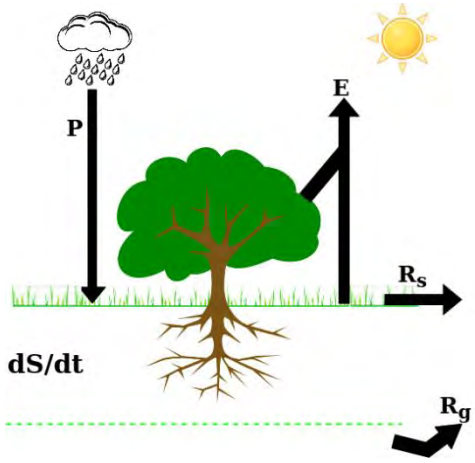


Figure 6: Diagram of the land water balance for a given surface soil layer; dS/dt represents the change in water content.

1.3 Measuring Soil moisture

Since the soil water balance is strongly affected by RZSM, accurate quantification of this component is essential in order to assess whether the available water can answer the plant needs or if it exceeds the plant demand and thus percolates below the root zone layer. The variability of RZSM can be evaluated using direct and indirect methods (Upadhyaya et al., 2021).

1.3.1 Direct measurements

The most common direct method of retrieving soil moisture in general and RZSM in particular is the gravimetric method which consists in weighing soil samples before and after over-drying at a temperature of 105 °C for 24-48 h. A detailed investigation of this method was described in (Reynolds, 1970). This method is simple, inexpensive and important to calibrate indirect measurement instruments. However, it is labor-intensive and time-consuming with regard to sample collection and drying. But the main disadvantage of this technique is the artificial change it brings to the experimental site. Actually, the gravimetric method is destructive because of the repeated sampling of a same site. It may damage the plant roots, modify the ground hydrological conditions and introduce variable drainage and infiltration characteristics.

Besides, RZSM can be directly measured using a lysimeter which is a device, typically a tank or a container. It defines a specific boundary to contain water in the soil and allows the measurement of the soil water balance or the volume of water percolating vertically or its quality. For a weighing lysimeter, the change in soil water storage is measured as a change in mass (Howell, 2004). However, the lysimeters are subject to many limitations. One main limitation is caused by the filling in of the soil into the lysimeter which can lead to a modification of the soil structure and might influence the condition of crops and soil life (Makkink, 1959).

1.3.2 Indirect measurements

Given the significant limitations affecting direct measurements, soil moisture is commonly observed by indirect measurement techniques or devices. Despite the good accuracy of the existing sensors, these techniques are hindered by a limited spatial coverage. Actually, the soil can't be digged at each point of the globe to install the sensors. The remote sensing technology overcomes this major limitation. However, the spatial and temporal resolutions can hamper soil moisture observation and need to be taken into account.

1.3.2.1 In-situ measurements

In-situ measurement techniques can provide both point-scale and larger impact measurements (decametric or hectometric scale). This section presents the most used techniques and is not a comprehensive review of all measurement techniques.

Neutron scattering

This technique was introduced and successfully used to indirectly measure soil moisture in the 1950s (Belcher et al., 1950; Gardner and Kirkham, 1952). Neutron probes are made up of a probe which contains a source of fast neutrons and a gauge that monitors the flux of slow neutrons scattered by the soil. In order to get root-zone measurements, the probe is lowered down an access tube until the intended depth of measurement. Neutrons with a high energy are scattered into the soil and are slowed by elastic collisions with nuclei of atoms. In soils, water is the major source of hydrogen atoms that can slow fast neutrons much more effectively than can any other element. Slow neutrons returning to the detector per unit time are counted. The density of the resultant cloud of slow neutrons is a function of the soil moisture content (Chanasyk and Naeth, 1996).

Dielectric sensors

These sensors are based on techniques which derive soil moisture from the dielectric property of the soil. These techniques mainly encompass Time Domain Reflectometry (TDR), capacitance technique and Frequency Domain Reflectometry (FDR) and are based on the fact that dielectric constant of soil is primarily related to its water content (Thomas, 1966; Cihlar and Ulaby, 1974; Hoekstra and Delaney, 1974; Topp et al., 1980). The dielectric constant is a measure of the response of the soil to an electromagnetic wave and is equal to a few units for dry matter and about 80 for free liquid water. Different sensors based on electromagnetic measurements of this constant have been developed by different companies (Vaz et al., 2013).

TDR sensors like the TDR 100 Campbell sensor emit an electromagnetic impulse and observe the response within an interval of time. Then, they convert the time taken by the wave to travel and to get reflected back to the receiver to distance unit, and display the information as a waveform. The volumetric soil moisture can be estimated using the dielectric constant of the soil (Abdullah et al., 2018). Besides, the capacitance technique relies on the fact that capacitance of soil is directly related to the dielectric constant and thus to water content in soil. Many capacitance soil moisture sensors consist of a probe and a pair of embedded electrodes. The Decagon 10HS and 5TE are two widely used capacitance sensors. The capacitance between the probes varies as a function of soil moisture (Selig et al., 1975). An oscillator applies a frequency between 50 and 150 MHz to the electrodes, which generates a resonant frequency whose magnitude depends on the dielectric constant of the soil. Volumetric soil moisture content is then estimated using the frequency and a calibration equation.

FDR offers an inexpensive measurement of soil water content and is an alternative to TDR. FDR sensors consist of short probes which make them geometrically more advantageous than TDR sensors. Different soil properties can be measured by selecting different frequencies since FDR sensors are sensitive to different physical and chemical soil properties in different frequency ranges based on the dielectric spectra of the soil (Xu et al., 2012).

Gamma attenuation technique

The gamma ray attenuation is a radioactive technique that can be used to determine soil water content. This method assumes that the scattering and absorption of gamma rays are linked to the density of matter in their path and that the specific gravity of a soil remains relatively constant as the wet density changes with increases or decreases in moisture. The gamma transmission technique measures changes in wet density to infer soil moisture (Zazueta and Xin, 1994). Unlike the direct method, this is a nondestructive technique with a fast response time (less than 1 min). However, it is expensive and difficult to use.

GNSS

As aforementioned, point-scale measurements are not relevant for all applications given the high temporal and spatial variability of soil moisture. Given that soil moisture data can be required at large scales, the Global Navigation Satellite Systems (GNSS) such as the Global Positioning System (GPS) receivers are a good alternative for soil moisture estimation through the analysis of the power variations of the GNSS signals recorded on the ground. Although GNSS was exclusively used to determine position, GPS receivers are sensitive to soil moisture. GPS stations are capable of providing a large network of observations with individual spatial scales of 10–40 m (Larson et al., 2008). The power of the GNSS signal is expressed as signal-to-noise ratio (SNR) which is equal to the ratio of the GNSS signal power to the measurement noise. This ratio is commonly used to assess the quality of the signal surrounding the GNSS station. The GNSS antenna simultaneously receives the direct and reflected signal from the GNSS satellite and land surface, respectively. The received signal is subject to an interference pattern which depends on the height difference between the GNSS antenna and the reflection point. It depends also on the elevation angle of the satellite and the GNSS frequency due to the motion of the satellite. Larson et al. (2008) compared reflection amplitudes at a GPS site in Uzbekistan to estimates of soil moisture from a land surface model over a 70-day period. Both estimates were consistent with rainfall events. The soil moisture estimates provided by GNSS and the land surface model increased when a rainfall event occurred and decreased over a period of around 10 days. This makes GNSS technology promising with consideration of the technical issues related to receiver/antenna differences.

1.3.2.2 Remote sensing observations

As aforementioned, soil moisture can be measured using direct or indirect in-situ techniques which are advantageous given their easy installation, their ability to measure soil moisture at different depths and their relative maturity. Despite their high accuracy, they are often costly and labor-intensive and sometimes destructive (e.g. gravimetric sampling) (Petropoulos et al., 2015). Besides, in-situ measurements cannot well represent the spatial distribution of soil moisture and are not suitable for continuous spatial and temporal coverage at regional and global scales (Rahimzadeh-Bajgiran et al., 2013).

In the last few decades, many advances in remote sensing (RS) techniques have been made to provide seamless soil moisture measurements from space. These techniques differ by the relationships governing the remotely-sensed signal and the soil moisture, the wavelength region of the electromagnetic spectrum used and the source of the electromagnetic energy (Table 1).

However, current remote sensing technology only provides SSM and no current satellite sensor is able to provide RZSM due to the limited penetration depth into the soil. SSM can be derived at different electromagnetic spectra ranging from the optical to the microwave ranges. Research is evolving to allow retrieval of RZSM. In this context, P-band sensors which are still under investigation are quite promising for soil moisture observation at deeper layers of soil (Shen et al., 2021).

Table 1. Overview of remote sensing techniques for soil moisture estimation (after Wang and Qu, 2009; Babaeian et al., 2019).

Spectrum domain		Properties observed	Advantages	Drawbacks
Optical		Soil reflection	- Fine spatial resolution and broad spatial coverage. - Potential for real-time applications (e.g., drones).	- Limited surface penetration depth (in the order of a few millimeters). - Signal contamination by clouds and vegetation. - Low temporal resolution.
Thermal infrared		Surface temperature	- Fine spatial resolution and broad spatial coverage. - Potential for real-time applications (e.g., Drone). - Strong correlation between SSM and surface land temperature.	- Limited surface characteristic depth (few millimeters). - Signal perturbed by clouds, meteorological conditions and vegetation biomass.
Microwave	Active	Backscatter coefficient Dielectric properties	- Moderate characteristic depth (around 5 cm). - High spatial	- Low temporal resolution - Signals perturbed by surface

			resolution and broad spatial coverage (global scale) - Backscatter is independent of solar illumination, clouds and atmospheric constituents → low atmospheric noise - Strong correlation between SM and backscattered power	roughness and vegetation biomass
	Passive	Brightness temperature Dielectric properties Soil temperature	- Moderate characteristic depth (around 5 cm) - High temporal resolution - Broad spatial coverage (global scale) - Brightness temperature not perturbed by atmospheric constituents and clouds.	- Low spatial resolution - Signals perturbed by surface roughness and vegetation biomass.

I- Satellite data

Optical RS

Reflectance-based methods can provide estimates of soil moisture at high spatial resolutions compared with other types of sensors such as microwave instruments. Despite the multitude of optical sensors that are currently in orbit, a limited body of literature exists on the use of optical observations to retrieve SSM (Petropoulos et al., 2015). Different studies have been conducted to infer the relationships between spectral reflectance and SSM (Ben-Dor et Banin, 1995; Chang et al.,

2001; Gao et al., 2013). The conclusions of many of those studies show that soil reflectance decreases with increasing soil moisture namely in the Short-Wave Infrared (SWIR) range (Moran et al., 1994, Weidong et al., 2002, Zhan et al., 2007).

Empirical and physical approaches have been proposed to estimate soil moisture from measured surface reflectance. Most of the empirical methods are based on the rationale of developing empirical spectral indices (WISOIL: Bryant et al, 2003; the Shortwave Angle Slope Index (SASI): Khanna et al., 2007; Normalized Soil Moisture Index (NSMI): Haubrock et al., 2008). However, these indices are very sensitive to the effects of the atmospheric water vapor. Also, empirical methods have proved reliable for soil moisture estimation under the conditions they were developed for but cannot be universally applied outside those conditions. This is due to the significantly varying soil composition that strongly affects spectral reflectance (Liu et al., 2009).

In addition to empirical approaches, physically-based models have been developed for soil moisture estimation (Bach and Mauser, 1994; Philpot, 2010; Bablet et al., 2018). Bach and Mauser (1994) describe a spectral extension to the VIS-SWIR of Ångström's model (Ångström, 1925) which accounts for light absorption in the water layer. Sadeghi et al (2015) proposed a model based on the Kubelka-Munk two-flux radiative transfer model (Kubelka and Munk, 1931) to estimate soil moisture while considering the effects of the absorption by soil water and soil particles and the scattering caused by the soil particles. Despite the promising results, its field of application is restrained since it can only be applied at some wavelengths. Bablet et al. (2018) developed a multilayer radiative transfer model of soil reflectance (MARMIT) to estimate soil moisture content. Bablet et al. (2020) proposed a laboratory experiment to assess surface and root zone soil moisture thanks to a spectrometer, two hyperspectral cameras, and the MARMITforSMC method which is based on the MARMIT model and was applied to each reflectance spectrum to produce high spatial resolution maps of soil moisture. Vertical profiles of soil moisture content were obtained with unprecedented spatial accuracy (~0.287 mm). A main drawback of optical observations is the limited surface information caused by clouds, water vapor and aerosols that can easily disturb the signals (De Jeu et al., 2008).

Thermal infrared RS

Thermal infrared (TIR) remote sensing measures the thermal emission of the Earth with an electromagnetic wave band between 3.5 and 14 μm (Curran, 1985). TIR techniques for SSM estimation primarily rely on land surface temperature (LST) measurements alone using the thermal inertia (T) method (Pratt and Ellyett, 1979; Verstraeten et al., 2006) or combined with vegetation indices (Claps and Laguardia, 2004; Carlson, 2007).

The thermal inertia (TI) method is straightforward and simple and can accomplish high accuracies in assessing soil moisture conditions. It relies on the fact that variations in soil moisture have a strong influence on the thermal properties of the soil, which is an intrinsic factor of soil surface temperature change (Wang and Qu, 2009). Verstraeten et al. (2006) found that thermal inertia i.e. resistance to temperature variation proportionally increases when soil water content increases, thereby reducing the diurnal temperature fluctuation range. The apparent thermal inertia (ATI) is a simple surrogate of TI and can be computed using MODIS albedo and LST products (Van doninck et al., 2011). The TIR-based methods have advantages of various spatial resolution satellite images available and link soil moisture to thermal inertia. However, the weak relationship between TIR

images and soil moisture in the densely vegetated areas impedes the applications of the TIR methods (Liu et al., 2020).

As aforementioned, several studies have applied the Richards' equation to retrieve root-zone soil moisture from surface soil moisture. Given the requirement of a precise definition of the soil physical characteristics which is not possible neither at the field scale nor at larger scales, simplifications of this approach based on multi-layer models have been proposed (Wagner et al., 2007). Ottlé and Vidal-Madjar (1994) proposed a two-layer model to indirectly retrieve surface soil moisture from TIR data. They showed that the use of a two-layer parameterization of the surface and the consideration of the vegetation and its evolution improve the daily simulation of soil moisture and of the water flows. Assimilating soil moisture inferred from thermal IR imagery in the model was overall the best performing option.

Microwave RS

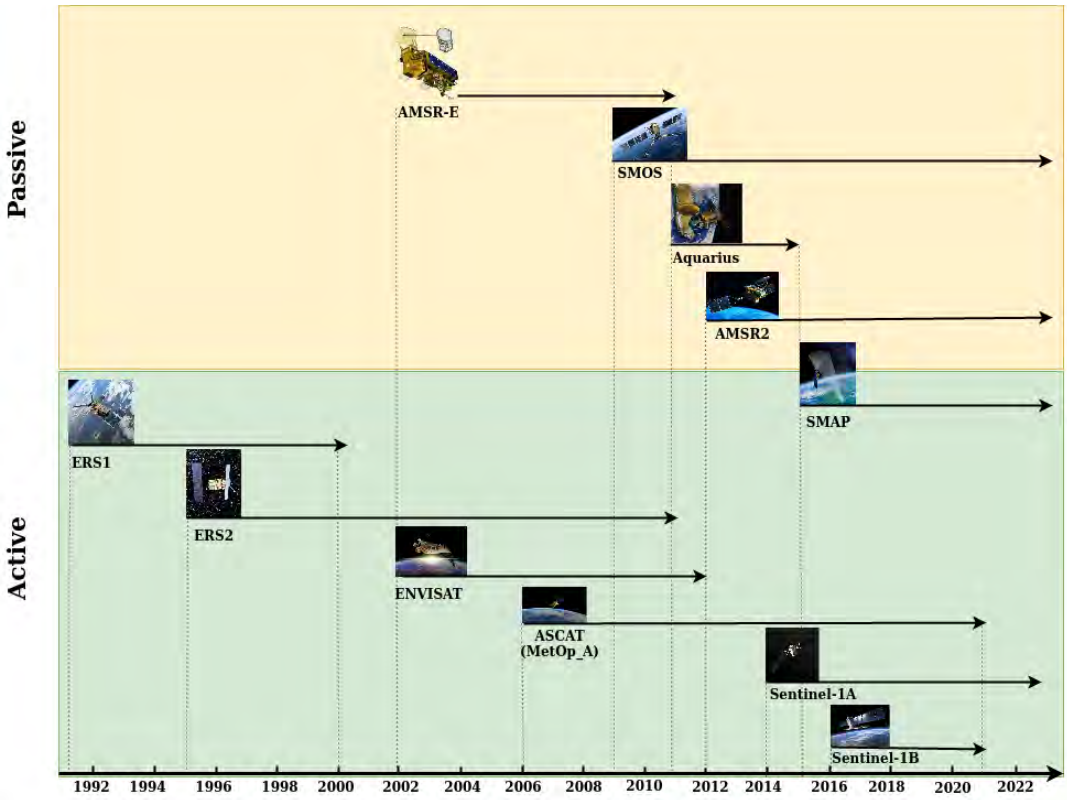


Figure 7: Some past and current microwave missions for SM retrieval

Active and passive microwave remote sensing missions (figure 7) have been extensively studied and proven promising for soil moisture monitoring at global and regional scales over the past decades (Jackson and Schmugge, 1989; Njoku and Entekhabi, 1996; Wigneron et al., 1998). At microwave frequencies, the most striking feature of the emission from the Earth's surface is the large contrast between water and land which is due to the large dielectric constant of water compared to most dry minerals (Schmugge et al., 2002). Since microwave sensors are sensitive to the dielectric properties of the soil, they are also highly sensitive to soil moisture.

- *Passive microwave:*

Passive microwave technology was demonstrated interesting for SSM monitoring and mapping over land surfaces (Njoku and Kong, 1977; Jackson et al., 1995; Wigneron et al., 2004). Passive microwave sensors measure the intensity of microwave emission from the soil, which is proportional to the product of the thermodynamic temperature of the soil and the surface emissivity called brightness temperature (TB).

L-band (1-2 GHz) radiometers have been massively studied over the last decades and are considered as an excellent tool to map soil moisture at global scale. Missions with spaceborne L-band radiometers include the European Space Agency (ESA) Soil Moisture and Ocean Salinity (SMOS) mission which was launched in 2009 (Kerr et al. 2010) and provides brightness temperature measurements at global scale with dual polarization and a broad range of incidence angles. Another mission is the Soil Moisture Active Passive (SMAP) which was launched in 2015 (Entekhabi et al. 2010) by the National Aeronautics and Space Administration (NASA). It had both passive and active microwave sensors onboard that provide global measurements of brightness temperature and backscatter at an incidence angle of 40°. Aquarius/SAC-D is also a space mission using this technology. It is equipped with the Aquarius instrument onboard which is a combined active/passive L-band microwave instrument designed to map the sea surface salinity (Le Vine et al., 2010). Despite its primary science objective, Aquarius was also used to retrieve SSM (Bindlish et al., 2015).

Jackson et al. (2010) assessed four Advanced Microwave Scanning Radiometer–Earth Observing System (AMSR-E) algorithms for soil moisture estimation using a seven-year record period of in-situ observations from four experimental networks in the U.S which had different climate classes. They found that all algorithms had similar correlation values ranging between 0.71 and 0.79 if site-specific corrections were applied. However, each algorithm had a different performance at each site.

Kerr et al. (2016) performed a global evaluation of different SMOS soil moisture products through comparisons against model simulations, other satellites and in situ measurements. Results showed that SMOS yielded consistent estimations and behaved very well when compared to other sensors and approaches. However, limitations were encountered namely the Radio Frequency Interference (RFI) which hindered the detection. A comparison against other satellite products also demonstrated the relevance of the SMOS observations over different eco climate regions and throughout the seasons.

Reichle et al. (2017) investigated the SMAP L4 soil moisture product through a validation against in-situ measurements from SMAP core validation sites. Unbiased RMSE (ubRMSE) values equal to 0.038 m^3/m^3 and 0.035 m^3/m^3 were recorded at the 9 km and 36 km scales, respectively. These performances meet the soil moisture accuracy requirement fixed at 0.04 m^3/m^3 (ubRMSE). Besides, results showed that this product outperformed model-only estimates. Reichle et al. (2017) better highlighted these findings through a validation against point-scale in-situ measurements from around 400 sparse network sites which cover a wide variety of climate and land cover conditions.

- Active microwave

Compared to radiometers, active microwave sensors can provide observations of backscatter at higher spatial resolutions. The magnitude of the SAR backscatter coefficient is linked to surface soil moisture through the contrast of the dielectric constants of bare soil and water. Active microwave sensors which consist of imaging (radars) and non-imaging sensors (altimeters and scatterometers), provide their own source of illumination and measure the difference in power between the transmitted and received electromagnetic radiation (Barrett et al., 2009). Approaches based on Synthetic aperture radars (SAR) are attractive for applications on watershed and field scale since SAR can reach a high spatial resolution and since backscatter signal is sensitive to SSM (Van Doninck et al., 2011). However, radar data are strongly affected by soil roughness and vegetation which makes the accurate inversion of backscatter to soil moisture difficult (Lakshmi, 2013; Verhoest et al., 2008).

The retrieval of surface soil moisture using the X-band SAR techniques was extensively studied mainly over bare areas. Baghdadi et al. (2012) assessed the potential of TerraSAR-X data for the estimation of soil moisture over bare soils using empirical models. The inversion of soil moisture from one and multi-incidence SAR data was tested. TerraSAR-X was demonstrated a reliable remote sensing tool for surface soil moisture estimation with an accuracy of about 3% (RMSE). However, this wavelength is constrained by the vegetation cover. Hence, several studies were conducted to estimate surface soil moisture at lower frequencies.

Lievens and Verhoest (2012) worked on the retrieval of surface soil moisture from a time series of multi-incidence HH and VV polarized RADARSAT-2 backscatter observations over a number of bare soil fields in The Netherlands. Two retrieval techniques were assessed namely the Integral Equation Model (IEM) which is a physically-based backscatter model and a change detection technique based on the rescaling of the SAR backscatter observations between dry and wet reference values over time periods with unchanged surface roughness. Large agreements between in situ measurements and radar backscatter were recorded across time and space, mainly over areas with medium surface roughness conditions.

The potential of artificial neural networks to estimate regional soil moisture from Sentinel-1 SAR data was investigated in (Paloscia et al., 2013). The model was trained using backscatter coefficients simulated from IEM and water cloud model (WCM) for different soil moisture, soil roughness and vegetation conditions. The neural network which aimed at predicting soil moisture based on SAR data and NDVI yielded soil moisture estimation accuracies between 2 and 5 vol % when validated against a database of in-situ measurements, SAR and optical data.

Wagner et al. (2013) investigated the retrieval of soil moisture using the Advanced Scatterometer (ASCAT) which is a C-band active microwave instrument on board of the Meteorological Operational (MetOp) satellites. They highlighted the attractivity of ASCAT to observe soil moisture given its wavelength, its high radiometric accuracy and its multiple-viewing capabilities. Good performances were recorded over some regions in Europe whereas limitations were observed over mountainous and some desert areas.

Tomer et al. (2015) proposed an algorithm based on the Cumulative Density Function (CDF) transformation of multi-temporal RADARSAT-2 backscatter coefficient to produce soil moisture values. A database of RADARSAT-2 images, SMOS L2 soil moisture products and in-situ soil moisture measurements over a semi-arid tropical region in South India was used to assess the

approach. The algorithm was proved able to estimate soil moisture with the advantage of not requiring any parameter calibrations. RMSE values between 0.02 m^3/m^3 and 0.06 m^3/m^3 were obtained when estimations were compared against in-situ data. A good agreement was also found with SMOS soil moisture data.

II- Unmanned aerial vehicles data

Unmanned aerial vehicles (UAV) also referred to as drones, unmanned aerial/aircraft systems (UAS) or remotely piloted aircraft systems (RPAS), can be considered as a low-cost alternative to conventional remote sensing platforms. These are actually data-gathering and transmitting aircrafts which are remotely controlled and are able to realize airborne operations. There are different types of UAVs which are commonly differentiated by the size of the vehicle, its altitude and the flight endurance. UAV remote sensing techniques have been shown very promising for several environmental monitoring applications. Sanchez-Azofeifa et al. (2017) investigated the reliability of these platforms for forest studies. They highlighted that UAVs can be well-suited for addressing current issues in remote sensing of tropical ecology and conservation since the low-altitude UAVs overcome significant constraints of high aerosol interference from water vapor and clouds. Unlike satellites which may entail costly solutions, UAVs can provide high-resolution and low-cost imagery to monitor active deforestation fronts and quantify ecosystem degradation for instance. Besides, UAVs can be reliable for agricultural applications. For instance, UAVs have been employed in order to map soil moisture using different sensors. Hassan-Esfahani et al. (2015) have used the AggieAir™ platform which provided UAV data that was fed to an artificial neural network (ANN) in order to estimate surface soil moisture. Multispectral and thermal images collected over an irrigated field at a 15 cm and 60 cm resolution were used as inputs to the ANN. The results showed the reliability of the ANN model to spatially estimate surface soil moisture at much finer spatial and temporal resolutions compared to conventional remote sensing technologies. A correlation coefficient equal to 0.88 was recorded for four dates in 2013 (16 May, 1 June, 9 June, and 17 June).

Besides, Lu et al. (2020) investigated the capability of UAVs to estimate soil moisture in a typical steppe namely the Loess Plateau of China. They confirmed that the average pixel brightness value of UAV visible images, which is defined by a computer to represent the brightness of images, could estimate the 0–10 cm soil moisture. Actually, the determination coefficient (R^2) between the in-situ and the estimated value of soil moisture was equal to 0.82 and 0.77 when the 0-10 cm soil moisture was at a stable value and larger than the stable value, respectively. These findings make UAVs a promising tool for soil moisture mapping in arid and semi-arid steppe in particular, and for steppe ecological research in general.

1.4 Root-Zone Soil moisture estimation

As highlighted in the previous sections, most of RS methods can only monitor skin and near-surface soil moisture. Given the importance of RZSM in better understanding of the agricultural and environmental processes it controls, close cross-disciplinary collaborations between the RS community and soil physicists and hydrologists have been established to link RZSM and remotely sensed skin and near-surface data (Babaeian et al., 2019). RZSM can be measured by in-situ sensors installed horizontally at a fixed depth or vertically along the soil profile (Walker et al., 2004; Francesca et al., 2010; Dobriyal et al., 2012). However, the installation of sensors can be a complex

task and might disturb the soil properties. This justifies the common approach of deriving RZSM from surface in-situ or RS soil moisture (Carranza et al., 2021).

1.4.1 Physically-based models

Land Surface and agronomical Models:

Land Surface Models (LSMs) are conceived to model surface and root zone soil moisture using physical and hydrological laws. These models can be considered as a promising tool for an enhanced representation of root-zone soil water dynamics relative to soil moisture proxy products (Crow et al., 2012). Actually, LSMs dynamically predict vertically-discretized profile soil moisture based on a complex representation of water flow within the soil column. LSMs require different forcing data as inputs namely precipitation, air temperature, air pressure, relative humidity, wind speed and solar radiation. Vertical soil water processes such as infiltration and drainage, depend on soil hydraulic properties which are linked to soil texture through pedo-transfer functions.

Different LSMs exist such as the Surface Externalisée (SURFEX) (Le Moigne et al., 2009), ORCHIDEE (Krinner et al., 2005), the Joint UK Land Environment Simulator (JULES) (Best et al., 2011), etc. However, Noah, the Catchment Land Surface Model (CLSM) and the Community Land Model (CLM) are the commonly used models. Noah and CLM are traditional land surface schemes that model soil moisture dynamics by solving a layer-based formulation of the standard diffusion and gravity drainage equations for unsaturated flow (Kumar et al., 2009). Noah uses four soil layers of increasing thicknesses of 10, 30, 60 and 100 cm and CLM uses ten unevenly spaced soil layers with thicknesses of 1.75, 2.76, 4.55, 7.5, 12.36, 20.38, 33.60, 55.39, 91.33, and 113.7 cm. An integrated RZSM product can be obtained by averaging the top three Noah layers and the top eight layers of CLM. CLSM is a non-traditional model since the vertical soil moisture profile is determined through deviations from the equilibrium soil moisture profile between the surface and the water table. Soil moisture is computed within both a 2-cm surface layer and a 1-m root-zone layer (Koster et al., 2000).

However, these models are constrained by the need for many inputs, the incompatibility between the relatively low spatial of surface and hydrological processes and the meteorological forcing errors mostly for rainfall that has the highest impact on soil moisture variability (Sabater et al., 2007).

Empirical surrogate models:

Wagner et al. (1999) proposed a simplified two-layer model to estimate the SWI of the root zone from European remote sensing (ERS) surface soil moisture data using an exponential filter. In this formulation, only one parameter (T) has to be calibrated. This parameter is called the characteristic time length and represents the timescale of soil moisture variations in units of time usually days (Ceballos et al., 2005). This parameter can be considered as a surrogate parameter for all the processes that impact the temporal dynamics of soil moisture such as the thickness of the soil layer, soil hydraulic properties, evaporation, run-off and vertical gradient of soil properties (Albergel et al., 2008). It is proportional to the ratio of the depth of the reservoir below the surface and a pseudo-diffusivity coefficient. The discrete formulation of the model can be expressed as follows:

$$SWI(t_n) = \frac{\sum_i^n ms(t_i) e^{-\frac{t_n - t_i}{T}}}{\sum_i^n e^{-\frac{t_n - t_i}{T}}}$$

Where ms(t_i) is scaled surface soil moisture estimated from remote sensing at time t_i by extrapolating and scaling the observed backscatter between the minimum and maximum values observed during the lifetime of the remote sensing instrument (wagner et al., 1999).

Stroud (1999) introduced a recursive formulation of the exponential filter which allowed an easier data manipulation compared to the original formulation. The potential of the exponential filter and its recursive formulation for RZSM estimation was highlighted in many studies (Albergel et al., 2008; Ford et al., 2014; Ceballos et al., 2005). Although SWI has the advantage of being independent of all land surface model or meteorological observations, the physical interpretation of the time constant T is challenging. Albergel et al. (2008) could not infer any significant relationship between this parameter and the main soil properties over France. The recursive formulation of the exponential filter can be written as follows:

$$SWI_{t_n} = SWI_{t_{n-1}} + K_n(ms(t_n) - SWI_{t_{n-1}})$$

Where SWI is the soil water index at time t_n, ms(t_n) is the scaled surface soil moisture at time t_n, K_n is the gain at time t_n, which occurs in [0, 1] and is equal to:

$$K_n = \frac{K_{n-1}}{K_{n-1} + e^{-\frac{t_n - t_{n-1}}{T}}}$$

For the initialization of the filter, gain K_1=1 and SWI_{t1}=ms(t_1) .

In order to tackle this limitation, Manfreda et al. (2014) developed the Soil Moisture Analytical Relationship (SMAR) model to estimate the soil moisture in the root zone based on the SSM. SMAR is derived from a simplified soil water balance equation for semiarid environments and establishes a relationship between the root zone and the surface soil moisture with a limited number of physically consistent parameters. In this model, the soil is assumed to be composed of a surface layer with a depth in the order of few centimeters and a layer below which is assumed equivalent to the rooting depth of vegetation. Infiltration is considered as the most representative process for the most significant water mass exchange between the two layers. Other processes such capillary rise are assumed negligible. In contrast to the SWI method, there are clear physical interpretations for the SMAR parameters which can be easily determined based on the soil texture and climate of the target location (Faridani et al., 2017).

1.4.2 Data assimilation

Data assimilation is a tool that combines observation data (from remote sensing or in-situ) and a dynamic model that includes the principles governing the system. It aims at providing a better estimate of the state of the system than data or model-only estimates (Zhang and Moore, 2015).

Data assimilation techniques rely on the relationship between surface and root zone through diffusion process, to propagate surface information to deeper soil layers (Entekhabi et al., 1994; Walker et al., 2001; Lü et al., 2010). Thus, the subsurface physics used in the LSMs is an important factor in determining the strength and validity of the downward propagation of surface information (Kumar et al., 2009). The Land Data Assimilation Systems (LDASs) can tackle the limitations of LSMs and provide a broader spatio-temporal coverage, a better consistency and accuracy of LSM estimates by using both in-situ and remote sensing soil moisture retrievals.

The RZSM estimates can be enhanced by assimilating remotely-sensed SM observations into a Soil Vegetation Atmosphere Transfer (SVAT) model. Reichle and Koster (2005) assimilated SSM retrievals at a global scale from the Scanning Multichannel Microwave Radiometer (SMMR) into a CLSM. Overall, the assimilation improved the average annual cycle of surface and root-zone soil moisture at specific locations. Validation against in-situ data showed that correlations for root-zone soil moisture were improved.

The Kalman filter (KF) and its extensions such as extended Kalman filter (EKF) and ensemble Kalman filter (EnKF) are sequential assimilation methods that have been extensively applied for soil moisture estimation. Walker et al. (2002) applied the KF data assimilation technique to a distributed three-dimensional soil moisture model in order to retrieve of profile soil moisture in a 6 ha catchment using near-surface soil moisture measurements. The EnKF is a widely-used method given its skill in handling non-linear systems and computational efficiency (Reichle et al., 2002; Crow and Wood, 2003). De Lannoy and Reichle (2016) assimilated L-band microwave brightness temperature observations using a spatially distributed EnKF and demonstrated that data assimilation improves both surface and root-zone soil moisture results over the open-loop (no assimilation) estimates in areas with limited vegetation and terrain complexity.

François et al. (2003) used an extended Kalman filter to assimilate soil moisture estimations provided by the European Space Agency (ESA) Remote Sensing Satellite (ERS) SAR in a two-layer hydrological model. The assimilation was conducted in the Orgeval agricultural river basin over two years. Results showed that this approach improved the Nash–Sutcliffe efficiency (NSE) for streamflow from 70 to 85% and demonstrated a higher sensitivity of the simulated flow to soil moisture in case of high soil moisture. The assimilation method was also able to correct for up to 10% errors in the input data such as potential evapotranspiration.

Heathman et al. (2003) used the Root Zone Water Quality Model (RZWQM) which is a physically-based and field-scale agricultural model, to study the feasibility of assimilating SSM for better estimation of RZSM. The validation against in situ data showed that the integration of data assimilation produced better model simulation results in the top 30cm layers than the model simulation without assimilation. Han et al. (2012) proposed an extension to RZWQM (Heathman et al., 2003) and applied field measured surface soil moisture to a point scale model.

Sabater et al. (2007) investigated a simplified one-dimensional variational data assimilation (1DVAR) technique to correct the modeled RZSM deficiencies of the ISBA model, using the observations of the surface soil moisture of the Surface Monitoring of the Soil Reservoir Experiment (SMOSREX). Given its lower computing time, the 1DVAR was considered a good alternative to the EnKF for the development of an operational data assimilation system that analyzes RZSM from SSM observations.

1.4.3 Data-driven methods

Given the amount of data required to parameterize physically-based models, errors in the models physics may propagate even if coupled with data assimilation techniques. Data-driven techniques such as Random Forests (RFs), ANNs and Support Vector Machines (SVMs) are increasingly being investigated for soil moisture estimation and have been proved reliable in many studies. Machine learning (ML) offers different methods capable of developing quantitative models without having assumptions on the inputs or on the investigated target.

ML is a branch of Artificial Intelligence (AI) that systematically applies algorithms to synthesize the underlying relationships among data and information (Awad and Khanna, 2015). AI is a field of computer science that allows computer to mimic the human behavior. In the context of AI, computers or machines in general achieve tasks according to stipulated rules and pre-established algorithms. AI is a superset of any computer program that requires human intelligence and comprises machine learning and deep learning (DL) (figure 8).

Machine learning characterizes the ability of a system to learn from problem-specific training data to automate the process of analytical model building and solve associated tasks (Janiesch et al., 2021). ML seeks to automatically learn meaningful relationships and patterns from examples and observations (bishop, 2016). ML techniques have been continuously improving through the implementation of more sophisticated learning algorithms and pre-processing approaches. Overall, the ultimate objective of ML is to predict future events that are unknown to the computer. Mitchell (1997) defined learning as follows: “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. Awad and Khanna (2015) highlighted that learning is a fundamental process to generalize a problem by acting on its historical experience. The training datasets define the experience and allow for largely accurate results on unseen tasks.

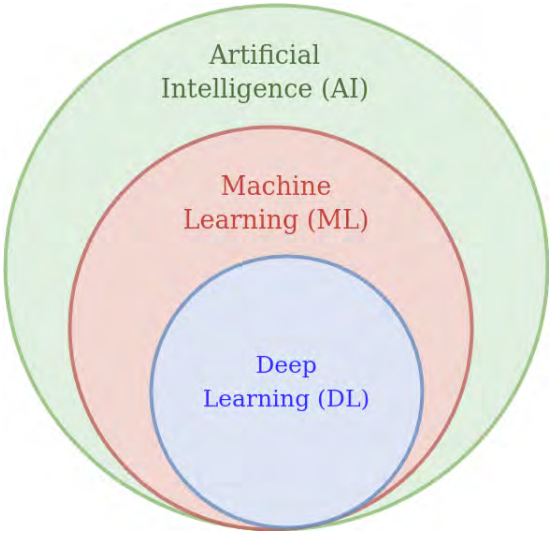


Figure 8: Venn diagram of the artificial intelligence and some of its integrated technologies.

As far as the terminology related to machine learning is concerned, we can cite supervised and unsupervised learning. Supervised learning encompasses learning techniques through which

machines predict the output based on labeled training data. Labeling data entails associating input data with the correct output. Actually, supervised algorithms deduce the relationship between the input data (the observations) and the output which is called target. The generalizability and predictive performance of the models which are trained using supervising learning depend among others on the size of the training dataset. As for unsupervised learning techniques, the algorithms are able to detect hidden patterns. This type of learning is well suited for image recognition or exploratory data analysis for instance. To mention but a few, dimensionality reduction and clustering algorithms are unsupervised.

Another interesting term in machine learning is “feature vector”. It is an n-dimensional numerical vector of explanatory variables given as an input to the problem. The feature vectors are often weighted to construct a predictor function that is used to assess the quality of the prediction (Awad and Khanna, 2015). The different steps of developing a ML algorithm can be summarized in figure 9. First, all of the relevant data subsets for the problem resolution are collected. Then, raw data is pre-processed i.e. converted to a useable format, cleaned by omitting corrupt data or filling gaps for instance and sampled such as redundancy is minimized and loss of information is avoided. The third step consists in transforming the data based on the considered machine learning algorithm. Data transformation encompasses for instance feature scaling which is an important step in numerous machine learning tasks mainly if the features have different value ranges (Bollegala, 2017). Not scaling features is likely to make the feature of the highest value range the dominating one. In order to achieve faster converge in many machine learning algorithms, scaling is a fundamental step. Normalization and standardization are the most commonly used scaling techniques. Normalization also called MinMax scaling, bounds the features values between an interval, usually [0,1] or [-1,1]. Standardization also called standard scaling, makes the data unitless by making the mean equal to 0 and the variance equal to 1. The transformed data is then split into training and test sets. The training dataset is fed to the algorithm and stored into a model based on a mapping between input and output. The performance of the algorithm is then evaluated through a test step. The kept-aside testing dataset, which has never been seen by the model in the training step, is used to assess the performance of the model based on metrics like accuracy or precision. If the model is underperforming, hyperparameters should be tuned until the accuracy is enhanced. Actually, hyperparameters are a set of parameters external to the model used in the training process and whose values cannot change during the training. Train-test split ratio, the optimization algorithm or the cost function are some examples of hyperparameters. On the other hand, parameters are internal to the model and are learned from the data during the training process. The values of the different parameters are continuously updated using an optimization algorithm during training, in contrast to hyperparameters that remain unchangeable. The weights and biased in a neural network are an example of parameters. Finally, an actual prediction task is executed by the validated model.

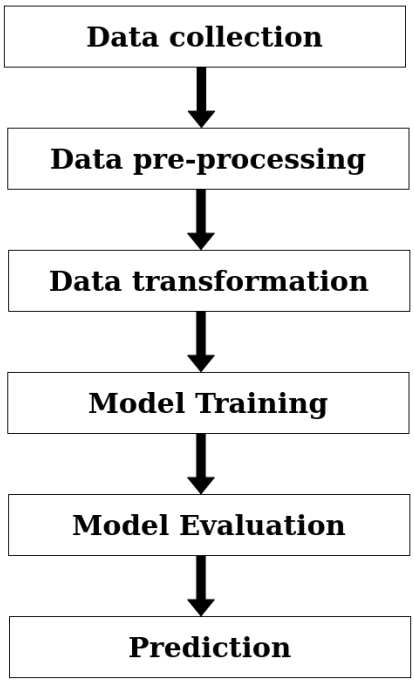


Figure 9: Machine learning workflow

Concerning root-zone soil moisture estimation, many machine learning algorithms have been shown reliable. For instance, Carranza et al. (2021) tested the random forest learning algorithm to extrapolate and interpolate RZSM on a daily timescale over a small agricultural catchment using in-situ measurements. Random forest is a classification ML algorithm which is made up of a set of decision trees that act as an ensemble. Carranza et al. (2021) demonstrated that the RF predictions have slightly higher accuracy for interpolation and similar accuracy for extrapolation in comparison with RZSM simulated from a process-based model combined with data assimilation. RFs outperform process-based models mainly in data-poor regions where soil hydraulic parameters are discontinuous or missing, since it is independent of all parameters required to estimate subsurface water flow.

Bordoni et al. (2018) implemented a SVM methodology to estimate soil moisture at different depths in a soil profile over a site in northern Italy, using only in-situ meteorological parameters. The support vector machine is a supervised learning technique that uses a set of labeled training data to produce input-output mapping functions which can be either a classification or a regression function (Wang, 2005). Two SVM models were developed in (Bordoni et al., 2018) such that the second model also considers parameters related to the antecedent meteorological conditions. The SVM model with predictors of meteorological data of a given day and of the antecedent meteorological conditions was proved to be particularly effective in estimating the time trends of soil moisture at different depths. Yu et al. (2012) used SVMs and the ensemble particle filter (EnPF) to develop a multi-layer soil moisture prediction model over a watershed in China and demonstrated that SVMs are statistically significant and resilient for both surface and root zone soil moisture prediction.

Another promising ML technique for nonlinear hydrological processes modeling, as reported by the American society of Civil Engineering (ASCE), is artificial neural networks (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000). Artificial neural

networks were developed from known models of biological nervous systems of living beings (da Silva, 2017). The processing units of an ANN are called neurons since they are simplified models of biological neurons (figure 10). They are nonlinear and perform simple functions to produce outputs based on activation functions. The implementation of an artificial neuron can be summarized in figure 10 that is based on the following equation:

$$u = \sum_{i=1}^{n} w_i \, . \, x_i - \theta$$

Such that: (x_1, x_2,.., x_n) are a set of values called features or input variables which are presented to the neuron. Each input is multiplied by a weight (w_1, w_2,.., w_n) that serves to assess the relevance of each input. A bias θ is substracted from the weighted sum of the inputs. The output (y) is the final value given by the neuron after the application of an activation function to the difference between the weighted sum of inputs and bias (u) such as:

$$y = f(u)$$

Where: y is the output of a neuron, f is the activation function.

The activation function aims at limiting the output within an acceptable range of values.

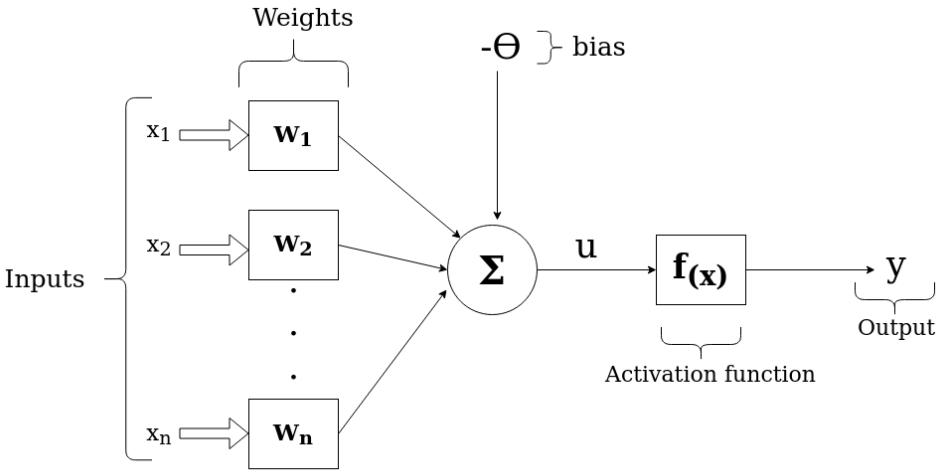


Figure 10: Structure of a neuron

An artificial neural network is made up of layers namely an input layer, one or more hidden layers and an output layer. The input layer receives the features which are usually scaled before the training step for a better precision. The hidden layers consist in hidden neurons responsible for inferring and extracting the input-output patterns associated with the considered problem. Finally, the output layer yields the final outputs. Depending on the neuron disposition and composition of the layers, the main architectures of ANN encompass for instance single-layer feedforward network, multilayer feedforward networks (MLP) and recurrent networks (da Silva, 2017) (figure 11).

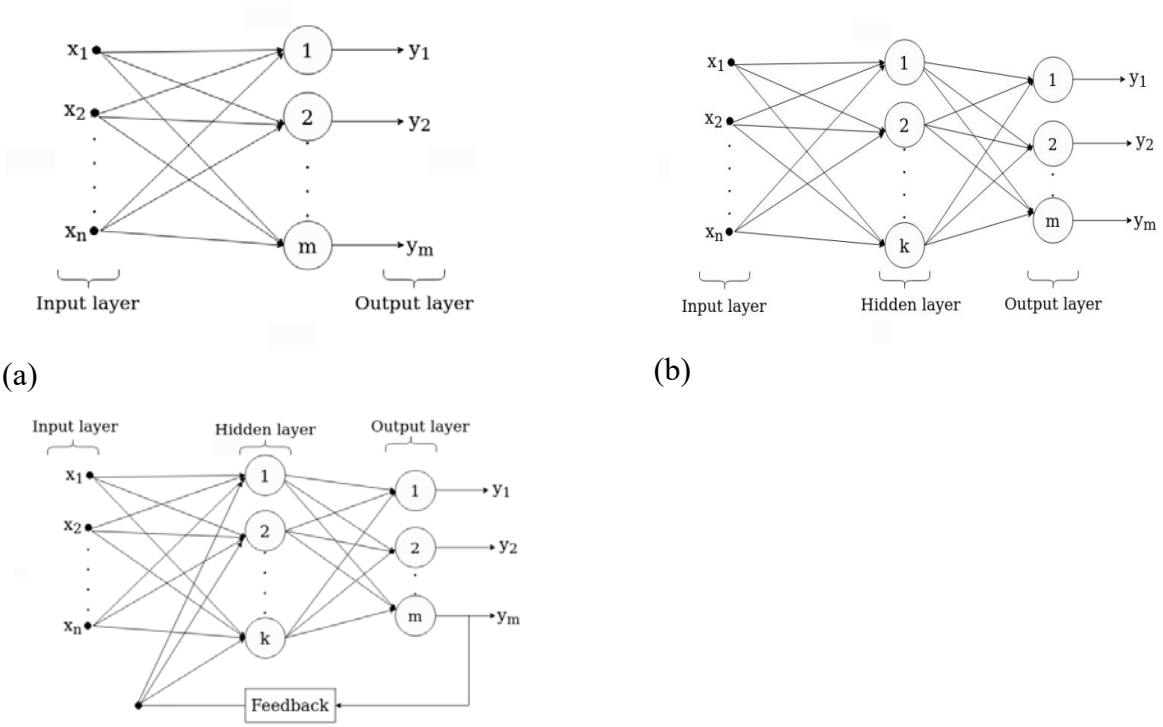


Figure 11: Examples of different architectures of artificial neural networks: (a) Single-layer feedforward network (b) Multilayer feedforward network (c) Recurrent neural network.

At a local scale, Al-Mukhtar (2016) evaluated three different types of ANN in order to model monthly RZSM in the upper reach of the Spree River catchment area (Germany) by using precipitation and antecedent soil moisture as features. Results of this study proved that the Layer recurrent network (LRN) and Feedforward (FF) networks are the most performing methods to model the nonlinear dynamic relationship such as that between precipitation and soil moisture. Moreover, results suggested that this method is a robust soil moisture predictor in this catchment. Elshorbagy and Parasuraman (2008) modeled soil moisture contents as a function of precipitation, air temperature, net radiation, and ground temperature in northern Alberta, Canada using ANN models. They showed that the ANN models outperformed a previously developed conceptual model for estimating the depth-averaged soil moisture content. However challenging due to the structure of the soil covers, predicting soil moisture using ANN is still achievable.

At a regional scale, Kornelsen and Coulibaly (2014) trained different ANN models to provide soil moisture at depths of 10, 20, and 50 cm using surface soil moisture observations and local meteorological information. They found that ANNs could well represent soil moisture as estimated by HYDRUS-1D, but performance was reduced compared to in-situ soil moisture observations outside the training conditions.

At a continental scale, Pan et al. (2017) underlined the ability of ANNs to achieve a high degree of flexibility providing good RZSM predictions over the U.S where the various climatic patterns and soil patterns caused little impact on the model performance in terms of timing and variability at a depth of 20 cm. The method was also used to generate RZSM using SMOS surface soil moisture data, and achieved a spatial soil moisture pattern comparable to that of Global Land Data

Assimilation System (GLDAS) Noah model with comparable performance to the SMOS surface soil moisture retrievals.

However, building a ML system has been requiring careful engineering in order to design a feature extractor that transforms raw data into a relevant internal representation or feature vector from which the learning subsystem often a classifier, could detect or classify patterns in the input (LeCun et al., 2015). Deep Learning (DL) is a sub-type of machine learning that requires deeper learning methods which transform the representation at one level, starting with the raw input, into a representation at a higher slightly more abstract level (LeCun et al., 2015). LeCun et al (2015) take the example of an image to explain DL. Actually, an image is represented by an array of pixel values. The first learning layer learns features that typically indicate the presence or not of edges at certain positions in the image. The second layer may spot particular arrangements of edges that make up motifs. The third layer may compile the motifs previously detected to reach recognizable objects. The core of DL algorithms is their ability to learn features from data using a general-purpose learning procedure (LeCun et al., 2015). Deep learning outperforms conventional algorithms of AI in many problems such as image and speech recognition.

Yu et al. (2021) proposed a hybrid CNN-GRU model to predict root-zone soil moisture with consideration of multi-scale spatiotemporal characteristics. Convolutional Neural Networks (CNN) were introduced by (LeCun, 1989) and designed at the beginning to address image recognition problems. As a deep learning technique, CNN has the advantage of a reduced complexity and a good computational efficiency. Soil moisture time series can for instance be transformed into meshes for the CNN to consider them as images since it is a technique relevant for image processing. Another type of neural networks is recurrent neural networks (RNN) which are suitable for work for time series data or sequence data. Unlike feed forward neural networks, RNNs are dynamic systems since they can handle temporal dependencies between data points and persist past information. This memory or self-feedback of neurons in the hidden layer(s) enables RNNs to save the information about previous inputs and use it to produce the next output. However, these networks are only efficient with short-term temporal dependencies and are highly impacted by gradient problems. The Long Short Term Memory (LSTM) is an enhanced variant of the RNN that overcomes this limitation of RNN namely the gradient vanishing/explosion when the sequence distance is long. LSTM works according to the same principle as RNN but a gate mechanism is introduced in these networks. Gated Recurrent Units (GRU) are another type of RNNs with memory cells and whose architecture are similar to LSTM but have a simpler structure and fewer internal gates, and thus simpler than LSTM (Shewalkar, 2018).The model developed in (Yu et al., 2021) was intended to predict RZSM at different depths over five sites in maize production areas in China and integrated the strong feature expression ability of CNN and the time series feature memory ability of GRU. RZSM prediction results on day 3 showed that the CNN-GRU model outperformed both CNN-based and GRU-based individual learners in terms of prediction accuracy and convergence rate. Also the predictions were improved with the increase of soil depth due to the greater soil moisture variability induced by evapotranspiration.

Yinglan et al. (2022) developed a convolutional long short-term memory (ConvLSTM) model to predict root-zone soil moisture based on remote sensing-based variables. ConvLSTM, a combination of a CNN and a LSTM, are suited for spatiotemporal sequence forecasting problems.

ConvLSTM are able to predict the future state of a given cell in the grid using the inputs and the past states of its local neighbors. ConvLSTM were first introduced by (Shi et al., 2015) as a solution for a precipitation nowcasting problem. Shi et al. (2015) extended the fully connected LSTM (FC-LSTM) to have convolutional structures in both the input-to-state and state-to-state transitions. A ConvLSTM layer is a recurrent layer, just like the LSTM, but internal matrix multiplications are exchanged with convolution operations. Data flow through the ConvLSTM cells and keep the input dimension 3D instead of being only a one-dimensional (1D) vector. Yinglan et al. (2022) used the Hydrus-1D model was used to generate large and spatiotemporal vertical soil moisture datasets for the ConvLSTM model training and validation. The fitting coefficients (R^2) recorded with the ConvLSTM model outperformed those achieved by the Global Land Data Assimilation System (GLDAS) products namely for deep layers.

In this study, we chose to focus on machine learning and more precisely on ANNs. Unlike physically-based methods, ANNs do not require an explicit definition of all of the physical and hydrological laws that govern the different variables involved in the system. They also require a one-time calibration to construct a relationship between the given inputs and the process of interest. Besides, different studies focused on mapping RZSM using ANNs but few have verified the feasibility of predicting RZSM at a large scale and to our knowledge no attempt was made to predict this variable at both a large scale and high spatial resolution. This observation advances the relevance and novelty of our work.

Chapter 2: Prediction of RZSM based on SSM using ANN

The following chapter resumes the methodology and results presented in the following paper:

Souissi, R., Al Bitar, A., and Zribi, M.: Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe, Water, 12, 3109, https://doi.org/10.3390/w12113109, 2020.

Published in Water journal.

2.1. Introduction

This chapter is our starting point towards a large-scale mapping of RZSM at high resolutions. As explained in the introduction, we start from only multi-location in-situ observations of SSM to feed an ANN model which will be in a later step hybridized through the addition physical process-related variables as will be detailed in the next chapter.

Given that RZSM is linked to SSM mainly through diffusion processes and evapotranspiration, an ANN model was developed to predict RZSM based on only SSM data in this chapter. Different aspects were explored to determine the reliability of this approach.

Prior to the evaluation of the reliability of the ANN model, different adjustments were applied to the model in order to obtain the best architecture. Different temporal sampling options were considered for the features construction. Actually, the different processes that govern the relationship between RZSM and SSM have variable time scales. Thus, different temporal steps namely hourly, daily and backward rolling averages over 10, 30 and 90 days were applied to SSM datasets. Besides, the importance of the scaling which was proved profitable before feeding data to the ANN in many studies (Priddy and Keller, 2005; Jayalakshmi and Santhakumaran, 2011), is highlighted through the different scaling techniques tested in this chapter. Also, different splitting strategies were assessed for the training, validation and test sets. This step helps assess the transferability of the method, the impact of the data density and quality.

Data from different soil moisture networks around the world which were provided by ISMN and which consist in a total of 346 soil moisture stations was used in this chapter. Soil moisture data were pre-processed by applying the quality flags provided by ISMN. Also, static variables such as land cover class, soil texture and climate class were downloaded for each station.

The disparity of soil textures and climate classes of the selected networks is a significant criterion to investigate the transferability of the approach. The contribution of each network is also explored through different splitting options of the training, validation and test sets. Transferability and contribution were quantified by the means of two indices termed $TranI_{Neti\text{-}Netj}$ and $ContI_{Neti\text{-}Netj}$, respectively.

The impact of the data quality on the predictions quality of fit was also evaluated. This analysis led to a statistical filtering approach to remove the underperforming soil moisture stations from training and validation. The genericness of the approach was also studied through an analysis of the RZSM predictions across climate classes. The interest and the limitations of the method were discussed in this study.

2.2 Conclusion

The main objective of this study was to study the feasibility of an ANN model to predict RZSM using only SSM information as input. Different regions around the world were considered and soil moisture data were collected from ISMN.

An ANN model which consists of three features namely backward rolling averages over 10, 30 and 90 days of SSM and applies standard scaling (SSCA) to scale features, was retained. Two indices namely contribution and transferability indices, were proposed in order to assess the transferability

of the approach and the contribution of each soil moisture network, respectively. Results showed that the training on stations that belong to a single and same network is not the best option. Some networks were revealed not representative of other networks such as networks 'FR-Aqui' and 'OZNET'. The French network is actually located in a forest unlike the other sites. The Australian network 'OZNET' lies within a River Catchment and covers different land covers and soil textures compared to other sites.

Besides, low quality data which are mostly observed over the stations of network 'SCAN' was proved to affect the performances of the model. The elimination of this network, although it is the densest network, improved the performances. Actually, the mean correlation and mean NSE were improved by 20.49% and 42.46% after removing 'SCAN' from the training and test sets, respectively. This result led us to apply a data filtering approach based on ECDF values. New training and test operations were conducted on the non eliminated stations after a data filtering at ECDF=0.65 which ensured good screening of underperforming stations and good sampling with respect to climate classes and soil properties. Data filtering allowed an enhancement of the performances such that a median, max, and min correlation of 0.77, 0.96, and 0.65, respectively, were recorded.

An analysis across climate classes confirmed the reliability of the method in regions of alternate wet and dry soil moisture cycles namely over stations which belong to the "Aw", the tropical savanna climate class. Also, stations which are characterized by a climate of group "C" group and which are distinguished by strong seasonal dynamics yielded good performance. This is the case of networks 'SMOSMANIA' and 'FR-Aqui' which hold agricultural areas such as the southwest plains in France where the knowledge of RZSM is of interest for sunflower and maize crops. However, it showed its limitations over regions where a surface/subsurface decoupling phenomena is observed such as regions of high evaporation rates. For instance, the stations characterized by a climate of group "B" ("BSk", "BWh", "BWk") and which cover desert areas showed the lowest performances. The decoupling phenomena is also linked to the vegetation type and root profile since the presence of a root system can redistribute the soil moisture from the lower to the upper layers. These findings motivate the next chapter where process-related information will be included in the ANN model to enhance the quality of prediction.

2.3 Article

water

MDPI

Article

Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe

Roïya Souissi *, Ahmad Al Bitar * and Mehrez Zribi

CESBIO—Centre d"Etudes Spatiales de la Biosphère, Université de Toulouse, CNES/CNRS/INRAE/IRD/UPS, 18 Avenue Edouard Belin, bpi 2801, CEDEX 09, 31401 Toulouse, France; mehrez.zribi@cesbio.cnes.fr

* Correspondence: roiya.souissi@cesbio.cnes.fr (R.S.); ahmad.albitar@cesbio.cnes.fr (A.A.B.); Tel.: +33-561-55-8537 (A.A.B.)

Received: 2 October 2020; Accepted: 3 November 2020; Published: 5 November 2020

Abstract: This paper explores the accuracy in using an artificial neural network (ANN) to estimate root-zone soil moisture (RZSM) at multiple worldwide locations using only in situ surface soil moisture (SSM) as a training dataset. The paper also addresses the transferability of the trained ANN across climatic and soil texture conditions. Data from the International Soil Moisture Network (ISMN) were collected for several networks with variable soil texture and climate classes. Several scaling, feature extraction, and training approaches were tested. An artificial neural network employing rolling averages (ANN_{RAV}) of SSM over 10, 30, and 90 days was developed. The results show that applying a standard scaling (SSCA) to the ANN input features improves the correlation, Nash–Sutcliffe efficiency (NSE), and root mean square error (RMSE) for 52%, 91%, and 87%, respectively, of the tested stations, compared to MinMax scaling (MMSCA). Different training sets are suggested, namely, training on data from all networks, data from one network, or data of all networks excluding one. Based on these trainings, new transferability (TranI) and contribution (ContI) indices are defined. The results show that one network cannot provide the best prediction accuracy if used alone to train the ANN. They also show that the removal of the less contributing networks enhances performance. For example, elimination of the densest network (SCAN) from the training enhances the mean correlation by 20.5% and the mean NSE by 42.5%. This motivates the implementation of a data filtering technique based on the ANN's performance. A median, max, and min correlation of 0.77, 0.96, and 0.65, respectively, are obtained by the model after data filtering. The performances are also analyzed with respect to the covered climatic regions and soil texture, providing insights into the robustness and limitations of the approach, namely, the need for complementary information in highly evaporative regions. In fact, the ANN using only SSM to predict RZSM has low performance when decoupling between the surface and root zones is observed. The application of ANN to obtain spatialized RZSM will require integrating remote sensing-based surface soil moisture in the future.

Keywords: soil moisture; root-zone soil moisture; artificial neural networks; ISMN

1. Introduction

Soil moisture is considered an important land parameter that stimulates interactions between the water and energy cycles, since it controls the partitioning of the mass and energy fluxes between land and the atmosphere [1]. Furthermore, soil moisture is integrated into several hydrological applications relevant to water resource decision-making [2]. Surface soil moisture (SSM) (0–5 cm) and root-zone soil moisture (RZSM) (30 cm–1 m), the two components of this variable, are both of interest. SSM is a key parameter that controls various processes in environmental systems [3] and is an important

driver of water and heat fluxes between land and the atmosphere. However, monitoring and forecast applications such as operational agriculture monitoring and crop yield forecasting [4] rely more on RZSM [5]. In addition, RZSM is of interest for short- and medium-range meteorological modelling and hydrological studies over vegetated areas [6]. Its knowledge is crucial for vegetation restoration, runoff and erosion processes [7], as well as climate change [8].

Soil moisture information can be retrieved through three main sources, namely, in situ measurements [9], model-based estimates, and satellite observations (for instance, the Soil Moisture and Ocean Salinity (SMOS) mission [10], the Soil Moisture Active Passive (SMAP) mission [11], the Advanced Microwave Scanning Radiometer (AMSR) [12], and the Advanced Scatterometer (ASCAT) [13]). In situ measurements are crucial for calibrating and validating the latter two [14,15]. Furthermore, land cover change or climate-related trends in the water cycle can be detected through long-term time series of in situ soil moisture (SM) observations [2]. The International Soil Moisture Network (ISMN) is one of the most exhaustive data hosting facilities, providing in situ soil moisture measurements collected from operational networks around the world [2]. Satellites cannot retrieve RZSM information because of the shallow penetration depth of spaceborne data, which is on the order of a few centimeters [16]. RZSM is nonlinearly related to SSM through different hydrological processes, such as diffusion processes [17]. Various computational techniques can be used to retrieve RZSM estimates based on weather forcing and surface information. Land surface models (LSMs) and agronomical models are among the most widely used methods (Surfex [18], ISBA [19], CLM [20], Aquacrop [21], SAFYE [22], etc.). However, the parameter identification and forcing data going into these models may be subject to errors and potentially lead to inaccurate tracking of the long-term evolution of soil moisture. This drawback advances the need for data assimilation techniques [23]. Nevertheless, land data assimilation systems (LDAS) may also propagate errors. Data-driven methods are suggested to overcome these drawbacks, including artificial neural networks (ANNs). ANNs have been widely used in the field of hydrology since the first hydrological implementation of ANN-based modelling by French et al. [24], and they have been used, among other applications, for soil moisture estimation [25]. ANNs are, first, not affected by the errors induced by a potential misconception of the physical relationships, as they do not require explicit configuration of these relationships [26]. In addition, ANNs require just a one-time calibration to be efficient with less heavy computational costs and provide instant estimations of soil moisture once instrument data are loaded [26]. The aforementioned advantages of ANNs, compared to other methods, explain several attempts to estimate RZSM based on surface information using ANNs. However, few studies have assessed the quality of RZSM estimations on a global scale [27].

The aim of this study is to investigate the ability of an ANN to predict RZSM based solely on in situ SSM information. This paper investigates the accuracy of the predicted RZSM over different soil moisture networks. A methodology is also suggested to determine the contribution of a given network to the global results and the transferability of the predictions across different networks. The different steps to reach this objective consist of (1) assessing the impact of the temporal parametrization of the input SSM, the scaling technique, and the impact of the training/validation/test sets; (2) evaluating the transferability and the contribution of a given network in the training process to determine the limitations; and (3) applying a data filtering technique to remove low-quality data.

2. Materials and Methods

2.1. In Situ Soil Moisture Datasets from ISMN

Several areas of the globe with different soil and climate parameters were considered. The measurements of soil moisture used in this study are provided by eight networks from ISMN. Figure 1 illustrates the location of the selected networks in the study.

Figure 1. The International Soil Moisture Network (ISMN) network distribution (adapted from the ISMN web data portal, scale 1 cm:1000 km).

The selected datasets (346 stations) are presented in Table 1 and fill the following criteria:

- Soil moisture data lie within the temporal range (January 2013–December 2019) to maximize common temporal coverage. Some stations do not have data that cover the whole temporal interval (absence of measurements, gaps generated after quality control) but are still selected as long as they fall into that period. The total number of considered records is 10,054,406 hourly values. The representativeness and size of the training dataset is an important criterion since ANNs are data-driven methods [27].
- A station is selected when soil moisture data are available at a depth of 5 cm for SSM and depths ranging between 30 and 60 cm for RZSM. Stations do not always have the same sensor installation and layout. Some stations have horizontal sensors ($depth_{from}$ = $depth_{to}$), whereas, for other stations, soil moisture sensors are disposed vertically ($depth_{from}$ <> $depth_{to}$). In the latter case, stations that fall into the interval [30, 60 cm] were chosen.
- A station is selected if it has at least 3000 hourly soil moisture values (cf. Sections 2.2.2 and 3.2).

Table 1. Overview of selected ISMN networks.

Network	Country	Number of Selected Stations	Selected RZSM Depth (cm)	SM Sensors	Length of Record (Hourly)
AMMA-CATCH	Benin, Niger	5 (3 in Benin +2 in Niger)	40	CS616	191,997
BIEBRZA-S-1	Poland	3	50	GS-3	11,401
CTP-SMTMN	China	54	40	EC-TM/5TM	716,139
HOBE	Denmark	29	55	Decagon-5TE	819,591
FR-Aqui	France	5	30, 34, 50	ThetaProbe ML2X	200,087
OZNET	Australia	19	30	Hydra Probe-CS616	519,938
SCAN	USA	209	50	Hydraprobe-Sdi-12/Ana	6,777,222
SMOSMANIA	France	22	30	ThetaProbe ML2X	818,031

Hourly values of SSM and RZSM at different depths (Table 1) along with their quality flags were extracted from the ISMN data portal. In addition, static variables such as soil texture, land cover, and climate classification were downloaded for each station.

The selected stations have different soil textures (Figure 2) and different climate classes according to the Köppen–Geiger climate classification (Figure 3). The heterogeneity of clay and sand percentages as well as climate classes will help us infer the potential impact of forgoing this information in the training process.

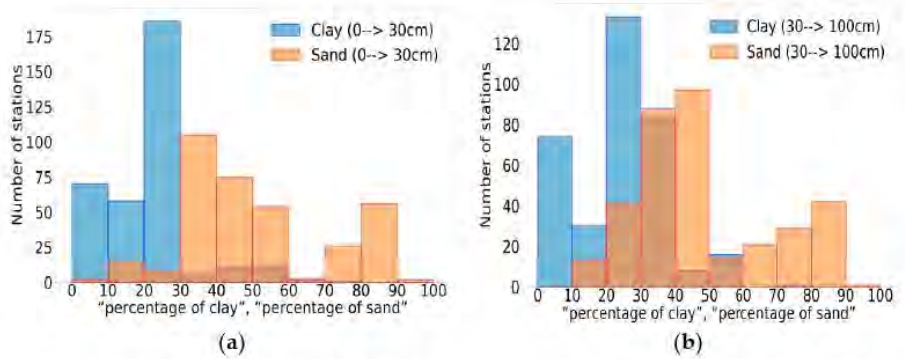


Figure 2. Clay/sand percentages for all of the stations. (**a**) Clay/sand percentages for the depth interval [0, 30 cm]. (**b**) Clay/sand percentages for the depth interval [30, 100 cm].

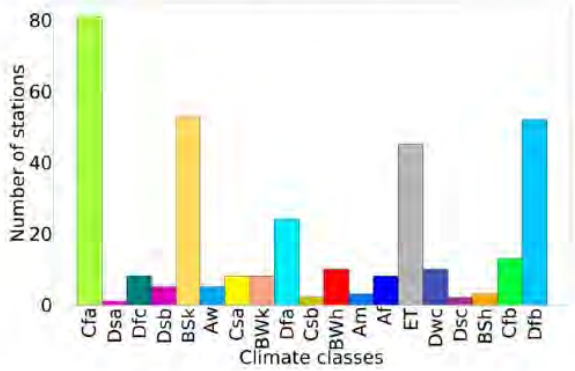


Figure 3. Climate class repartition for the SM stations (the color code is the same as that used in the updated world map of the Köppen–Geiger climate classification [28]).

2.2. *Methods*

2.2.1. Configuration of the Artificial Neural Network

ANNs can be single or multilayered. The multilayer perceptron (MLP), which is a multilayer feed-forward ANN, is one of the most commonly used ANNs and is considered as the most popular in water resources. A multilayer perceptron is a variant of the original model proposed by Rosenblatt in the 1950s and it has one or more hidden layers between its input and output layers. The neurons are organized in layers such that neurons of the same layer are not interconnected and that the connections are directed from lower to upper layers [29]. Each neuron returns an output based on a weighted sum of all inputs and according to a nonlinear function called the transfer or activation function [30]. The input layer, made up of SSM values, is connected to the hidden layer(s), which is made up of hidden neurons. The final estimates of the ANN are given by an activation function associated with the final layer called the output layer, using a sum of the weighted outputs of the hidden neurons.

Under the assumption of an ANN with one hidden layer, the whole process can be summarized by the following equation:

$$Y = f_2(\sum_{j=1}^{L} w_j f_1(\sum_{i=1}^{N} X_i w_{ij} + b_1) + b_2) \quad (1)$$

where Y is the output of the ANN and f_1 and f_2 are the activation functions of the hidden layer and the output layer, respectively. w_{ij} and w_j are the weights given to the neurons in the input layer and hidden layer, respectively. b_1 and b_2 are the biases of the input layer and hidden layer, respectively. L and N are the number of hidden neurons and inputs, respectively. Figure 4 includes a simplified diagram of a fully connected ANN with one hidden layer.

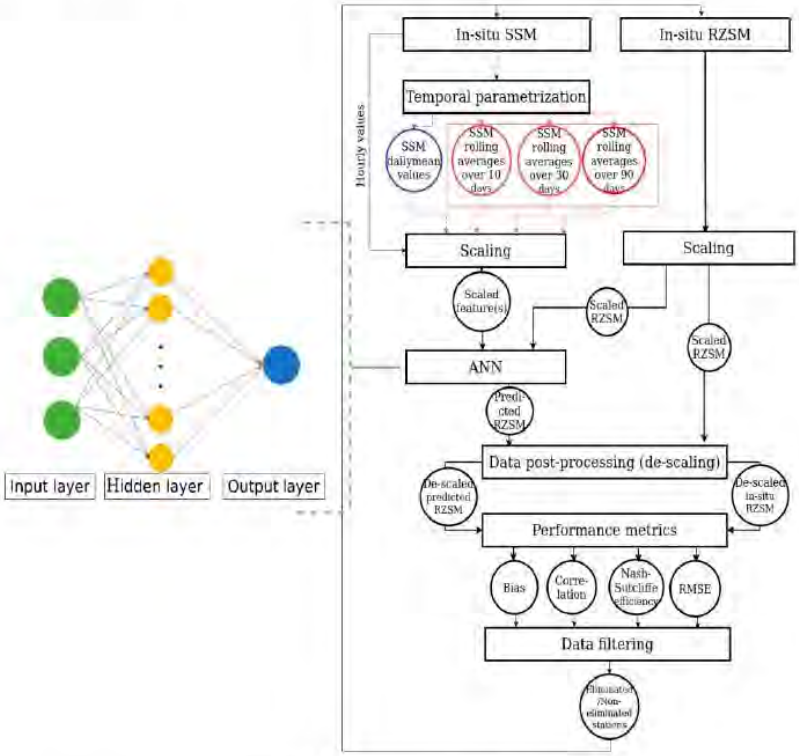


Figure 4. Data processing scheme.

Considering that problems with two hidden layers are rarely encountered and even if the corresponding ANN configuration can represent functions regardless of shape [31], we tested a one- and two-hidden layer ANN architecture. For the number of hidden neurons, a small number leads to underfitting, which may lead to inaccurate detection of complicated signals within the data [32]. In contrast, too many hidden neurons lead not only to overfitting that makes the information contained in the training set insufficient to train all of the hidden neurons but also to a longer training time [32]. Given this information and based on preliminary analysis of the output performances in terms of root mean square error (RMSE) not shown in this paper, an ANN architecture of one hidden layer with 20 hidden neurons was adopted for the remainder of the paper. A tangent sigmoid function was selected as the activation function of the hidden layer due to its anti-symmetry feature, which may accelerate the learning process [27]. A linear function was associated with the output layer. This can be justified by the experiments in [33], where they show that MLPs made up of one input layer, one hidden layer with a nonlinear transfer function, and one output layer with a linear transfer function can

approximate any function with a finite number of discontinuities. A quadratic cost function is used as a loss function, and stochastic gradient descent (SGD) is used as the optimization algorithm.

2.2.2. Features and Scaling

The input and target datasets are preprocessed such that only dates with both SSM and RZSM measurements available are kept. All other dates not filling this condition are dropped. The observed proxy variable (dielectric constant) by in situ instruments is in some cases corrected with soil temperature. Since our objective in these exercises is to test the capacity of SSM to predict RZSM, surface temperature was not considered in the feature construction. Land cover and climate conditions have a high impact on the variability of SSM, and RZSM is mainly linked to SSM through diffusion in porous media and evapotranspiration. These processes present variable specific time scales based on soil properties. Based on that, different temporal configurations were assessed for ANN input features:

- ANN_H: A one-feature ANN such as the feature is the hourly values of SSM.
- ANN_D: A one-feature ANN such as the feature is the daily mean values of SSM.
- ANN_{RAV}: A three-feature ANN such as the three features is the SSM backward rolling average values over 10, 30, and 90 days.

The target dataset (i.e., the RZSM dataset) is truncated for each station; for example, the first value fitted in the neural network is the 2160th available hourly RZSM value (applying the rolling average over 90 days on SSM requires the truncation of input and target data at the 2160th value, which corresponds to 90 days of hourly soil moisture retrievals). The target and input data are then scaled to fall into the same range of values. Non-scaling training was performed, and two scaling methods were tested:

- SSCA (Standard scaling): Standard scaling or Z-score normalization transforms the distribution of a dataset such that the mean and standard deviation of the observations are 0 and 1, respectively, using Equation (2):

$$Z_{norm} = \frac{x - \bar{x}}{\sigma_x} \tag{2}$$

where Z_{norm} is the normalized data, x is the input, $\bar{x}$ is the mean, and σ_x is the standard deviation of the input data [32].

- MMSCA (MinMax scaling): This scaling scheme constrains the range of each input feature or each output of a neural network. This is usually performed by rescaling the features or outputs from one range of values to a new range of values. Generally, the features are rescaled to lie within a range of 0 to 1 or from −1 to 1. The rescaling is often accomplished by using a linear interpolation formula such as [34]:

$$x_i' = \left(max_{target} - min_{target}\right) \times \left[\frac{x_i - min_{value}}{max_{value} - min_{value}}\right] + min_{target} \tag{3}$$

where x_i' is the scaled data, x_i is the input, max_{target} and min_{target} are the new maximum and minimum values, respectively, and max_{value} and min_{value} are the original maximum and minimum values of the input data, respectively.

The data are scaled and more precisely standardized before the training step. The result vector leaving the ANN (i.e., the vector of predicted RZSM) is in the standardized format and has to be "de-standardized". The same goes for the standardized in situ RZSM [32]. Subsequently, performance metrics are computed.

2.2.3. Training and Test Configuration

As mentioned above, one of the objectives of this paper is to assess the genericness of the model. Another point to investigate, at this level, is the training set and assess the impact of its density and its data quality, for instance. For this, different configurations were considered and termed as follows:

- ANN-TOT refers to a training/test approach where 70% of the whole global dataset (70% of the stations of all networks) forms the training set, the remaining 30% of the global dataset consists of a validation set, and the test set is made up of the whole dataset.
- ANN-Net_i refers to a training/test approach where 70% of the values belonging to the stations of a given network (Net_i) form the training set, the remaining 30% of values remaining in Net_i serve as a validation set, and the test set is made up of the whole dataset.
- ANN-(TOT-Net_i) refers to a training/test approach where 70% of the whole global dataset minus the values of a given network (Net_i) form the training set, the remaining 30% of the global dataset minus measurements of Net_i serve as a validation set, and the test set is made up of the whole dataset.

2.2.4. Performance Indicators

Individual Station Performance Metrics

The model is assessed through the following performance metrics: bias, Pearson correlation coefficient, Nash–Sutcliffe efficiency (NSE) (Equation (4)), and RMSE. The final step of the processing (Figure 4) consists of the comparison of the actual values of RZSM with the predicted values and outputting individual performance metrics of each station.

$$\mathrm{NSE} = 1 - \frac{\sum_1^N \left(\mathrm{RZSM}_{\mathrm{insitu}} - \mathrm{RZSM}_{\mathrm{predicted}}\right)^2}{\sum_1^N \left(\mathrm{RZSM}_{\mathrm{insitu}} - \overline{\mathrm{RZSM}_{\mathrm{insitu}}}\right)^2} \tag{4}$$

where N is the length of the SM dataset of the considered station.

In addition to the individual performance metrics generated for each station, performance metrics are also generated for all the stations per network.

Skill Indices

Different skill indices are computed to assess the transferability and the contribution of a given network to the training process. First, the performance differences between ANN-TOT and ANN-Net_i are assessed through a coefficient termed $TranI_{Neti}$ (Transferability Index), which is based on the correlation values yielded by each test network (Net_j) (Equation (5)).

$$\mathrm{TranI}_{\mathrm{Neti-Netj}} = 100 \times \frac{\mathrm{corr}_{\mathrm{ANN-Neti}}(\mathrm{Net_j}) - \mathrm{corr}_{\mathrm{ANN-TOT}}(\mathrm{Net_j})}{\mathrm{corr}_{\mathrm{ANN-TOT}}(\mathrm{Net_j})} \tag{5}$$

Subsequently, the contribution of a given network can be assessed when the performance results yielded by ANN-(TOT-Net_i) are compared with those yielded by ANN-TOT. Consequently, we computed the coefficient $ContI_{Neti}$ (Contribution Index) (Equation (6)).

$$\mathrm{ContI}_{\mathrm{Neti-Netj}} = 100 \times \frac{\mathrm{corr}_{\mathrm{ANN-(TOT-Neti)}}\left(\mathrm{Net_j}\right) - \mathrm{corr}_{\mathrm{ANN-TOT}}\left(\mathrm{Net_j}\right)}{\mathrm{corr}_{\mathrm{ANN-TOT}}\left(\mathrm{Net_j}\right)} \tag{6}$$

Both indices are based on correlation values. This choice can be justified by the importance of this indicator, which is often used in the assessment of level agreement between soil moisture products [35]. The correlation is sensitive to both the skill of retrievals with regard to short-term soil

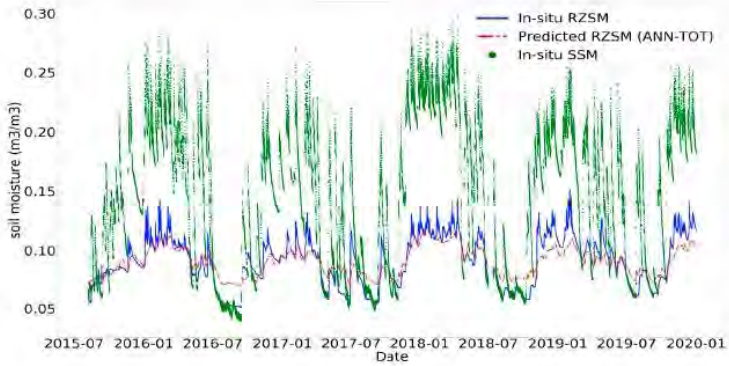
moisture anomalies and their ability to capture typical soil moisture seasonal cycling [36]. Moreover, selection of the SSCA removes the bias.

2.2.5. Data Filtering

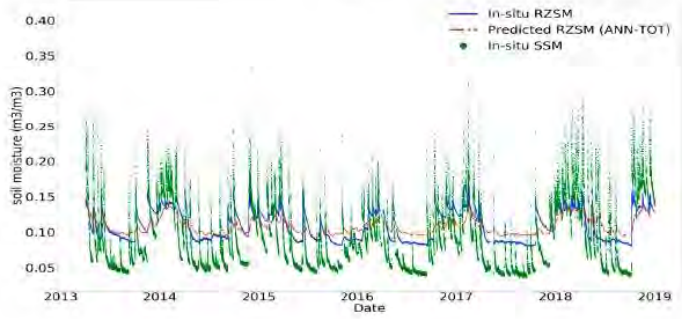
A filtering method was developed to detect underperforming stations and eliminate them from the training dataset. The filtering is based on setting q_{th} quantiles of the correlation values yielded by each test station using the ANN-TOT approach. Once the training/test process is over and the performance metrics for each station are retrieved, a loop runs through the stations one by one and selects those whose correlation is less than the q_{th} quantile of correlation. The training/test process is then reconducted such as the training set is made up of 70% of the non-eliminated stations, the validation set is made up of the remaining 30% of the non-eliminated stations, and the test set is formed by both eliminated and non-eliminated stations. This operation is repeated q times. This new training/test approach is hereafter referred to as ANN-TOT-Qual-Stat ("Qual" represents quality since this method aims to improve the quality of results).

3. Results and Discussion

Figure 5 illustrates the RZSM outputs of the ANN model through two selected examples over French networks. The time series shown below present in situ RZSM in blue, ANN-predicted RZSM (with ANN-TOT) in red, and in situ SSM in green over the stations "Hillan2" (network "FR-Aqui") and "Lezignan-Corbieres" (network "SMOSMANIA"). We can see that RZSM predictions follow up the evolution of in situ RZSM with almost a positive bias during dry events and a negative bias during wet events. Some fake peaks are sometimes generated after an abrupt increase or decrease in SSM.



(a)



(b)

Figure 5. In situ SSM, in situ RZSM, and predicted RZSM times series. (**a**) Station "Hillan2" ("FR-Aqui"). (**b**) Station "Lezignan-Corbieres" ("SMOSMANIA").

3.1. Impact of Scaling

The three scaling schemes presented in the methods section were tested using the different training approaches (cf. Section 2.2.2). Figure 6 displays the statistical distributions as histogram plots yielded by the three scaling schemes for the training approach ANN-TOT.

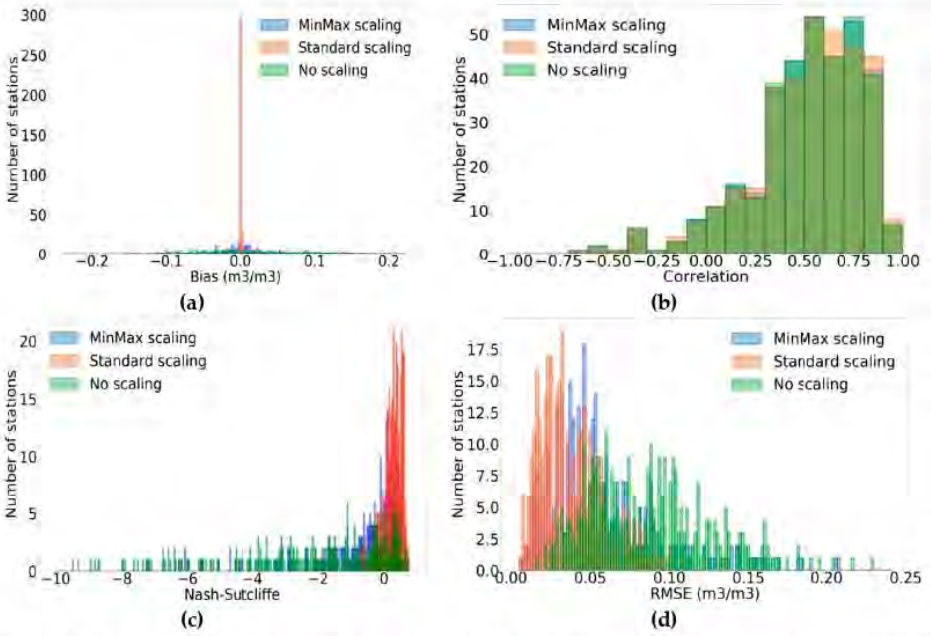


Figure 6. Performance metrics for all of the stations with the different scaling schemes; blue—MMSCA; red—SSCA; green—no scaling (training approach: ANN-TOT). (**a**) Bias. (**b**) Correlation. (**c**) NSE (Nash–Sutcliffe efficiency; NSE values less than 10 were replaced by −10 for better readability). (**d**) RMSE.

The results highlight the importance of scaling to improve the performance metrics, given the poor performance when the data are unscaled (very negative NSE reaching −285.76, high RMSE values with an average value of 0.0872 m^3/m^3). This confirms the statement that the application of preprocessing transformations to the input data is always profitable in practice before presenting data to the neural network [37] and that scaling techniques enhance the reliability of the trained network [38]. The outputs are likewise post-processed to obtain the required output values. It is, then, more relevant to only compare MMSCA with SSCA.

- Bias is considerably reduced with the application of SSCA. This is expected, as the SSCA method by construction tends to eliminate bias. These values ranged between −0.002 and 0.002 m^3/m^3 for SSCA, whereas MMSCA yielded bias values between −0.105 and 0.196 m^3/m^3.
- Correlation values are quite similar for the two scaling methods. An insignificant difference of less than 0.001 for correlation values is obtained by MMSCA and SSCA for approximately 60% of the stations (206 stations). Approximately 52% of the stations (181 stations) have higher correlation values with SSCA, approximately 6% of the stations (23 stations) have the same correlation values for both scaling methods, and the remaining stations (142 stations) have higher correlation values with MMSCA.
- RMSE values are also improved with SSCA in comparison with MMSCA mainly due to the enhancement of bias correction. Approximately 87% of the stations (302 stations) show lower RMSE values with SSCA, approximately 7% of the stations (25 stations) have invariable RMSE values, and the remaining stations (19 stations) have better RMSE values with MMSCA. The maximum

decrease (and thus, improvement) in RMSE is recorded for the "Reynolds Homestead" station ("SCAN" network) with SSCA such that the decrease is equal to 0.145 m^3/m^3. RMSE values yielded by SSCA and no scaling are consistent with previous results advanced in [27] for RZSM estimates at a depth of 50 cm in the case of the "SCAN" network. Actually, the authors in [27] used linear rescaling to compare ANN-simulated soil moisture (generated by SMOS data) to the reference datasets (GLDAS-1/Noah output). The ANN-simulated RZSM values were bias-corrected to match the mean and standard deviation of the reference set. The authors in [27] obtained a mean RMSE of 0.054 m^3/m^3 following bias correction against a mean RMSE of 0.082 m^3/m^3 without bias correction. In our case, for the network "SCAN", SSCA gives a mean RMSE equal to 0.042 m^3/m^3 against a mean RMSE of 0.090 m^3/m^3 without scaling. For SSCA, RMSE is equal to the unbiased root mean square error (ubRMSE) since bias is eliminated by construction. In fact, the relation between these two metrics is as follows:

$$RMSE^2 = ubRMSE^2 + bias^2 \quad (7)$$

- NSE values are drastically improved when the SSCA is applied. Approximately 91% of the stations (315 stations) have better NSE values. The best improvements are recorded for stations "PrairieView#1" and "GuilarteForest", which belong to the network "SCAN", such as NSE differences (SSCA-MMSCA), which are equal to 86.827 and 85.483, respectively. The difference in behavior between correlation and NSE can be explained by the fact that NSE is a function of RMSE (Equation (8)). Given that RMSE is considerably reduced for most stations with SSCA, NSE is improved.

$$NSE = 1 - \frac{RMSE^2}{\sigma_0^2} \quad (8)$$

where the symbol "o" refers to the observation, i.e., in situ RZSM.

While the results in terms of correlation are close, the enhancement in bias correction justifies the choice of SSCA as the scaling method. For this reason, it is adopted for the rest of the paper.

3.2. Impact of the Temporal Information

The three-temporal preprocessing approaches for feature extraction, presented in the methods section, yield close results with slightly better results for the backward rolling average (ANN_{RAV}) (Figure 7). The mean correlation is equal to 0.509, 0.511, and 0.561 with the hourly, daily mean, and rolling average SSM values, respectively. Similarly, the mean NSE is equal to 0.260, 0.263, and 0.325, and the mean RMSE is equal to 0.0392, 0.0391, and 0.0359 m^3/m^3 with the hourly, daily mean, and rolling average SSM values, respectively. In light of the results above, the backward rolling average approach (ANN_{RAV}) is adopted for the rest of the paper.

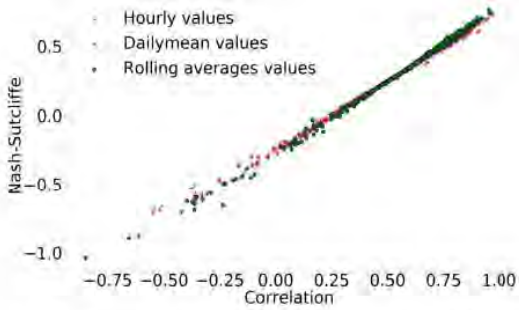


Figure 7. Correlation and NSE scatter plots (training approach: ANN-TOT); blue cross—ANN_H; red star—ANN_D; green circle—ANN_{RAV}.

3.3. Impact of the Training Approach

To assess the transferability of the trained ANN across networks, we suggested the training approach ANN-Net$_i$, which corresponds to training over one network. Table 2 presents the TranI$_{Neti}$ values as introduced in Section 2.2.4. Columns indicate the training approach, and rows specify the network on which the test was done. A positive cell value means that ANN-Net$_i$ outperforms ANN-TOT and vice versa for negative values.

Table 2. Transferability index (TranI) for the selected networks.

Training Test	ANN-AMMA-CATCH	ANN-BIEBRZA-S-1	ANN-CTP-SMTMN	ANN-FR-Aqui	ANN-HOBE	ANN-OZNET	ANN-SCAN	ANN-SMOSMANIA
AMMA-CATCH	**+1.12%**	**+0.10%**	**+0.61%**	**+0.61%**	0%	0%	−1.02%	**+0.51%**
BIEBRZA-S-1	−0.66%	**+3.53%**	−2.21%	−0.55%	−0.55%	−3.31%	−1.88%	**+0.99%**
CTP-SMTMN	−0.88%	−3.62%	**+0.77%**	−0.33%	**+0.33%**	**+0.11%**	−0.99%	−0.21%
FR-Aqui	**+0.46%**	−3.56%	−1.26%	**+2.53%**	−1.49%	−3.1%	−2.76%	−2.07%
HOBE	−2.40%	−1.49%	−1.03%	−1.83%	**+0.34%**	−0.92%	−1.26%	−0.34%
OZNET	−5.03%	−6.42%	−1.51%	−5.28%	−0.50%	**+1.26%**	−1.89%	−3.02%
SCAN	−1.5%	−1.39%	−1.07%	−1.07%	−0.43%	−0.64%	**+0.11%**	−1.28%
SMOSMANIA	**+0.57%**	−1.82%	**+0.11%**	−0.57%	**+1.82%**	−1.25%	−3.65%	**+3.53%**

The first result that can be drawn when comparing ANN-TOT with ANN-Net$_i$ is that the latter gives slightly better results when the test network is Net$_i$, i.e., the model works better for a given network when the training is solely processed on that network. The positive TranI$_{Neti}$ coefficients displayed in the diagonal element of Table 2 demonstrate that.

Further observations can be drawn from Table 2. As expected, ANN-Neti performs worse than ANN-TOT when applied to the networks on which the training has not been performed. The training approach ANN-BIEBRZA-S-1 (i.e., training processed on the BIEBRZA-S-1 network) displays the maximum performance loss compared to ANN-TOT (average loss of −1.83%). Actually, the "BIEBRZA-S-1" network has only three usable stations for our study (i.e., which satisfy the conditions established in Section 2.1), which contain little data (Table 1). The stations of this network have high organic carbon content (39.4%), as provided by ISMN based on the Harmonized World Soil Database v1.1 by IIASA, unlike the rest of the stations where organic carbon content <10%. Besides, the grassland site of "BIEBRZA-S-1" network is located on an intensively mowed, drained meadow with semi-organic soil (muck-peat soil). There, the surface soil layers featured a strong annual cycle with a maximum amplitude of around 60 vol. % [39]. These observations may explain the behavior of the BIEBRZA-S-1 network.

The "OZNET" test network delivered the worst performance compared to the other test networks when the training was run on the other networks. Figure 8 displays the correlation and NSE values for the stations of the "OZNET" network with different training approaches. The behavior of the "OZNET" test network may be explained by the climate specificities of this region of the world, which are characterized by reversed seasons compared to the Northern Hemisphere.

Moreover, some networks are not representative of other networks, i.e., ANN-Net$_i$ performs worse than ANN-TOT for the Net$_j$ test network and vice versa (ANN-Net$_j$ performs worse than ANN-TOT for test network Net$_i$). If we separately consider the "OZNET" and "FR-Aqui" test networks, we see that ANN-TOT gives better correlation values than ANN-FR-Aqui and ANN-OZNET. Actually, the FR-Aqui network is situated in southwestern France (Figure 1), and its sites cover "the Les Landes" forest of the Bordeaux-Aquitaine region with one additional site (Parcmeteo) in Bordeaux city. The soil texture in the "Les Landes" forest is mainly sandy and characterized by the presence of dark organic matter to a depth of 30 cm. The "OZNET" network lies within the Murrumbidgee River Catchment in Australia. The soil texture in the top layer is predominantly silty loam, loamy sand, and sandy loam. The study area of network "OZNET" covers farms of flood irrigation and dryland cropping (Coleambally Irrigation Area (CIA)) and pastures of grazing.

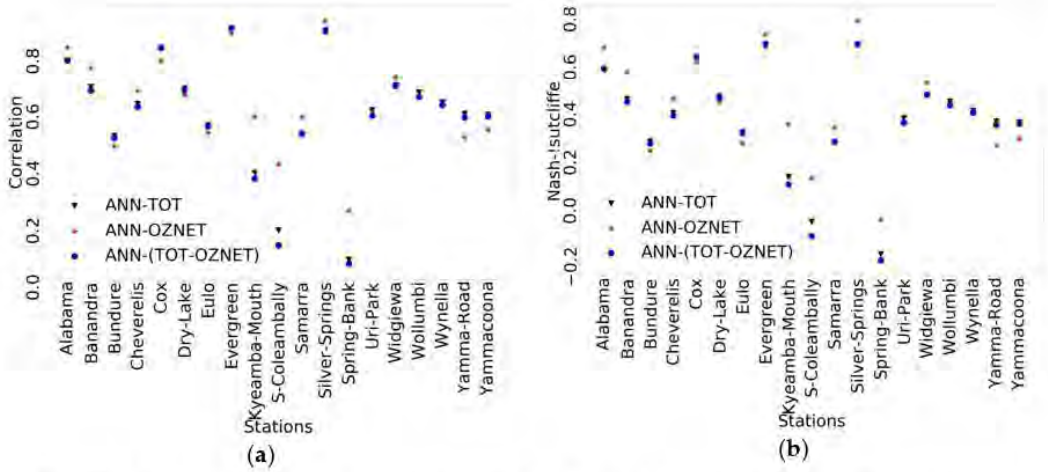


Figure 8. Performance metrics for the stations of the "OZNET" network with the different training approaches. (**a**) Correlation. (**b**) NSE.

In this paragraph, the ContI indices obtained from the ANN-TOT and ANN-(TOT-Net$_i$) setups are presented. The aim of ContI is to help assess the potential influence of a given network Net$_i$. Table 3 presents the ContI values as introduced in Section 2.2.4. Columns indicate the training approach, and rows specify the network on which the test was performed. A positive cell value indicates that ANN-(TOT-Net$_i$) outperforms ANN-TOT and vice versa for negative values. The first observation that can be drawn from Table 3 is the positive impact the extraction of the "SCAN" network from the training process would have on all of the test networks except for "OZNET" (loss of −0.13% against ANN-TOT) and "SCAN" (loss of −0.53% against ANN-TOT). This is an interesting case study since "SCAN" is the densest network (Table 1). The negative impact induced by the elimination of the "SCAN" network from the training process on the "OZNET" network can be explained by the climate classification of the stations of both networks. Actually, 7 stations of the "OZNET" network have a common climate class ("cfa") with approximately 30% of the stations of the "SCAN" network (66 stations). The remaining 12 stations of the "OZNET" network share the climate class ("Bsk") with approximately 20% of the stations of the "SCAN" network (41 stations).

Table 3. Contribution index (ContI) for the selected networks.

Training Test	ANN-(TOT-AMMA-CATCH)	ANN-(TOT-BIEBRZA-S-1)	ANN-(TOT-CTP-SMTMN)	ANN-(TOT-FR-Aqui)	ANN-(TOT-HOBE)	ANN-(TOT-OZNET)	ANN-(TOT-SCAN)	ANN-(TOT-SMOSMANIA)
AMMA-CATCH	−0.20%	−0.10%	−0.31%	−0.20%	0%	0%	**0.92%**	0%
BIEBRZA-S-1	−0.44%	−0.44%	−0.66%	−0.22%	−0.44%	−0.33%	−0.33%	−0.11%
CTP-SMTMN	0%	0%	−0.33%	**0.11%**	0%	0%	**0.66%**	**0.22%**
FR-Aqui	−0.46%	−0.35%	−0.46%	−0.58%	−0.12%	−0.12%	**1.61%**	−0.12%
HOBE	−0.11%	−0.11%	−0.23%	−0.11%	−0.23%	−0.11%	**0.34%**	**0.11%**
OZNET	0%	−0.13%	−0.38%	0%	−0.13%	−0.38%	−0.13%	**0.25%**
SCAN	0%	0%	**0.11%**	0%	0%	0%	−0.53%	0%
SMOSMANIA	−0.12%	−0.23%	−0.81%	0%	0%	**0.12%**	**2.77%**	**0.69%**

In addition, the aforementioned observation demonstrates the impact of data quality on performance. Although "SCAN" is the densest network, its elimination refines the results (positive ContI values). Figure 9 helps assess the data quality of the "SCAN" network. When considering the training approach ANN-TOT, negative values of NSE and correlation are yielded by approximately 19% (41 stations) and 7% (16 stations) of the stations belonging to "SCAN". NSE and correlation values less than 0.5 are obtained for approximately 80% (166 stations) and 40% (87 stations) of the stations in the "SCAN" network. Similarly, with the training approach ANN-SCAN, negative values of NSE and correlation are recorded for approximately 18% (38 stations) and 8% (18 stations) of "SCAN"

stations. NSE and correlation values less than 0.5 are given by 85% (179 stations) and 40% (85 stations) of its stations.

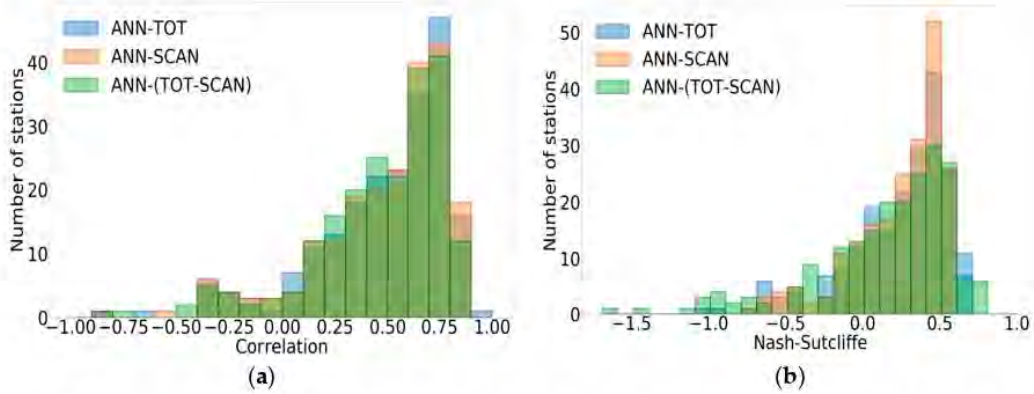


Figure 9. Performance metrics for the "SCAN" test network with ANN-TOT, ANN-SCAN, and ANN-(TOT-SCAN). (**a**) Correlation. (**b**) NSE.

Examining "SCAN" stations one by one shows that station "LyeBrook" (2042) gives the lowest NSE and correlation values: −1.037 and −0.849. A closer look into the soil moisture time series of this station (Figure 10) reveals, on the one hand, data gaps that were well identified in the ISMN quality flag and, on the other hand, constantly low SSM values over a long period of time. Many phenomena may be behind the registration of a constant value over time, such as frost periods and longer sensor dropouts [40]. These constant values lead to an inaccurate prediction of RZSM by the ANN that automatically predicts constant RZSM values in the period overlapping with constant SSM values.

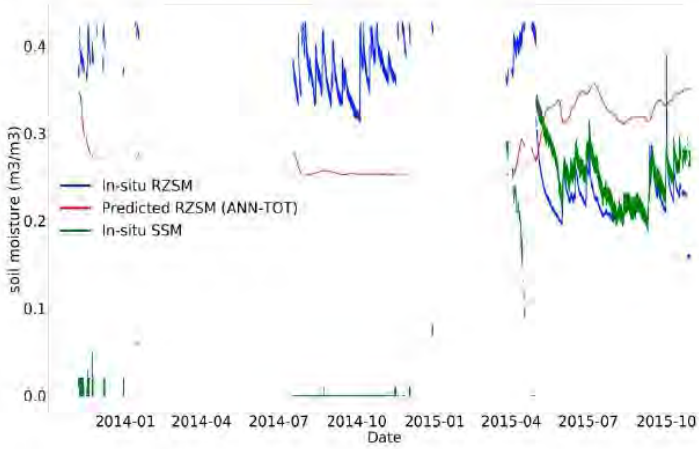


Figure 10. In situ SSM, in situ RZSM, and predicted RZSM of station "LyeBrook" ("SCAN" network).

From another perspective, the "SCAN" test network is not influenced by the elimination of the other networks from the training dataset (0% loss), except for "CTP-SMTMN" and "SCAN". This can be explained, first, by the density of the "SCAN" network, which represents 67.4% of the whole dataset. This dominant proportion makes the weight of other networks such as "BIEBRZA-S-1" (0.11% of the whole dataset) or "AMMA-CATCH" (1.9% of the whole dataset) not relevant against the density of the "SCAN" network in the training approach ANN-TOT. The elimination of the "CTP-SMTMN" network (7.12% of the whole dataset), i.e., the application of the training approach ANN-(TOT-CTP-SMTMN), leads to worse results compared with ANN-TOT. Actually, the "CTP-SMTMN" network stations are

either located in the "ET" (tundra) (83% of the stations, i.e., 45 stations) or "Dwc" (9 stations) climate classes. Both climate classes are solely covered by this network. This shows the importance of a good sampling of climate classes to perform accurate estimates.

As a conclusion from the results above, the "SCAN" network was removed from both the training and the test datasets (Figure 11). The mean correlation and mean NSE are improved by 20.49% and 42.46%, respectively. Correlation values less than 0.5 are yielded by 31.21% of the stations (108 out of a total of 346 stations) and 15.33% of the stations (21 out of a total of 137 stations) before the elimination of "SCAN" and after the elimination of "SCAN", respectively. Correlation values greater than 0.7 are recorded for 40.75% of the stations (141 out of a total of 346 stations) and 55.47% of the stations (76 out of a total of 137 stations) before the elimination of "SCAN" and after the elimination of "SCAN", respectively. Negative NSE values are obtained by 14.45% of the stations (50 out of a total of 346 stations) and 6.57% of the stations (9 out of a total of 137 stations) before the elimination of "SCAN" and after the elimination of "SCAN", respectively. NSE values greater than 0.5 are recorded for 33.53% of the stations (116 out of a total of 346 stations) and 53.28% of the stations (73 out of a total of 137 stations) before the elimination of "SCAN" and after the elimination of "SCAN", respectively.

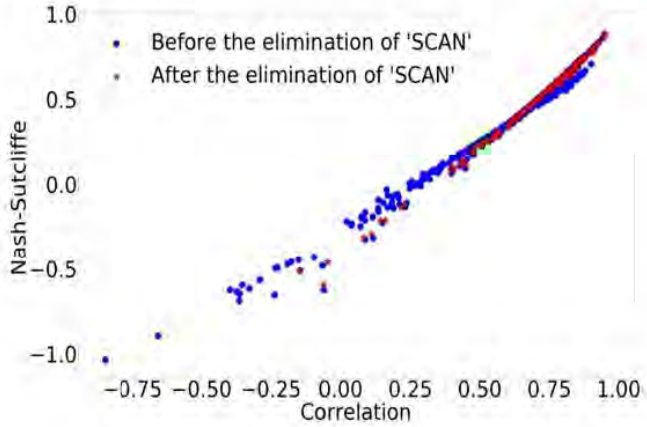


Figure 11. Correlation and NSE scatter plots. Blue circle—before the elimination of the "SCAN" network; red star—after eliminating the "SCAN" network.

3.4. Data Filtering

As previously described in Section 2.2.5, the developed filtering method is intended to identify the underperforming stations and remove them from the training process. The method is a straightforward exclusion using q_{th} quantiles of the correlation vector given by the test stations when ANN-TOT is adopted. Table 4 presents the number of eliminated stations (ES) and non-eliminated stations (NES) in accordance with the value of q.

Table 4. Number of eliminated stations (ES) and non-eliminated stations (NES) based on q_{th} quantiles.

q	Number of ES	Number of NES
0.9	308	38
0.8	275	71
0.75	254	92
0.65	224	122
0.5	170	176
0.4	141	205
0.3	105	241
0.2	71	275
0.1	38	308

After selecting the stations to remove, the training approach ANN-TOT-Qual-Stat is run as described in Section 2.2.5, and performance metrics are yielded for all of the stations based on the value of the q_{th} quantile. Figure 12 shows that the poorest performances (negative NSE and negative correlation values) are recorded for the stations that were eliminated from the training process (regardless of q). Such a result is expected. More importantly, for the non-eliminated stations, q value 0.1 yields better performance metrics than the rest of q values until the level where the correlation is equal to 0.5 and NSE is equal to 0. Beyond that level, q values yield quite similar performance metrics with a slight enhancement for q = 0.9 (a maximum correlation of 0.963 against 0.955 with q = 0.9 and q = 0.1, respectively, and a maximum NSE of 0.922 against 0.809 for q = 0.9 and q = 0.1, respectively). The maximum correlation value is recorded for the station "Nalohou-Mid" ("AMMA-CATCH" network) with both q = 0.9 and q = 0.1. The correlation value yielded for the same station before the application of this data filtering technique is equal to 0.856. Similarly, the maximum NSE value is obtained by the station "Nalohou-Mid" with both q values, whereas it is equal to 0.593 before the application of the data filtering method. This station has a tropical savanna climate and is characterized by strong seasonal dynamics that the ANN model manages to capture.

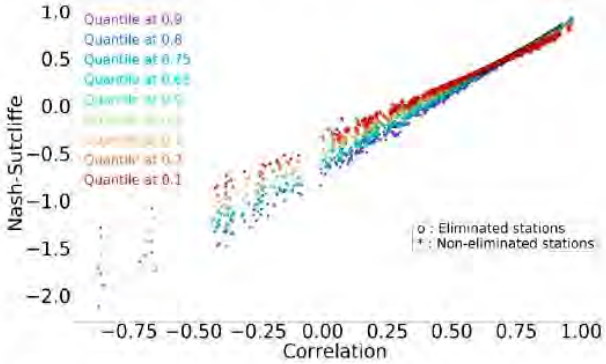


Figure 12. Correlation and NSE values after data filtering.

Table 5 presents the improvement rate in correlation, NSE, and RMSE for the eliminated and non-eliminated stations based on the q_{th} quantile. Of the non-eliminated stations, 60.98% (q = 0.4) to 73.68% (q = 0.9) show better correlation values. A total of 67.85% (q = 0.1) to 100% (q = 0.9) of the non-eliminated stations give better NSE values, and 39.94% (q = 0.1) to 100% (q = 0.9) yield better RMSE.

Table 5. Improvement rates in the individual performance metrics for the eliminated stations (ES) and non-eliminated stations (NES) based on the q_{th} quantiles.

Q	Number of ES	Number of NES	Correlation	NSE	RMSE
0.9	308	38	48.7% of ES 73.68% of NES	28.57% of ES 100% of NES	34.41% of ES 100% of NES
0.8	275	71	44.72% of ES 63.38% of NES	26.18% of ES 97.18% of NES	36.72% of ES 97.18% of NES
0.75	254	92	47,24% of ES 70.65% of NES	24.8% of ES 95.65% of NES	17.71% of ES 88.04% of NES
0.65	224	122	41.07% of ES 63.93% of NES	19.19% of ES 88.53% of NES	11.16% of ES 78.69% of NES
0.5	170	176	47.06% of ES 66.48% of NES	14.71% of ES 88.07% of NES	10.59% of ES 73.86% of NES
0.4	141	205	41.13% of ES 60.98% of NES	14.18% of ES 78.05% NES	7.09% of ES 63.41% of NES
0.3	105	241	39.05% of ES 66.39% of NES	13.33% of ES 78% of NES	7.62% of ES 60.17% of NES
0.2	71	275	25.35% of ES 60% of NES	11.26% of ES 73.45% of NES	0% of ES 50.18% of NES
0.1	38	308	23.68% of ES 63.31% of NES	13.16% of ES 67.85% of NES	0% of ES 39.94% of NES

3.5. Impact of Climate and Soil Texture

To investigate the model's genericness and transferability, the model's predictions are analyzed across climatic regions and soil texture. For this exercise, data filtering is applied with a value of q equal to 0.65, ensuring good screening of underperforming stations while providing good coverage of the climate classes and soil properties. The new training was run on 70% of the previously detected NES (122 stations), 30% of the remaining NES were used for validation, and the test was run on all of the NES. Figure 13a,b present the correlation distribution with respect to the climate classes and the percentage of subsurface clay. Clearly, data filtering with a threshold of 0.65 leads to under-sampling in some cases, namely, for climate classes "Af", "Am", "Bwk", and "Csb" and for the clay fraction interval (30%, 40%), where only one sample was available. Figure 13a provides relevant insights into the impact of climate regions on the results. It is clear that the stations belonging to "Aw", the tropical savanna climate class, yield the best correlation values. This observation can be explained by the strong seasonal dynamics and the presence of wet/dry cycles for the stations of this particular class. Group "B" ("BSk", "BWh", "BWk") includes desert areas where the link between SSM and RZSM is weaker than elsewhere because of the evaporation rates and sporadic rainfall, which reduce the link between observed SSM and RZSM and thus, the performances. This result is consistent with [35], who worked on the assessment of the level of agreement between different LSM products. In fact, they obtained low correlations in the deserts that have, by definition, low mean precipitation and a correspondingly low precipitation variance. Reference [35] confirmed that model agreement should be largest in regions with large variations in precipitation forcing because a larger precipitation variance suggests a larger variation in the moisture storage that all of the models can more easily capture. Moreover, an equilibrium and a regression approach were applied in [41] to establish a relationship between SSM and RZSM. They confirmed that errors in the RZSM estimations are encountered more for the first approach during periods of high surface evaporation or intermittent rainfall and when there is significant evapotranspiration. This shows the limitations of predicting the RZSM from SSM only, and in these specific conditions, it is of interest to include evapotranspiration-related observation variables as input features to the ANN model. Evapotranspiration was identified as a primary variable to predict RZSM in [42]. They showed that an ANN model trained with the dataset of soil moisture profiles generated by the HYDRUS-1D model using meteorological data from the lower Great Lakes region and tested on the same region was sensitive to evapotranspiration because of its role in extracting moisture from the soil. While their results may be dependent on the model's physical assumptions and uncertainty in inputs, in our results using a statistical model with no a priori assumptions, we reach the same conclusions. In Figure 13a, the "C" group, which covers areas of good quality data (mainly "SMOSMANIA" and "FR-Aqui") and is distinguished by dry/wet cycles, yields good performance. These regions are of interest because they hold agricultural areas in Europe, such as the southwest plains in France, where the knowledge of RZSM is of interest for sunflower and maize crops. Mechanistic or physical modeling of the water movement in the soil in the current state of knowledge is governed by the Richards equation. These approaches are very dependent on soil hydrodynamic parameters. Several parsimonious approaches were utilized to counter these drawbacks. Reference [6] used the recursive formulation of the exponential filter [43,44] to retrieve the root-zone soil moisture index (SWI_m) from the in situ SSM of the SMOSREX network in France and the SIM model outputs. The seasonal and interannual variability of SWI_m were also captured after the optimization of the characteristic time length of the filter (T_{opt}). Reference [6] found that over the tested sites, no link could be established between soil texture and the characteristic time length T and highlighted that there is a potential climatic effect that may exist but requires further investigation. The exponential model can be assimilated in fluid mechanics to apply mass conservation equations to an emptying bucket with a transfer function. Alternatively, the ANN does not require assumptions on the model structure (non-linearity is addressed by increasing the complexity of the Neural Network), and because of this, it can be considered more suitable to study its transferability. On the other hand, CDF matching [45] and ANN [27,42] are two statistical methods that have been used to derive RZSM

from SSM. While CDF matching determines the RZSM from SSM by correcting the SSM probability density function to match the observed RZSM, ANNs do not require a priori knowledge of probabilities. As such, they provide a more general framework and the trained ANN model can be applied outside of the training dataset. However, they have some drawbacks as they require a larger dataset than CDF matching to determine the network weighting coefficients. If not available, a risk of overfitting can exist. In the current study, this risk is not present considering the large number of available SSM and RZSM datasets. Nonetheless, as shown in this paper, the results cannot be completely generalized in areas of high evaporation, for instance. Figure 13a also presents the performance for the "Dfa" class, which covers northern areas characterized by harsh cold winters. The presence of frost events may explain a weaker link between SSM and RZSM and thus, weaker correlations. Reference [35] obtained low average correlation values between the different LSM products in high northern latitudes and explained that by the differences in the parameterization of snow and frozen soil for each product. Overall, the performances across climate conditions obtained in our paper are coherent with the results over the continental United States in [27]. In fact, the authors in [27] developed several ANN models to retrieve RZSM at depths of 20 and 50 cm using data from sites located in the continental United States. Each ANN model used a combination of soil texture, SSM, and cumulative values of air temperature, surface soil temperature, rainfall, and snowfall for the input features. Reference [27] confirmed that the retained soil moisture sites could not be considered representative of all soil and climate conditions at a global scale and showed that the ANNs were effective at retrieving RZSM at a depth of 20 cm with a correlation coefficient above 0.7 in most cases, whereas they were less effective at predicting RZSM at 50 cm. This can be explained by surface–subsurface decoupling. Reference [41,46] showed that for a given surface zone depth, the deeper the profile is, the less the correlation between surface and profile soil moisture. Reference [47,48] also confirmed that this surface–subsurface decoupling, controlled by the soil's hydraulic properties, may occur in coarse-textured and stratified soils as well as dry conditions. Reference [6] also showed that soil depth or thickness is the main factor impacting RZSM retrieval. This exposes a second limitation in addition to the impact of evapotranspiration mentioned above. Figure 13b shows that the low clay fractions present a larger dispersion of correlation in comparison with percentages greater than 30%. In our case, the result can be explained by the small number of stations having such clay percentages. In general, no direct relation between soil texture and model performances can be concluded, which is in agreement with [6].

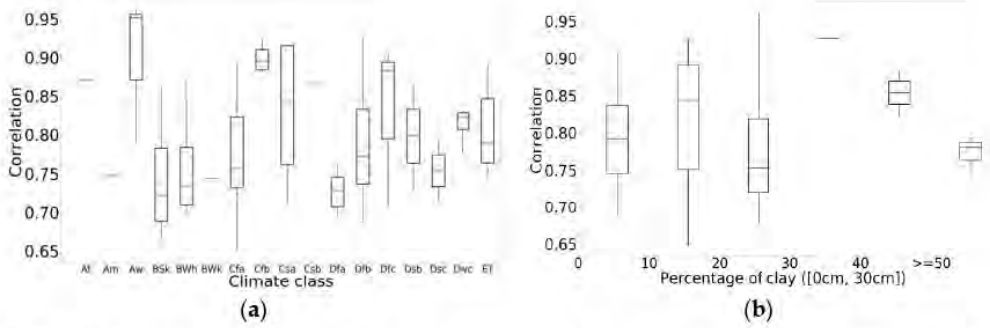


Figure 13. Correlation boxplots after the application of the data filtering technique (q = 0.65) with respect to climate classes (**a**) and subsurface soil clay percentage (**b**).

4. Conclusions

Throughout this study, we developed an ANN model to estimate RZSM based only on in situ SSM information in several regions across the globe. The main conclusion of the study is that an ANN of 1 hidden layer and 20 hidden neurons can provide accurate predictions of RZSM, provided that a specific ANN configuration is considered. For instance, testing two scaling approaches, we found that SSCA provided the best results, as it minimizes the bias by construction and improves the correlation,

Nash–Sutcliffe, and RMSE when compared with MMSCA. Moreover, a neural network of three features employing rolling averages of SSM over 10, 30, and 90 days is recommended over a single SSM estimate. We assessed the transferability of the trained ANN across observation networks and the contribution of each network to the model learning skills by developing two new indices (TranI and ContI). As expected, the results show that training with data from a single network cannot provide the best predictions. More interestingly, our experiment showed the impact of moderate to low-quality data on the performance of the model through the example of the network "SCAN", which, while being the densest (67.4% of the whole global dataset), deteriorated the performance of the model. Based on this, we applied a statistical filtering method to eliminate underperforming stations. We analyzed the model performances across climate classes and soil textures. The results showed that the model performs best in regions with alternating wet and dry conditions, while performances were lower over very dry areas with high evaporation rates and sporadic rainfall. This has been depicted by several studies, and the results suggest that to enhance the accuracy over these regions, input features related to the evapotranspiration process need to be added. We did not find relevant results about the impact of soil texture on the model performance, but this should be further investigated with spatial data. We also identified that decoupling between the surface and subsurface deteriorated the predictions of RZSM over 50 cm in depth. In the future, SSM from Earth Observation (EO) will be tested with the current approach to provide spatially distributed RZSM over different climate regions.

Author Contributions: Methodology, A.A.B., R.S. and M.Z.; software, R.S.; writing—original draft preparation, R.S.; writing—review and editing, A.A.B. and M.Z.; supervision, M.Z. and A.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: Souissi's PhD is co-funded by the ERANET RET_SIF ANR and French National Space Agency CNES Thesis programme.

Acknowledgments: This work was supported by the French National Research Agency (ANR) and the French National Space Agency (CNES). The authors thank the International Soil Moisture Network (ISMN) and the supporting networks for the availability of soil moisture data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brocca, L.; Ciabatta, L.; Massari, C.; Camici, S.; Tarpanelli, A. Soil Moisture for Hydrological Applications: Open questions and New Opportunities. *Water* **2017**, *9*, 140. [CrossRef]
2. Dorigo, W.A.; Wagner, W.; Hohensinn, R.; Hahn, S.; Paulik, C.; Xaver, A.; Gruber, A.; Drusch, M.; Mecklenburg, S.; van Oevelen, P.; et al. The International Soil Moisture Network: A data hosting facility for global in situ soil moisture measurements. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1675–1698. [CrossRef]
3. Haubrock, S.-N.; Chabrillat, S.; Lemmnitz, C.; Kaufmann, H. Surface soil moisture quantification models from reflectance data under field conditions. *Int. J. Remote Sens.* **2008**, *29*, 3–29. [CrossRef]
4. Mladenova, I.E.; Bolten, J.D.; Crow, W.; Sazib, N.; Reynolds, C. Agricultural Drought Monitoring via the Assimilation of SMAP Soil Moisture Retrievals Into a Global Soil Water Balance Model. *Front. Big Data* **2020**, *3*, 10. [CrossRef]
5. Manfreda, S.; Brocca, L.; Moramarco, T.; Melone, F.; Sheffield, J. A physically based approach for the estimation of root-zone soil moisture from surface measurements. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 1199–1212. [CrossRef]
6. Albergel, C.; Rüdiger, C.; Pellarin, T.; Calvet, J.-C.; Fritz, N.; Froissard, F.; Suquia, D.; Petitpa, A.; Piguet, B.; Martin, E. From near-surface to root-zone soil moisture using an exponential filter: An assessment of the method based on in-situ observations and model simulations. *Hydrol. Earth Syst. Sci.* **2008**, *12*, 1323–1337. [CrossRef]
7. Gao, X.; Wu, P.; Zhao, X.; Zhang, B.; Wang, J.; Shi, Y. Estimating the spatial means and variability of root-zone soil moisture in gullies using measurements from nearby uplands. *J. Hydrol.* **2013**, *476*, 28–41. [CrossRef]
8. Han, E.; Merwade, V.; Heathman, G.C. Application of data assimilation with the Root Zone Water Quality Model for soil moisture profile estimation in the upper Cedar Creek, Indiana. *Hydrol. Process.* **2012**, *26*, 1707–1719. [CrossRef]

9. Lekshmi, S.U.S.; Singh, D.N.; Baghini, M.S. A critical review of soil moisture measurement. *Measurement* **2014**, *54*, 92–105.
10. Kerr, Y.H.; Waldteufel, P.; Wigneron, J.-P.; Delwart, S.; Cabot, F.; Boutin, J.; Escorihuela, M.-J.; Font, J.; Reul, N.; Gruhier, C.; et al. The SMOS Mission: New Tool for Monitoring Key Elements of the Global Water Cycle. *Proc. IEEE* **2010**, *98*, 666–687. [CrossRef]
11. Entekhabi, D.; Njoku, E.G.; O'Neill, P.E.; Kellogg, K.H.; Crow, W.T.; Edelstein, W.N.; Entin, J.K.; Goodman, S.D.; Jackson, T.J.; Johnson, J.; et al. The Soil Moisture Active Passive (SMAP) Mission. *Proc. IEEE* **2010**, *98*, 704–716. [CrossRef]
12. Owe, M.; de Jeu, R.; Holmes, T. Multisensor historical climatology of satellite-derived global land surface moisture. *J. Geophys. Res. Earth Surf.* **2008**, 113. [CrossRef]
13. Wagner, W.; Hahn, S.; Kidd, R.; Melzer, T.; Bartalis, Z.; Hasenauer, S.; Figa-Saldaña, J.; de Rosnay, P.; Jann, A.; Schneider, S.; et al. The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications. *Metz* **2013**, *22*, 5–33. [CrossRef]
14. Ceballos, A.; Scipal, K.; Wagner, W.; Martínez-Fernández, J. Validation of ERS scatterometer-derived soil moisture data in the central part of the Duero Basin, Spain. *Hydrol. Process.* **2005**, *19*, 1549–1566. [CrossRef]
15. Wagner, W.; Naeimi, V.; Scipal, K.; de Jeu, R.; Martínez-Fernández, J. Soil moisture from operational meteorological satellites. *Hydrogeol. J.* **2007**, *15*, 121–131. [CrossRef]
16. Wagner, W.; Blöschl, G.; Pampaloni, P.; Calvet, J.-C.; Bizzarri, B.; Wigneron, J.-P.; Kerr, Y. Operational readiness of microwave remote sensing of soil moisture for hydrologic applications. *Hydrol. Res.* **2007**, *38*, 1–20. [CrossRef]
17. Sabater, J.M.; Jarlan, L.; Calvet, J.-C.; Bouyssel, F.; De Rosnay, P. From Near-Surface to Root-Zone Soil Moisture Using Different Assimilation Techniques. *J. Hydrometeor.* **2007**, *8*, 194–206. [CrossRef]
18. Masson, V.; Le Moigne, P.; Martin, E.; Faroux, S.; Alias, A.; Alkama, R.; Belamari, S.; Barbu, A.; Boone, A.; Bouyssel, F.; et al. The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. *Geosci. Model Dev.* **2013**, *6*, 929–960. [CrossRef]
19. Noilhan, J.; Mahfouf, J.-F. The ISBA land surface parameterisation scheme. *Glob. Planet. Chang.* **1996**, *13*, 145–159. [CrossRef]
20. Oleson, W.; Lawrence, M.; Bonan, B.; Flanner, G.; Kluzek, E.; Lawrence, J.; Levis, S.; Swenson, C.; Thornton, E.; Dai, A.; et al. *Technical Description of version 4.0 of the Community Land Model (CLM)*; NCAR: Boulder, CO, USA, 2010. [CrossRef]
21. Raes, D.; Steduto, P.; Hsiao, T.C.; Fereres, E. AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description. *Agron. J.* **2009**, *101*, 438–447. [CrossRef]
22. Battude, M.; Al Bitar, A.; Brut, A.; Tallec, T.; Huc, M.; Cros, J.; Weber, J.-J.; Lhuissier, L.; Simonneaux, V.; Demarez, V. Modeling water needs and total irrigation depths of maize crop in the south west of France using high spatial and temporal resolution satellite imagery. *Agric. Water Manag.* **2017**, *189*, 123–136. [CrossRef]
23. Pleim, J.E.; Xiu, A. Development of a Land Surface Model. Part II: Data Assimilation. *J. Appl. Meteor.* **2003**, *42*, 1811–1822. [CrossRef]
24. Tanty, R.; Desmukh, T.S. MANIT BHOPAL Application of Artificial Neural Network in Hydrology—A Review. *IJERT* **2015**, *V4*, IJERTV4IS060247. [CrossRef]
25. Elshorbagy, A.; Parasuraman, K. On the relevance of using artificial neural networks for estimating soil moisture content. *J. Hydrol.* **2008**, *362*, 1–18. [CrossRef]
26. Kolassa, J.; Reichle, R.H.; Liu, Q.; Alemohammad, S.H.; Gentine, P.; Aida, K.; Asanuma, J.; Bircher, S.; Caldwell, T.; Colliander, A.; et al. Estimating surface soil moisture from SMAP observations using a Neural Network technique. *Remote Sens. Environ.* **2018**, *204*, 43–59. [CrossRef] [PubMed]
27. Pan, X.; Kornelsen, K.C.; Coulibaly, P. Estimating Root Zone Soil Moisture at Continental Scale Using Neural Networks. *J. Am. Water Resour. Assoc.* **2017**, *53*, 220–237. [CrossRef]
28. Peel, M.C.; Finlayson, B.L.; McMahon, T.A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1633–1644. [CrossRef]
29. Ramchoun, H.; Amine, M.; Idrissi, J.; Ghanou, Y.; Ettaouil, M. Multilayer Perceptron: Architecture Optimization and Training. *IJIMAI* **2016**, *4*, 26. [CrossRef]
30. Oyebode, O.; Stretch, D. Neural network modeling of hydrological systems: A review of implementation techniques. *Nat. Resour. Modeling* **2019**, *32*, e12189. [CrossRef]

31. Heaton, J. *Introduction to Neural Networks with Java*; Heaton Research, Inc.: St. Louis, MO, USA, 2008; ISBN 9781604390087.
32. Chai, S.-S.; Walker, J.; Makarynskyy, O.; Kuhn, M.; Veenendaal, B.; West, G. Use of Soil Moisture Variability in Artificial Neural Network Retrieval of Soil Moisture. *Remote Sens.* **2009**, *2*, 166–190. [CrossRef]
33. Yonaba, H.; Anctil, F.; Fortin, V. Comparing Sigmoid Transfer Functions for Neural Network Multistep Ahead Streamflow Forecasting. *J. Hydrol. Eng.* **2010**, *15*, 275–283. [CrossRef]
34. Bishop, C.M.; Bishop, P. *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, UK, 1995; ISBN 9780198538646.
35. Koster, R.D.; Guo, Z.; Yang, R.; Dirmeyer, P.A.; Mitchell, K.; Puma, M.J. On the Nature of Soil Moisture in Land Surface Models. *J. Clim.* **2009**, *22*, 4322–4335. [CrossRef]
36. Crow, W.T.; Miralles, D.G.; Cosh, M.H. A Quasi-Global Evaluation System for Satellite-Based Surface Soil Moisture Retrievals. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 2516–2527. [CrossRef]
37. Priddy, K.L.; Keller, P.E. *Artificial Neural Networks: An Introduction*; SPIE Press: Bellingham, WA, USA, 2005; ISBN 9780819459879.
38. Jayalakshmi, T.; Santhakumaran, A. Statistical Normalization and Back Propagation for Classification. *IJCTE* **2011**, 89–93. [CrossRef]
39. Dabrowska-Zielinska, K.; Musial, J.; Malinska, A.; Budzynska, M.; Gurdak, R.; Kiryla, W.; Bartold, M.; Grzybowski, P. Soil Moisture in the Biebrza Wetlands Retrieved from Sentinel-1 Imagery. *Remote Sens.* **2018**, *10*, 1979. [CrossRef]
40. Dorigo, W.A.; Xaver, A.; Vreugdenhil, M.; Gruber, A.; Hegyiová, A.; Sanchis-Dufau, A.D.; Zamojski, D.; Cordes, C.; Wagner, W.; Drusch, M. Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network. *Vadose Zone J.* **2013**, *12*, vzj2012.0097. [CrossRef]
41. Jackson, T.J.; Hawley, M.E.; O'Neill, P.E. Preplanting Soil Moisture Using Passive Microwave Sensors1. *Jawra J. Am. Water Resour. Assoc.* **1987**, *23*, 11–19. [CrossRef]
42. Kornelsen, K.C.; Coulibaly, P. Root-zone soil moisture estimation using data-driven methods. *Water Resour. Res.* **2014**, *50*, 2946–2962. [CrossRef]
43. Wagner, W.; Lemoine, G.; Rott, H. A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data. *Remote Sens. Environ.* **1999**, *70*, 191–207. [CrossRef]
44. Stroud, P.D. *A Recursive Exponential Filter For Time-Sensitive Data*; Los Alamos National Laboratory: Los Alamos, Mexico, 1999.
45. Gao, X.; Zhao, X.; Brocca, L.; Pan, D.; Wu, P. Testing of observation operators designed to estimate profile soil moisture from surface measurements. *Hydrol. Process.* **2019**, *33*, 575–584. [CrossRef]
46. Arya, L.M.; Richter, J.C.; Paris, J.F. Estimating profile water storage from surface zone soil moisture measurements under bare field conditions. *Water Resour. Res.* **1983**, *19*, 403–412. [CrossRef]
47. Walker, J.P.; Willgoose, G.R.; Kalma, J.D. Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: Simplified Kalman filter covariance forecasting and field application. *Water Resour. Res.* **2002**, *38*, 37-1–37-13. [CrossRef]
48. Hirschi, M.; Mueller, B.; Dorigo, W.; Seneviratne, S.I. Using remotely sensed soil moisture for land–atmosphere coupling diagnostics: The role of surface vs. root-zone soil moisture variability. *Remote Sens. Environ.* **2014**, *154*, 246–252. [CrossRef]

Chapter 3: Prediction of RZSM based on SSM and process-related inputs using ANN

The following chapter resumes the methodology and results presented in the following paper:

Souissi, R., Zribi, M., Corbari, C., Mancini, M., Muddu, S., Tomer, S. K., Upadhyaya, D. B., and Al Bitar, A.: Integrating process-related information into an artificial neural network for root-zone soil moisture prediction, Hydrol. Earth Syst. Sci., 26, 3263–3297, https://doi.org/10.5194/hess-26-3263-2022, 2022.

Published in Hydrology and Earth System Sciences(HESS) journal.

3.1 Introduction

RZSM is linked to SSM through a nonlinear relationship controlled by different hydrological processes like infiltration and evapotranspiration. According to the previous chapter, the approach based on the prediction of RZSM based on only SSM information has limitations. In order to complexify the method and eventually enhance the quality of predictions, we considered the option of developing another type of ANN namely a Convolutional Neural Network (CNN). The developed CNN model takes 60-day sequences of SSM as input and predicts RZSM. Different hyperparameters and parameters were tested and the compared. However, the CNN model was not conclusive when compared against the performances yielded by the MLP model. This complexification option was discarded. Instead, the approach was hybridized by adding physical process-related features into the ANN.

In this chapter, different ANN models were developed such that the ANN features include SSM and one process-related variable. An ANN model which includes SSM and a combination of process-related features was also developed.

The infiltration process was considered in this chapter by including a SWI that was computed using a recursive exponential filter. Soil evaporation was also taken into account in order to better estimate RZSM namely in areas of high evaporation rates. This process was modeled by the means of an evaporation efficiency variable whose computation was based on a remote-sensing potential evapotranspiration. The impact of the addition of LST in the model which is linked to evaporation, was also explored. Vegetation dynamics were also considered. They were not modeled in this study but only inferred from remote-sensing NDVI.

All of the ANN models were trained on the stations provided by ISMN and which were identified of good quality data after the data filtering step in the previous chapter. In a first time, the different ANN models were tested on the rest of the ISMN stations that were considered in the last chapter. Additional tests were conducted on stations external to the ISMN database namely on stations over Tunisia, Italy and India. This step helped assess the robustness of the ANN models.

The individual impact of each process-related feature on the RZSM prediction accuracy was assessed through a climatic analysis. The impact of the joint use of the process-related features in the most complex ANN model was also studied.

3.2 Conclusion

This chapter investigated the impact of the addition of process-related variables in ANN models on the quality of RZSM predictions. Different ANN models which are made up of different features were developed and intercompared. A global analysis was conducted for the climate regions covered by the considered stations around the globe. A global map with the most relevant process-related variables for each climate class was produced. The objective of this step was to identify the most relevant variable that has to be added with SSM in the ANN model in order to obtain the best estimations of RZSM.

Overall, the process-related variable when included in the ANN model helped improve the accuracy of the predictions but over some areas their use was not beneficial. Results suggested that the most

relevant feature for arid areas with high evaporation rates such as bare areas of Africa, the Middle East and Australia was evaporation efficiency. NDVI was proved to be the most informative variable over agricultural regions namely over the internal part of continental Europe and near the Mediterranean basin and in the Great Plains region in the USA and over transition zones. The classification was not as reliable over all climate regions due to the generalization of the climatic analysis results to areas not considered in this study. Over regions with climate class 'Dfc' (cold dry without a dry season, cold summer climate), the evaporation efficiency was found the most relevant instead of temperature. Overall, the ANN model which included SSM, NDVI, SWI and evaporation efficiency was the most performing. Actually, the correlation value obtained with this model increased for 84.06 %, 61.29 % and 62.07 % of the training, validation and test stations when compared to the reference model of the first axis (ANN_SSM), respectively. In addition, RMSE was minimized for 62.32 %, 54.84 % and 54.02 % of the training, validation and test stations with this model compared to reference model (ANN_SSM), respectively.

The study was also focused on the evaluation of the robustness of the approach through additional tests over external sites in central Tunisia, India and Italy. The same finding regarding the positive impact of the process-related features was observed over Tunisia. The mean correlation across the Tunisian stations significantly increased from 0.44 when only SSM was considered to 0.8 when all process-related features were combined with SSM. In fact, the Tunisian site is characterized by a semiarid environment with sporadic rainfall events and high evaporation rates. This finding corroborated the reliability of our hybrid approach based on an association of a data-driven method with process-related variables. However, the change in correlation after the addition of process-related features in India and Italy namely NDVI, was nonsignificant and could be linked to the cloudy conditions and thus the noisy MODIS products.

3.3 Article

Hydrol. Earth Syst. Sci., 26, 3263–3297, 2022
https://doi.org/10.5194/hess-26-3263-2022

Hydrology and Earth System Sciences

Integrating process-related information into an artificial neural network for root-zone soil moisture prediction

Roiya Souissi[1], Mehrez Zribi[1], Chiara Corbari[2], Marco Mancini[2], Sekhar Muddu[3], Sat Kumar Tomer[4], Deepti B. Upadhyaya[3,4], and Ahmad Al Bitar[1]

[1]CESBIO – Centre d'Etudes Spatiales de la Biosphère, Université de Toulouse, CNES/CNRS/INRAE/IRD/UPS, Toulouse, France
[2]Department of Civil and Environmental Engineering (DICA), Polytechnic University of Milan, 20133 Milan, Italy
[3]Department of Civil Engineering, Indian Institute of Science, Bangalore 560012, India
[4]Satyukt analytics Pvt Ltd, Sanjay Nagar Main Rd, MET Layout, Bengaluru, Karnataka 560094, India

Correspondence: Roiya Souissi (roiya.souissi@cesbio.cnes.fr)

Received: 11 February 2022 – Discussion started: 17 February 2022
Revised: 30 May 2022 – Accepted: 10 June 2022 – Published: 28 June 2022

Abstract. Quantification of root-zone soil moisture (RZSM) is crucial for agricultural applications and the soil sciences. RZSM impacts processes such as vegetation transpiration and water percolation. Surface soil moisture (SSM) can be assessed through active and passive microwave remote-sensing methods, but no current sensor enables direct RZSM retrieval. Spatial maps of RZSM can be retrieved via proxy observations (vegetation stress, water storage change and surface soil moisture) or via land surface model predictions. In this study, we investigated the combination of surface soil moisture information with process-related inferred features involving artificial neural networks (ANNs). We considered the infiltration process through the soil water index (SWI) computed with a recursive exponential filter and the evaporation process through the evaporation efficiency computed based on a Moderate Resolution Imaging Spectroradiometer (MODIS) remote-sensing dataset and a simplified analytical model, while vegetation growth was not modeled and was only inferred through normalized difference vegetation index (NDVI) time series. Several ANN models with different sets of features were developed. Training was conducted considering in situ stations distributed in several areas worldwide characterized by different soil and climate patterns of the International Soil Moisture Network (ISMN), and testing was applied to stations of the same data-hosting facility. The results indicate that the integration of process-related features into ANN models increased the overall performance over the reference model level in which only SSM features were considered. In arid and semiarid areas, for instance, performance enhancement was observed when the evaporation efficiency was integrated into the ANN models. To assess the robustness of the approach, the trained models were applied to observation sites in Tunisia, Italy and southern India that are not part of the ISMN. The results reveal that joint use of surface soil moisture, evaporation efficiency, NDVI and recursive exponential filter represented the best alternative for more accurate predictions in the case of Tunisia, where the mean correlation of the predicted RZSM based on SSM only sharply increased from 0.443 to 0.801 when process-related features were integrated into the ANN models in addition to SSM. However, process-related features have no to little added value in temperate to tropical conditions.

1 Introduction

Soil moisture is a major land parameter integrated into several agricultural, hydrological and meteorological applications (Koster et al., 2004; Paris Anguela et al., 2008). This essential climate variable (ECV) consists of two components, namely, surface soil moisture (SSM) (0–5 cm) and root-zone soil moisture (RZSM). RZSM corresponds to the soil moisture in the region in which the main vegetation rooting network is developing. Its definition varies depending on vegetation type and pedoclimatic conditions. The importance of RZSM is mainly highlighted in agricultural applications

through vegetation stress and water needs and in carbon and nitrogen cycles, as RZSM influences biogeochemical activities in soil (Martínez-Espinosa et al., 2021). RZSM is nonlinearly related to SSM through different hydrological processes, such as diffusion processes. RZSM may be extracted by evaporation at the surface through root extraction or by capillary rises (Calvet et Noilhan, 2000). SSM quantification is achieved through three main sources: in situ measurements, model estimates and remote-sensing-based products. Microwave remote-sensing technologies involving sensors such as Soil Moisture and Ocean Salinity (SMOS) (Kerr et al., 2010), Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010), Advanced Microwave Scanning Radiometer (AMSR) (Owe et al., 2008) and Advanced Scatterometer (ASCAT) (Wagner et al., 2013) have been employed to retrieve SSM at coarse resolutions. Current satellite sensors can only provide surface soil moisture information because of the shallow penetration depth of spaceborne data (on the order of a few centimeters) (Wagner et al., 2007). Fine-spatial-resolution synthetic aperture radar (SAR) data can also be applied in synergy with optical data to retrieve soil moisture (Zribi et al., 2011; Hajj et al., 2014; Dorigo et al., 2011), but again for surface soil moisture. The International Soil Moisture Network (ISMN) is an exhaustive data-hosting facility focused on soil moisture data and associated ancillary information. The ISMN provides in situ soil moisture measurements collected from operational soil moisture networks worldwide (Dorigo et al., 2011). Various models can be adopted to estimate RZSM, such as land surface models (Surfex) (Masson et al., 2013), Interaction Sol-Biosphère-Atmosphère (ISBA) (Noilhan and Mahfouf, 1996), the Community Land Model (CLM; Oleson et al., 2010) or the Joint UK Land Environment Simulator (JULES) (Best et al., 2011) or dedicated crop models such as Aquacrop (Raes et al., 2009) or Simple Algorithm For Yield Estimate (SAFYE) (Battude et al., 2017). While these models provide the advantage of physical process-based estimates, these estimates depend on the availability and accuracy of ancillary information. Model predictions are often enhanced by the implementation of data assimilation techniques, such as the land data assimilation system (LDAS) (Sabater et al., 2007; Entekhabi et al., 2020).

Data-driven methods such as artificial neural networks (ANNs) have also been commonly applied in hydrology as detailed, for instance, by the ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2020) and in Tanty et al. (2015). One of their advantages is that these models do not require an explicit model structure to accurately represent the involved hydrological processes but instead construct a relationship between the given inputs and the process of interest. Therefore, ANNs are regarded as dynamic input–output mapping models heavily relying on the provided training data relevant to target values (Pan et al., 2017). Moreover, ANNs only require a one-time calibration to provide soil moisture estimations once instrument data are loaded and thus generate relatively low computational costs (Kolassa et al., 2018). These advantages explain the approach to estimate RZSM based on surface information with ANNs in various methodologies (Pan et al., 2017; Grillakis et al., 2021; Souissi et al., 2020). In this paper, we do not address ANN applications as a model twin where the ANN model is trained on the target for mimicking purposes and subsequently generates predictions while requiring a short computation time or fewer input simplifications. Here, we are instead interested in the adoption of ANNs as independent models trained on in situ observations. Within this context, Pan et al. (2017) successfully applied an ANN as a model for shallow 20 cm root-zone soil moisture prediction with a global correlation coefficient of 0.7. Grillakis et al. (2021) proposed employing an ANN as a means of calibrating and regionalizing the time constant of a recursive exponential filter, which was thereafter applied at the regional scale. A combined implementation of a Bayesian probabilistic approach and an ANN to infer RZSM at different depths from optical unmanned aerial vehicle (UAV) acquisitions via local training was also applied (Hassan-Esfahani et al., 2017). Multitemporal averaged features to predict RZSM based on only SSM and to investigate the transferability of a trained ANN across different climatic conditions globally were proposed in Souissi et al. (2020). Temporal information can be considered in ANNs through recurrent neural networks (RNNs), long short-term memory (LSTM) architectures (Liu et al., 2021), 1D convolutional neural networks (CNNs), or multitemporal averaging. In Souissi et al. (2020), median, maximum and minimum correlation values of 0.77, 0.96 and 0.65 were, respectively, reported across training, validation and test datasets. The use of climatic variables such as precipitation and surface temperature and intrinsic surface properties such as soil texture and land cover has also been considered in ANNs (Liu et al., 2021). The choice of variables depends not only on the data availability, but also on the objectives. Finally, ANN-based approaches pertain to the more general term of machine learning approaches, and within this framework, the random forest approach has been applied to root-zone soil moisture prediction (Carranza et al., 2021). The aforementioned studies have investigated the application of multiple information sources to predict root-zone soil moisture. The input features are commonly curated for quality, and correlation analysis is conducted to determine the useful inputs, while physical processes are not considered. In this paper, we introduce process-related features based on simplified analytical models representing the major processes contributing to root-zone soil moisture dynamics. In this work, RZSM refers to a point observation of water content at a depth ranging between 30 and 55 cm. We investigate the impact of the application of different process-related variables on the precision of RZSM predictions as well as the robustness of our approach. (1) We start from a previously developed ANN model (Souissi et al., 2020), and we extend the feature list to include NDVI time series, surface soil tem-

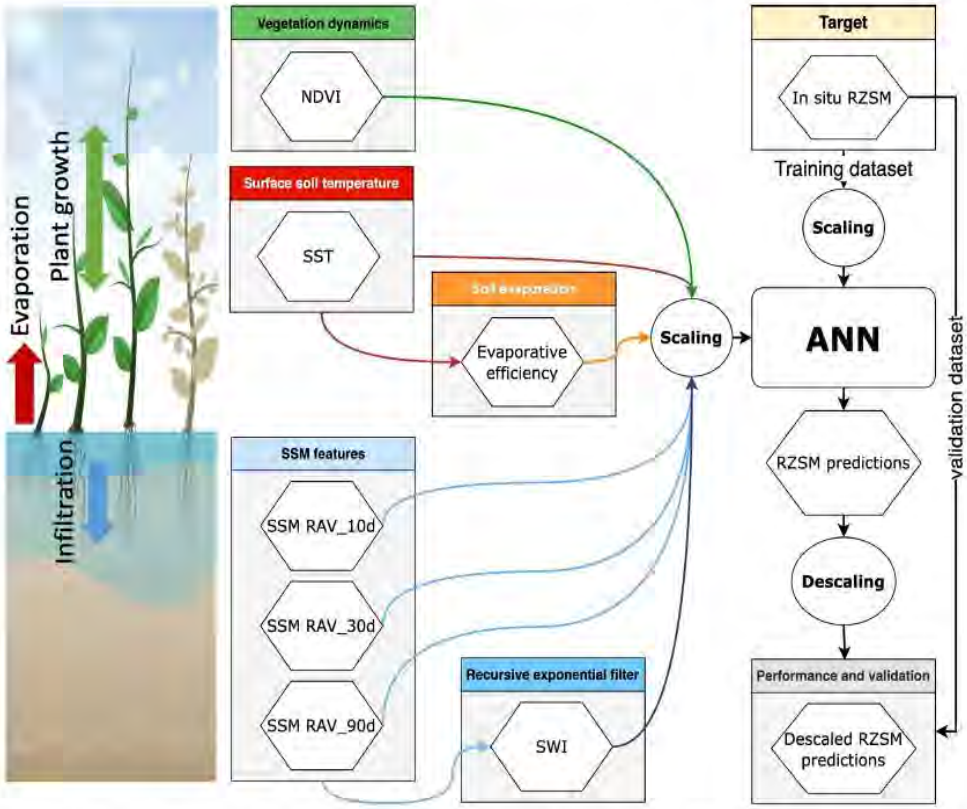


Figure 1. Overview of the processing configuration showing the components of the model: the tested models are variations of this ANN with a different combination of inputs (see Table 1). The scaling and descaling are applied to each dataset separately.

perature and process-related variables, namely, the soil water index given by a recursive exponential filter and remote-sensing-based evaporation efficiency. (2) The robustness of the approach is assessed through additional tests involving stations not included in the ISMN database in Tunisia, Italy and southern India. (3) Climatic analysis is conducted to infer the most indicative process-related features for each climate pattern.

2 Materials and methods

The proposed methodology entails the construction of several ANN models with both direct (SSM, surface temperature and NDVI) and intermediate sets of features (soil water index and evaporation efficiency) computed based on simplified analytical models. An overview of the processing configuration is shown in Fig. 1. Standard scaling is applied to each dataset separately so that the different inputs fall into the same range of values. Then the ANN outputs are descaled to make the comparison with actual values of RZSM possible.

This approach results in a combination of ANN models (Table 1). Each model has one or more process-related features in addition to three SSM features which correspond to backward rolling averages of in situ SSM computed over 10, 30 and 90 d. All the ANN model hyperparameters remain the same except the number of input features.

The model with the simplest starting point is ANN_SSM based on Souissi et al. (2020). The most complex model includes the full set of inputs. Intercomparison of the model performance provides information on the added value of each input. All input features are scaled, and training is performed on each of these features based on scaled in situ RZSM data retrieved from the ISMN. The RZSM model predictions are validated against an independent set of observations.

2.1 Datasets

2.1.1 ISMN soil moisture data

The first training and test operations were conducted on eight ISMN networks previously considered in Souissi et al. (2020). Figure 2 shows the distribution of the considered soil moisture networks with different soil textures and climatic parameters (see Appendix B). For each station, the RZSM observation point is located between 30 and 55 cm (Table 2). For each soil moisture hourly acquisition, the ISMN provides quality flags. Quality flags can be marked as "C" (exceeding the plausible geophysical range), "D" (questionable/dubious), "M" (missing), or "G" (good) (Dorigo et

Table 1. ANN model configurations with the respective input variables; [a]: rolling averages of SSM over 10 d; [b]: rolling averages of SSM over 30 d; [c]: rolling averages of SSM over 90 d; [d]: number of parameters of the ANN model.

Model features	SSM_10d_RAV[a]	SSM_30d_RAV[b]	SSM_90d_RAV[c]	SST	NDVI	SWI	EVAP	Nb[d]
ANN_SSM	X	X	X					101
ANN_SSM_TEMP	X	X	X	X				121
ANN_SSM_NDVI	X	X	X		X			121
ANN_SSM_EXP-FILT-T5	X	X	X			X		121
ANN_SSM_EVAP-EFF-B60	X	X	X				X	121
ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5	X	X	X		X	X	X	161

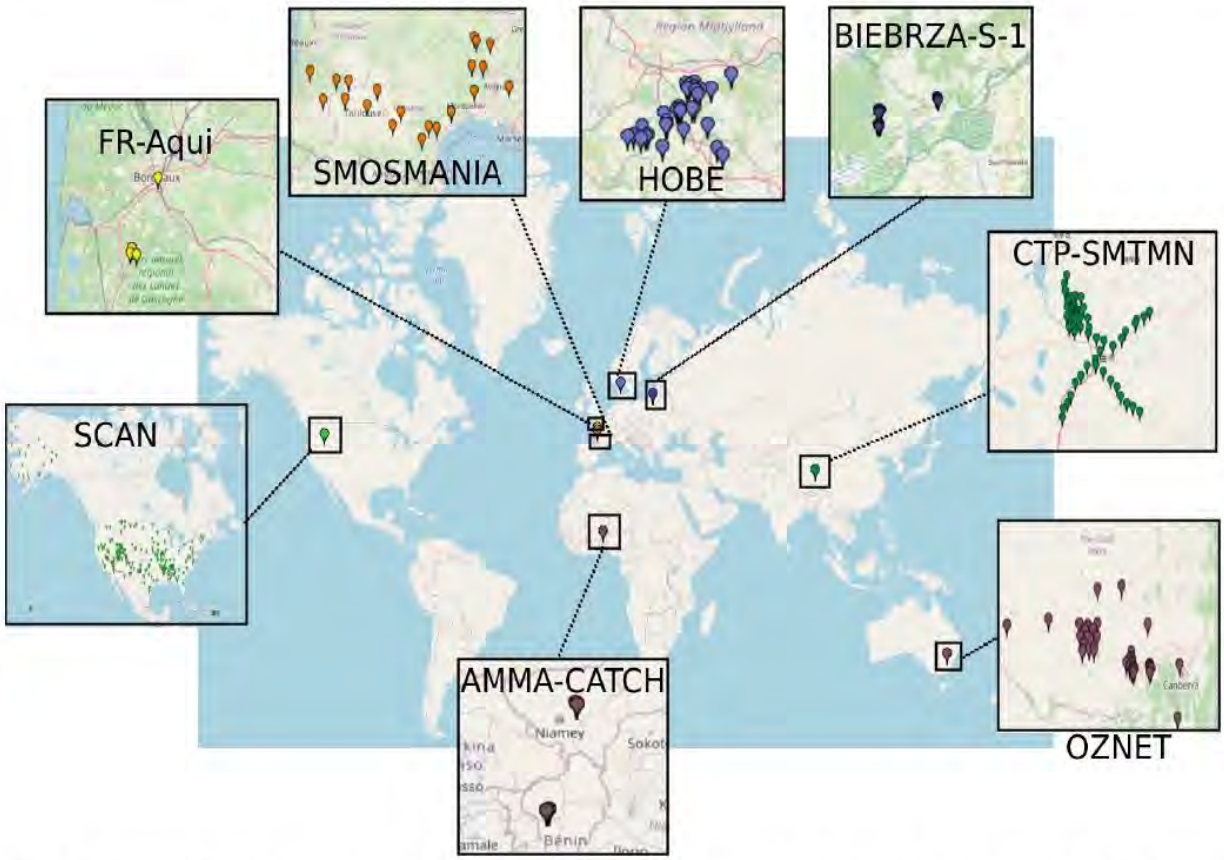


Figure 2. International Soil Moisture Network (ISMN) network distribution (adapted from the ISMN web data portal: https://www.geo.tuwien.ac.at/insitu/data_viewer/, last access: 7 April 2022; scale: 1 cm = 1000 km).

al., 2011). Category "D" has subset flags, namely, "D01" for which in situ soil temperature < 0 °C, "D02" that flags points at which in situ air temperature < 0 °C, as well as "D03" that also flags areas where the Global Land Data Assimilation System (GLDAS) soil temperature < 0 °C. In our study, only soil moisture data whose quality flag is marked "G" were retained.

2.1.2 External soil moisture data

The external networks are considered to assess the transferability and robustness of the approach. The trained models are run for predictions only over these sites. They have been selected to cover semiarid, moderate and tropical semiarid climates.

- *Tunisian site*: the Merguellil site is located in central Tunisia (9°54′ E, 35°35′ N). This site is characterized by a semiarid climate with highly variable rainfall patterns (average equal to 300 mm yr^{-1}), very dry summer seasons, and wet winters. The Merguellil site represents an agricultural region where croplands, namely, olive groves and cereal fields, prevail (Zribi et al., 2021). At this study site, a network of continuous ThetaProbe sta-

Table 2. Overview of the considered ISMN and external networks.

Network	Country	Number of selected stations	Selected RZSM depth (cm)	SM sensors
AMMA-CATCH	Benin, Niger	5 (3 in Benin and 2 in Niger)	40	CS616
BIEBRZA-S-1	Poland	3	50	GS-3
CTP-SMTMN	China	54	40	EC-TM/5TM
HOBE	Denmark	29	55	Decagon-5TE
FR-Aqui	France	5	30, 34, 50	ThetaProbe ML2X
OZNET	Australia	19	30	Hydra Probe-CS616
SCAN	USA	209	50	Hydraprobe-Sdi-12/Ana
SMOSMANIA	France	22	30	ThetaProbe ML2X

tions installed at bare soil locations provided moisture measurements at depths of 5 and 40 cm. All measurements were calibrated against gravimetric estimations. Data were obtained from the Système d'Information Environmental (SIE) web application catalog (SIE, 2021).

- *Italian site*: the Landriano site is located in northern Italy (Pavia Province, Lombardy Region). This station is located in a maize field, which was monitored in 2006 and from 2010 to 2011 (Masseroni et al., 2014). The average rainfall in Pavia Province is 650–700 mm, the climate is classified as Cfa (see Appendix A) and the field is irrigated by the border method with an average irrigation amount of approximately 100 to 200 mm per application with one to two applications per season due to the presence of a shallow groundwater table. Soil moisture measurements were performed with time domain reflectometer (TDR) soil moisture sensors. Five TDR soil moisture sensors were installed along a profile at depths of 5, 20, 35, 50 and 70 cm.

- *Indian site*: the Berambadi watershed is located in Gundalpet Taluk, Chamarajanagara district, in the southern part of Karnataka state in India and covers an area of approximately 84 km^2. The average rainfall is equal to 800 mm yr^{-1}, and the climate is classified as Aw (see Appendix A). Hydrological variables have been intensively monitored since 2009 in the Berambadi watershed by the Environmental Research Observatory ORE BVET and AMBHAS Observatory. The soil moisture levels at the surface (5 cm) and root zone (50 cm) are monitored with a HydraProbe sensor at different agricultural sites across the watershed, and in the current study, four stations were chosen.

2.1.3 Surface soil temperature

In addition to in situ soil moisture, the ISMN optionally includes meteorological and soil variables that are available over specific time periods. Values of the in situ surface soil temperature among these variables can be employed as a useful indicator of the soil moisture data quality. The soil temperature was provided in degrees Celsius, and the plausible values range from −60 to 60 °C. Regarding soil moisture data, surface soil temperature data were also provided with quality flags (Dorigo et al., 2011). However, the drawback is that this variable is not available in all networks, which is the case with the AMMA-CATCH network.

2.1.4 Normalized difference vegetation index

We considered the remote-sensing-based normalized difference vegetation index (NDVI) to infer vegetation dynamics. We extracted this index from the Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices product (MOD13Q1 version 6). MODIS Vegetation Indices data are generated at 16 d intervals and a 250 m spatial resolution as a level-3 product. This product provides two primary vegetation layers. The first vegetation layer is the NDVI, which is referred to as the continuity index of the existing National Oceanic and Atmospheric Administration-Advanced Very High Resolution Radiometer (NOAA-AVHRR)-derived NDVI. The algorithm chooses the best available pixel value from all the acquisitions over the 16 d period. The criteria considered are low cloud coverage, low viewing angle, and the highest NDVI value (Huete et al., 1999). To obtain daily NDVI values, we conducted linear interpolation of the 16 d product.

2.1.5 Potential evapotranspiration

Similarly, we assessed the impact of considering a remote-sensing-based evaporation efficiency, which is initially defined as the ratio of actual to potential soil evaporation, on RZSM prediction. The computation details of this variable will be given later (see Sect. 2.2.2). We employed the remote-sensing-based potential evapotranspiration (PET) to compute the evaporation efficiency. We extracted the PET from the MOD16A2 Evapotranspiration/Latent Heat Flux version 6

product, which is an 8 d composite dataset produced at a 500 m pixel resolution. The algorithm used for this product collection is based on the logic of the Penman–Monteith equation, which employs inputs of daily meteorological reanalysis data along with MODIS remote-sensing data products such as vegetation property dynamics, albedo and land cover. The MOD16A2 product provides layers for the composite evapotranspiration (ET), latent heat flux (LE), potential ET (PET) and potential LE (PLE). The pixel values for the PET layer include the sum of all 8 d within the composite period (Running et al., 2017). To obtain daily PET values, we performed a linear interpolation over the 8 d product, and then we divided the interpolated value by 8.

2.2 Methods

2.2.1 Recursive exponential filter

Two ANN models presented in Table 1 contained extra knowledge on infiltration process information based on the outputs of the recursive exponential filter (Stroud, 1999) as a feature. The recursive exponential filter was first introduced by Wagner et al. (1999) to estimate the soil water index (SWI) from surface soil moisture. SWI is computed as follows:

$$\mathrm{SWI}_{t_n} = \mathrm{SWI}_{t_{n-1}} + K_n\left(\mathrm{ms}\,(t_n) - \mathrm{SWI}_{t_{n-1}}\right), \tag{1}$$

where SWI_{t_n} is the soil water index at time t_n, $\mathrm{ms}(t_n)$ is the scaled surface soil moisture at time t_n (scaled between maximum and minimum values), K_n is the gain at time t_n, which occurs in [0, 1] and is given by

$$K_n = \frac{K_{n-1}}{K_{n-1} + e^{-\frac{\left(t_n - t_{n-1}\right)}{T}}}, \tag{2}$$

and T is a time constant and is the only required tuning parameter to compute the recursive exponential filter.

For the initialization of the filter, gain $K_1 = 1$ and $\mathrm{SWI}_{(t1)} = \mathrm{ms}(t_1)$.

Regarding T values, we considered an empirical list ([1, 3, 5, 7, 10, 13, 15, 20, 40, 60]), which was partly inspired by Paulik et al. (2014) ($T \in$ [1, 5, 10, 15, 20, 40, 60, 100]). Given the list of T values, recursive exponential filter outputs were computed for all of the stations (346 stations) given each T value. Based on the correlation values between the in situ RZSM values and the recursive exponential filter-based RZSM pre-estimates, we established the optimal time variable T, hereafter referred to as T_{best}, for each station.

2.2.2 Evaporation efficiency

An ANN model with evaporation efficiency input was also developed. This variable, which is defined as the ratio of the actual to potential soil evaporation, was first introduced in Noilhan and Planton (1989), Jacquemin and Noilhan (1990) and Lee and Pielke (1992) and thereafter readapted in Merlin et al. (2010) to include the soil thickness. In our work, we use a modified evaporation efficiency formulation based on the third model developed in Merlin et al. (2010), which can be expressed as follows (see Appendix C):

$$\beta = \left[\frac{1}{2} - \frac{1}{2}\cos\left(\pi\theta/\theta_{\mathrm{max}}\right)\right]^{P^*}, \tag{3}$$

where β is evaporation efficiency and θ is the water content in the soil layer of thickness L. θ_{max} is the maximum soil moisture at each station. P^* is a parameter computed as follows:

$$P^* = \frac{\mathrm{PET}}{2B}. \tag{4}$$

P^*, a proxy of parameter P (see Appendix C), represents an equilibrium state controlled by retention forces in the soil, which increase with the thickness L of considered soil and by evaporative demands at the soil surface. PET is the potential evapotranspiration extracted from the MODIS 500 m 8 d product (MOD16A2).

The soil evaporation efficiency computed by model 3, developed in Merlin et al. (2010), decreases when PET increases. Retention force and evaporative demand make the term P increase (replaced by P^*), as if an increase in potential evaporation LE_{p} (here replaced by PET) at the soil surface would make the retention force in the soil greater.

Merlin et al. (2010) tested this approach at two sites in southwestern France using in situ measurements of actual evaporation, potential evaporation, and soil moisture at five different depths collected in summer. Model 3 was able to represent the soil evaporation process with a similar accuracy to the classical resistance-based approach for various soil thicknesses up to 100 cm. Merlin et al. (2010) affirm that the parameterization of P as a function of LE_{p} (here PET) indicates that β cannot be considered a function of soil moisture alone since it also depends on potential evaporation. Moreover, the effect of potential evaporation on β appears to be equivalent to that of soil thickness on β. This equivalence is physically interpreted as an increase in retention forces in the soil in reaction to an increase in potential evaporation.

2.2.3 Artificial neural network implementation

The multilayer perceptron (MLP), which is a multilayer feedforward ANN, is one of the most widely applied ANNs, mainly in the field of water resources (Abrahart and See, 2007). The multilayer perceptron contains one or more hidden layers between its input and output layers. Neurons are organized in layers such that the neurons of the same layer are not interconnected and that any connections are directed from the lower to upper layers (Ramchoun et al., 2016). Each neuron returns an output based on the weighted sum of all inputs and according to a nonlinear function referred to as the transfer or activation function (Oyebode and Stretch,

Table 3. Proportion of the stations whose performance enhances using the ANN models enriched with process-related features compared to model ANN_SSM ([a]: % of stations at which the correlation improves over the model ANN_SSM level; [b]: % of stations at which RMSE improves over the model ANN_SSM level).

Model	Training stations		Validation stations		Test stations	
	% of stations (corr ↑)[a]	% of stations (RMSE ↓)[b]	% of stations (corr ↑)[a]	% of stations (RMSE ↓)[b]	% of stations (corr ↑)[a]	% of stations (RMSE ↓)[b]
ANN_SSM_NDVI	65.82	44.3	45.71	40.0	55.22	40.3
ANN_SSM_TEMP	49.4	25.3	55.56	38.89	59.35	42.99
ANN_SSM_EXP-FILT-T5	64.56	36.71	60.61	42.42	63.68	50.25
ANN_SSM_EVAP-EFF-B60	54.55	28.57	52.94	41.18	52.33	48.19
ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5	84.06	62.32	61.29	54.84	62.07	54.02

2019). The input layer, consisting of SSM values and/or other process-related variables, is connected to the hidden layer(s), which comprises hidden neurons. The final ANN-derived estimates of the ANN are given by an activation function associated with the final layer denoted as the output layer, based on the sum of the weighted outputs of the hidden neurons.

We started with the ANN model developed in Souissi et al. (2020), whose architecture consists of one hidden layer of 20 hidden neurons, a tangent sigmoid function as the activation function of the hidden layer, a quadratic cost function as the loss function and the stochastic gradient descent (SGD) technique as the optimization algorithm. This model was developed to estimate RZSM based on only in situ SSM information. SSM was not applied as a feature of hourly values but was employed in the form of three features, namely, SSM rolling averages over 10, 30 and 90 d. Additional ANN models were developed to study, through each model, the impact of the application of the NDVI, SWI, evaporation efficiency and surface soil temperature as features. A model combining surface soil moisture, NDVI, evaporation efficiency and the recursive exponential filter was further considered. These ANN models were trained and validated on the 122 ISMN stations considered to be of good quality after a data-filtering step as detailed in Souissi et al. (2020). Training of the above ANN models was conducted considering 70 % of these 122 stations. Thirty percent was reserved for validation, and testing was conducted at the rest of the stations. So, in summary, 122 stations were considered for the training/validation of the ANN models and 224 stations if all input data available were used for testing. In a second step, tests were conducted on data external to the ISMN database, namely, on sites of Tunisia, Italy and India. The trained models over the ISMN are used only in prediction mode over these sites. The data for SSM in addition to the other features are used as inputs, and RZSM is predicted in outputs.

3 Results

3.1 Exponential filter characteristic time length

A large proportion of the stations attained an optimal time constant (T_{best}) value equal to 60 d, which suggests an abnormally long infiltration time. These stations belong to the SCAN network and exhibit an RZSM acquisition depth of 50 cm, in contrast to other networks such as SMOSMANIA, for instance, where RZSM is retrieved at 30 cm. The high values correspond to correlation with seasonal dynamics rather than infiltration processes. This depth could explain the anomalously long infiltration time. This is consistent with Paulik et al. (2014), in which the average T value with the highest correlation (T_{best}) increased with increasing depth of the in situ observations.

For comparison purposes, Paulik et al. (2014) found that 23.98 % of the stations achieved $T_{best} = 5$ d, while 21.58 % of the stations achieved $T_{best} \geq 60$ d (60 or 100 d).

Albergel et al. (2008) considered an average T_{best} value of 6 d for the SMOSMANIA network. This value represented the average T_{best} value for all stations belonging to the SMOSMANIA network. In our case, the average T_{best} value for all stations of the SMOSMANIA network reached 9 d. In this study, an average T_{best} value could be established for each station or each network. However, this is not relevant to our work because we aim to evaluate maps of remote-sensing data in the next steps, and thus we did not compute T_{best} at each location. We fixed the value of T to 5 d as a median infiltration time.

3.2 Intercomparison of the ANN models

The distribution histograms for training, validation and test stations (Fig. 3) show that the integration of the considered process-related features improved the prediction accuracy in certain cases compared to the reference. Time series of good and less good quality of fit were provided in Appendix E for training, validation and test stations using the reference model ANN_SSM and the most complex ANN model.

In terms of the NDVI, 65.82 %, 45.71 % and 55.22 % of the stations attained better correlation values with

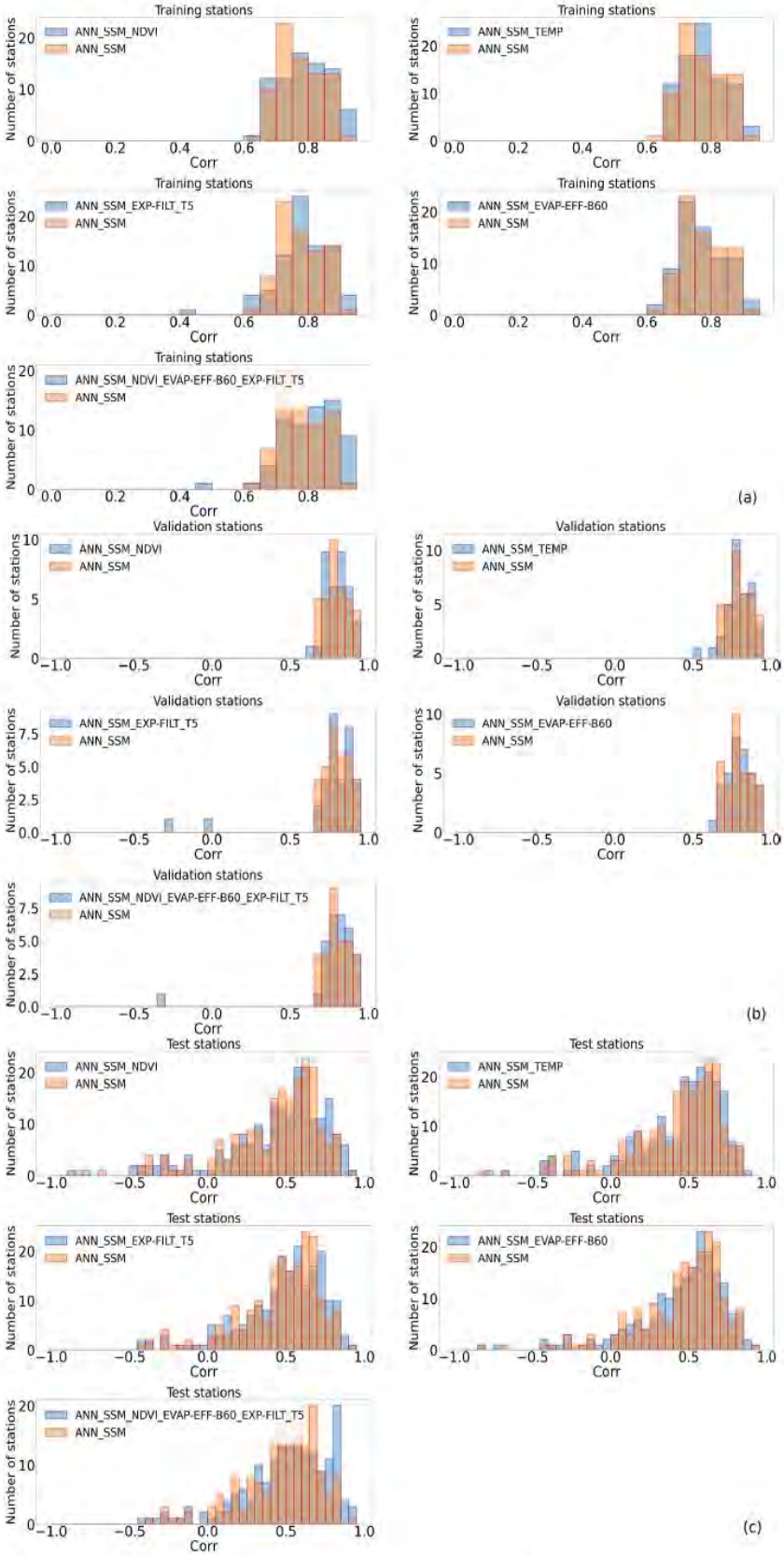


Figure 3. Correlation histograms of all tested ANN models compared to ANN_SSM (**a**) on training stations (**b**), on validation stations (**c**) and on test stations (see Appendix D for RMSE histograms).

ANN_SSM_NDVI than those obtained with ANN_SSM for the training, validation and test stations, respectively. Root mean square error (RMSE) decreased for 44.3 %, 40.0 % and 40.3 % of the stations with ANN_SSM_NDVI compared to model ANN_SSM for the training, validation and test stations, respectively (Table 3).

In regard to the ANN_SSM_TEMP model that integrates the soil surface temperature, 49.4 %, 55.56 % and 59.35 % of the training, validation and test stations exhibited higher correlation values than those obtained with the ANN_SSM model, respectively. RMSE decreased with ANN_SSM_TEMP compared to model ANN_SSM for 25.3 %, 38.89 % and 42.99 % of the training, validation and test stations, respectively.

64.56 %, 60.61 % and 63.68 % of the training, validation and test stations attained better correlations than those obtained with model ANN_SSM, respectively. In addition, RMSE decreased for 36.71 %, 42.42 % and 50.25 % of the training, validation and test stations with ANN_SSM_EXP-FILT-T5 compared to model ANN_SSM, respectively.

Regarding the evaporation efficiency, we considered different values of the fitting parameter B (Eq. 4) such that B remained within the [50, 60] interval. This parameter can be fitted using different variables, such as the wind speed or relative humidity. Comparisons based on the correlation values provided by the different models for each B value indicated that the performance was insensitive to the B value. Thus, we fixed the B value to $60\,\mathrm{W\,m^{-2}}$. Comparison of models ANN_SSM and ANN_SSM_EVAP-EFF-B60 revealed that 54.55 %, 52.94 % and 52.33 % of the training, validation and test stations attained higher correlation values with the latter model, respectively. RMSE was reduced for 28.57 %, 41.18 % and 48.19 % of the training, validation and test stations with ANN_SSM_EVAP-EFF-B60 compared to model ANN_SSM, respectively.

Finally, we investigated the impact of the joint application of the NDVI, recursive exponential filter ($T = 5\,\mathrm{d}$) and evaporation efficiency ($B = 60\,\mathrm{W\,m^{-2}}$) in the ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 model. The surface soil temperature was not included, as its effect is included in the evaporation process. At 84.06 %, 61.29 % and 62.07 % of the training, validation and test stations, the correlation value obtained with this model was higher than that obtained with the ANN_SSM model, respectively. In addition, RMSE was minimized for 62.32 %, 54.84 % and 54.02 % of the training, validation and test stations with ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 compared to model ANN_SSM, respectively.

Considering model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5, only one training station had a decrease in correlation by more than 0.1, namely, station "Lind#1" (network "SCAN") compared to the reference model ANN_SSM. All inputs were not available at the same dates, which implied a significant reduction in data points (see Appendix F). The decrease in correlation and increase in RMSE did not exceed 0.1 and $0.01\,\mathrm{m^3\,m^{-3}}$, respectively, for the rest of the stations of lower performance metrics with the most complex ANN (Table 4).

Similarly for validation stations, only one station had a decrease in correlation above 0.1, namely, station "PineNut" (network SCAN), with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5. This decrease can be also explained because of data shortage (see Appendix F). The decrease in correlation and increase in RMSE did not exceed 0.1 and $0.01\,\mathrm{m^3\,m^{-3}}$, respectively, for the rest of the stations of lower performance metrics with the most complex ANN (Table 4).

Regarding the test stations, correlation decreased by more than 0.1 and RMSE increased by more than $0.01\,\mathrm{m^3\,m^{-3}}$ with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 compared to model ANN_SSM, detected for only two stations. Both stations, namely, stations "S-Coleambally" and "Widgiewa", which belong to network "OZNET", significantly lose in data volume when process-related variables are integrated into ANN and more precisely because of NDVI data availability (see Appendix F). For the rest of the test stations, correlation decreased and RMSE increased simultaneously by less than 0.1 and $0.01\,\mathrm{m^3\,m^{-3}}$, respectively, with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 (Table 4).

Always in terms of the general performance of model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5, about 75 % of the stations have an RMSE of less than $0.05\,\mathrm{m^3\,m^{-3}}$, and around half of the stations have an RMSE of less than $0.04\,\mathrm{m^3\,m^{-3}}$. This accuracy is consistent, for instance, with the target value in the SMAP (Entekhabi et al., 2010) and SMOS (Kerr et al., 2010) missions, which is equal to $0.04\,\mathrm{m^3\,m^{-3}}$, and also with the average sensor accuracy adopted by Dorigo et al. (2013), which is equal to $0.05\,\mathrm{m^3\,m^{-3}}$. Overall, the most complex model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 can successfully characterize the soil moisture dynamics in the root zone since half of the stations have a correlation value of greater than 0.7. Pan et al. (2017) developed different ANN models to estimate RZSM at depths of 20 and 50 cm over the continental USA using surface information. They found that half of the stations have an RMSE of less than $0.06\,\mathrm{m^3\,m^{-3}}$ and that more than 70 % of stations have a correlation above 0.7 when predicting RZSM at 20 cm. However, the developed ANN was less effective in RZSM prediction at 50 cm, which is also in accordance with Kornelsen and Coulibaly (2014). In our study, the densest soil moisture network is SCAN, located in the USA. Soil moisture was predicted at a depth of 50 cm over this network. Around half of the stations have a correlation value of above 0.6 and an RMSE of less than $0.04\,\mathrm{m^3\,m^{-3}}$ after the integration of process-related inputs. Pan et al. (2017) suggest that the use of only time-dependent variables may not be sufficient for the ANN models to accurately predict RZSM and suggest adding soil texture data.

Table 4. Proportion of the stations whose correlation decreases using the ANN models enriched with process-related features compared to model ANN_SSM ($\Delta_{corr} = corr_{ANN\_SSM\_X} - corr_{ANN\_SSM}$; X denotes one or a combination of process-related variables)

Model	Training stations		Validation stations		Test stations	
	% of stations corr ↓ and $0.05 < \Delta_{corr} < 0.1$	% of stations corr ↓ and $\Delta_{corr} > 0.1$	% of stations corr ↓ and $0.05 < \Delta_{corr} < 0.1$	% of stations corr ↓ and $\Delta_{corr} > 0.1$	% of stations corr ↓ and $0.05 < \Delta_{corr} < 0.1$	% of stations corr ↓ and $\Delta_{corr} > 0.1$
ANN_SSM_NDVI	3.8	0	2.86	0	9.95	5.97
ANN_SSM_TEMP	0	1.2	0	2.78	4.67	3.27
ANN_SSM_EXP-FILT-T5	6.33	1.27	3.03	9.09	6.97	3.48
ANN_SSM_EVAP-EFF-B60	10.39	1.3	0	2.94	6.74	5.7
ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5	4.35	1.45	6.45	3.23	9.2	6.9

3.3 Robustness of the approach

To further assess the robustness of our approach, which involves RZSM prediction using the different ANN models with different features, we predicted RZSM at sites not previously considered in previous parts of the study. The selected stations are located in the Kairouan Plain, a semiarid region in central Tunisia, the Landriano site located in the north of Italy, and the Berambadi watershed located in Gundalpet Taluk, southern India. In the case of Tunisia, model ANN_SSM yielded moderate to low-precision predictions, as highlighted by the performance metrics listed in Table 5. The time series (see Appendix G) show that the RZSM predictions followed the SSM seasonality, which was reflected by the false peaks generated in the RZSM predictions whenever a sharp increase or decrease occurred in the SSM values. This observation was also found in Souissi et al. (2020). Actually, the Kairouan Plain is characterized by a semiarid environment where rainfall events infrequently occur and the level of evaporation is high. The reference model ANN_SSM shows its limitations in accurately predicting RZSM in areas with no alternate wet and dry cycles.

However, the consideration of additional features, namely, the NDVI, evaporation efficiency and SWI in the ANN models, resulted in good agreement between the in situ and predicted RZSM values (Fig. 4). The correlation values were improved by 60.04 %, 169.5 %, 112.02 %, 80.23 % and 53.7 % at stations Barrouta-160, Hmidate_163, Barrage_162, Bouhajla_164 and P12, respectively, with the ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 model over ANN_SSM model values. Similarly, RMSE values were reduced (Table 5). As shown in Fig. 4, the most complex ANN model is able to capture the variations of RZSM. This finding highlights the added value of our hybrid approach based on an association of a machine learning method with process-related variables. Instead of injecting uncertain information into physical models, such as soil properties, we used a nonparametric method related to physical processes without using forcing data that may be subject to errors and potentially lead to inaccurate tracking of the long-term evolution of soil moisture.

A second comparison can be conducted between the quality of fit of these independent datasets and training datasets. Actually, the climate class of the Tunisian stations is Bsh (see Appendix A). At the training stage, no station falls into climate class Bsh (see Appendix A). However, some training stations fall under a similar climate class, which is Bsk (see Appendix B). Table 5 presents correlation and RMSE values for these training stations and Tunisian sites with both models ANN_SSM and ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5. For all training stations, performance metrics are slightly enhanced, with the most complex ANN model compared to the reference model ANN_SSM, except for stations GrouseCreek, Harmsway and Lind#1, whose performance decreases. Overall, the range of correla-

Table 5. Performance metrics of models ANN_SSM and ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 at training stations of climate "Bsk" and Tunisian stations of climate "Bsh".

Model Station	ANN_SSM		ANN_SSM_NDVI_EVAP -EFF-B60_EXP-FILT-T5	
	Correlation	RMSE	Correlation	RMSE
Training stations (climate class Bsh)				
Banandra (OZNET)	0.701	0.05	0.764	0.046
DRY-LAKE (OZNET)	0.674	0.031	0.692	0.03
CPER (SCAN)	0.691	0.032	0.695	0.032
EPHRAIM (SCAN)	0.758	0.051	0.791	0.046
GrouseGreek (SCAN)	0.818	0.033	0.802	0.035
HarmsWay (SCAN)	0.705	0.034	0.622	0.038
Lind#1 (SCAN)	0.605	0.055	0.483	0.022
External test stations (Tunisia)				
Barrouta_160	0.463	0.021	0.714	0.016
Hmidate_163	0.318	0.019	0.834	0.011
Barrage_162	0.416	0.035	0.864	0.019
Bouhajla_164	0.435	0.016	0.733	0.01
P12	0.581	0.047	0.861	0.029

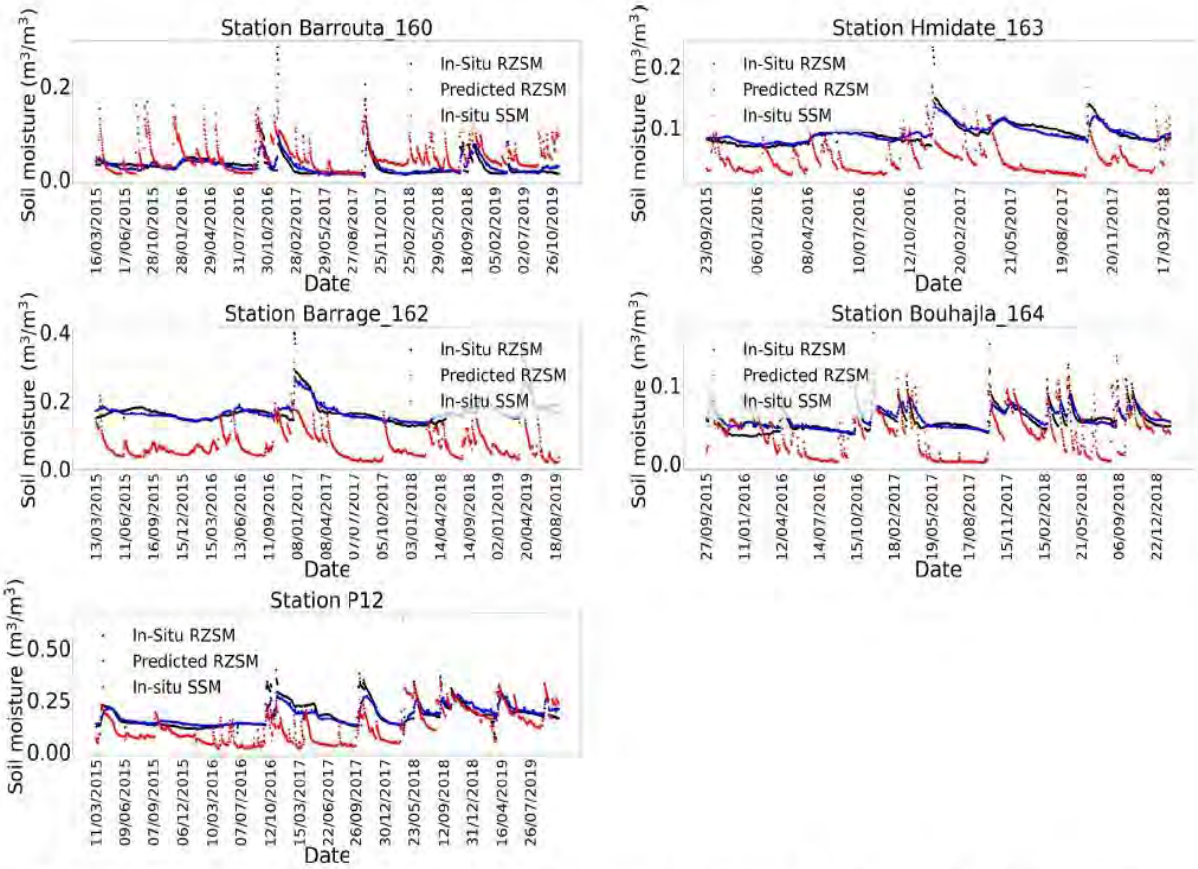


Figure 4. In situ SSM, in situ RZSM, and predicted RZSM series at the stations in the Kairouan Plain (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 (see Appendix G for a larger figure format).

tion values is similar for training and external validation stations with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5, and RMSE is greatly reduced for the Tunisian stations compared to the training stations. Given the results on unseen datasets, namely, on Tunisia, the performance of the most complex ANN model is good as it is able to generalize the patterns present in the training dataset.

At the southern Indian stations, the ANN_SSM model yielded good agreement even without the integration of process-related features (Table 6). These features added little to nonsignificant improvement. The same observation was made at the Italian site. The application of multiple features performed the best under arid conditions, e.g., in Tunisia. In the tropical and temperate climate regions, this was not the case. The presence of clouds in the MODIS NDVI and potential evapotranspiration products could explain this observation at sites of southern India and northern Italy. In southern India, for instance, the maximum variability in soil moisture occurred during the monsoon season, which is characterized by a large amount of clouds. Moreover, the coarse resolution of the MODIS NDVI product makes it sometimes not adapted to the considered site. Chen et al. (2016) investigated the impact of sample impurity and landscape heterogeneity on crop classification using coarse-spatial-resolution MODIS imagery. They showed that the sample impurity such as mixed crop types in a specific sample, compositional landscape heterogeneity, which is the richness and evenness of land cover types in a landscape, and configurational heterogeneity, which is the complexity of the spatial structure of land cover types in a specific landscape, are sources of uncertainty affecting crop area mapping when using coarse-spatial-resolution imagery. High-resolution NDVI from sensors like Sentinel-2 could have been used in this exercise to mitigate the spatial resolution issue; however, MODIS data were privileged in order to provide NDVI and PET from the same sensor.

4 Discussion

Climate analysis of the results yielded by the different models indicated that, among all the models, the climate class with the highest mean correlation change rate (Fig. 5) was class BWk (see Appendix A), which regroups desert areas where the link between SSM and RZSM is weak due to high evaporative rates. Class Dfa (see Appendix A), which includes areas experiencing harsh and cold winters, also yielded a high mean correlation change rate ($> 100\,\%$). Similarly, at stations of this climate type, the link between the surface and root zone is poor. In regard to class Cfa (see Appendix A), in which more than $80\,\%$ of the total stations belong to the SCAN network, the high mean correlation change rate could be explained by the surface–subsurface decoupling phenomena detected within this network, as previously reported in Souissi et al. (2020). The model with the largest number of stations with improved predictions over the ANN_SSM model predictions was ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5. Actually, the coupled use of process-related features in the ANN models exerted a greater impact on the prediction accuracy than that exerted by the one-at-a-time application of these features. In model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5, the three process-based features jointly employed seemed to counterbalance the weight of the three SSM features. In this model, the process-related features were equally represented versus the SSM information depicted by three features. The redundancy of the considered SSM information could explain the limited impact of the one-at-a-time addition of process-related features.

In addition, Karthikeyan and Mishra (2021) demonstrated that, at root depths beyond 20 cm, the importance of SSM was notably lower than that at the 20 cm depth, signifying decorrelation between surface and deeper SM values, which is in accordance with the findings in Souissi et al. (2020), and it was further revealed that vegetation exhibits a higher importance than that of the meteorological predictors and precipitation. Kornelsen and Coulibaly (2014) indicated that evapotranspiration is the most important meteorological input for the prediction of soil moisture in the root zone with the MLP, which reflects the importance of the water vapor flux in soil moisture state determination.

$$\text{mean\_corr\_change\_rate} = \text{mean}\left(\frac{\text{corr}_{\text{ANN\_SSM\_X}} - \text{corr}_{\text{ANN\_SSM}}}{\text{corr}_{\text{ANN\_SSM}}} \cdot 100\right) \quad (5)$$

where X denotes a process-related variable ($X \in$ ["NDVI", "EXP-FILT-T5", "EVAP-EFF-B60", "TEMP"]).

The world map illustrated in Fig. 6 shows the best-performing ANN models based on the mean correlation change rate (Eq. 5). We assumed that the results in a given area of a specific climate class could be extended to other areas of the same climate class even if we did not consider the data for these areas. The climate classes without at least one station were marked in black and labeled with "NO DATA".

In arid areas such as the eastern and western sides of the USA with high evaporation rates, ANN_SSM_EVAP-EFF-B60 was the best-performing model. Similarly, in bare areas of Africa, the Middle East and Australia where the Bwh climate class prevailed (arid desert hot climate; see Appendix A), the evaporation efficiency was the best informative variable.

In the internal part of continental Europe and near the Mediterranean basin, the NDVI was the most relevant indicator for RZSM estimation, where agricultural fields dominated. Similarly, the Great Plains region in the USA was deeply affected by the NDVI, as this region is a cultivated area. In Nordic areas characterized by the ET climate class (Polar Tundra climate, see Appendix A) and mainly covered by grassland and shrubland areas according to ESA CCI land cover maps.

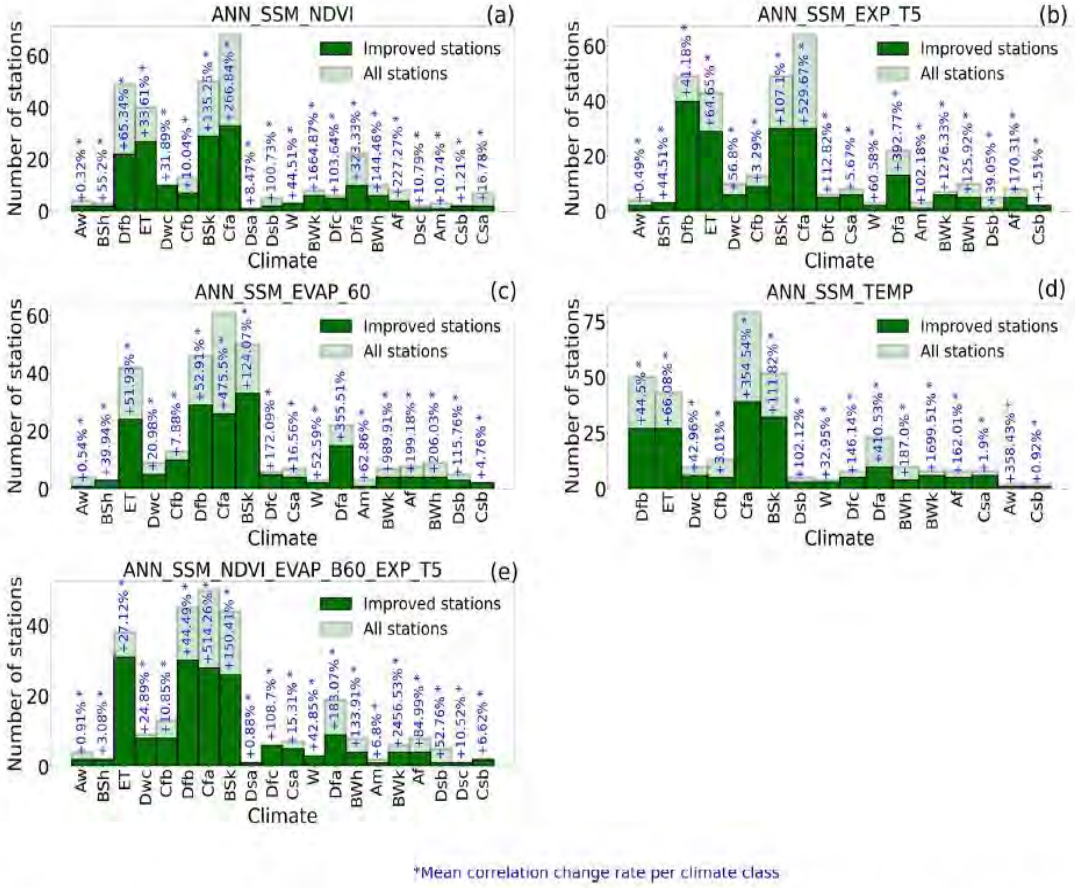


Figure 5. Climate classification of the stations performing better with models **(a)** ANN_SSM_NDVI, **(b)** ANN_SSM_EXP-FILT-T5, **(c)** ANN_SSM_EVAP-EFF-B60, **(d)** ANN_SSM_TEMP and **(e)** ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 compared to model ANN_SSM (dark green corresponds to stations whose correlation improved with complexified models, light green corresponds to total stations and rate in blue corresponds to the mean correlation change rate per climate class).

Table 6. Performance metrics of models ANN_SSM, ANN_SSM_NDVI and ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 at the sites in southern India and northern Italy.

Model station	ANN_SSM		ANN_SSM_NDVI		ANN_SSM_NDVI_EVAP-EFF_B60_EXP-FILT-T5	
	Correlation	RMSE	Correlation	RMSE	Correlation	RMSE
India						
Madyanahundi	0.813	0.04	0.78	0.042	0.744	0.044
Bheemanbidu	0.76	0.046	0.784	0.044	0.763	0.046
Beechanalli2	0.825	0.038	0.787	0.04	0.743	0.044
Beechanalli1	0.713	0.024	0.713	0.024	0.633	0.025
Italy						
Landriano	0.861	0.038	0.827	0.041	0.841	0.038

In Nordic areas characterized by the ET climate class, the soil temperature was the most important root-zone soil moisture indicator, mainly because of the freeze–thaw events encountered in these regions. In tropical savannah wet areas (class Aw; see Appendix A), the ANN_SSM_TEMP model was the best-performing model.

This classification definitely suffered from limitations mainly provoked by the generalization of the climatic analysis results to areas not considered in this study. For instance,

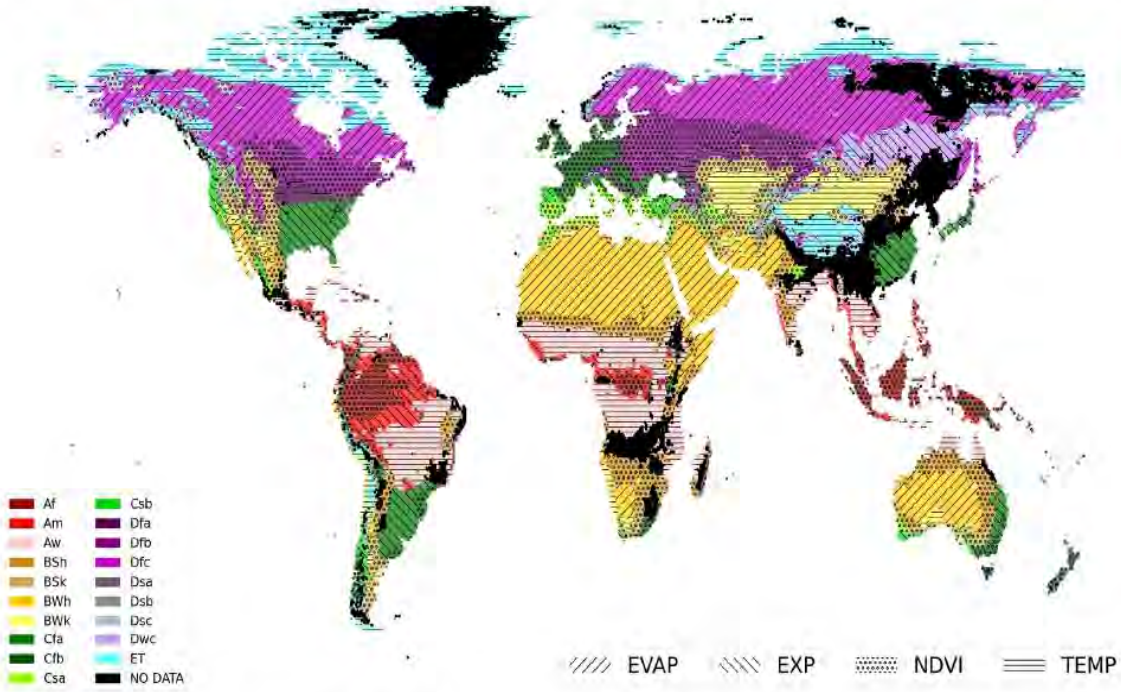


Figure 6. World map of the best-performing ANN models per climate class based on the mean correlation change rate; colors correspond to climate classes (see Appendix A), and hatches correspond to the most contributive input to the predictions, namely, EVAP (evaporation efficiency), EXP (exponential filter SWI), NDVI, and TEMP (surface soil temperature).

in regions of climate class Dfc (cold dry without a dry season, cold summer climate; see Appendix A), we expect the temperature to serve as the most relevant indicator instead of the evaporation efficiency.

5 Conclusion

In this study, we developed several ANN models to estimate RZSM based either solely on in situ SSM information or on a group of process-related features in addition to SSM, namely, the soil water index computed with a recursive exponential filter, evaporation efficiency, NDVI and surface soil temperature. Different regions across the globe with distinct land cover and climate patterns were considered. The main conclusion of this study was that the consideration of more features in addition to SSM information could enhance the accuracy of RZSM predictions mainly in regions where the link between SSM and RZSM is weak.

In arid areas with high evaporation rates, the most informative feature was the evaporation efficiency. In regions with agricultural fields, the NDVI was, for example, the most relevant indicator to predict RZSM. Overall, the best-performing model included the surface soil moisture, NDVI, SWI and evaporation efficiency as features. The robustness of the approach was further assessed through additional tests considering external sites in central Tunisia, India and Italy. Similarly, the process-related features exerted a positive impact on the prediction accuracy when combined with surface soil moisture in the case of Tunisia. The mean correlation across the five Tunisian stations sharply increased from 0.44 when only SSM was considered to 0.8 when all process-related features were combined with SSM. In India and Italy, the correlations were already high with the reference model ANN_SSM. The change in correlation after the addition of process-related features, namely, NDVI, is about −0.04, which is nonsignificant and is potentially because of the cloudy conditions in India and the noisy MODIS products. Also, the crop heterogeneity and sample impurity make MODIS NDVI products not adapted to all sites.

As a research perspective, datasets can be separated into clusters corresponding to major climate classes and/or soil types. More analysis can be conducted in this direction to eventually make connections between the different inputs and climate/soil configurations.

Future work will examine the ability of the developed model to estimate RZSM across larger areas based on remote-sensing global soil moisture products. The use of remote-sensing-derived soil moisture products may yield lower correlations with the reference model ANN_SSM, which potentially implies further improvement when process-related features are added.

Appendix A: Climate classes (Köppen classification)

Af	Tropical rainforest
Am	Tropical monsoon
As	Tropical savanna dry
Aw	Tropical savanna wet
BWk	Arid desert cold
BWh	Arid desert hot
BWn	Arid desert with frequent fog
BSk	Arid steppe cold
BSh	Arid steppe hot
BSn	Arid steppe with frequent fog
Csa	Temperate dry hot summer
Csb	Temperate dry warm summer
Csc	Temperate dry cold summer
Cwa	Temperate dry winter, hot summer
Cwb	Temperate dry winter, warm summer
Cwc	Temperate dry winter, cold summer
Cfa	Temperate without a dry season, hot summer
Cfb	Temperate without a dry season, warm summer
Cfc	Temperate without a dry season, cold summer
Dsa	Cold dry summer, hot summer
Dsb	Cold dry summer, warm summer
Dsc	Cold dry summer, cold summer
Dsd	Cold dry summer, very cold winter
Dwa	Cold dry winter, hot summer
Dwb	Cold dry winter, warm summer
Dwc	Cold dry winter, cold summer
Dwd	Cold dry winter, very cold winter
Dfa	Cold dry without a dry season, hot summer
Dfb	Cold dry without a dry season, warm summer
Dfc	Cold dry without a dry season, cold summer
Dfd	Cold dry without a dry season, very cold winter
ET	Polar tundra
EF	Polar eternal winter
W	Water

Appendix B: Climate and soil texture properties of the soil moisture stations

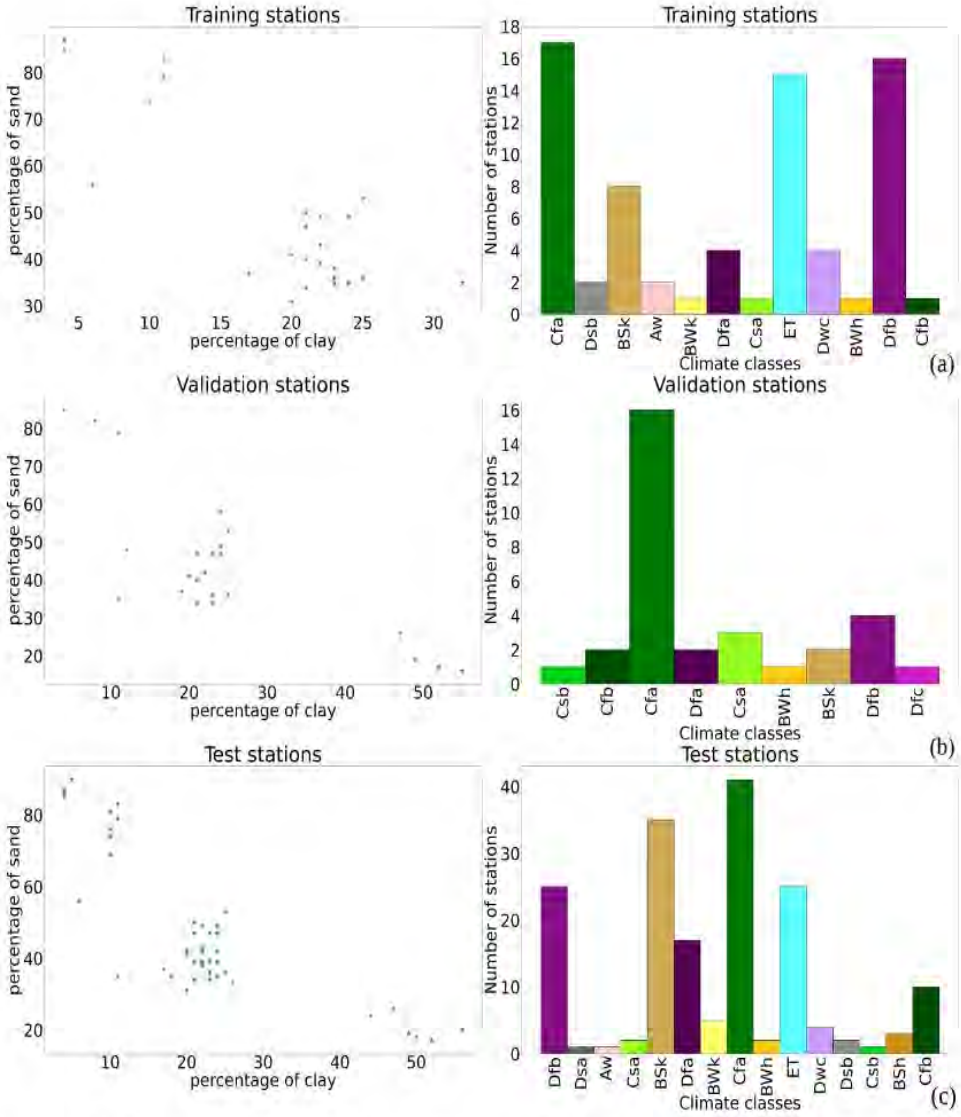


Figure B1. Climate and soil texture for **(a)** training stations, **(b)** validation stations and **(c)** test stations.

Appendix C: Evaporation efficiency

Evaporation efficiency (Sect. 2.2.2): the standard equations to compute evaporation efficiency (β_3) in Merlin et al. (2010) are as follows:

$$\beta_3 = \left[\frac{1}{2} - \frac{1}{2}\cos\left(\pi\theta_L/\theta_{max}\right)\right]^P \text{ for } \theta_L \le \theta_{max},$$
$$\beta_3 = 1 \text{ for } \theta_L > \theta_{max}, \quad \text{(C1)}$$

where θ_L is the water content in the soil layer of thickness L. P is a parameter computed as follows:

$$P = \left(\frac{1}{2} + A_3\frac{L - L_1}{L_1}\right)\frac{LE_p}{B_3}. \quad \text{(C2)}$$

θ_{max} is the soil moisture at saturation. LE_p is the potential evaporation. L_1 is the thinnest represented soil layer, and A_3 (unitless) and B_3 (W m^{-2}) are the two best-fit parameters a priori depending on the soil texture and structure, respectively.

Appendix D: Intercomparison of ANN models based on RMSE

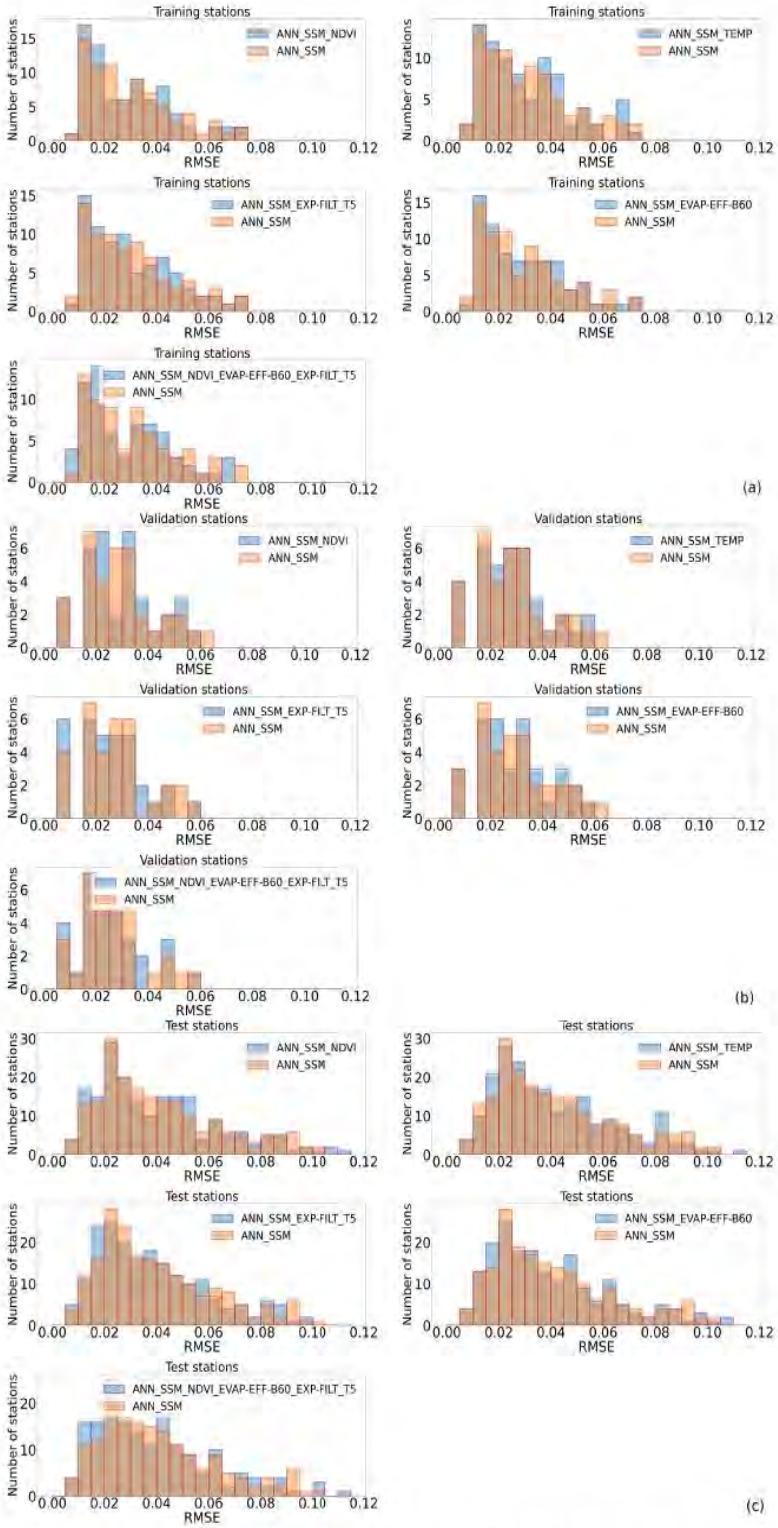


Figure D1. RMSE histograms of all tested ANN models compared to ANN_SSM (**a**) on training stations, (**b**) on validation stations and (**c**) on test stations.

Appendix E: Time series of good and less good quality of fits

E1 Training stations

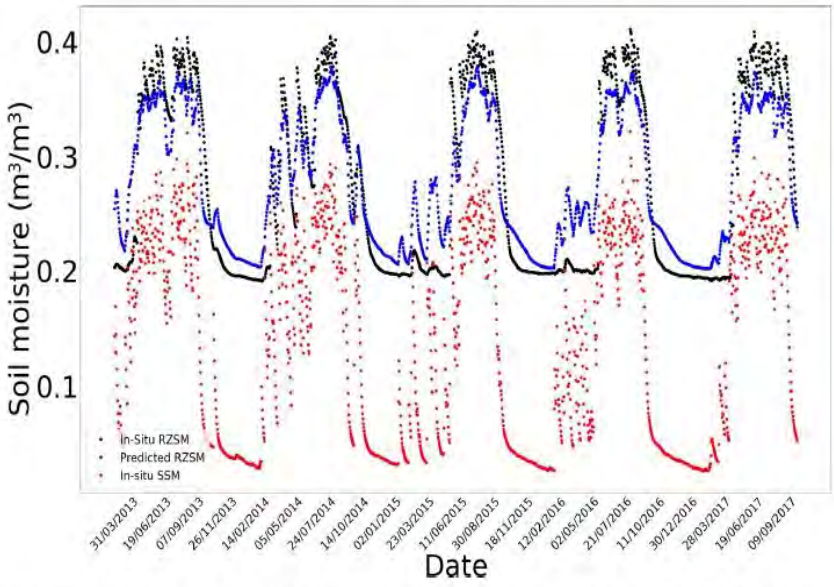


Figure E1. In situ SSM, in situ RZSM, and predicted RZSM series at station Beloufoungou Mid (AMMA-CATCH) with model ANN_SSM.

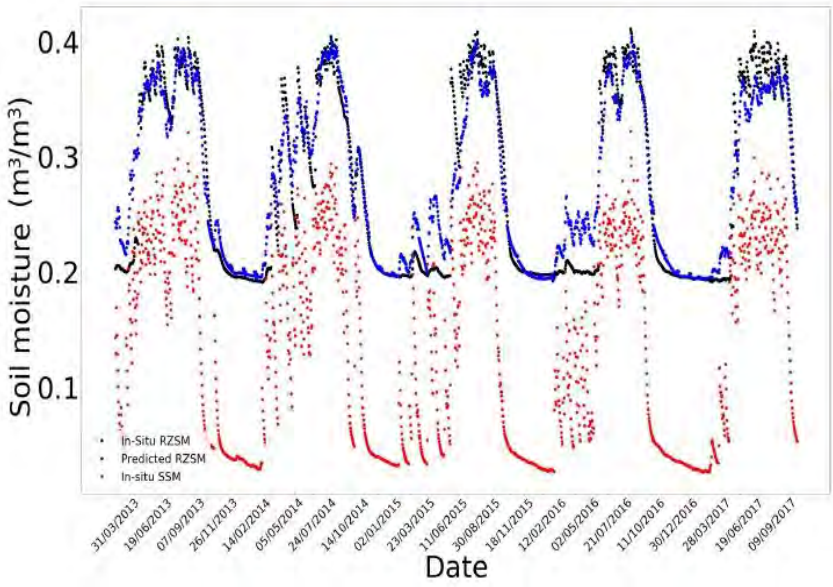


Figure E2. In situ SSM, in situ RZSM, and predicted RZSM series at station Beloufoungou Mid (AMMA-CATCH) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

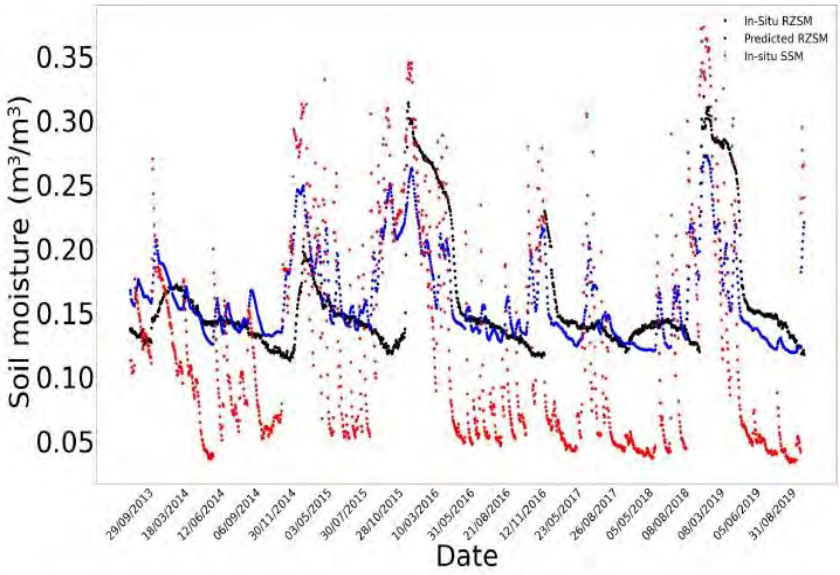


Figure E3. In situ SSM, in situ RZSM, and predicted RZSM series at station HarmsWay (SCAN) with model ANN_SSM.

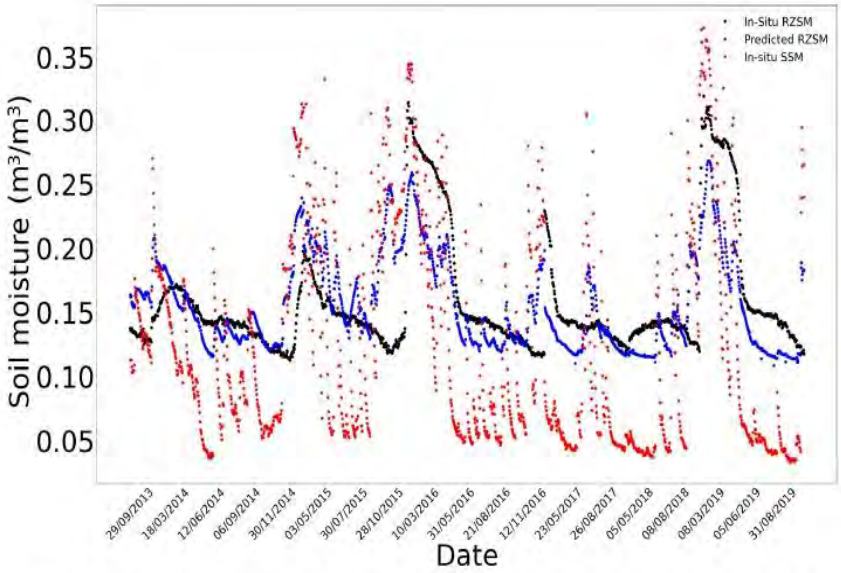


Figure E4. In situ SSM, in situ RZSM, and predicted RZSM series at station HarmsWay (SCAN) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

E2 Validation stations

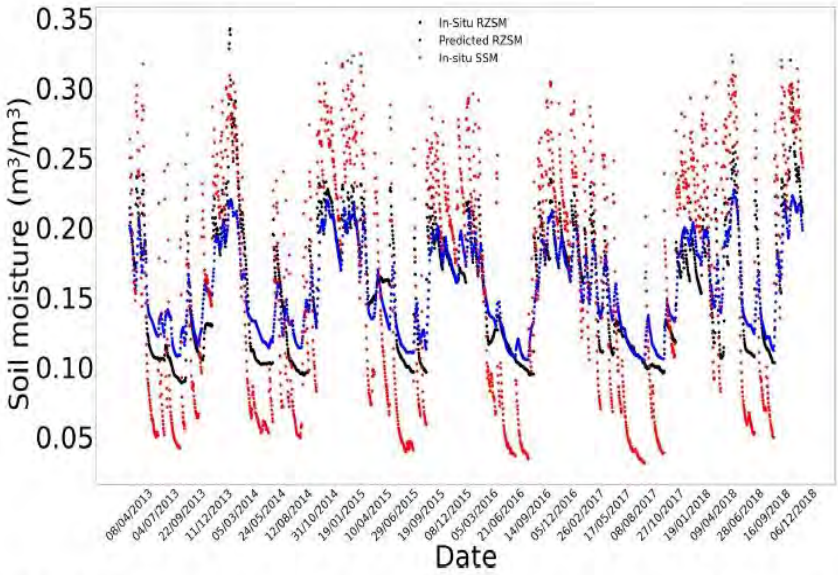


Figure E5. In situ SSM, in situ RZSM, and predicted RZSM series at station Cabrieres D'Avignon (SMOSMANIA) with model ANN_SSM.

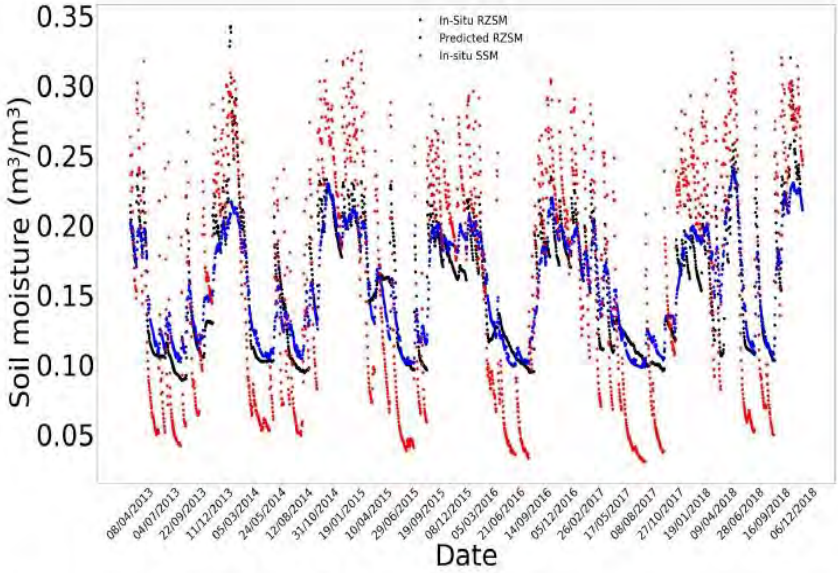


Figure E6. In situ SSM, in situ RZSM, and predicted RZSM series at station Cabrieres D'Avignon (SMOSMANIA) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

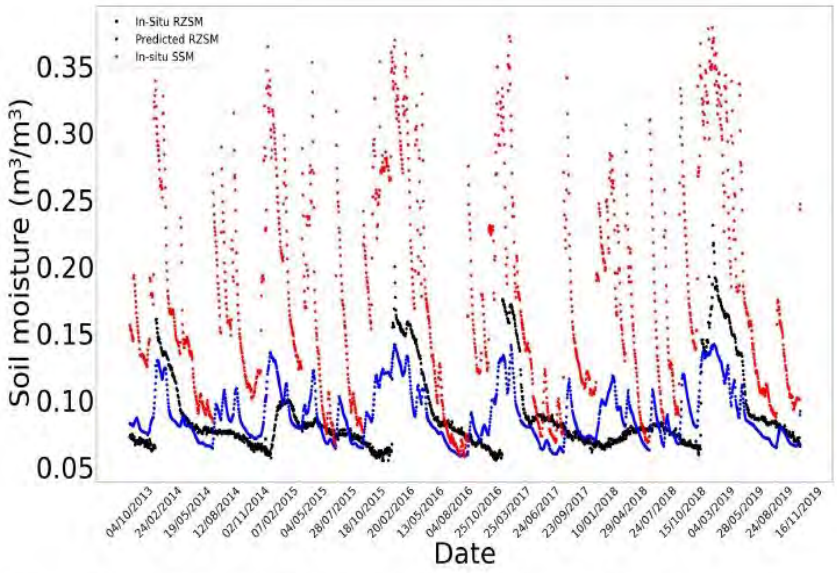


Figure E7. In situ SSM, in situ RZSM, and predicted RZSM series at station Nephi (SCAN) with model ANN_SSM.

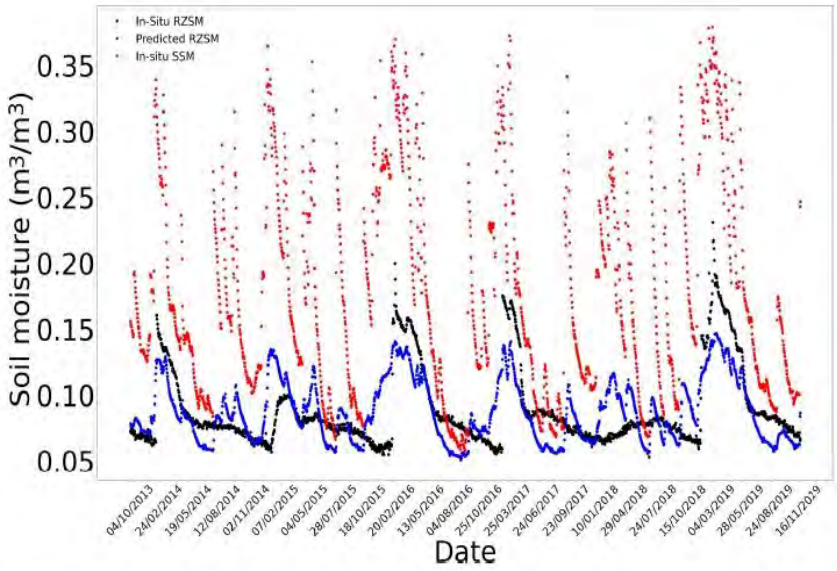


Figure E8. In situ SSM, in situ RZSM, and predicted RZSM series at station Nephi (SCAN) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

E3 ISMN test stations

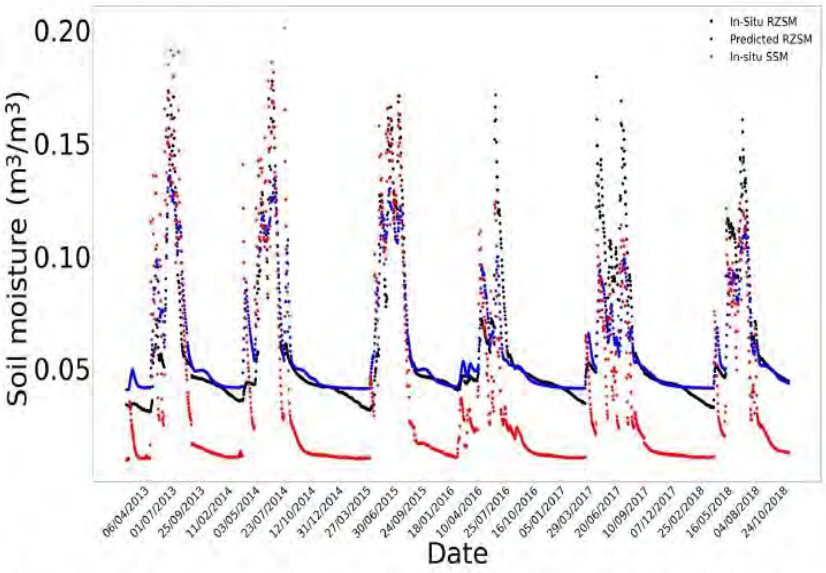


Figure E9. In situ SSM, in situ RZSM, and predicted RZSM series at station Wankama (AMMA-CATCH) with model ANN_SSM.

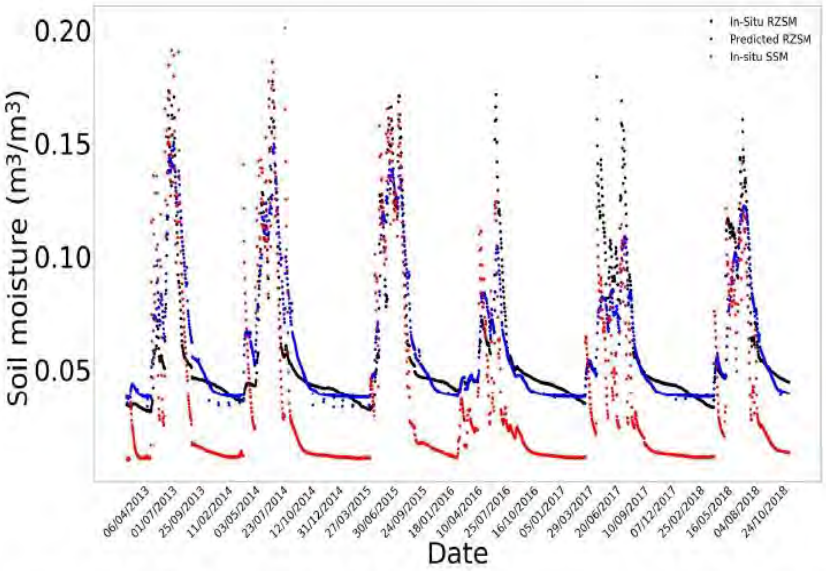


Figure E10. In situ SSM, in situ RZSM, and predicted RZSM series at station Wankama (AMMA-CATCH) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

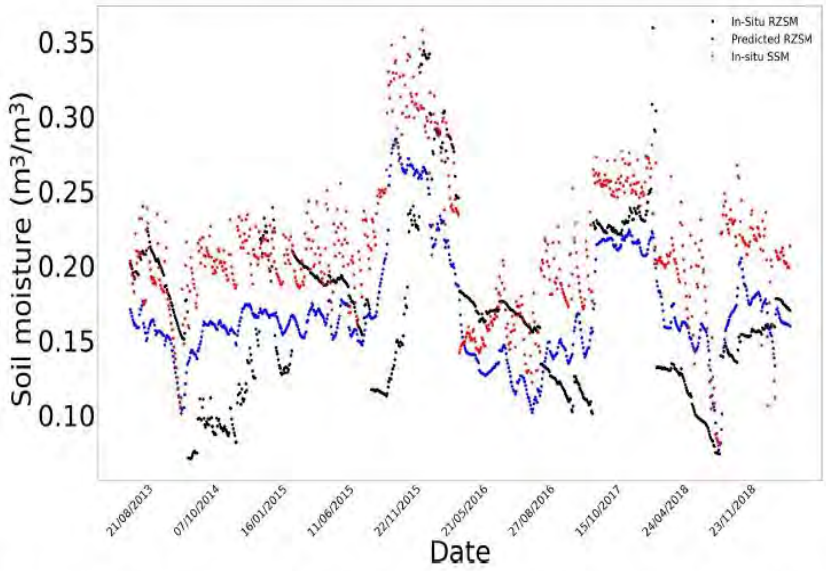


Figure E11. In situ SSM, in situ RZSM, and predicted RZSM series at station 2.04 (HOBE) with model ANN_SSM.

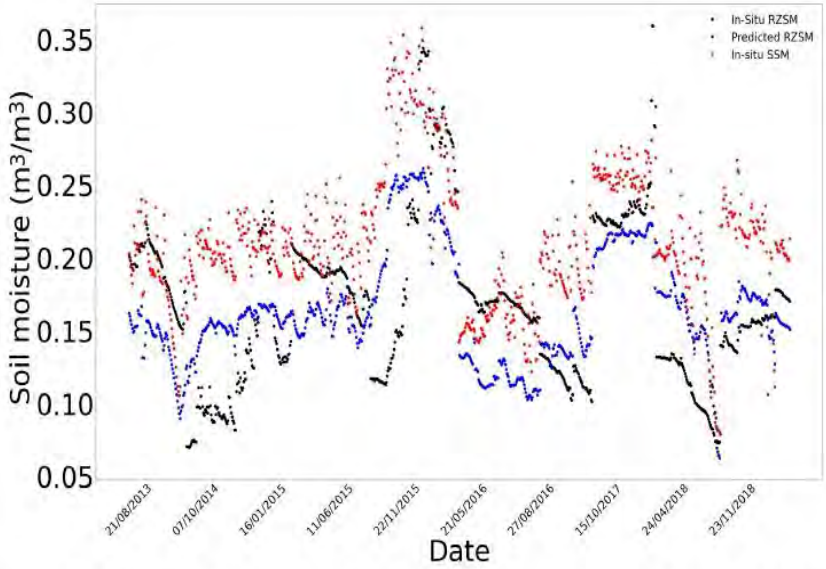


Figure E12. In situ SSM, in situ RZSM, and predicted RZSM series at station 2.04 (HOBE) with model ANN_SSM_ NDVI_EVAP-EFF-B60_EXP-FILT-T5.

Appendix F: Worst-performing examples of model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5

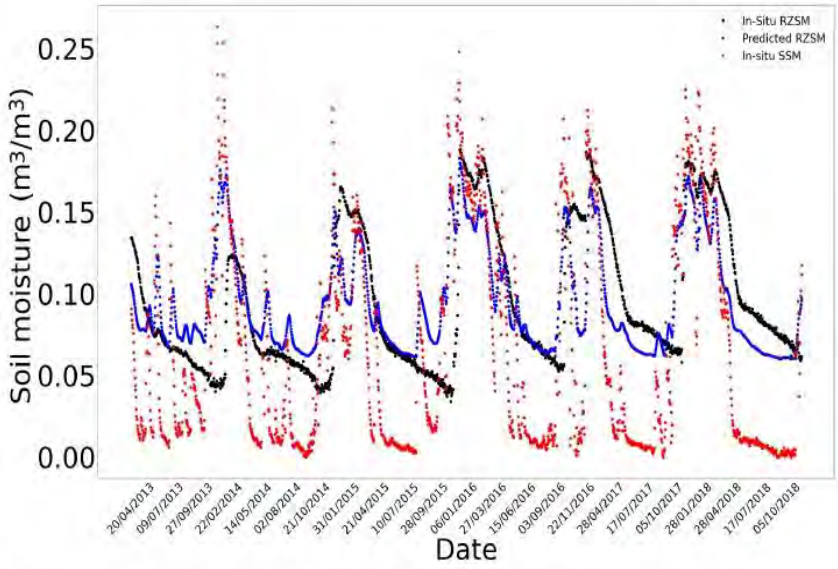


Figure F1. In situ SSM, in situ RZSM, and predicted RZSM series at station Lind#1 with model ANN_SSM.

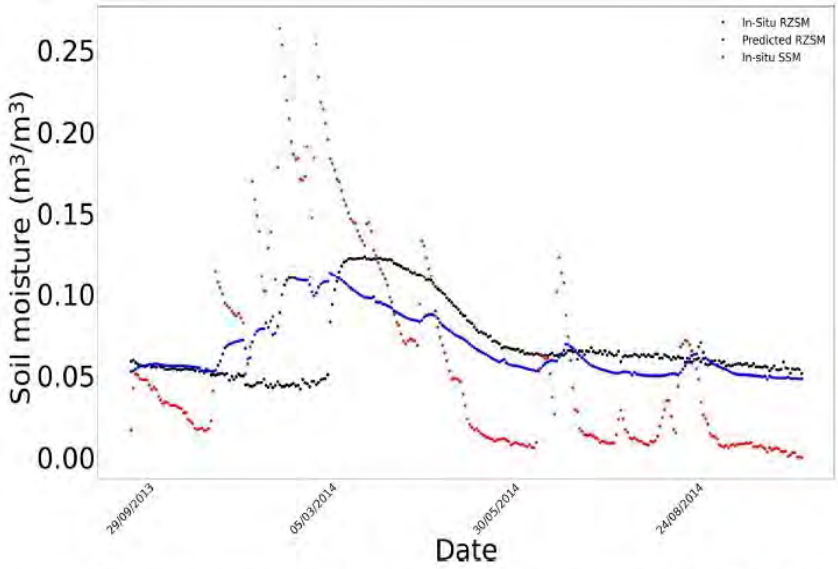


Figure F2. In situ SSM, in situ RZSM, and predicted RZSM series at station Lind#1 with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

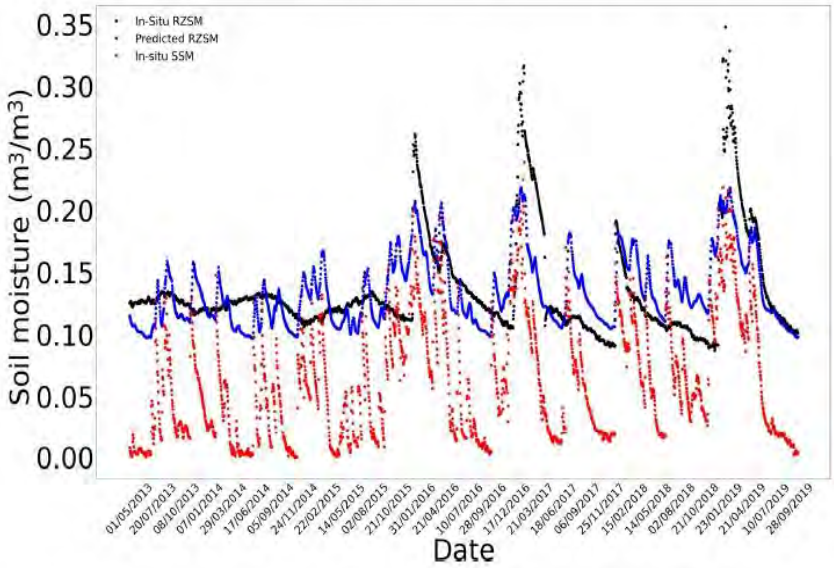


Figure F3. In situ SSM, in situ RZSM, and predicted RZSM series at station PineNut with model ANN_SSM.

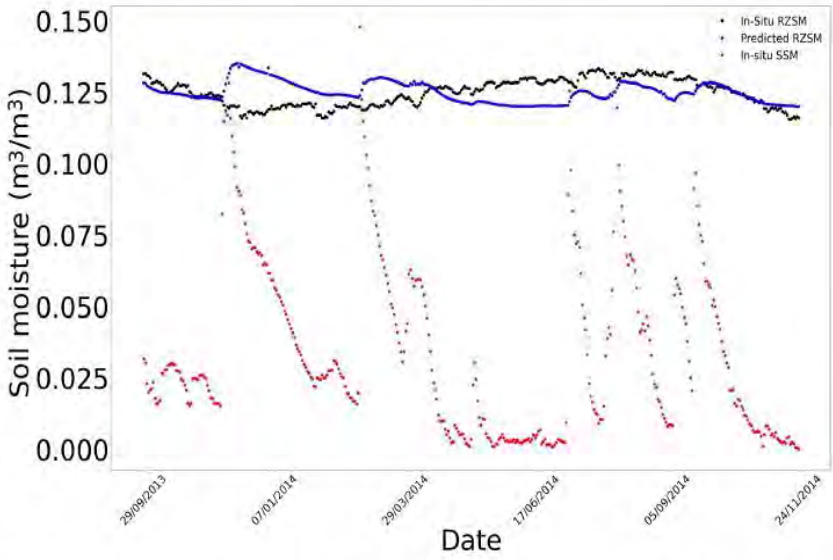


Figure F4. In situ SSM, in situ RZSM, and predicted RZSM time series at the stations in the Kairouan Plain (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5 (see Appendix G for a larger figure format).

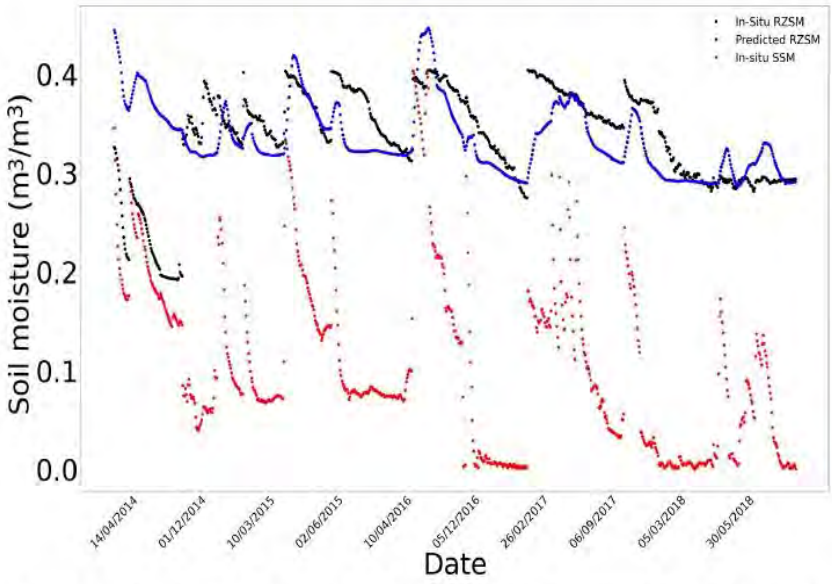


Figure F5. In situ SSM, in situ RZSM, and predicted RZSM series at station S-Coleambally with model ANN_SSM.

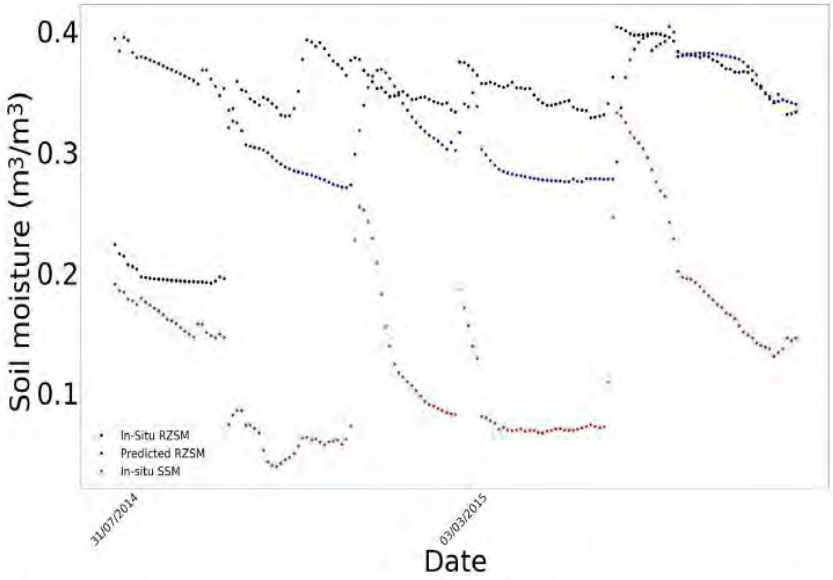


Figure F6. In situ SSM, in situ RZSM, and predicted RZSM series at station S-Coleambally with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

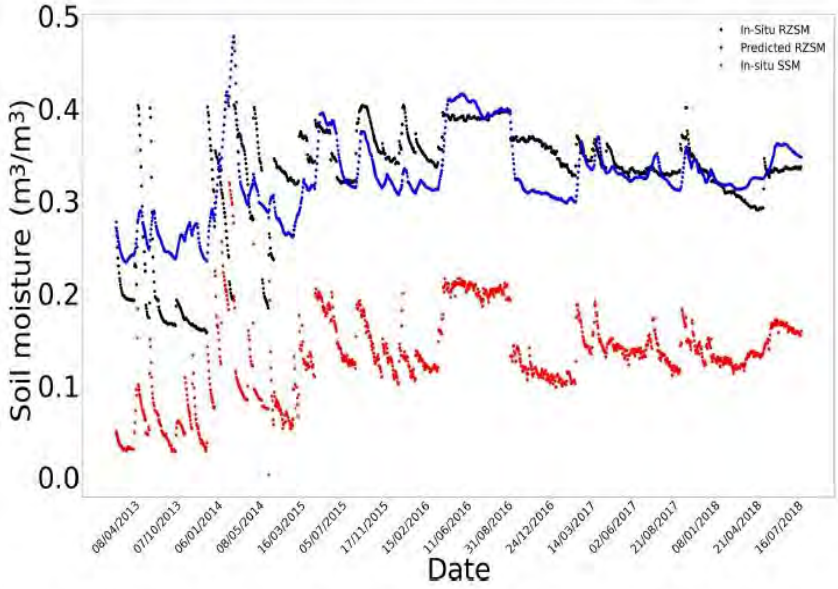


Figure F7. In situ SSM, in situ RZSM, and predicted RZSM series at station Widgiewa with model ANN_SSM.

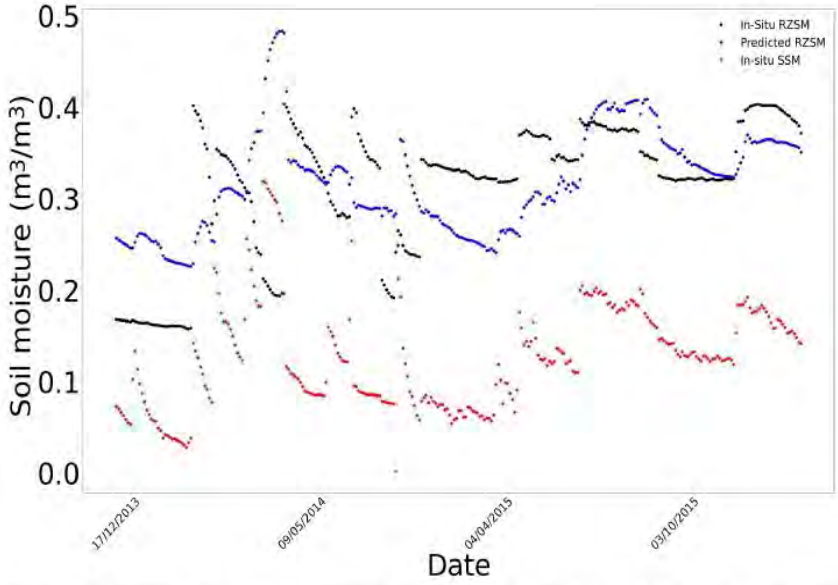


Figure F8. In situ SSM, in situ RZSM, and predicted RZSM series at station Widgiewa with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

Appendix G: Soil moisture time series over Tunisia

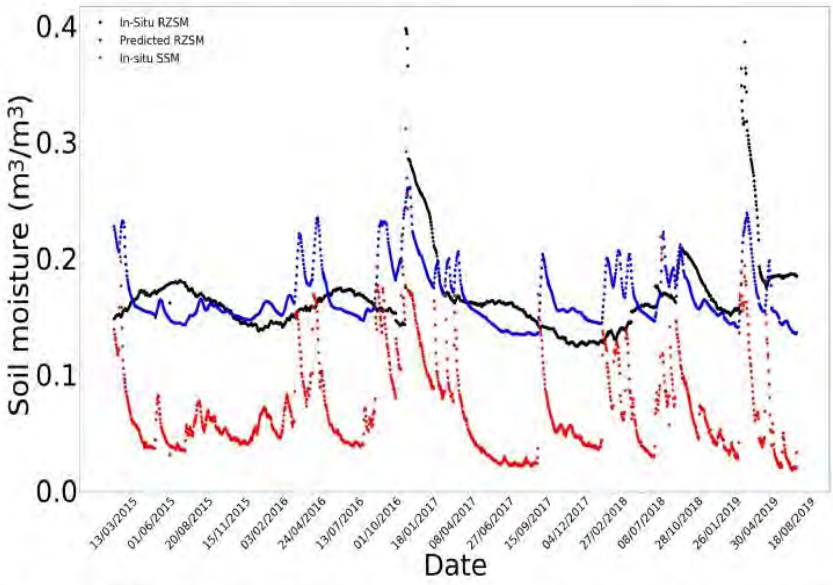


Figure G1. In situ SSM, in situ RZSM, and predicted RZSM series at station Barrage-162 (Tunisia) with model ANN_SSM.

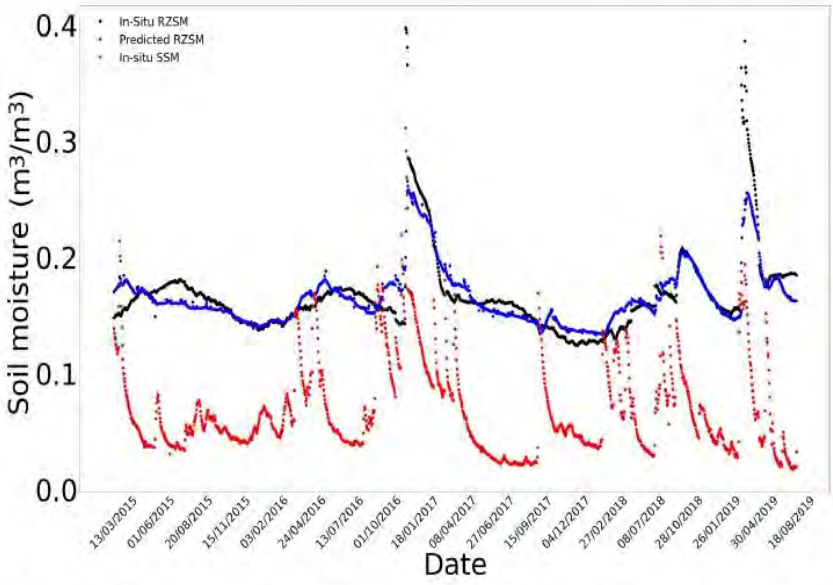


Figure G2. In situ SSM, in situ RZSM, and predicted RZSM series at station Barrage-162 (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

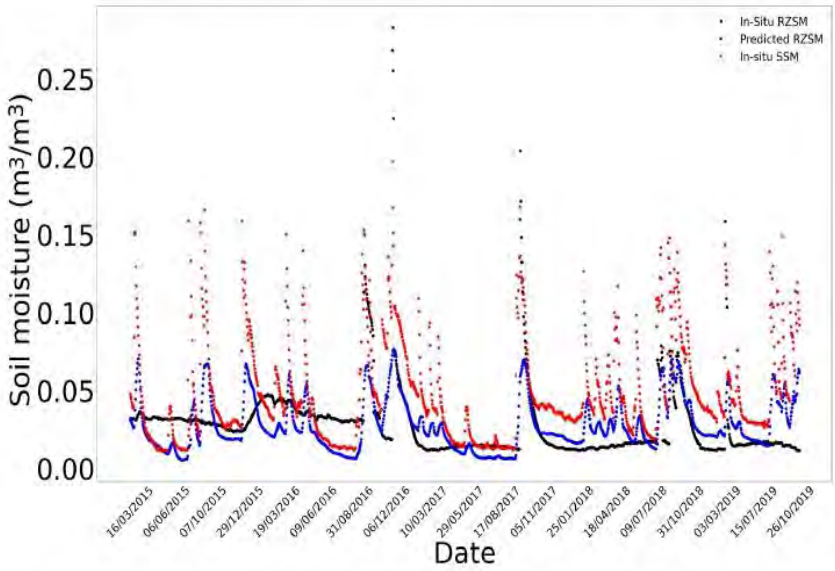


Figure G3. In situ SSM, in situ RZSM, and predicted RZSM series at station Barrouta_160 (Tunisia) with model ANN_SSM.

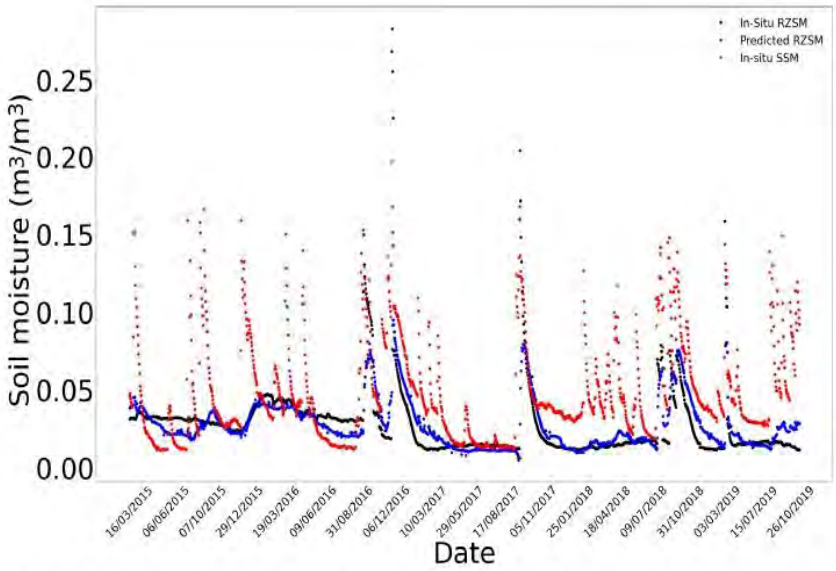


Figure G4. In situ SSM, in situ RZSM, and predicted RZSM series at station Barrouta_160 (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

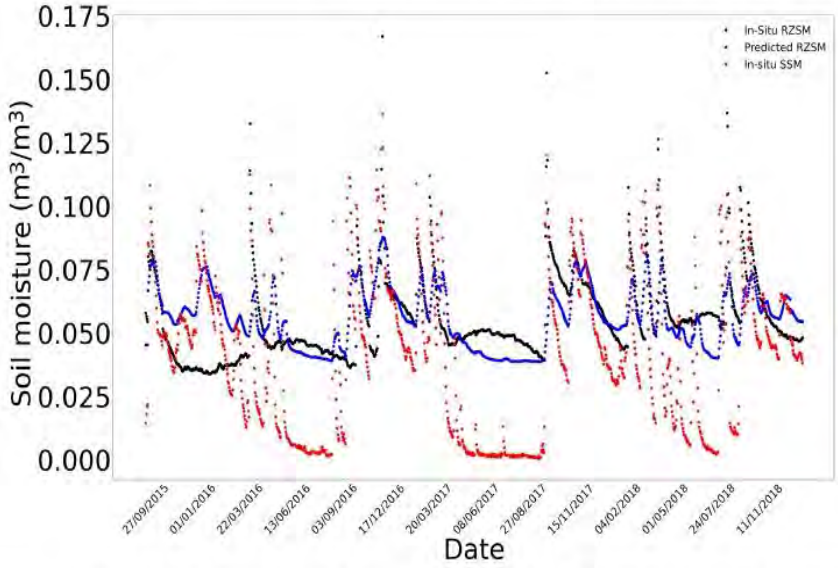


Figure G5. In situ SSM, in situ RZSM, and predicted RZSM series at station Bouhajla_164 (Tunisia) with model ANN_SSM.

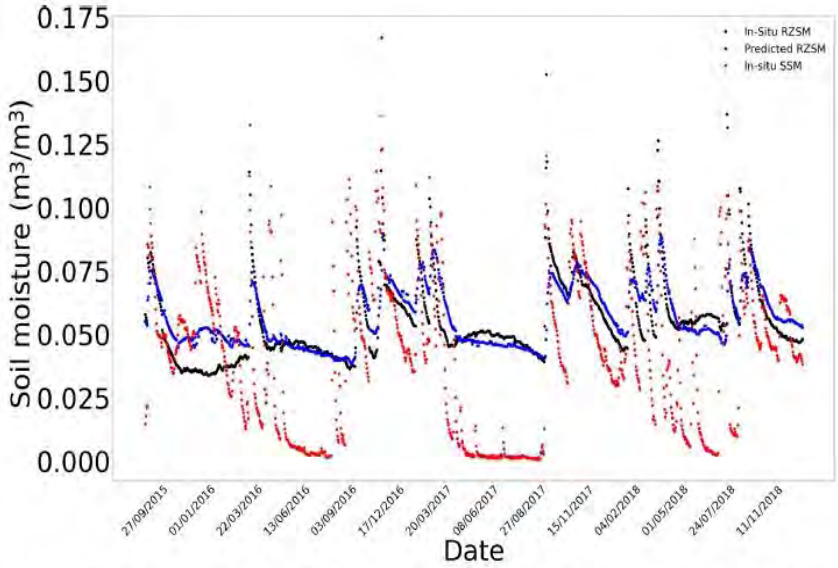


Figure G6. In situ SSM, in situ RZSM, and predicted RZSM series at station Bouhajla_164 (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

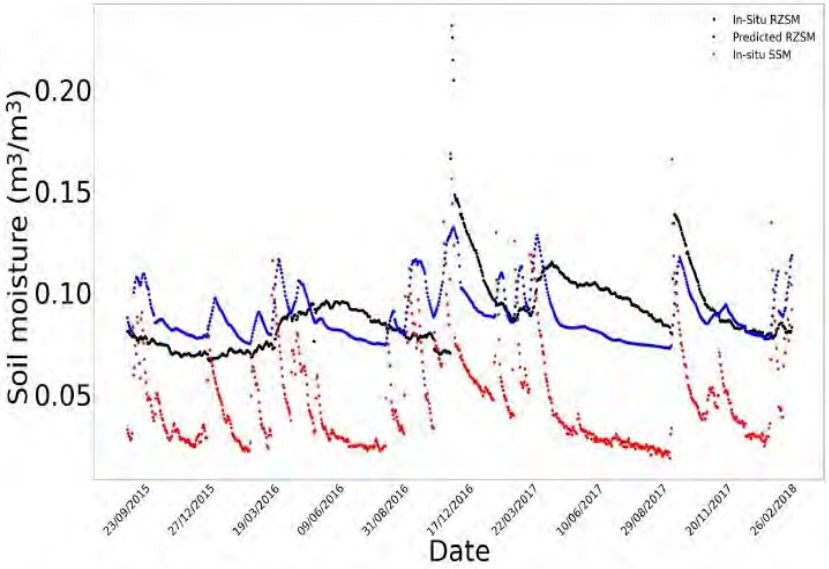


Figure G7. In situ SSM, in situ RZSM, and predicted RZSM series at station Hmidate_163 (Tunisia) with model ANN_SSM.

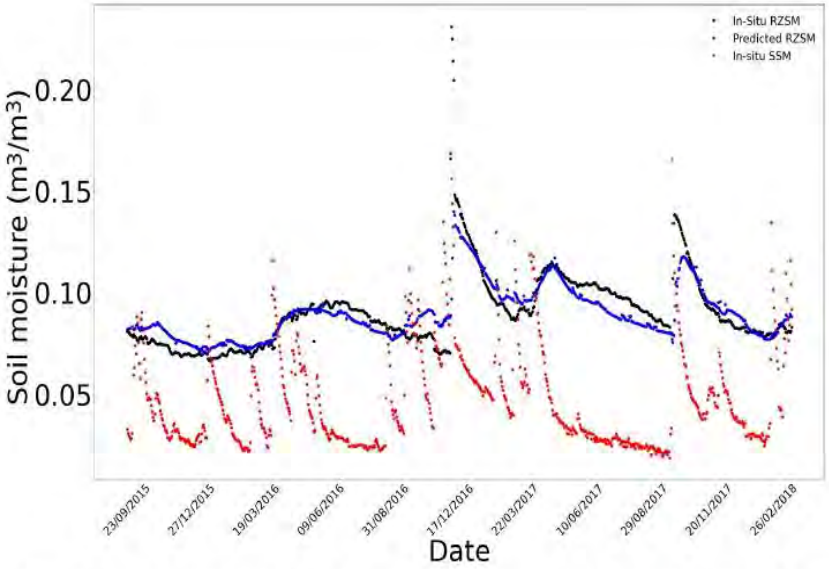


Figure G8. In situ SSM, in situ RZSM, and predicted RZSM series at station Hmidate_163 (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T5.

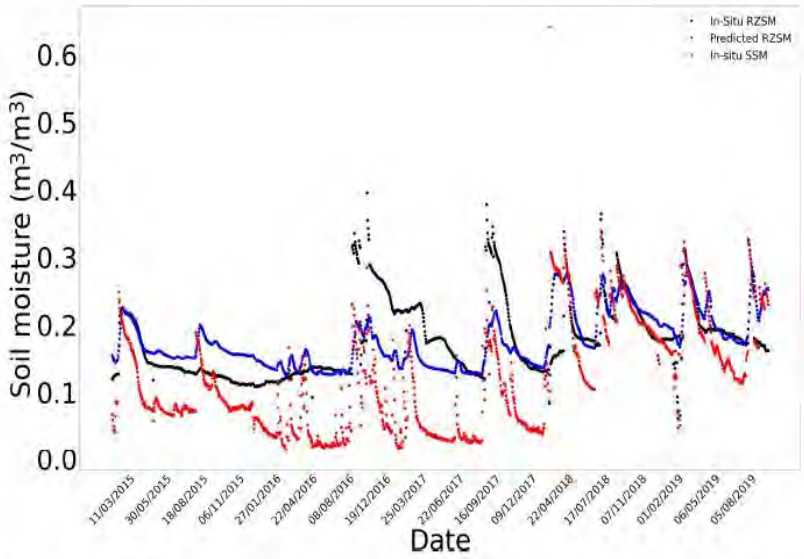


Figure G9. In situ SSM, in situ RZSM, and predicted RZSM series at station P12 (Tunisia) with model ANN_SSM.

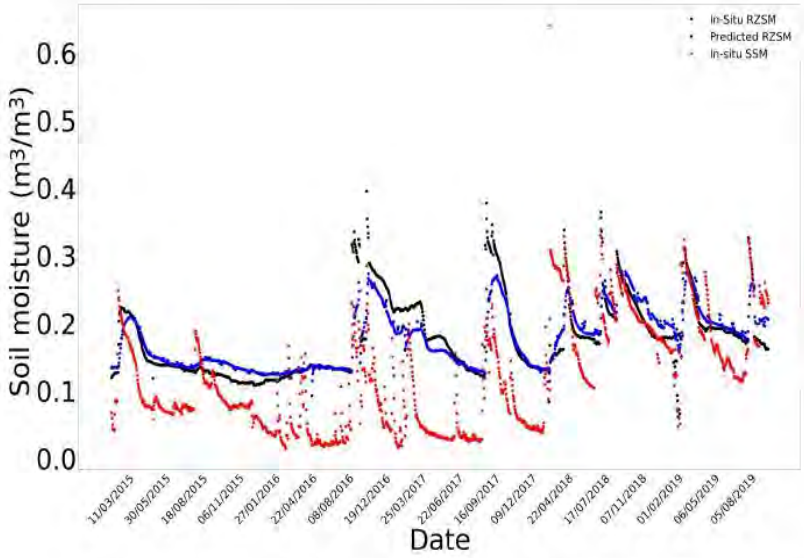


Figure G10. In situ SSM, in situ RZSM, and predicted RZSM series at station P12 (Tunisia) with model ANN_SSM_NDVI_EVAP-EFF-B60_EXP-FILT-T.

Code availability. The code is not publicly accessible because this is in the requirements of some projects.

Data availability. In situ soil moisture data and in situ surface temperature data can be publicly accessed from the International Soil moisture data hosting facility (Dorigo et al.,2021). MODIS NDVI data are available at https://doi.org/10.5067/MODIS/MOD13Q1.006 (Didan, 2015) and MODIS potential evapotranspiration data are available at https://doi.org/10.5067/MODIS/MOD16A3.006 (Running et al., 2017).

Author contributions. The methodology was established by AAB, RS and MZ. The original draft was written by RS. The manuscript was reviewed and edited by all the authors. Soil moisture data over the Italian site and the site description were provided by CC and MM. Soil moisture data over the Indian site and the site description were provided by SM, SKM and DBU. This PhD thesis is under the supervision of MZ and AAB. All the authors have read and agreed to the published version of the manuscript.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The authors were financially supported by the French National Research Agency (ANR) and the French National Space Agency (CNES) for this PhD thesis. The authors also thank the International Soil Moisture Network (ISMN) and supporting networks for providing the soil moisture data. The authors are also grateful for the external soil moisture data provided by the teams of Centre d'Etudes Spatiales de la Biosphère (CESBIO) and Institut National Agronomique de Tunisie (INAT) over Merguellil site in Tunisia, the teams of the Polytechnic University of Milan over Italian site and the teams of the Indian Institute of Science Indian over Indian sites.

Financial support. This research has been supported by the Agence Nationale de la Recherche (grant nos. RET-SIF-ERANETMED–ANR-17-NMED-0004-01 and SMARTIES-PRIMA–ANR-NMED).

Review statement. This paper was edited by Philippe Ackerer and reviewed by two anonymous referees.

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11, 1563–1579, https://doi.org/10.5194/hess-11-1563-2007, 2007.

Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., and Martin, E.: From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations, Hydrol. Earth Syst. Sci., 12, 1323–1337, https://doi.org/10.5194/hess-12-1323-2008, 2008.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology: Artificial Neural Networks in Hydrology, II, Hydrol. Appl., 5, 124–137, https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(124), 2000.

Battude, M., Al Bitar, A., Brut, A., Tallec, T., Huc, M., Cros, J., Weber, J.-J., Lhuissier, L., Simonneaux, V., and Demarez, V.: Modeling water needs and total irrigation depths of maize crop in the south west of France using high spatial and temporal resolution satellite imagery, Agr. Water Manage., 189, 123–136, https://doi.org/10.1016/j.agwat.2017.04.018, 2017.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, Geosci. Model Dev., 4, 677–699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Calvet, J.-C. and Noilhan, J.: From Near-Surface to Root-Zone Soil Moisture Using Year-Round Data, J. Hydrometeorol., 1, 393–411, https://doi.org/10.1175/1525-7541(2000)001<0393:FNSTRZ>2.0.CO;2, 2000.

Carranza, C., Nolet, C., Pezij, M., and van der Ploeg, M.: Root zone soil moisture estimation with Random Forest, J. Hydrol., 593, 125840, https://doi.org/10.1016/j.jhydrol.2020.125840, 2021.

Chen, Y., Song, X., Wang, S., Huang, J., and Mansaray, L. R.: Impacts of spatial heterogeneity on crop area mapping in Canada using MODIS data, ISPRS J. Photogram. Remote Sens., 119, 451–461, https://doi.org/10.1016/j.isprsjprs.2016.07.007, 2016.

Didan, K.: MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V006, NASA EOSDIS Land Processes DAAC [data set], https://doi.org/10.5067/MODIS/MOD13Q1.006, 2015.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675–1698, https://doi.org/10.5194/hess-15-1675-2011, 2011.

Dorigo, W. A., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A. D., Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network, Vadose Zone J., 12, 1–21, https://doi.org/10.2136/vzj2012.0097, 2013.

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Bitelli, M., Blöschl, G., Bogena, H., Brocca, L., Calvet, J.-C., Camarero, J. J., Capello, G., Choi, M., Cosh, M. C., van de Giesen, N., Hajdu, I., Ikonen, J., Jensen, K. H., Kanniah, K. D., de Kat, I., Kirchengast, G., Kumar Rai, P., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J. P., Osenga, E., Palecki, M. A., Pellarin, T., Petropoulos, G. P., Pfeil, I., Powers, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Varlagin, A., Vreugdenhil, M., Walker, J., Wen, J., Wenger, F., Wigneron, J. P., Woods, M., Yang, K., Zeng, Y., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., and Sabia, R.: The International Soil Moisture Network: serving Earth system science for over a decade, Hydrol. Earth Syst. Sci., 25, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021, 2021.

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

Entekhabi, D., Nakamura, H., and Njoku, E. G.: Retrieval of soil moisture profile by combined remote sensing and modeling, in: Retrieval of soil moisture profile by combined remote sens-

ing and modeling, De Gruyter, 485–498, ISBN 9783112319307, 2020.

Grillakis, M. G., Koutroulis, A. G., Alexakis, D. D., Polykretis, C., and Daliakopoulos, I. N.: Regionalizing Root-Zone Soil Moisture Estimates From ESA CCI Soil Water Index Using Machine Learning and Information on Soil, Vegetation, and Climate, Water Resour. Res., 57, e2020WR029249, https://doi.org/10.1029/2020WR029249, 2021.

Hajj, M., Baghdadi, N., Belaud, G., Zribi, M., Cheviron, B., Courault, D., Hagolle, O., and Charron, F.: Irrigated Grassland Monitoring Using a Time Series of TerraSAR-X and COSMO-SkyMed X-Band SAR Data, Remote Sens., 6, 10002–10032, https://doi.org/10.3390/rs61010002, 2014.

Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., and Mckee, M.: Spatial Root Zone Soil Water Content Estimation in Agricultural Lands Using Bayesian-Based Artificial Neural Networks and High- Resolution Visual, NIR, and Thermal Imagery, Irrig. Drain., 66, 273–288, https://doi.org/10.1002/ird.2098, 2017.

Huete, A., Justice, C., and Van Leeuwen, W.: MODIS vegetation index (MOD13), Algorithm Theor. Basis Doc., https://www.researchgate.net/profile/Phillip-Stroud/publication/242230998_A_Recursive_Exponential_Filter_For_Time-Sensitive_Data/links/00b49538f4fa1be826000000/A-Recursive-Exponential-Filter-For-Time-Sensitive-Data.pdf (last access: 27 June 2022), 1999.

Jacquemin, B. and Noilhan, J.: Sensitivity study and validation of a land surface parameterization using the HAPEX-MOBILHY data set, Bound.-Lay. Meteorol., 52, 93–134, https://doi.org/10.1007/BF00123180, 1990.

Karthikeyan, L. and Mishra, A. K.: Multi-layer high-resolution soil moisture estimation using machine learning over the United States, Remote Sens. Environ., 266, 112706, https://doi.org/10.1016/j.rse.2021.112706, 2021.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S. E., Drinkwater, M. R., Hahne, A., Martín-Neira, M., and Mecklenburg, S.: The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle, Proc. IEEE, 98, 666–687, https://doi.org/10.1109/JPROC.2010.2043032, 2010.

Kolassa, J., Reichle, R. H., Liu, Q., Alemohammad, S. H., Gentine, P., Aida, K., Asanuma, J., Bircher, S., Caldwell, T., Colliander, A., Cosh, M., Holifield Collins, C., Jackson, T. J., Martínez-Fernández, J., McNairn, H., Pacheco, A., Thibeault, M., and Walker, J. P.: Estimating surface soil moisture from SMAP observations using a Neural Network technique, Remote Sens. Environ., 204, 43–59, https://doi.org/10.1016/j.rse.2017.10.045, 2018.

Kornelsen, K. C. and Coulibaly, P.: Root-zone soil moisture estimation using data-driven methods, Water Resour. Res., 50, 2946–2962, https://doi.org/10.1002/2013WR014127, 2014.

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C. T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C.-H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., and Yamada, T.: Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305, 1138–1140, https://doi.org/10.1126/science.1100217, 2004.

Lee, T. J. and Pielke, R. A.: Estimating the Soil Surface Specific Humidity, J. Appl. Meteorol. Clim., 31, 480–484, https://doi.org/10.1175/1520-0450(1992)031<0480:ETSSSH>2.0.CO;2, 1992.

Liu, Y., Chen, D., Mouatadid, S., Lu, X., Chen, M., Cheng, Y., Xie, Z., Jia, B., Wu, H., and Gentine, P.: Development of a Daily Multilayer Cropland Soil Moisture Dataset for China Using Machine Learning and Application to Cropping Patterns, J. Hydrometeorol., 22, 445–461, https://doi.org/10.1175/JHM-D-19-0301.1, 2021.

Martínez-Espinosa, C., Sauvage, S., Al Bitar, A., Green, P. A., Vörösmarty, C. J., and Sánchez-Pérez, J. M.: Denitrification in wetlands: A review towards a quantification at global scale, Sci. Total Environ., 754, 142398, https://doi.org/10.1016/j.scitotenv.2020.142398, 2021.

Masseroni, D., Corbari, C., and Mancini, M.: Validation of theoretical footprint models using experimental measurements of turbulent fluxes over maize fields in Po Valley, Environ. Earth Sci., 72, 1213–1225, https://doi.org/10.1007/s12665-013-3040-5, 2014.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A.: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes, Geosci. Model Dev., 6, 929–960, https://doi.org/10.5194/gmd-6-929-2013, 2013.

Merlin, O., Bitar, A. A., Rivalland, V., Béziat, P., Ceschia, E., and Dedieu, G.: An Analytical Model of Evaporation Efficiency for Unsaturated Soil Surfaces with an Arbitrary Thickness, J. Appl. Meteorol. Clim., 50, 457–471, https://doi.org/10.1175/2010JAMC2418.1, 2010.

Noilhan, J. and Mahfouf, J.-F.: The ISBA land surface parameterisation scheme, Global Planet. Change, 13, 145–159, https://doi.org/10.1016/0921-8181(95)00043-7, 1996.

Noilhan, J. and Planton, S.: A Simple Parameterization of Land Surface Processes for Meteorological Models, Mon. Weather Rev., 117, 536–549, https://doi.org/10.1175/1520-0493(1989)117<0536:ASPOLS>2.0.CO;2, 1989.

Oleson, W., Lawrence, M., Bonan, B., Flanner, G., Kluzek, E., Lawrence, J., Levis, S., Swenson, C., Thornton, E., Dai, A., Decker, M., Dickinson, R., Feddema, J., Heald, L., Hoffman, F., Lamarque, J.-F., Mahowald, N., Niu, G.-Y., Qian, T., Randerson, J., Running, S., Sakaguchi, K., Slater, A., Stockli, R., Wang, A., Yang, Z.-L., Zeng, X., and Zeng, X.: Technical Description of version 4.0 of the Community Land Model (CLM), NCAR/U-CAR, https://doi.org/10.5065/D6FB50WZ, 2010.

Owe, M., de Jeu, R., and Holmes, T.: Multisensor historical climatology of satellite-derived global land surface moisture, J. Geophys. Res., 113, F01002, https://doi.org/10.1029/2007JF000769, 2008.

Oyebode, O. and Stretch, D.: Neural network modeling of hydrological systems: A review of implementation techniques, Nat. Resour. Model., 32, e12189, https://doi.org/10.1111/nrm.12189, 2019.

Pan, X., Kornelsen, K. C., and Coulibaly, P.: Estimating Root Zone Soil Moisture at Continental Scale Using Neural Networks, J. Am. Water Resour. Assoc., 53, 220–237, https://doi.org/10.1111/1752-1688.12491, 2017.

Paris Anguela, T., Zribi, M., Hasenauer, S., Habets, F., and Loumagne, C.: Analysis of surface and root-zone soil moisture dynamics with ERS scatterometer and the hydrometeorological model SAFRAN-ISBA-MODCOU at Grand Morin watershed (France), Hydrol. Earth Syst. Sci., 12, 1415–1424, https://doi.org/10.5194/hess-12-1415-2008, 2008.

Paulik, C., Dorigo, W., Wagner, W., and Kidd, R.: Validation of the ASCAT Soil Water Index using in situ data from the International Soil Moisture Network, Int. J. Appl. Earth Obs. Geoinform., 30, 1–8, https://doi.org/10.1016/j.jag.2014.01.007, 2014.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E.: AquaCrop – The FAO Crop Model to Simulate Yield Response to Water: II. Main Algorithms and Software Description, Agron. J., 101, 438–447, https://doi.org/10.2134/agronj2008.0140s, 2009.

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., and Ettaouil, M.: Multilayer Perceptron: Architecture Optimization and Training, Int. J. Interact. Multimed. Artific. Intel., 4, 26–30, https://doi.org/10.9781/ijimai.2016.415, 2016.

Running, S., Mu, Q., and Zhao, M.: MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500 m SIN Grid V006.2017, NASA EOSDIS Land Processes DAAC [data set], https://doi.org/10.5067/MODIS/MOD16A2.006, 2017.

Sabater, J. M., Jarlan, L., Calvet, J.-C., Bouyssel, F., and De Rosnay, P.: From Near-Surface to Root-Zone Soil Moisture Using Different Assimilation Techniques, J. Hydrometeorol., 8, 194–206, https://doi.org/10.1175/JHM571.1, 2007.

SIE: SIE portal (Système d'Information Environnemental), https://sie.cesbio.omp.eu/, last access: 8 December 2021.

Souissi, R., Al Bitar, A., and Zribi, M.: Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe, Water, 12, 3109, https://doi.org/10.3390/w12113109, 2020.

Stroud, P. D.: A Recursive Exponential Filter For Time-Sensitive Data, Los Alamos national Laboratory, LAUR-99-5573, https://modis.gsfc.nasa.gov/data/atbd/atbd_mod13.pdf (last access: January 2022), 1999.

Tanty, R., Desmukh, T. S., and Bhopal, M.: Application of Artificial Neural Network in Hydrology – A Review, Int. J. Eng. Tech. Res., 4, 184–188, https://doi.org/10.17577/IJERTV4IS060247, 2015.

Wagner, W., Lemoine, G., and Rott, H.: A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data, Remote Sens. Environ., 70, 191–207, https://doi.org/10.1016/S0034-4257(99)00036-X, 1999.

Wagner, W., Blöschl, G., Pampaloni, P., Calvet, J.-C., Bizzarri, B., Wigneron, J.-P., and Kerr, Y.: Operational readiness of microwave remote sensing of soil moisture for hydrologic applications, Hydrol. Res., 38, 1–20, https://doi.org/10.2166/nh.2007.029, 2007.

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., de Rosnay, P., Jann, A., Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang, Y., Blöschl, G., Eitzinger, J., and Steinnocher, K.: The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications, Meteorol. Z., 22, 5–33, https://doi.org/10.1127/0941-2948/2013/0399, 2013.

Zribi, M., Chahbi, A., Shabou, M., Lili-Chabaane, Z., Duchemin, B., Baghdadi, N., Amri, R., and Chehbouni, A.: Soil surface moisture estimation over a semi-arid region using ENVISAT ASAR radar data for soil evaporation evaluation, Hydrol. Earth Syst. Sci., 15, 345–358, https://doi.org/10.5194/hess-15-345-2011, 2011.

Zribi, M., Foucras, M., Baghdadi, N., Demarty, J., and Muddu, S.: A New Reflectivity Index for the Retrieval of Surface Soil Moisture From Radar Data, IEEE J. Select. Top. Appl. Earth Obs. Remote Sens., 14, 818–826, https://doi.org/10.1109/JSTARS.2020.3033132, 2021.

Chapter 4: RZSM spatial maps at large scale and 1km resolution

This chapter is the final step of this PhD and encompasses the work presented in the following paper:

Souissi, R., Al Bitar, A., Corbari, C., Mancini, M., and Zribi, M.: Root-Zone soil moisture over Continental Europe using machine learning, *Submitted to International Journal of Applied Earth Observation and Geoinformation,* 2022.

4.1 Introduction

RZSM has been extensively studied at local or regional scales. However, few attempts were made to map RZSM at large scales. The approaches which are currently suggested to answer the need of large-scale RZSM are hindered by the coarse spatial resolution at which RZSM is predicted. In order to address the needs of some applications namely agriculture, a high spatial resolution of at least 1 km is required for reliable estimates of RZSM. This chapter sets out a methodology to map RZSM at 1 km spatial resolution over Continental Europe which is a large area with disparate climate and soil patterns. The work is based on the model developed in the previous chapter which includes SSM, NDVI, SWI and evaporation efficiency. No training was done in this step; the model was run only in prediction mode. The objective is to assess the feasibility of spatially generalizing a locally-trained ANN model.

As a first step, high resolution RZSM maps were produced using for the SSM features the Copernicus product SSM1km which is based on C-band Sentinel-1 data to generate the three SSM features required by the ANN model. The produced maps were validated against the ERA5-Land reanalysis RZSM product and against in-situ measurements which are provided by ISMN over four European soil moisture networks. In order to further assess the quality of the produced maps, additional RZSM maps were derived from other SSM products of coarser spatial resolutions. Daily RZSM maps were generated at 36 km spatial resolution using the SMAP level 3 SSM product (SMAP_L3_SM_P). Similarly, these maps underwent a large-scale and local validations against the ERA5-Land RZSM product and in-situ data, respectively. Finally, daily RZSM maps were also generated by our ANN model at 9 km spatial resolution using the ERA5-Land reanalysis SSM product and validated like the previous maps.

4.2 Conclusion

One major result of this work consists in the major impact of the quality and the temporal resolution of the used SSM inputs has on the quality of the RZSM predictions. Maps at high, intermediate and coarse spatial resolutions were produced using the ANN model developed in the previous chapter. This model encompasses SSM, NDVI, SWI and evaporation efficiency features. A large-scale validation was conducted against the reanalysis RZSM product provided by ERA5-Land. A local comparison was also done against in-situ RZSM data provided by ISMN over four European soil moisture networks. The produced RZSM maps at 1 km resolution were consistent with validation data over many areas. This was illustrated in the seasonal correlation maps comparing the predictions to the ERA5-land reanalysis product and also by correlation and RMSE boxplots which compare the predictions to in-situ data. However, the C-band SSM information which is represented by the SSM1km product, was proved inadequate over complex sceneries such as forests, irrigated areas, areas with freeze/thaw events, etc. This finding is in accordance with previous studies which investigated the quality of this product and found the same limitations. Despite the drawbacks, this product has the advantage of a high spatial resolution which is necessary of agricultural applications.

When it comes to coarse spatial resolutions, The SSM product provided by the SMAP mission (SMAP_L3_SM_P) yielded slightly more accurate predictions. The seasonal correlation maps were enhanced mainly over agricultural areas. However, RZSM prediction maps produced at both 1 km and 36 km spatial resolutions were less accurate then the ERA5-Land reanalysis RZSM product when compared against in-situ data. This finding has led us to test the ANN model with the ERA5-land reanalysis SSM product at 9 km spatial resolution. Overall, the predictions improved compared to the previous ANN model outputs. This can be seen through the enhanced correlation and RMSE values. Most interestingly, the prediction accuracy of the maps at 9 km spatial resolution was better than that of the ERA5-land RZSM product namely over network SMOSMANIA in France. Despite the limitations detected with the high resolution RZSM predictions and the coarse resolution hampering the rest of the maps, the results confirm that a data-driven approach like ANN can be very reliable for RZSM estimation. The use of better quality SSM source products is a promising option for enhanced RZSM estimates.

4.3 Article

Root-Zone soil moisture over Continental Europe using machine learning

Roiya Souissi[1,*], Ahmad Al Bitar[1,*], Chiara Corbari[2], Marco Mancini[2], Mehrez Zribi[1]

[1] CESBIO—Centre d'Etudes Spatiales de la Biosphère, Université de Toulouse, CNES/CNRS/INRAE/IRD/UPS

[2] Department of Civil and Environmental Engineering (DICA), Polytechnic University of Milan, 20133 Milano, Italy

[*] *Correspondence to*: Roiya Souissi (roiya.souissi@univ-tlse3.fr), Ahmad Al Bitar (ahmad.albitar@gmx.com)

Abstract

Root zone soil moisture (RZSM) is a land variable of great importance for different applications. No current remote sensing technique can directly retrieve the root zone component but many efforts were deployed to map RZSM at large scales. However, the available products provide RZSM at coarse spatial resolutions and thus are not adapted for agricultural applications which require at least sub-kilometric resolutions. The main objective of this study is to produce spatial RZSM maps at 1km resolution over continental Europe using an artificial neural network (ANN) model which is based on surface soil moisture (SSM) and process-related variables. Daily RZSM maps were produced at subkilometric spatial resolution such that the SSM features were computed using the Copernicus Surface Soil Moisture 1km Version 1 product (SSM1km). Besides, evaporation efficiency was included in the model in order to account for the evaporation process. Soil water index (SWI) was computed by a recursive exponential filter using SSM information to quantify infiltration. Remote sensing Normalized difference vegetation index (NDVI) was used to characterize the plant growth. The ANN model is run only in prediction mode, i.e. with no prior calibrations. The training has been already conducted in a previous study such that the training dataset included in-situ SSM information provided by the International soil Moisture Network (ISMN) and remote sensing-based features over different areas of the world. The quality of the produced RZSM maps at 1km ($RZSM_{ANN\_SSM1km}$) was assessed at large scale through a comparison against the ERA5-land reanalysis RZSM product ($RZSM_{ERA5}$) and at a local scale through a comparison against in-situ measurements provided by ISMN. The performance of the model was acceptable over many areas but was hindered by complex contexts heavily impact the C-band SSM product. The same ANN model was also used to map RZSM at coarser resolutions. Additional RZSM maps at 36km ($RZSM_{ANN\_SMAP36km}$) and 9km ($RZSM_{ANN\_ERA5\_9km}$) spatial resolutions were generated using the SMAP_L3_SM_P product provided by The Soil Moisture Active Passive (SMAP) mission and the ERA5-Land reanalysis SSM product, respectively. Seasonal correlation maps between $RZSM_{ANN\_SMAP36km}$ and $RZSM_{ERA5}$ clearly show an enhancement compared to those between $RZSM_{ANN\_SSM1km}$ and $RZSM_{ERA5}$. Besides, the $RZSM_{ANN\_ERA5\_9km}$ product outperformed $RZSM_{ERA5}$ over some areas mainly over network SMOSMANIA in France. The added value and limitations of the model were discussed with regards to the different challenges of the temporal availability of the SSM1km product, the land cover, the freeze/thaw conditions, etc. Better SSM inputs could improve the predictions and thus allow global mapping of RZSM.

Keywords: Root-zone soil moisture, Artificial Neural Networks, Subkilometric resolution, Sentinel-1, SMAP, ERA5-land, ISMN.

1 Introduction

Large-scale soil moisture mapping is necessary to better manage the water resources allocated for hydrology, meteorology and agriculture mainly in the current drought and climate change contexts (Berg and Sheffield, 2018). To mention but a few, soil moisture retrieval improves the assessment of available plant water and minimizes the risks of environmentally damaging human activities. Moreover, skillful quantification of root-zone soil moisture (RZSM) not only helps detect plant water stress and thus drought events, but also improves weather forecasting and climate projections through enhanced latent heat fluxes (Dirmeyer et al., 2006).

Remote sensing techniques have been proved efficient in retrieving surface soil moisture at high spatial and temporal resolutions. Several satellite missions provide surface soil moisture (SSM) estimations at different spatial resolutions. For instance, active and passive microwave sensors provide this information at regional to global scales (Paloscia et al. 2001). Coarse resolution (25–50 km) soil moisture estimates are useful in support of numerical weather prediction, climate monitoring and flood forecasting (Lopez et al. 2020). They can be retrieved by the Soil Moisture Active Passive (SMAP) mission (Entekhabi et al., 2010) at 36km resolution (level 3) and 9 km (level 3 enhanced). Besides, the Soil Moisture and Ocean Salinity (SMOS) mission (Kerr et al., 2001) provides level 3 SSM estimates at 25km spatial resolution (Al Bitar et al., 2017). Synthetic Aperture Radar (SAR) instruments, such as the C-band in Sentinel-1, provide SSM at high spatial and temporal resolutions (Zribi et al., 2011). In this context, different algorithms based on machine learning or change detection techniques have been developed to allow soil moisture mapping at field scale or 1 km spatial resolution (Tomer et al. 2015, Tomer et al. 2016, El-Hajj et al., 2017, Bauer-Marschallinger et al., 2019, Nativel et al., 2022).

However, no current remote sensing technique can directly retrieve RZSM since the soil representative depth impacting satellite instruments does not go beyond a few centimeters except for dry sandy soils. Actually, the soil moisture retrieval depth is approximately equal to 5cm at L-band. However, P-band may provide RZSM in the future since a widely-held view is that this moisture retrieval depth increases with wavelength. Accordingly, P-band (~40-cm wavelength/0.75 GHz) is under investigation for soil moisture observation over deeper layers of soil (Shen et al., 2021). RZSM can be also estimated through proxy information namely vegetation water stress. Vegetation canopy temperature and evaporative fraction (EF) have been used to estimate soil moisture (Hain et al., 2009) but these methods require surface flux data in addition to micrometeorological data, where accurate surface flux data is not explicitly available (Akuraju et al., 2021).

RZSM information can be directly collected from in-situ sensors that are installed either vertically or horizontally in the soil (Dobriyal et al., 2012). The International Soil Moisture Network (ISMN) encompasses comprehensive surface and root-zone soil moisture databases provided by operational soil moisture networks worldwide (Dorigo et al., 2011). Since direct in-situ measurements are not always available, different analytical methods can be applied to estimate RZSM measurements. These methods are based on theoretical or empirical relations between environmental variables controlling RZSM state (Carranza et al., 2021). Land surface models (LSM) such as Interaction Sol-Biosphère-Atmosphère (ISBA) (Noilhan and Mahfouf, 1996) and the Community Land Model (CLM; Oleson et al., 2010) are also used for estimating root zone soil moisture. Other methodologies include water budgets in crop models such as Aquacrop (Raes et al., 2009) that was successfully used to estimate RZSM. Such models are often coupled with data assimilation techniques (Kumar et al., 2009, Lievens et al. 2016, Reichle et al., 2019) since these models are affected by the accuracy of ancillary information.

Data-driven methods that include Machine Learning (ML) techniques have been widely used in soil hydrology in the last couple of decades and more particularly in deriving RZSM from surface information.

The different ML algorithms build mathematical models based on training sets and covariates to extract information from data. Furthermore, they are tuned to handle diverse and large volumes of datasets, which may be relevant for large scale studies or for operational water management (Carranza et al., 2021). For instance, artificial neural networks have been applied to predict RZSM (Kornelsen and Coulibaly, 2014; Pan et al., 2017; Souissi et al., 2020; Souissi et al., 2022). Some studies have demonstrated that ANNs can achieve good RZSM estimates at local scales using surface measurements (Elshorbagy et al., 2010). While analytical solutions entail some assumptions to the physical model, data-driven approaches tend to construct a relation between inputs and outputs. Kornelsen and Coulibaly (2014) trained different surrogate ANN models with the data of different soil moisture profiles generated by HYDRUS-1D model over the lower Great Lakes region. They found that the ANNs were able to well represent the soil moisture dynamics of the independent testing sites from the same region, when the HYDRUS-1D estimates were close to the observations. The ability of ANN models to accurately predict RZSM over large areas and at fine resolutions is still understudied. Pan et al. (2017) used ANN models to generate RZSM in North America using SMOS level 3 soil moisture data (nominal resolution of 43 km, which correspond to 86% of the signal (Al Bitar et al., 2012)), and achieved a spatial soil moisture pattern comparable to that of Global Land Data Assimilation System Noah model with comparable performance to the SMOS surface soil moisture retrievals. Souissi et al. (2022) developed several ANN models to estimate RZSM based either solely on in situ SSM information or on a group of process-related features in addition to SSM namely the soil water index computed with a recursive exponential filter, evaporation efficiency and NDVI. Different regions across the globe with distinct land cover and climate patterns were considered. Overall, good agreement between in-situ RZSM and predictions was recorded.

The main objective of this study is to assess the utility of spatially generalizing a locally-trained ANN model. It aims at assessing the feasibility of producing spatially-coherent RZSM maps based on local training datasets. To do so, we propose a method to map RZSM at 1 kilometer resolution over a large area such as Continental Europe using an ANN model whose features are SSM backward rolling averages over 10, 30 and 90 days computed using the Copernicus Surface Soil Moisture 1km Version 1 product (SSM1km), remote sensing-based evaporation efficiency, SWI computed using a recursive exponential filter and NDVI datasets from MODIS. This ANN model was previously developed and trained using in-situ SSM information from ISMN and remote sensing-based features over different areas of the world in (Souissi et al., 2022). (1) We produced RZSM maps at 1km resolution, hereafter called $RZSM_{ANN\_SSM1km}$, that we validated at the European scale through a comparison against ERA5-land RZSM reanalysis datasets ($RZSM_{ERA5}$), and locally through a comparison against in-situ measurements provided by ISMN over four European soil moisture networks. (2) We assessed the impact of using multi-source input SSM information (radar, microwave and reanalysis) on the quality of the predictions. RZSM maps at different spatial resolutions, depending on the SSM source, were produced to highlight the advantages and limitations of each source.

2 Materials and Methods

2.1 Datasets

2.1.1 Kilometric resolution SSM data

The Copernicus Surface Soil Moisture 1km Version 1 product (SSM1km) was used to compute the SSM features of the ANN model in order to generate spatial maps of RZSM at 1km spatial resolution. The SSM1km product is obtained from Sentinel-1 radar backscatter images (level-1 data), acquired in Interferometric Wade Swath (IW) mode and VV-polarization and jointly provided by the European Space Agency (ESA) and the European Commission (EC). Daily images at 1km of relative surface soil moisture (in percent saturation) were created from this raw satellite data (figure 1(a)).

The SSM1km retrieval algorithm is based on the TU Wien Change Detection Model (Bauer-Marschallinger et al., 2017) which derives SSM from the backscatter coefficient σ^0. This model interprets changes in backscatter as changes in soil moisture. Other surface properties such as the soil roughness and the vegetation structure are considered as static parameters. To estimate SSM, the actual backscatter value σ^0 (θ, t) at time t and observation angle θ is normalized to a reference angle Θ and linearly scaled between dry and wet reference values corresponding to minimum and maximum backscattering coefficients, yielding relative surface soil water saturation SSM(t) (Bauer-Marschallinger et al., 2019).

Copernicus SSM1km is available for the European continent every 2-4 days, with the inclusion the Sentinel-1B data starting in October 2016, depending on the individual location in relation to the non-uniform coverage pattern, depending on the individual location (Bauer-Marschallinger et al., 2017). However, the Sentinel1-B satellite sensor is unavailable since 23 December 2021 which has led to a significant degradation (of about 50%) to the daily observational coverage. The SSM data are available in the Copernicus Global Land Service (CGLS) data portal. The CGLS provides a large number of harmonized and co-formatted bio-geophysical products in Near-Real-Time (NRT) at global scale (Bauer-Marschallinger et al., 2017).

Pixel-wise backward rolling averages, over 10, 30 and 90 days of the SSM1km datasets were computed and transformed into spatial maps, since the ANN model requires three SSM features corresponding to these rolling averages. For a given date and pixel, the SSM_RAV10d is equal to the mean of the available SSM values over the last 10 days (no-data values are ignored). The same goes for SSM_RAV30d and SSM_RAV90d (figure 1).

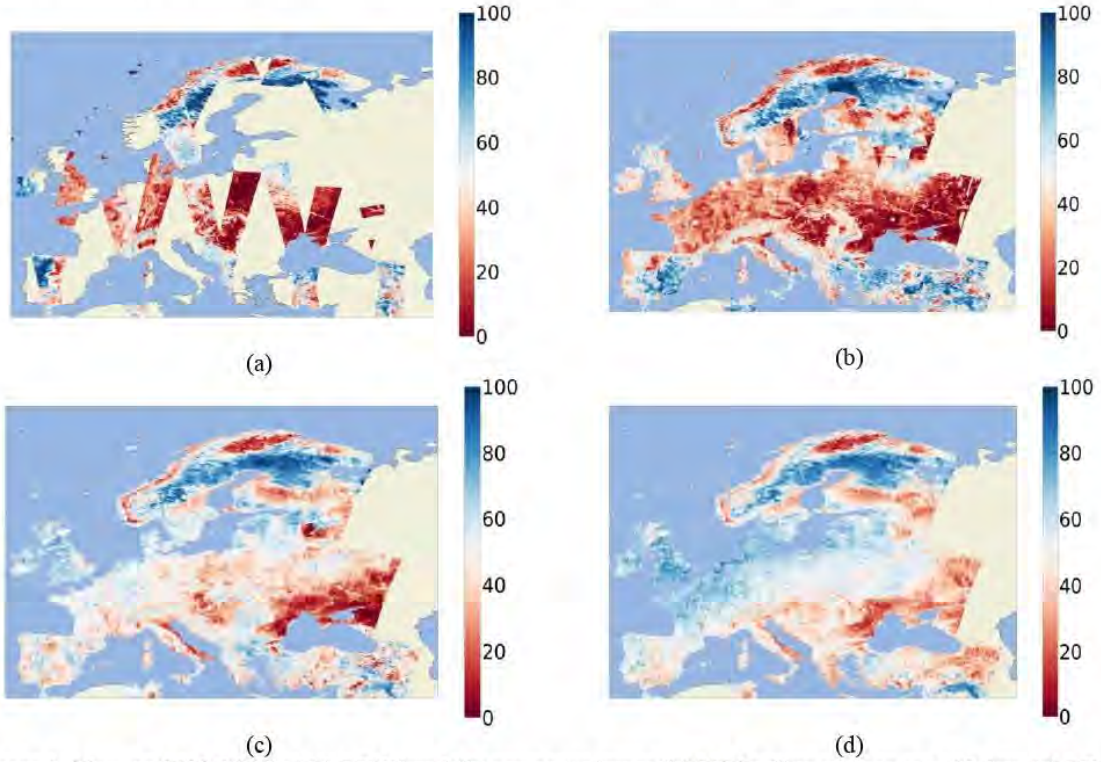


Figure 1: Maps on 2020-04-05 of (a) SSM1km daily composite image (b) SSM rolling average over 10 days (c) SSM rolling average over 30 days (d) SSM rolling average over 90 days (plots are projected onto Plate Carrée and all soil moisture is in relative units (%) , pale yellow areas are not covered).

2.1.2 Coarse resolution SSM data

The Soil Moisture Active Passive (SMAP) mission was launched by the National Aeronautics and Space Administration (NASA) in 2014 to study the surface soil water. SMAP has an active instrument (radar) and a passive one (a radiometer) on board. The SMAP observations are acquired at a fixed angle (40°) in dual polarization with a 39 km x 47 km nominal resolution. While the radiometer provides "passive" estimates with its coarse spatial resolution, the radar analyzes the "active" backscatter obtained from a Synthetic Aperture Radar (SAR) technology at 3 km spatial resolution. The SAR stopped operations 3 months after launch due to failure. The combination of the two datasets creates the final product, joining the penetrating capacity of the "passive" technology with the high spatial resolution of the "active" one. SMAP level3 soil moisture data from passive sensor at 36km (SMAP_L3_SM_P) were downloaded from the NASA Earthdata portal. Daily composite SSM maps were derived from the AM (descending) and PM (ascending) overpass of SMAP_L3_SM_P.

2.1.3 ERA5-land data

ERA5 is the fifth generation ECMWF reanalysis for the global climate and weather for the past 4 to 7 decades. ERA5 provides hourly estimates for a large number of atmospheric, ocean-wave and land-surface variables. Four main subsets exist, namely hourly and monthly products, on pressure levels as well as on single levels (atmospheric, ocean-wave and land surface quantities). ERA5-Land is a replay of the land component of the ERA5 climate reanalysis which is forced by meteorological fields from ERA5 (table 1). The outputs are provided on an hourly frequency and the fields are masked over oceans. Daily means of volumetric RZSM, volumetric SSM and 24-hour accumulated precipitation values were downloaded from the «Daily statistics calculated from ERA5 data » application that allows users to compute and download daily statistics of different variables from a number of hourly ERA5-land datasets (Muñoz-Sabater et al., 2021). As far as soil moisture information is considered, the ECMWF Integrated Forecasting System (IFS) has a four-layer representation of soil: Layer 1: 0 - 7cm, Layer 2: 7 - 28cm, Layer 3: 28 - 100cm, Layer 4: 100 - 289cm. The first layer (0 - 7cm) was selected to extract SSM information that will be used further to compare the different RZSM predictions. As for RZSM, the third layer (28-100 cm) was selected since we are interested in the root zone ranging between 30 and 55cm (Souissi et al., 2022).

The ERA5-land data is subset to the selected rectangular spatial region of interest and sampled at the selected frequency. For our study, the selected region of interest is continental Europe (-11°E,35°S, 50°W,72°N). The data is then aggregated to a daily frequency using the selected statistic and returned to the user in a single netCDF file.

Table 1: Data description of dataset "ERA5-Land hourly data from 1950 to present" (Copernicus Climate Change Service, 2019)

Data type	Gridded
Projection	Regular latitude-longitude grid
Horizontal coverage	Global
Horizontal resolution	0.1° x 0.1°; Native resolution is 9 km
Temporal coverage	January 1950 to present
Temporal resolution	Hourly

Precipitation data were used in the discussion part (see section 4) to highlight one of the limitations of SAR based SSM. Precipitation maps over Continental Europe were generated from the precipitation data of the ERA5-Land hourly dataset. We extracted the Total precipitation (tp) variable which is equivalent to the accumulated liquid and frozen water that falls to the Earth's surface. It is the sum of large-scale precipitation and convective precipitation. The units of this parameter are depth in meters of water equivalent. Total precipitation data were download for each day at 00:00 UTC. Actually, the total precipitation for a given day

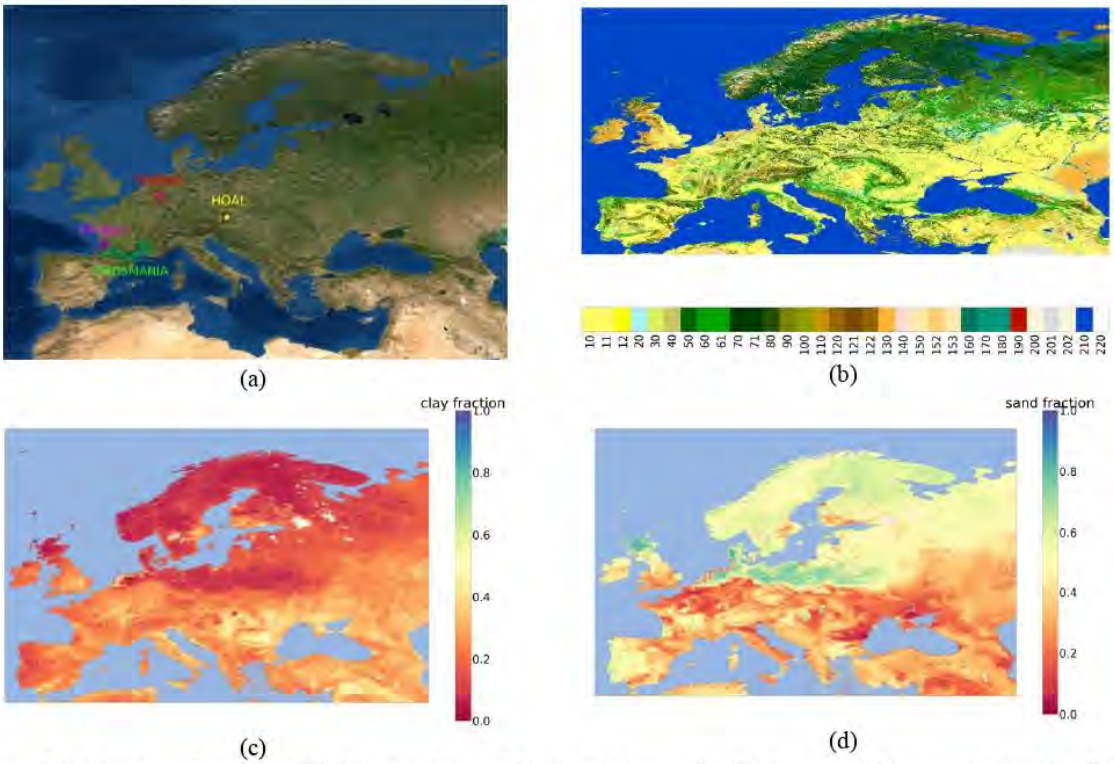
'd' is equal to the total precipitation at day 'd+1' at 00:00 UTC since it represents the accumulated flux over the previous 24 hours (equation 1).

$$tp_d = tp_{(d+1)\ 00{:}00UTC} \quad (1)$$

where d is the day for which the average flux is being computed.

2.1.4 In-situ soil moisture data

The International Soil Moisture Network (ISMN) is an exhaustive centralized data hosting facility which provides global soil moisture data (Dorigo et al., 2011). ISMN aims primarily at collecting ground soil moisture measurements from different data organizations and making them available in a harmonized format through a centralized free and open web portal (https://ismn.earth/en/, last access: 21 July 2022). In addition to this main objective, ISMN also integrates advanced quality control methods (Dorigo et al., 2013), provides additional metadata and ancillary variables as well as software code to users. Moreover, the ISMN has substantially grown in terms of networks, stations, and datasets (Dorigo et al., 2021). In-situ SSM information from eight ISMN networks of different climates and soil types were used to train the ANN model used in this study (Souissi et al., 2022). Moreover, RZSM data from four ISMN networks over Europe (figure 2(a)) were used in this paper to validate the RZSM predictions. The data used in the validation step were not seen or used during the training step.



(a) (b) (c) (d)

Figure 2: (a) In-situ soil moisture ISMN networks used for the validation of RZSM maps (b) 1 km resolution ESA CCI land cover map over Europe for year 2020 (see appendix B for land cover classes) (c) Clay fraction map of the top 5cm of the soil at 1km resolution created from the SoilGrid250m database (d) Sand fraction of the top 5cm of the soil at 1km resolution created from the SoilGrid250m database

2.1.5 Moderate Resolution Imaging Spectroradiometer (MODIS) data

Vegetation dynamics were considered in the architecture of the ANN model through a remote sensing-based normalized difference vegetation index (NDVI) feature. The model also takes account of the evaporation

process through the use of an evaporation efficiency based on potential evapotranspiration (PET). NDVI and PET are provided by Moderate Resolution Imaging Spectroradiometer (MODIS) products with a spatial resolution of 250 m and 500 m, respectively. NDVI and PET were extracted from the MOD13Q1 version 6 product with a revisit frequency of 16 days and the MOD16A2 version 6 each 8 days, respectively. According to the criteria set out by Huete et al. (1999), the NDVI value yielded by the MODIS product corresponds to the best pixel value of all acquisitions over the 16-day time range where the cloud coverage and view angle are low. Besides, the considered potential evapotranspiration value is equal to the sum of PET values encompassed by the 8-day window, as mentioned in Running et al. (2017).

In order to produce daily maps of NDVI and PET, linear interpolation was performed on the 16-day and 8-day products, respectively. The maps were downscaled to the spatial resolution of 1km such that it matches the SSM1km product resolution using the GDAL nearest neighbor resampler.

PET was then used to compute evaporation efficiency as described in the third model developed by (Merlin et al., 2010). In our work, a modified formulation (equation 2) was used and is further detailed in (Souissi et al., 2022). Evaporation efficiency can be expressed as follows:

$$\beta = [\frac{1}{2} - \frac{1}{2}\cos(\pi\theta/\theta_{max})]^{P*} \qquad (2)$$

where: - β is evaporation efficiency

- θ is the water content in the soil layer of a given thickness (here the surface layer is considered).

- θ_{max} is equal the maximum soil moisture value for each site.

- P^* is a parameter that can be computed as follows:

$$P^* = \frac{PET}{2B} \quad (3)$$

-PET is the potential evapotranspiration provided by the MOD16A2 product.

2.1.6 Ancillary data

In order to rescale RZSM outputs, sand and clay fractions were used to compute the soil moisture at saturation and the wilting point (see section 2.2). Clay and sand maps (figure 2(c), (d)) over continental Europe are available at a 250m resolution from SoilGrids version 2.0 product (Poggio et al., 2021). SoilGrids is a system for global digital soil mapping that uses state-of-the-art machine learning methods to map the spatial distribution of soil properties across the globe. The outputs of SoilGrids are global soil property maps at six standard depth intervals at a spatial resolution of 250 meters. All SoilGrids maps can be acquired through the Web Coverage Service (WCS) which is the most convenient way of obtaining spatial subsets of the various quantities. WCS is a standard issued by the Open Geospatial Consortium (OGC) which is designed to remotely accede to raster maps. WCS functions over the HTTP protocol and is supported by Python that ensures access to WCS through the OWSLib library.
The CCI land cover (LC) maps were used to evaluate the quality of RZSM predictions based on the land cover class. We generated the LC map over continental Europe for the year 2020 at a 1km spatial resolution (figure 2(b)).

2.2 Methods

The proposed methodology consists in predicting RZSM over continental Europe (figure 3) using the most complex ANN model in terms of number of features, which was developed in (Souissi et al., 2022). The features used in this model are SSM and process-related features. The model was trained and validated using

in-situ SSM information from ISMN, SWI which was computed using a recursive exponential filter and based on in-situ SSM as well as remote sensing-based variables namely NDVI and evaporation efficiency. Training was conducted on stations of eight soil moisture networks such that different climate regions and soil textures are encompassed. The detailed methodology and results can be found in (Souissi et al., 2022). In this paper, the block 'ANN model test' (figure 3) is implemented in order to achieve the RZSM maps block (figure 3). The only difference between the sub-blocks of the 'ANN model test' block is the SSM information source and thus the spatial resolution (SSM1km, SMAP_L3_SM_P, ERA5-Land). Each sub-block consists of the test datasets that are injected into the ANN model to produce a RZSM map which spatial resolution depends on the selected SSM product.

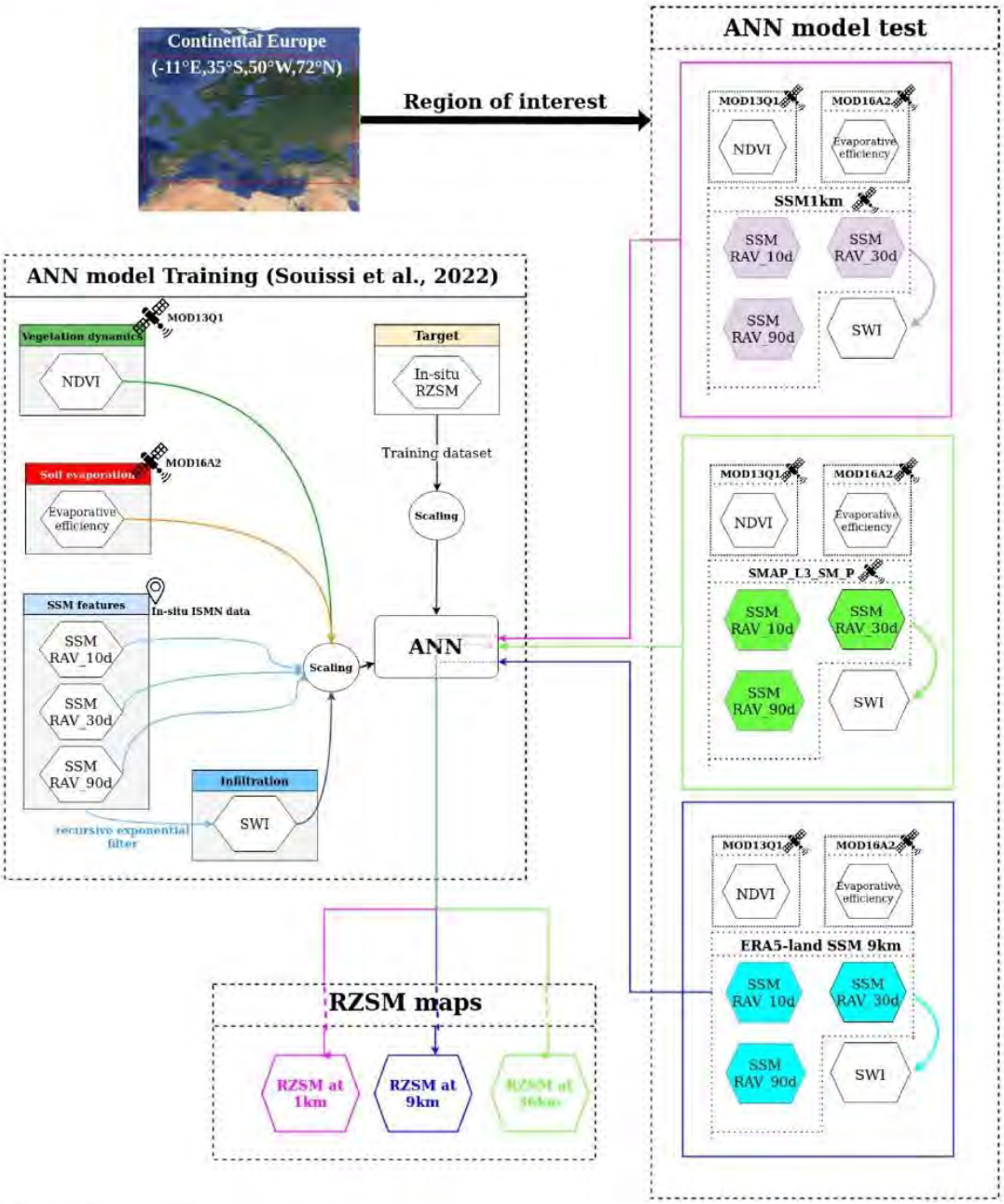


Figure 3: RZSM maps production steps

In this paper, the used ANN model is a feed-forward neural network commonly known as multilayer perceptron (MLP). MLP is considered as a machine learning technique which has been shown reliable several applications such as hydrology (Taver et al., 2015, Abrahart and See, 2007). It is arranged as a stack of layers namely an input layer, at least one hidden layer and an output layer. All layers are composed of neurons such that those of the first layer are called features and correspond to the input variables of the model. A weighted sum of the inputs and a bias are injected to each neuron of the hidden layer through the activation function. In order to minimize the error function called loss function, the model internally adjusts the weights between the neurons.
The architecture of our model consists of an input layer of SSM and process-related features, one hidden layer of 20 hidden neurons and an output layer. The activation function of the hidden layer is a tangent sigmoid function, the loss function is a quadratic cost function and the optimization algorithm is a stochastic gradient descent as already implemented in (Souissi et al., 2022).

Three SSM features are considered in this model, namely backward moving average of SSM using three temporal windows of 10, 30 and 90 days. Three process-related features are also used, namely evaporation efficiency (EVAP), SWI and NDVI. The model has been already trained using data from different areas of the world with different climatic and soil characteristics as detailed in (Souissi et al., 2022). In this study, the model is used only in prediction mode i.e. with no prior calibration.

ANN_SSM1km refers to the ANN model that uses the SSM1km product to compute the SSM features. ANN_SMAP36km refers to the ANN model that uses the SMAP_L3_SM_P product to compute the SSM features and ANN_ERA5-9km is the ANN model that uses the ERA5-land SSM reanalysis product to compute the SSM features (table 2).

Table 2: ANN model configurations with the respective input variables; [a]: rolling averages of SSM over 10 days; [b]: rolling averages of SSM over 30 days; [c]: rolling averages of SSM over 90 days; [*]: the SSM source product

Model / **Features**	SSM1km*			SMAP_L3_SM_P*			ERA5-9km*			NDVI	SWI	EVAP
	SSM_10d_RAV[a]	SSM_30d_RAV[b]	SSM_90d_RAV[c]	SSM_10d_RAV[a]	SSM_30d_RAV[b]	SSM_90d_RAV[c]	SSM_10d_RAV[a]	SSM_30d_RAV[b]	SSM_90d_RAV[c]			
ANN_SSM1km	X	X	X							X	X	X
ANN_SMAP36km				X	X	X				X	X	X
ANN_ERA5-9km							X	X	X	X	X	X

The time series of the different features were timestamp-matched such that only valid data for all inputs are kept. The RZSM predictions are scaled by the soil moisture at saturation and permanent wilting point.

The soil moisture at saturation is computed as in (Cosby et al., 1984):

$$\theta_{sat} = 0.489 - 0.126 * f_{sand} \quad (4)$$

Where f_{sand} is equal to the sand fraction (value between 0 and 1).

The Permanent wilting point (PWP) , defined as the minimum soil moisture at which a plant wilts and can no longer recover its turgidity, can be computed based on (Saxton et al., 1986) as follows:

$$PWP = \frac{15^{\frac{1}{B}}}{A} \quad (5)$$

Where: $A = exp(-4.396 - 0.0715\,(\%\,\text{clay}) - 4.880 * 10^{-4} * (\%\text{sand})^2 - 4.285 * 10^{-5} * (\%\text{sand})^2 * (\%\text{clay}))$ (3.1)

$$B = -3.140 - 0.00222 * (\%\text{clay})^2 - 3.484 * 10^{-5} * (\%\text{sand})^2 * (\%\text{clay}) \quad (3.2)$$

% sand, %clay are the percentages of sand and clay, respectively

3. Results

3.1 Illustration for RZSM predictions

In this part, the impact of the considered SSM product on the quality of the RZSM maps is investigated. More statistically-representative results can be found in the next paragraph. Two dates were chosen to compare RZSM maps yielded by SSM1km with those given by SMAP_L3_SM_P. The first considered date occurs in spring, a season in which quantification of RZSM is necessary while the second occurs in the autumn. In a first time, 10-day averaged SSM maps provided by both products can be compared (figure 4). The raw SSM maps as provided by both products are not shown here, given the presence of gaps mainly for the S1 product. The SMAP_L3_SM_P map is gapped over Northern Europe in spring due to the freeze-thaw mask, included in this product, in contrast to SSM1km which does not have a similar mask. In April (figure 4(a), (b)), SSM1km overestimates SSM over South Spain compared to SMAP and underestimates it over the UK. The discrepancy between both products is more amplified in autumn (figure 4(c), (d)). SSM1km overestimates SSM mainly over Northern Spain, France, North Italy and a big proportion of central Europe.

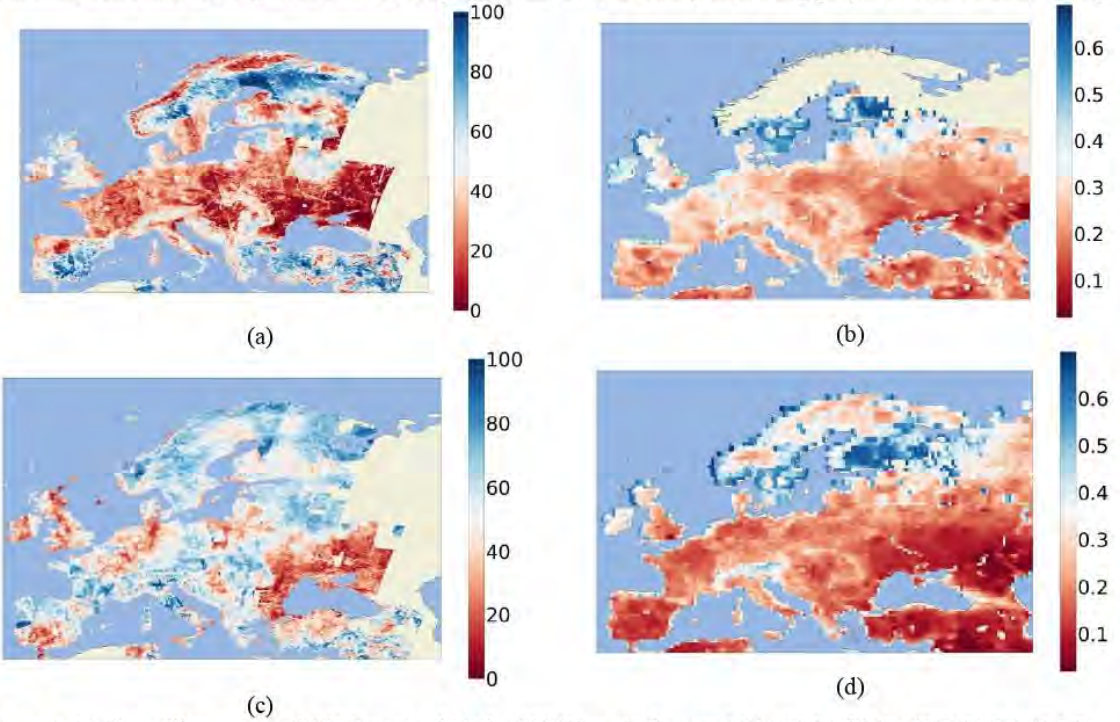


Figure 4: 10-day rolling averaged SSM maps of (a) the SSM1km product on 2020-04-04 (b) the SMAP_L3_SM_P product on 2020-04-04 (c) the SSM1km product on 2020-09-28 (d) the SMAP_L3_SM_P product on 2020-09-28

As for the predictions, the RZSM maps at 1km tend to mistakenly identify regions as wet whereas they should be dry (figure 5(a)) and vice versa (figure 5(c)). The same ANN model was used to predict in a first time RZSM at a 1km resolution using the SSM1km product, and at a second time RZSM at 36km resolution using the SMAP_L3_SM_P. Since RZSM predictions are highly impacted by SSM, they also show different patterns. In spring (figure 5(a), (b)), $RZSM_{ANN\_SSM1km}$ is overestimated mainly over western Spain and underestimated over France and Central Europe, compared to $RZSM_{ANN\_SMAP36km}$. In autumn (figure 5(c), (d)), $RZSM_{ANN\_SSM1km}$ is mostly overestimated over northern Spain and underestimated over the UK, compared to $RZSM_{ANN\_SMAP36km}$. These observations led us to inspect the quality of the SSM remote sensing products.

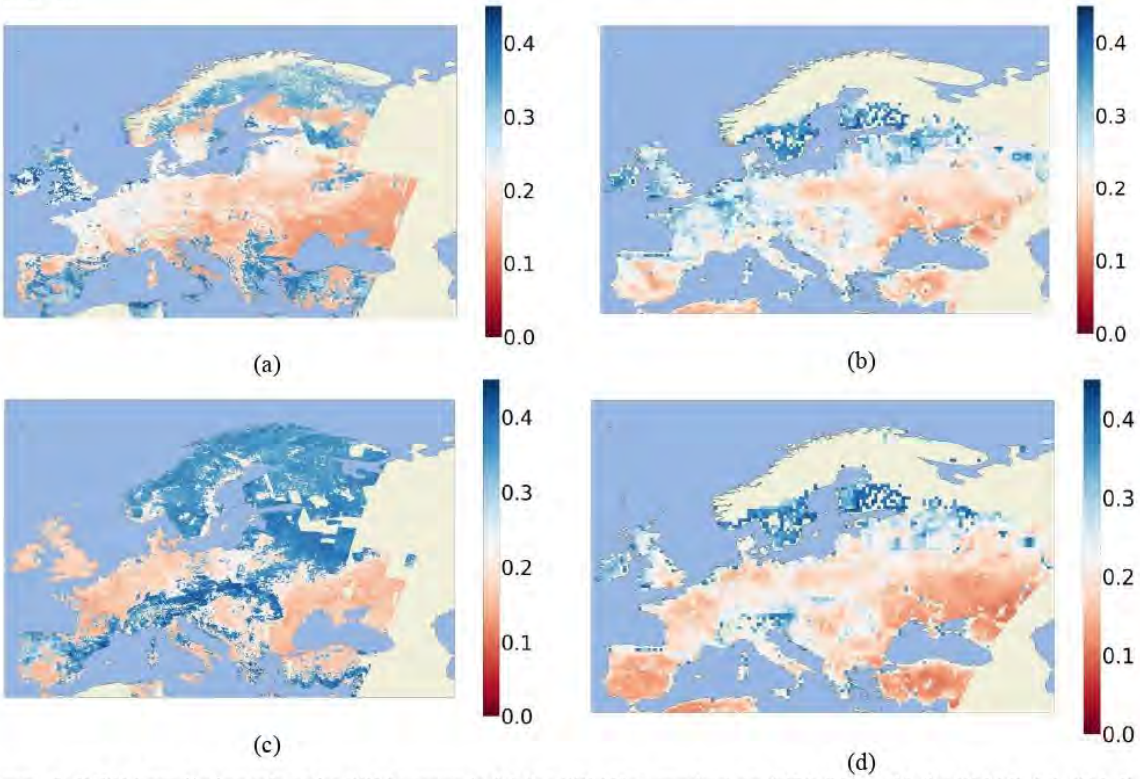


Figure 5: RZSM prediction when the SSM source is (a) the SSM1km product on 2020-04-04 (b) the SMAP_L3_SM_P product on 2020-04-04 (c) the SSM1km product on 2020-09-28 (d) the SMAP_L3_SM_P product on 2020-09-28.

3.2 Correlation maps between RZSM products

In order to assess the quality of $RZSM_{ANN\_SSM1km}$, we produced seasonal correlation maps between $RZSM_{ERA5}$ and $RZSM_{ANN\_SSM1km}$ (figure 6). Correlation values were computed pixel-wise for each season. $RZSM_{ERA5}$ was oversampled to the 1km grid using the GDAL library nearest neighbor resampler. Nordic areas show the lowest correlation values due to the challenging landscape (forest covers) and freeze-thaw events. Over agricultural areas in Eastern Europe, the lowest correlations are recorded in spring and summer which correspond to the growing seasons of most crops.

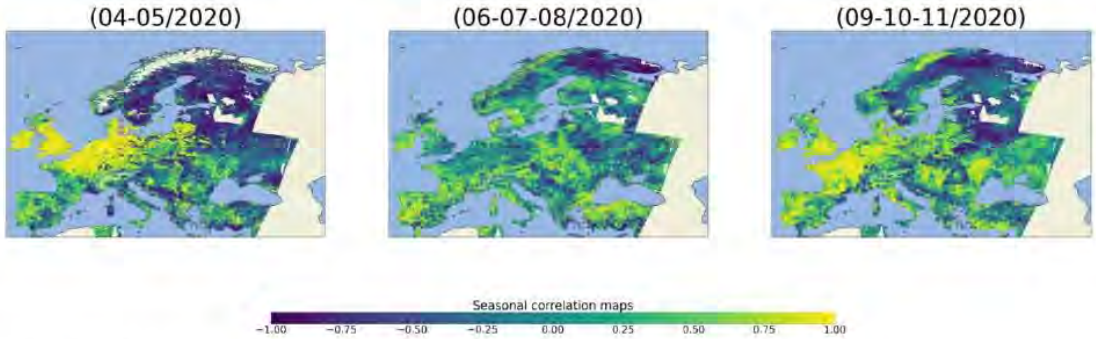
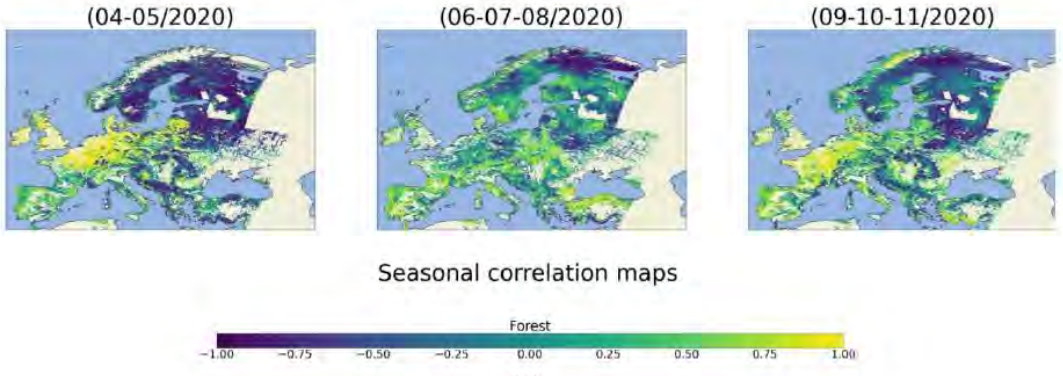


Figure 6: Seasonal correlation maps of $RZSM_{ANN\_SSM1km}$ and $RZSM_{ERA5}$

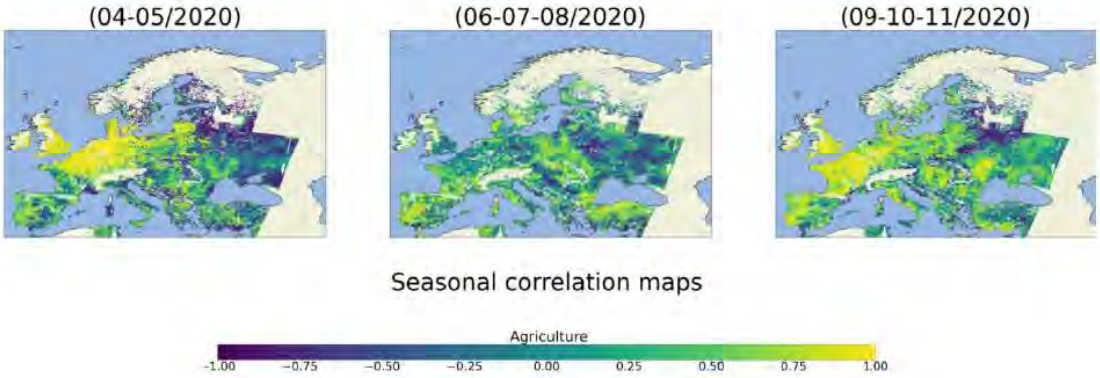
Besides, the performance of the model based on land cover was assessed through seasonal correlation maps for the two main land cover classes namely: **Forest** (CCI LC classes: 50-100, 160, 170) and **Agriculture** (CCI LC classes: 10-40).

Table 3 shows negative to low correlation values over the majority of forested areas. This is in accordance with the findings of (Bazzi et al., 2019) who assessed the SSM1km product over network SMOSMANIA for the year 2017. In forested areas, the product tended to overestimate SSM. Actually, the backscattered signal in C-band over forests is highly affected by the forest canopy because the penetration of the SAR signal to the ground surface is very low. High backscattering coefficients that are observed in the forest could induce high soil moisture estimations in SSM1km and thus an overestimation of SSM. Bauer-Marschallinger et al. (2019) also showed that the SSM1km product is not capable of estimating SSM over dense vegetation like forests.

Figure 7(b) and table 3 show, as aforementioned, that the best correlation values are recorded outside of the crops growing season. Actually, Bauer et al. (2018) showed that S-1 and ASCAT SSM retrievals are very akin, but one major difference is the absence of a dynamical vegetation correction in the Sentinel-1 algorithm. A comparison against SSM datasets observed by the Advanced Scatterometer (ASCAT) instrument further indicated an interference of vegetation dynamics with the SSM signal during the growing season. They also suggested that implementing a vegetation correction in the SSM1km algorithm would potentially improve the overall signal quality over vegetated areas.



(a)



(b)

Figure 7: Seasonal correlation maps over (a) Forests (b) Agricultural areas

Table 3: Performance of forest/agriculture pixels based on seasonal correlation values ($RZSM_{ANN\_SSM1km}$ vs. $RZSM_{ERA5}$)

Season	% pixels corr<0	% pixels 0<corr<0.5	% pixels corr>=0.5
Forest			
04-05	72.07	13.68	14.25
06-07-08	47.82	32.54	19.63
09-10-11	61.94	22.51	15.55
Agriculture			
04-05	45.77	21.53	32.7
06-07-08	42.02	33.72	24.26
09-10-11	24.19	30.95	44.86

Additional correlation maps were produced in order to compare $RZSM_{ANN\_SMAP36km}$ with $RZSM_{ERA5}$ (figure 8).Table 4 shows that $RZSM_{ANN\_SMAP36km}$ is more accurate than $RZSM_{ANN\_SSM1km}$. The use of the SMAP_L3_SM_P product gives better correlation values compared to SSM1km (Table 4). Ayres et al. (2021) compared more than twelve thousand SMAP soil moisture measurements to in-situ soil moisture measurements throughout the US, half of which are forested. For the forested sites, SMAP achieved a reasonable level of accuracy (unbiased RMSD: 0.06 m^3/m^3 or 0.053 m^3/m^3 after accounting for random representativeness errors) indicating SMAP is sensitive to changes in soil moisture in forest ecosystems. Also, $RZSM_{ANN\_SMAP36km}$ is more accurate for the vegetation land cover compared to $RZSM_{ANN\_SSM1km}$.

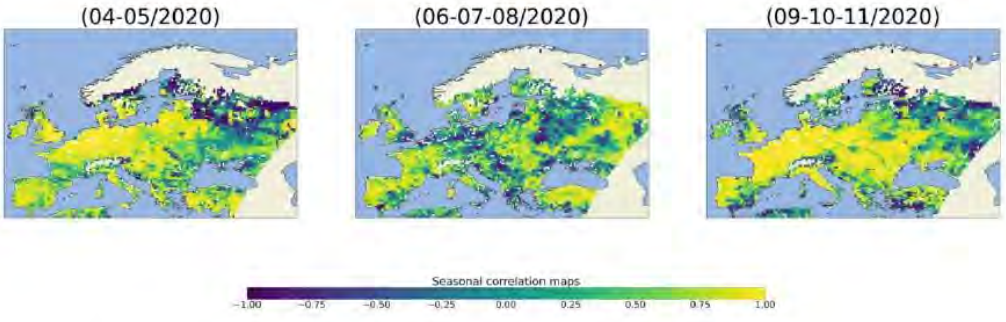


Figure 8: Seasonal correlation maps of $RZSM_{ANN\_SMAP36km}$ and $RZSM_{ERA5}$

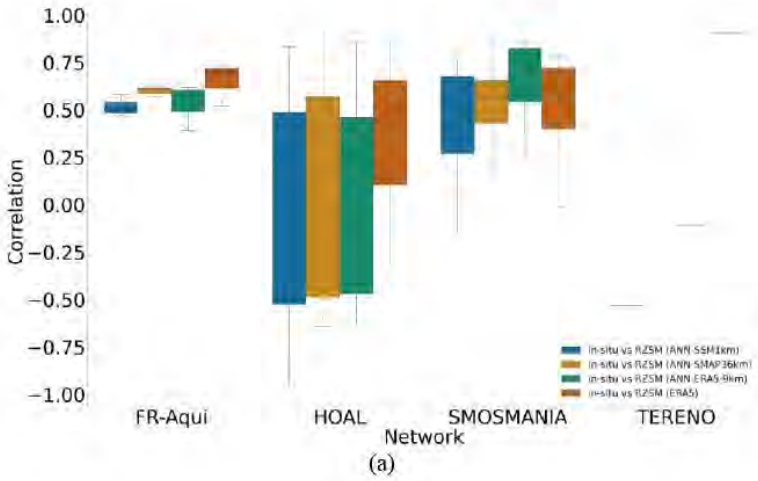
Table 4: Performance of forest/agriculture pixels based on seasonal correlation values ($RZSM_{ANN\_SMAP36km}$ vs. $RZSM_{ERA5}$)

Season	% pixels corr<0	% pixels 0<corr<0.5	% pixels corr>=0.5
Forest			
04-05	34.43	15.21	50.36
06-07-08	28.18	24.66	47.16
09-10-11	29.69	19.69	50.62
Agriculture			
04-05	19.46	19.94	60.59
06-07-08	28.61	22.76	48.63
09-10-11	16.58	17.37	66.05

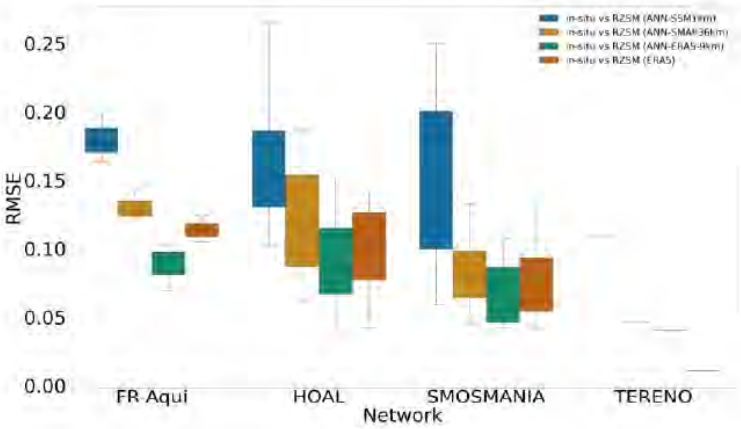
3.3 Comparison between RZSM product and in-situ data

Four in-situ soil moisture networks were selected to validate RZSM predictions. In-situ RZSM data over those networks are provided by ISMN which is a data-hosting facility of soil moisture data all over the world (Dorigo et al., 2011). RZSM data were downloaded over the year 2020 (stations with no data over 2020 are excluded). Actually, in-situ measurements are relevant to characterize the soil moisture dynamics at a point or field scale. Spatial representativeness must be considered when comparing in-situ data which correspond to point measurements and remote sensing data which cover a wider area determined by the satellite footprint. Comparisons of the remotely sensed SM data, here 1km data, to point-like ground data are often somewhat troublesome due to the scale mismatch (Bauer-Marschallinger et al., 2019). This leads to a non negligible representativeness error.

The RZSM time series of the satellite pixels overlapping the in-situ stations have been statistically compared to RZSM ground measurements. Performance metrics, namely Pearson correlation and RMSE, were computed over the stations of the 4 soil moisture networks considering the different RZSM estimates namely $RZSM_{ERA5}$, $RZSM_{ANN\_SSM1km}$, $RZSM_{ANN\_SMAP36km}$ and $RZSM_{ANN\_ERA5\text{-}9km}$ (figure 9). $RZSM_{ANN\_ERA5\text{-}9km}$ corresponds to the RZSM predictions using the ANN model that computes SSM features using ERA5-land reanalysis SSM product.



(a)



(b)

Figure 9: Comparison of RZSM predictions using different SSM information sources and the $RZSM_{ERA5}$ against in-situ measurements (a) Person correlation (b) RMSE

Table 4: Overall summary of statistics from time series analysis of different RZSM estimates versus in-situ RZSM. Cells show the mean Pearson correlation and RMSE values per network.

Network / Model	Correlation				RMSE			
	ANN_SSM1km	ANN_SMAP36km	ANN_ERA5-9km	ERA5	ANN_SSM1km	ANN_SMAP 36km	ANN_ERA 5-9km	ERA 5
FR-Aqui	0.515	0.599	0.531	0.654	0.18	0.130	0.088	0.114
HOAL	0.013	0.129	0.091	0.388	0.163	0.120	0.091	0.0100
SMOSMANIA	0.41	0.484	0.651	0.533	0.147	0.081	0.078	0.087
TERENO	-0.533	0.402	-0.108	0.904	0.11	0.046	0.04	0.01

Figure 9 shows that the RZSM predictions that are based on SSM1km data are the least accurate. However, the same ANN model yields more accurate RZSM predictions when SMAP_L3_SM_P data are used to compute the SSM features. The model ANN_ERA5-9km is even more accurate and its predictions $RZSM_{ANN\_ERA5\text{-}9km}$ outperform $RZSM_{ERA5}$ in the case of network SMOSMANIA for instance.

When it comes to network SMOSMANIA, the SSM1km data were previously compared against in-situ data in literature (Bauer-Marschallinger et al., 2018; Bauer-Marschallinger et al., 2021). During spring and summer, the average soil moisture levels have been shown mismatching compared to in-situ SSM. The SSM1km product tended to underestimate the in-situ data in spring and to overestimate it in summer. Bauer-Marschallinger et al. (2021) recorded a mean Spearman rho and RMSD values equal to 0.37 and 0.07 m^3/m^3 from time series analyses of SSM1km versus in-situ measurements over stations of network SMOSMANIA. In our case, a mean Pearson correlation and RMSE values of 0.41 and 0.147 m^3/m^3 were recorded when comparing $RZSM_{ANN\_SSM1km}$ against in-situ measurements, respectively (table 4), which happen to be the worst performance statistics. $RZSM_{ANN\_ERA5\text{-}9km}$ is the most accurate product since it recorded the highest mean correlation (0.651) and least RMSE (0.078 m^3/m^3). More specifically, TERENO network is located in Western Germany and is characterized by a wet context. Bauer-Marschallinger et al. (2021) recorded a mean Spearman rho and RMSD values equal to 0.53 and 0.08 m^3/m^3 from time series analyses of SSM1km versus

in-situ measurements over stations of network TERENO. Besides, network FR-Aqui is situated in southwestern France and most of its sites cover "Les Landes" forest of the Bordeaux-Aquitaine region. The forest land cover may explain the least accurate $RZSM_{ANN\_SSM1km}$ compared to $RZSM_{ANN\_SMAP36km}$ and $RZSM_{ERA5}$. Bauer-Marschallinger et al. (2021) highlighted that complex scattering mechanisms related to phenological processes such as leaf emergence occur over deciduous broadleaf forest and complicate the soil moisture estimation. This effect was observed for the FR-Aqui station parcmeteo.

4. Discussion

In this paper, RZSM maps at different spatial resolutions were generated over continental Europe using an ANN model with different features and SSM information sources. Although the main aim was to generate maps at 1km spatial resolution, coarser maps were also generated (at 9km and 36km) to highlight the impact of the quality of the input data.

As shown in the results section, the model RZSM predictions greatly depend on the considered SSM datasets, which is expected from a data-driven approach in which SSM is the central information. The C-band information given by the SSM1km product is prone to to various limitations. For instance, its relatively limited revisit time makes it miss rainfall events and thus soil moisture peaks (Bauer-Marschallinger et al., 2021). Figure 10 shows the 24-hour accumulated precipitation on 2020-06-05 as well as the SSM1km data on the same date (see Appendix A). The overpasses of Sentinel-1 don't cover the areas where strong rainfall events occurred. Table 5 presents the different SSM values given by the different SSM products over a pixel where an important rainfall event occurred. Values of rolling averages of SSM1km over 10, 30, 90 days are provided in relative values since the SSM1km product is in relative units. Those values were scaled using the minimum and maximum values of SSM datasets from the SMAP_L3_SM_P product. This latter classifies the pixel as a wet pixel, which is in accordance with the rainfall event, whereas SSM1km product classifies it as dry.

Figure 10: Maps on 2020-06-05 of the ERA5-land 24-hour accumulated precipitation and SSM1km

Table 5: Comparison between the SSM values of the different SSM products over the same pixel after a rainfall event.

Row	Column	tp (mm)	SMAP_L3_SM_P (m^3/m^3)	SSM10d		SSM1m		SSM3m	
779	741	93.51	0.537	%	m^3/m^3	%	m^3/m^3	%	m^3/m^3
				30	0.215	36	0.257	37	0.264

Gomis-Cebolla et al. (2022) assessed the quality of the SSM1km product as well as different SMAP and SMOS products over three case studies in the Mediterranean Bio-geographical region of the Iberian Peninsula (Hozgarganta (southern Spain), Ceira (western Portugal) and Carraixet (eastern Spain)). A relative temporal agreement was found for all products except SSM1km which had different temporal dynamics.

During dry months, the SSM1km product showed a high overestimation of SSM in Carraixet which has a semi-arid climate (climate class 'BSk'). The poor statistics recorded with the SSM1km product were attributed to two main factors. In temperate regions where the vegetation cover has a leaf area index larger than 0.6, the sensitivity of C-band backscatter to soil moisture was likely to be reduced. For dry surfaces with low topography and comparable electromagnetic properties, the soil roughness was the most dominating surface properties effect on the measured radar backscatter coefficients. The signal can be considered as a measure of the surface roughness at a scale comparable to the sensor wavelength (Marticorena et al., 2006).

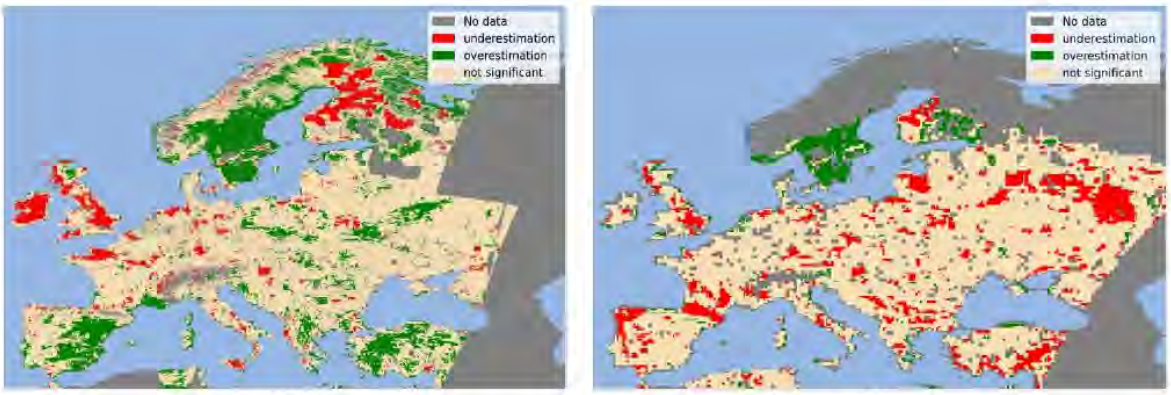
Figure 11 shows the difference between monthly $RZSM_{ANN\_SSM1km}$ and $RZSM_{ERA5}$ maps and similarly between monthly $RZSM_{ANN\_SMAP36km}$ and $RZSM_{ERA5}$ maps. Only absolute difference values which are greater than 0.1 m^3/m^3 are shown. When compared to reference $RZSM_{ERA5}$, RZSM overestimations, underestimations and non-significant differences are mapped in green, red and yellow, respectively (not covered areas are mapped with grey). The significant differences are more recurrent in the case of $RZSM_{ANN\_SSM1km}$ compared to $RZSM_{ANN\_SMAP36km}$.

The eastern part of the Iberian Peninsula which is characterized by a semi-arid climate is subject to permanent overestimations of RZSM (see appendix C). Wagner et al. (2022) have shown a subsurface scattering, over this region, which could be important in dry season and is neglected in proposed change detection algorithm.

Model ANN _SMAP36km doesn't yield an overestimation over that area. This can be explained by the fact that Sentinel-1 carries a C-band SAR which is more sensitive to vegetation and surface roughness than the L-band SMAP radiometer (Calvet et al., 2011). Bauer-Marschallinger et al. (2018) detected signal patterns indicating irrigation activities in SSM1km time series over network RHEMEDUS, which is located in Spain and showed that is likely to be completely missed by satellite data when the irrigation is applied to a small area. Paciolla et al. (2020) also found irrigation impact patterns in Northern and southern Italy. However, this remains a weak argument to explain the very large overestimation patterns in Spain that cover 100 of kilometers even though irrigation signals are detected in the SSM1km from S1.
Furthermore, model ANN_SSM1km tends to overestimate RZSM over Nordic regions as already highlighted in the results section due to the highly dynamic freeze/thaw processes and complex land cover. These regions are flagged more effectively with SMAP_L3_SM_P whereas SSM1km is not trustworthy because no specific mask or flag exists for such conditions.

Overestimations of RZSM compared to ERA5-land reanalysis RZSM datasets are also recorded around the Provence vineyards area (France) potentially because of the overestimations affecting SSM1km data. Similarly, RZSM overestimations can be detected around Bordeaux as well as Languedoc Roussillon vineyards (figure 11). In this context, Bazzi et al. (2019) found that over cells with predominant vineyards, SSM1km tends to overestimate SSM because of the high backscattering signal reflected from metals that are usually present in vineyards. Baghdadi et al. (2006) also showed that the vineyard parcels with metal stakes within the study site (a Mediterranean vineyard site near Bordeaux in southwestern France) have a stronger radar signal than parcels with wooden stakes.

Figure 11 also highlights a significant underestimation of RZSM over UK all over the year (see appendix C). This underestimation is less remarkable in the case of $RZSM_{ANN\_SMAP36km}$. Actually, the change detection algorithm of the SSM1km product is not very efficient in contexts of soil moisture high variability.



(a) (b)

Figure 11: RZSM monthly difference map (July 2020) of $RZSM_{ERA5}$ and (a) $RZSM_{ANN\_SSM1km}$ (b) $RZSM_{ANN\_SMAP36km}$; green: product (a) or (b) overestimates $RZSM_{ERA5}$, red: product (a) or (b) underestimates $RZSM_{ERA5}$; yellow: the correlation difference is not very significant (<0.1); gray: uncovered areas

Based on these findings, microwave-based SSM information seemed to correct some issues arising from the use SSM1km. However, RZSM maps at coarse resolutions are not in accordance with the needs of agricultural applications. In this context, the use of fusion products would be of interest (El Hajj et al., 2014; Tomer et al., 2016; El Hajj et al., 2017). The results also showed that data-driven approaches and more specifically ANN are still reliable for RZSM estimation. This has been highlighted through the comparison between $RZSM_{ERA5}$ and $RZSM_{ANN\_ERA5\text{-}9km}$ which suggested a prediction quality enhancement with the latter, mainly over SMOSMANIA network.

5 Conclusion

In this study, we assessed the ability of an ANN model to map RZSM at 1km spatial resolution over continental Europe. The considered ANN model has been previously trained over different regions across the globe with distinct land cover and climate classes. In this paper, the model is only used in test mode to predict RZSM based on remote sensing-retrieved SSM information and process-related variables namely SWI, evaporation efficiency and NDVI. RZSM maps at 1km spatial resolution were produced using the SSM1km product as a source to compute the three SSM features required by the model. One main finding was that the quality of the SSM input information and its temporal repetitiveness play a major role in the accuracy of the predictions which was assessed at the map scale using the ERA5-land reanalysis RZSM product and locally using in-situ RZSM measurements. The limitations of C-band based SSM i.e. SSM1km product were highlighted over complex sceneries like forested areas, areas with freeze/thaw events, during the growing season of crops, with irrigation, etc. This result was further investigated by producing RZSM maps at coarse resolution. The SMAP_L3_SM_P product was used to compute the SSM features of the ANN model and map RZSM at a spatial resolution of 36km. The accuracy of the predictions was further enhanced compared to model ANN_SSM1km but the estimations are hampered by the coarse spatial resolution. However, both ANN configurations yielded less accurate RZSM predictions than the ERA5-land RZSM reanalysis datasets. This has led us to produce RZSM maps at 9km resolution using the ERA5-land reanalysis SSM product as a source to compute the SSM features of the ANN model. Overall, the quality of the predictions was improved compared to the ANN models based on the previous SSM source products. The quality enhancement was more obvious when model ANN_ERA5-9km was used, since it outperformed the ERA5-land RZSM reanalysis product.

As a research perspective, the impact of the use of better quality remote sensing SSM products can be assessed. Disaggregation approaches and fusion techniques can also be considered in order to provide more precise SSM information to the ANN model. Despite the relative limitations of the RZSM product inferred

from Sentinel-1 SSM data, this product should greatly improve with the launches of the other satellites of the Sentinel-1 constellation or through the fusion with data from other constellations such as Radarsat.

Acknowledgments

The PhD thesis of R. Souissi was financed by the ERANET RET-SIF project and CNES, and complementary financing was provided by the PRIMA Program SMARTIES project. The authors thank the International Soil Moisture Network (ISMN) and supporting networks for providing the soil moisture data.

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11, 1563–1579, https://doi.org/10.5194/hess-11-1563-2007, 2007.

Akuraju, V. R., Ryu, D., and George, B.: Estimation of root-zone soil moisture using crop water stress index (CWSI) in agricultural fields, GIScience & Remote Sensing, 58, 340–353, https://doi.org/10.1080/15481603.2021.1877009, 2021.

Al Bitar, A., Leroux, D., Kerr, Y. H., Merlin, O., Richaume, P., Sahoo, A., and Wood, E. F.: Evaluation of SMOS Soil Moisture Products Over Continental US Using the SCAN/SNOTEL Network, IEEE Trans. Geosci. Remote Sensing, 50, 1572–1586, https://doi.org/10.1109/TGRS.2012.2186581, 2012.

Al Bitar, A., Mialon, A., Kerr, Y. H., Cabot, F., Richaume, P., Jacquette, E., Quesney, A., Mahmoodi, A., Tarot, S., Parrens, M., Al-Yaari, A., Pellarin, T., Rodriguez-Fernandez, N., and Wigneron, J.-P.: The global SMOS Level 3 daily soil moisture and brightness temperature maps, Earth Syst. Sci. Data, 9, 293–315, https://doi.org/10.5194/essd-9-293-2017, 2017.

Ayres, E., Colliander, A., Cosh, M., A. Roberti, J., Simkin, S., and Genazzio, M. A.: Validation of SMAP Soil Moisture at Terrestrial National Ecological Observatory Network (NEON) Sites Show Potential for Soil Moisture Retrieval in Forested Areas, https://doi.org/10.36227/techrxiv.14681298.v1, 2021.

Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., and Tuller, M.: Ground, Proximal, and Satellite Remote Sensing of Soil Moisture, Rev. Geophys., 57, 530–616, https://doi.org/10.1029/2018RG000618, 2019.

Baghdadi, N., Holah, N., Dubois-Fernandez, P., Dupuis, X., and Garestier, F.: Evaluation of polarimetric L- and P-bands RAMSES data for characterizing Mediterranean vineyards, 32, 380–389, https://doi.org/10.5589/m07-001, 2006.

Battude, M., Al Bitar, A., Brut, A., Tallec, T., Huc, M., Cros, J., Weber, J.-J., Lhuissier, L., Simonneaux, V., and Demarez, V.: Modeling water needs and total irrigation depths of maize crop in the south west of France using high spatial and temporal resolution satellite imagery, Agr. Water Manage., 189, 123–136, https://doi.org/10.1016/j.agwat.2017.04.018, 2017.

Bauer-Marschallinger, B., Cao, S., Schaufler, S., Paulik, C., Naeimi, V., and Wagner, W.: 1km Soil Moisture from Downsampled Sentinel-1 SAR Data: Harnessing Assets and Overcoming Obstacles., 17330, 2017.

Bauer-Marschallinger, B., Schaufler, S., Navacchi, C., Validation report Surface Soil Moisture collection 1km version 1 (I1.20). Copernicus Global Land Operations “Vegetation and Energy” ”CGLOPS-1” Framework Service Contract N°199494 (JRC). Available online:

https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_VR_SSM1km-V1_I1.20.pdf (accessed on 19 July 2022).

Bauer-Marschallinger, B., Freeman, V., Cao, S., Paulik, C., Schaufler, S., Stachl, T., Modanesi, S., Massari, C., Ciabatta, L., Brocca, L., and Wagner, W.: Toward Global Soil Moisture Monitoring With Sentinel-1: Harnessing Assets and Overcoming Obstacles, IEEE Trans. Geosci. Remote Sensing, 57, 520–539, https://doi.org/10.1109/TGRS.2018.2858004, 2019.

Bauer-Marschallinger, B., Massart, S., Quality assessment report update 2021 Surface Soil Moisture collection 1km version 1.0 (I1.00). Copernicus Global Land Operations "Vegetation and Energy" "CGLOPS-1" Service Contract N° 941115 - ISP- 2021 (JRC). Available online: https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/CGLOPS1_QAR2021_SSM1km-V1_I1.00.pdf (accessed on 19 July 2022).

Bazzi, H., Baghdadi, N., El Hajj, M., Zribi, M., and Belhouchette, H.: A Comparison of Two Soil Moisture Products S ^{2} MP and Copernicus-SSM Over Southern France, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, 12, 3366–3375, https://doi.org/10.1109/JSTARS.2019.2927430, 2019.

Berg, A. and Sheffield, J.: Climate Change and Drought: the Soil Moisture Perspective, Curr Clim Change Rep, 4, 180–191, https://doi.org/10.1007/s40641-018-0095-0, 2018.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, Geosci. Model Dev., 4, 677–699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Calvet, J.-C., Wigneron, J.-P., Walker, J., Karbou, F., Chanzy, A., and Albergel, C.: Sensitivity of Passive Microwave Observations to Soil Moisture and Vegetation Water Content: L-Band to W-Band, 49, 1190–1199, https://doi.org/10.1109/TGRS.2010.2050488, 2011.

Carranza, C., Nolet, C., Pezij, M., and van der Ploeg, M.: Root zone soil moisture estimation with Random Forest, Journal of Hydrology, 593, 125840, https://doi.org/10.1016/j.jhydrol.2020.125840, 2021.

Copernicus Climate Change Service: ERA5-Land hourly data from 2001 to present, https://doi.org/10.24381/CDS.E2161BAC, 2019.

Cosby, B. J., Hornberger, G. M., Clapp, R. B., and Ginn, T. R.: A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils, Water Resour. Res., 20, 682–690, https://doi.org/10.1029/WR020i006p00682, 1984.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface, Bulletin of the American Meteorological Society, 87, 1381–1398, https://doi.org/10.1175/BAMS-87-10-1381, 2006.

Dobriyal, P., Qureshi, A., Badola, R., and Hussain, S. A.: A review of the methods available for estimating soil moisture and its implications for water resource management, Journal of Hydrology, 458–459, 110–117, https://doi.org/10.1016/j.jhydrol.2012.06.021, 2012.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a

data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675–1698, https://doi.org/10.5194/hess-15-1675-2011, 2011.

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

Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Bitelli, M., Blöschl, G., Bogena, H., Brocca, L., Calvet, J.-C., Camarero, J. J., Capello, G., Choi, M., Cosh, M. C., van de Giesen, N., Hajdu, I., Ikonen, J., Jensen, K. H., Kanniah, K. D., de Kat, I., Kirchengast, G., Kumar Rai, P., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J. P., Osenga, E., Palecki, M. A., Pellarin, T., Petropoulos, G. P., Pfeil, I., Powers, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Varlagin, A., Vreugdenhil, M., Walker, J., Wen, J., Wenger, F., Wigneron, J. P., Woods, M., Yang, K., Zeng, Y., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., and Sabia, R.: The International Soil Moisture Network: serving Earth system science for over a decade, Hydrol. Earth Syst. Sci., 25, 5749–5804, https://doi.org/10.5194/hess-25-5749-2021, 2021.

El Hajj, M., Baghdadi, N., Belaud, G., Zribi, M., Cheviron, B., Courault, D., Hagolle, O., and Charron, F.: Irrigated Grassland Monitoring Using a Time Series of TerraSAR-X and COSMO-SkyMed X-Band SAR Data, Remote Sens., 6, 10002–10032, https://doi.org/10.3390/rs61010002, 2014.

El Hajj, M., Baghdadi, N., Zribi, M., and Bazzi, H.: Synergic Use of Sentinel-1 and Sentinel-2 Images for Operational Soil Moisture Mapping at High Spatial Resolution over Agricultural Areas, Remote Sensing, 9, 1292, https://doi.org/10.3390/rs9121292, 2017.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application, Hydrol. Earth Syst. Sci., 14, 1943–1961, https://doi.org/10.5194/hess-14-1943-2010, 2010.

Entekhabi, D., Nakamura, H., and Njoku, E. G.: Retrieval of soil moisture profile by combined remote sensing and modeling, in: Retrieval of soil moisture profile by combined remote sensing and modeling, De Gruyter, 485–498, ISBN 9783112319307, 2020.

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

Gomis-Cebolla, J., Garcia-Arias, A., Perpinyà-Vallès, M., and Francés, F.: Evaluation of Sentinel-1, SMAP and SMOS surface soil moisture products for distributed eco-hydrological modelling in Mediterranean forest basins, Journal of Hydrology, 608, 127569, https://doi.org/10.1016/j.jhydrol.2022.127569, 2022.

Hain, C. R., Mecikalski, J. R., and Anderson, M. C.: Retrieval of an Available Water-Based Soil Moisture Proxy from Thermal Infrared Remote Sensing. Part I: Methodology and Validation, Journal of Hydrometeorology, 10, 665–683, https://doi.org/10.1175/2008JHM1024.1, 2009.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Martinuzzi, J., Font, J., and Berger, M.: Soil moisture retrieval from space: the Soil Moisture and Ocean Salinity (SMOS) mission, 39, 1729–1735, https://doi.org/10.1109/36.942551, 2001.

Kornelsen, K. C. and Coulibaly, P.: Root-zone soil moisture estimation using data-driven methods, Water Resour. Res., 50, 2946–2962, https://doi.org/10.1002/2013WR014127, 2014.

Kumar, S. V., Reichle, R. H., Koster, R. D., Crow, W. T., and Peters-Lidard, C. D.: Role of Subsurface Physics in the Assimilation of Surface Soil Moisture Observations, Journal of Hydrometeorology, 10, 1534–1547, https://doi.org/10.1175/2009JHM1134.1, 2009.

Marticorena, B., Kardous, M., Bergametti, G., Callot, Y., Chazette, P., Khatteli, H., Le Hégarat-Mascle, S., Maillé, M., Rajot, J.-L., Vidal-Madjar, D., and Zribi, M.: Surface and aerodynamic roughness in arid and semiarid areas and their relation to radar backscatter coefficient: SURFACE ROUGHNESS IN ARID AREAS, J. Geophys. Res., 111, n/a-n/a, https://doi.org/10.1029/2006JF000462, 2006.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Nativel, S., Ayari, E., Rodriguez-Fernandez, N., Baghdadi, N., Madelon, R., Albergel, C., and Zribi, M.: Hybrid Methodology Using Sentinel-1/Sentinel-2 for Soil Moisture Estimation, Remote Sensing, 14, 2434, https://doi.org/10.3390/rs14102434, 2022.

Noilhan, J. and Mahfouf, J.-F.: The ISBA land surface parameterisation scheme, Global Planet. Change, 13, 145–159, https://doi.org/10.1016/0921-8181(95)00043-7, 1996.

Lievens, H., De Lannoy, G. J. M., Al Bitar, A., Drusch, M., Dumedah, G., Hendricks Franssen, H.-J., Kerr, Y. H., Tomer, S. K., Martens, B., Merlin, O., Pan, M., Roundy, J. K., Vereecken, H., Walker, J. P., Wood, E. F., Verhoest, N. E. C., and Pauwels, V. R. N.: Assimilation of SMOS soil moisture and brightness temperature products into a land surface model, Remote Sensing of Environment, 180, 292–304, https://doi.org/10.1016/j.rse.2015.10.033, 2016.

Lopez, T., Al Bitar, A., Biancamaria, S., Güntner, A., and Jäggi, A.: On the Use of Satellite Remote Sensing to Detect Floods and Droughts at Large Scales, Surv Geophys, 41, 1461–1487, https://doi.org/10.1007/s10712-020-09618-0, 2020.

Oleson, W., Lawrence, M., Bonan, B., Flanner, G., Kluzek, E., Lawrence, J., Levis, S., Swenson, C., Thornton, E., Dai, A., Decker, M., Dickinson, R., Feddema, J., Heald, L., Hoffman, F., Lamarque, J.-F., Mahowald, N., Niu, G.-Y., Qian, T., Randerson, J., Running, S., Sakaguchi, K., Slater, A., Stockli, R., Wang, A., Yang, Z.-L., Zeng, X., and Zeng, X.: Technical Description of version 4.0 of the Community Land Model (CLM), NCAR/UCAR, https://doi.org/10.5065/D6FB50WZ, 2010.

Paloscia, S., Macelloni, G., Pampaloni, P., Ruisi, R., and Santi, E.: CHEERS Experiment in Tuscany: a comparison between L-band and C- and X-bands capability in measuring soil moisture, in: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No.01CH37217), IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium, Sydney, NSW, Australia, 22–24, https://doi.org/10.1109/IGARSS.2001.976045, 2001.

Pan, X., Kornelsen, K. C., and Coulibaly, P.: Estimating Root Zone Soil Moisture at Continental Scale Using Neural Networks, J. Am. Water Resour. Assoc., 53, 220–237, https://doi.org/10.1111/1752-1688.12491, 2017. Paciolla, N., Corbari, C., Al Bitar, A., Kerr, Y., and Mancini, M.: Irrigation and Precipitation Hydrological Consistency with SMOS, SMAP, ESA-CCI, Copernicus SSM1km, and AMSR-2 Remotely Sensed Soil Moisture Products, Remote Sensing, 12, 3737, https://doi.org/10.3390/rs12223737, 2020.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Reichle, R. H., Liu, Q., Koster, R. D., Crow, W. T., De Lannoy, G. J. M., Kimball, J. S., Ardizzone, J. V., Bosch, D., Colliander, A., Cosh, M., Kolassa, J., Mahanama, S. P., Prueger, J., Starks, P., and Walker, J. P.: Version 4 of the SMAP Level-4 Soil Moisture Algorithm and Data Product, J. Adv. Model. Earth Syst., 11, 3106–3130, https://doi.org/10.1029/2019MS001729, 2019.

Saux-Picart, S., Ottlé, C., Decharme, B., André, C., Zribi, M., Perrier, A., Coudert, B., Boulain, N., Cappelaere, B., Descroix, L., and Ramier, D.: Water and energy budgets simulation over the AMMA-Niger super-site spatially constrained with remote sensing data, Journal of Hydrology, 375, 287–295, https://doi.org/10.1016/j.jhydrol.2008.12.023, 2009.

Saxton, K. E., Rawls, W. J., Romberger, J. S., and Papendick, R. I.: Estimating Generalized Soil-water Characteristics from Texture, Soil Science Society of America Journal, 50, 1031–1036, https://doi.org/10.2136/sssaj1986.03615995005000040039x, 1986.

Shen, X., Walker, J. P., Ye, N., Wu, X., Boopathi, N., Yeo, I.-Y., Zhang, L., and Zhu, L.: Soil Moisture Retrieval Depth of P- and L-Band Radiometry: Predictions and Observations, IEEE Trans. Geosci. Remote Sensing, 59, 6814–6822, https://doi.org/10.1109/TGRS.2020.3026384, 2021.

Souissi, R., Al Bitar, A., and Zribi, M.: Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe, Water, 12, 3109, https://doi.org/10.3390/w12113109, 2020.

Souissi, R., Zribi, M., Corbari, C., Mancini, M., Muddu, S., Tomer, S. K., Upadhyaya, D. B., and Al Bitar, A.: Integrating process-related information into an artificial neural network for root-zone soil moisture prediction, Hydrology and Earth System Sciences, 26, 3263–3297, https://doi.org/10.5194/hess-26-3263-2022, 2022.

Taver, V., Johannet, A., Borrell-Estupina, V., and Pistre, S.: Feed-forward *vs* recurrent neural network models for non-stationarity modelling using data assimilation and adaptivity, Hydrological Sciences Journal, 60, 1242–1265, https://doi.org/10.1080/02626667.2014.967696, 2015.

Tomer, S., Al Bitar, A., Sekhar, M., Zribi, M., Bandyopadhyay, S., Sreelash, K., Sharma, A. K., Corgne, S., and Kerr, Y.: Retrieval and Multi-scale Validation of Soil Moisture from Multi-temporal SAR Data in a Semi-Arid Tropical Region, Remote Sensing, 7, 8128–8153, https://doi.org/10.3390/rs70608128, 2015.

Tomer, S., Al Bitar, A., Sekhar, M., Zribi, M., Bandyopadhyay, S., and Kerr, Y.: MAPSM: A Spatio-Temporal Algorithm for Merging Soil Moisture from Active and Passive Microwave Remote Sensing, Remote Sensing, 8, 990, https://doi.org/10.3390/rs8120990, 2016.

Wagner, W., Lindorfer, R., Melzer, T., Hahn, S., Bauer-Marschallinger, B., Morrison, K., Calvet, J.-C., Hobbs, S., Quast, R., Greimeister-Pfeil, I., and Vreugdenhil, M.: Widespread occurrence of anomalous C-band backscatter signals in arid environments caused by subsurface scattering, Remote Sensing of Environment, 276, 113025, https://doi.org/10.1016/j.rse.2022.113025, 2022.

Wagner, W., Noll, J., Borgeaud, M., and Rott, H.: Monitoring soil moisture over the Canadian Prairies with the ERS scatterometer, IEEE Trans. Geosci. Remote Sensing, 37, 206–216, https://doi.org/10.1109/36.739155, 1999.

Zribi, M., Chahbi, A., Shabou, M., Lili-Chabaane, Z., Duchemin, B., Baghdadi, N., Amri, R., and Chehbouni, A.: Soil surface moisture estimation over a semi-arid region using ENVISAT ASAR radar data for soil evaporation evaluation, Hydrol. Earth Syst. Sci., 15, 345–358, https://doi.org/10.5194/hess-15-345-2011, 2011.

Appendix A: Limitations of product SSM1km (rainfall events)

Figure A1: Data on 2020-05-10 of ERA5-land precipitation and SSM1km

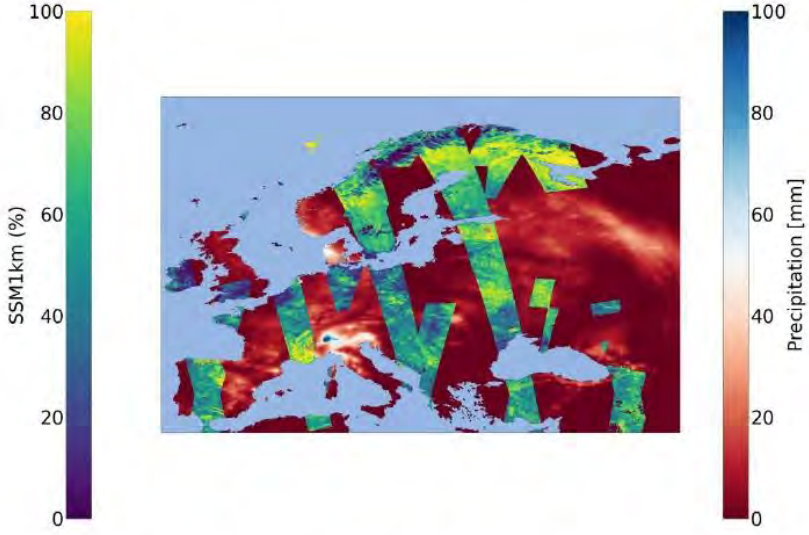
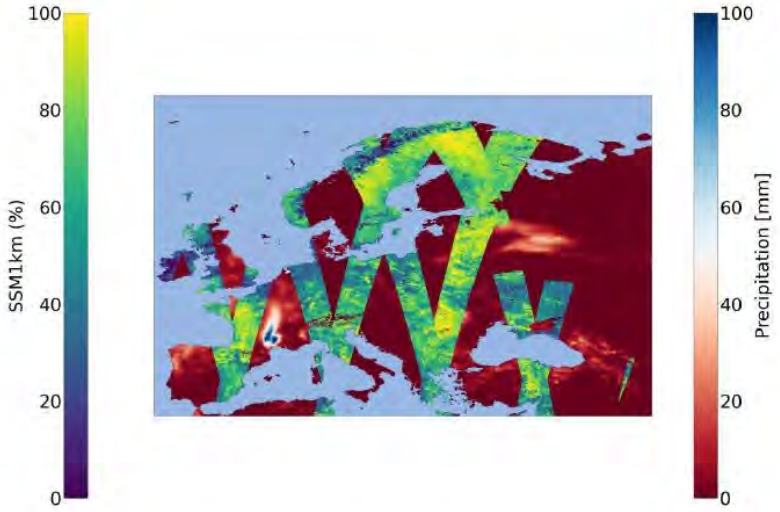


Figure A2: Data on 2020-06-04 of ERA5-land precipitation and SSM1km



839

840 Figure A3: Data on 2020-06-12 of ERA5-land precipitation and SSM1km

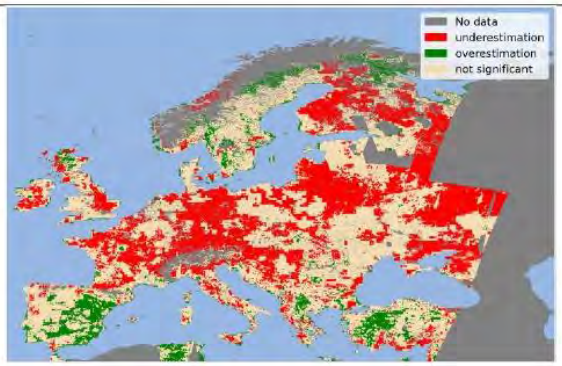
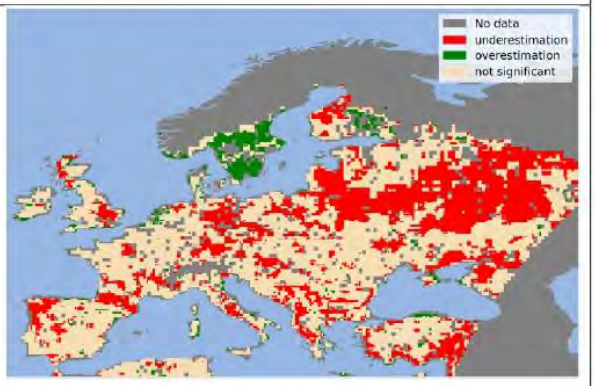
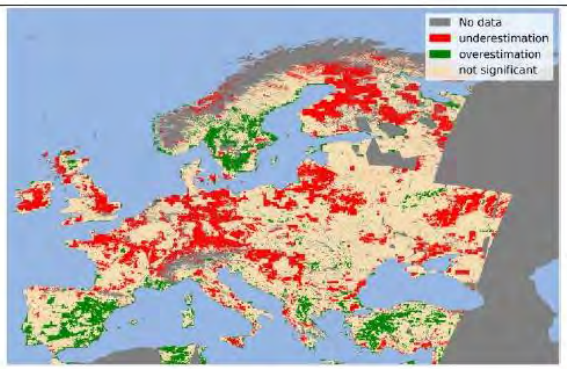
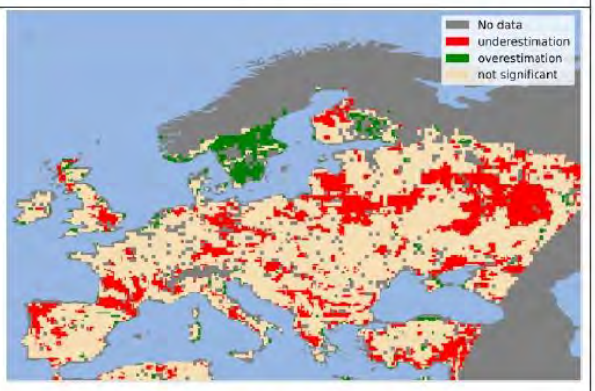
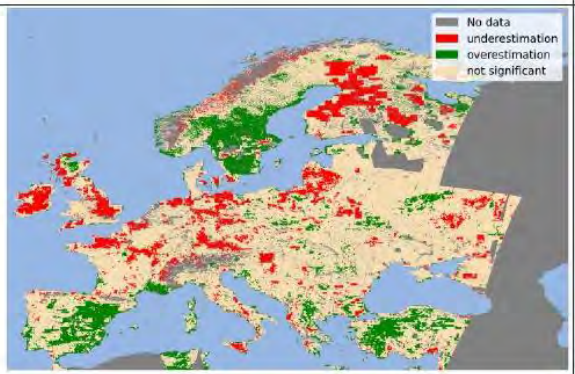
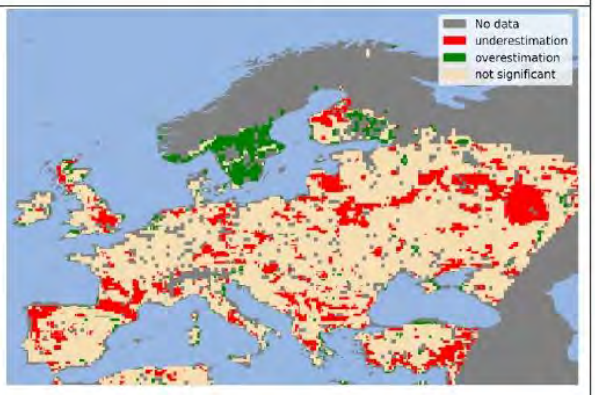
841 **APPENDIX B: CCI Land cover classes** (screenshot from CCI-LC_Maps_Legend document available at:
842 http://maps.elie.ucl.ac.be/CCI/viewer/, last access: 07 July 2022)

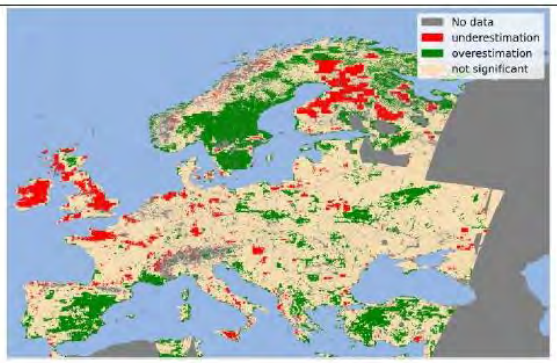
VALUE	LABEL	COLOR
0	No Data	
10	Cropland, rainfed	
11	Herbaceous cover	
12	Tree or shrub cover	
20	Cropland, irrigated or post-flooding	
30	Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	
40	Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	
50	Tree cover, broadleaved, evergreen, closed to open (>15%)	
60	Tree cover, broadleaved, deciduous, closed to open (>15%)	
61	Tree cover, broadleaved, deciduous, closed (>40%)	
62	Tree cover, broadleaved, deciduous, open (15-40%)	
70	Tree cover, needleleaved, evergreen, closed to open (>15%)	
71	Tree cover, needleleaved, evergreen, closed (>40%)	
72	Tree cover, needleleaved, evergreen, open (15-40%)	
80	Tree cover, needleleaved, deciduous, closed to open (>15%)	
81	Tree cover, needleleaved, deciduous, closed (>40%)	
82	Tree cover, needleleaved, deciduous, open (15-40%)	
90	Tree cover, mixed leaf type (broadleaved and needleleaved)	
100	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	
110	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	
120	Shrubland	
121	Evergreen shrubland	
122	Deciduous shrubland	
130	Grassland	
140	Lichens and mosses	
150	Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	
152	Sparse shrub (<15%)	
153	Sparse herbaceous cover (<15%)	
160	Tree cover, flooded, fresh or brakish water	
170	Tree cover, flooded, saline water	
180	Shrub or herbaceous cover, flooded, fresh/saline/brakish water	
190	Urban areas	
200	Bare areas	
201	Consolidated bare areas	
202	Unconsolidated bare areas	
210	Water bodies	
220	Permanent snow and ice	

843

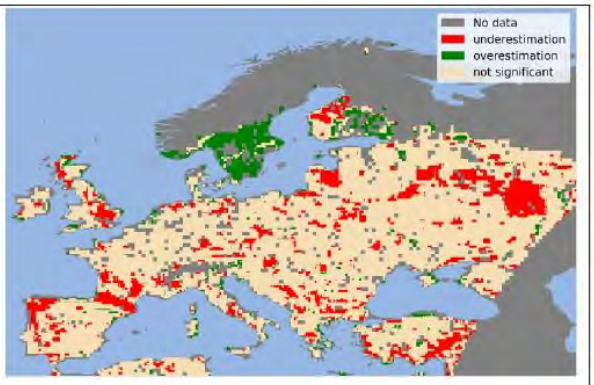
844

APPENDIX C: RZSM monthly difference maps

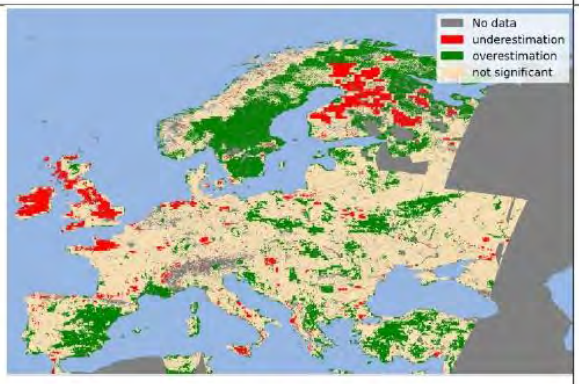
$RZSM_{ANN\_SSM1km} - RZSM_{ERA5}$	$RZSM_{ANN\_SMAP36km} - RZSM_{ERA5}$
 **04-2020**	 **04-2020**
 **05-2020**	 **05-2020**
 **06-2020**	 **06-2020**



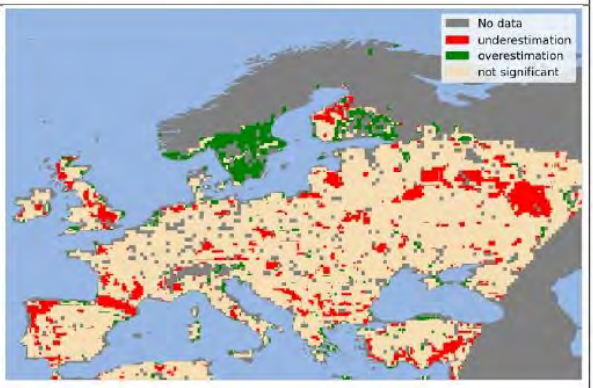
07-2020



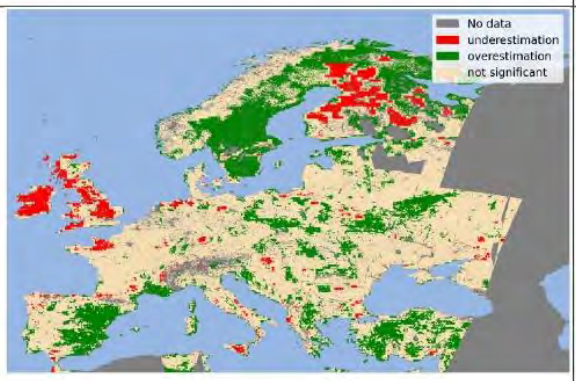
07-2020



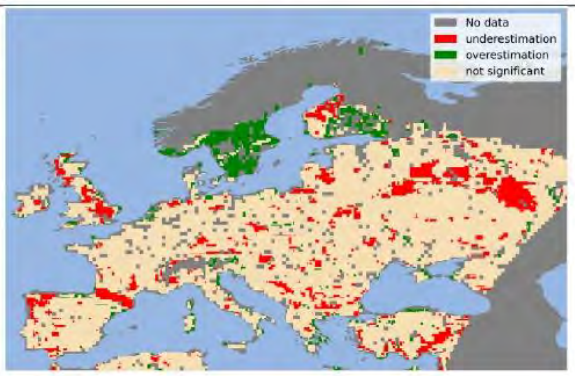
08-2020



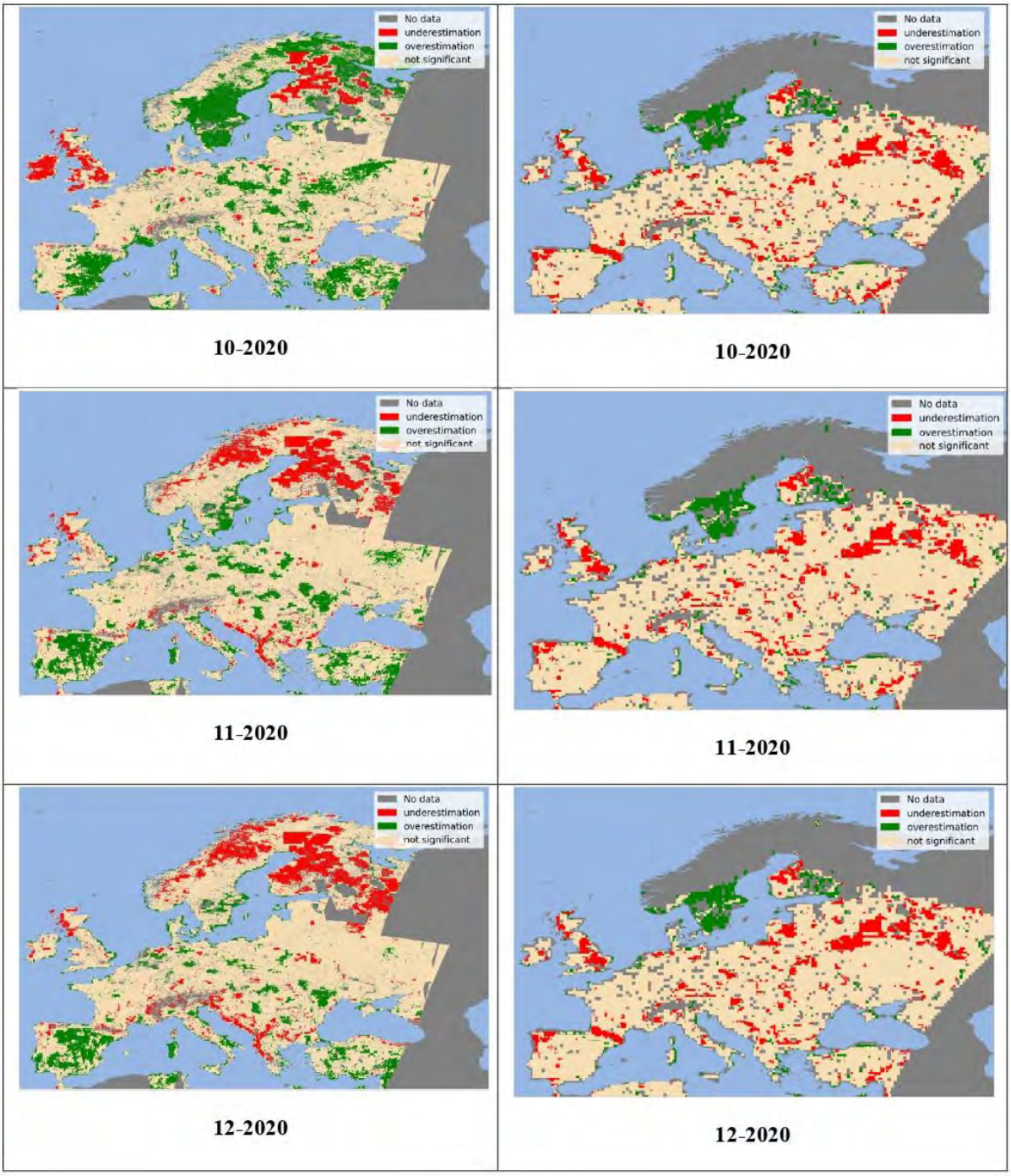
08-2020



09-2020



09-2020


No data
underestimation
overestimation
not significant
10-2020
10-2020
11-2020
11-2020
12-2020
12-2020

846

847

General conclusion and perspectives

Classical representations of the water cycle leave out the anthropogenic effects. Actually, only 15% of the water cycle diagrams depict human interaction with water (Abbott et al., 2019). Crop use of soil moisture is one form of human impact on the water cycle. In the context of agriculture, soil moisture is a key variable in crop health monitoring and yields prediction for instance. The second component of this variable, i.e. root-zone soil moisture is of considerable interest since plants draw water from the soil profile.

This PhD was mainly centered on the prediction of RZSM at large scales and subkilometric resolution. A first step consisted in predicting RZSM while linking it to surface moisture based on a large database from the global database ISMN. This reasoning relies on the fact that both soil moisture components are interconnected through diffusion processes for instance. At this stage, an ANN model was developed and trained on in-situ SSM information provided by the ISMN over soil moisture stations of different characteristics. Different experiments were conducted to determine the best ANN architecture. The accuracy and transferability of the approach were assessed as well as the contribution of each soil moisture network. Results demonstrated the reliability of the method since a median, maximum and minimum correlation values equal to 0.77, 0.96 and 0.65 were recorded after applying a statistical filtering that aimed at eliminating low-quality stations from training and validation. In some other contexts and more specifically in areas where a surface/subsurface decoupling phenomenon occurs such as in arid regions, the approach was proved lacking. In an effort to better capture these specific conditions and better account for more complex conditions, we considered the option of enriching the model with process-related variables. Based on the previous results, different ANN models were developed such that their features include SSM and a process-related variable. The most complex ANN model was fed with SSM features and a combination of process-related variables. The considered variables are SWI which was computed using a recursive exponential filter, evaporation efficiency and NDVI. The different models were trained on the stations identified of good-quality in the first axis which cover broad contexts in terms of types of soil and climate. Each model was tested on an ensemble of ISMN stations not seen in the training and validation steps. Analysis across climate classes showed that the using more features besides SSM could boost the performance of the approach. More specifically, the consideration of evaporation efficiency in areas of high evaporation rates was proved beneficial. Over agricultural fields and transition zones like the Sahel zone or the Australian transition zone, NDVI was revealed the most relevant variable for RZSM prediction. The most reliable model was the most complex model in which RZSM is predicted based on three SSM features and three process-related variables namely NDVI, evaporation efficiency and SWI. The robustness of the methodology was also evaluated through additional tests applied on stations which are not covered by the ISMN database namely stations over central Tunisia, India and Italy. The quality of predictions increased significantly in the case of Tunisia when the most complex ANN model was used. However, performance slightly changed over India and Italy potentially due to the presence of clouds and the crop heterogeneity that affect the MODIS products.

Since local-scale RZSM predictions are not relevant with regards to agricultural applications, we extended the application at a large scale. The most complex ANN model which was

developed in the second axis, was used in prediction mode in order to produce large scale maps of RZSM at 1 km spatial resolution. The model was tested over continental Europe using remote sensing features. Different SSM remote sensing products were employed to compute the three SSM features required by the model. Maps of RZSM at 1 km spatial resolution were generated using C-band Sentinel-1 SSM product with 1 km spatial resolution and temporal repetitiveness of 2-4 days, depending on location, since the launch of Sentinel-1B in October 2016 and until its shutdown in December 2021 (SSM1km). The locally-trained neural network model was able to track RZSM variations with a reasonable accuracy over several areas of Europe. The predictions were validated through a comparison against the ERA5-Land reanalysis RZSM product and a comparison against in-situ data. Moreover, the use of rolling averages of SSM as inputs to the model helped us overcome the input data gaps problem and produce daily RZSM maps.

Nevertheless, the SSM1km product exhibited some limitations that impacted the accuracy of RZSM predictions. Additional RZSM maps were produced by the ANN model at different spatial resolutions based on the considered SSM information source. Although most of the limitations raised by the use of the SSM1km product were satisfactorily addressed with the L-band SMAP product which is provided with a 36km spatial resolution and 3-day temporal resolution, the predictions were not as accurate when compared with in-situ RZSM information. The model was also applied using ERA5-land reanalysis SSM datasets and outperformed the ANN models based on remotely sensed SSM datasets. It also improved on the accuracy of the ERA5-land reanalysis RZSM datasets which has an hourly native temporal resolution.

Throughout this PhD, we demonstrated the feasibility of mapping RZSM at large scales and subkilometric resolutions using a locally-trained ANN model. This step provides the foundation for the feasibility of RZSM prediction at global scale and subkilometric resolution. In the following, we present perspectives for improvement of the approach. The perspectives can be related to the choice of the method, data, etc. In this PhD, we chose to work with a machine learning method and more precisely artificial neural networks because of their reliability and promising capabilities, compared to physical methods or data assimilation as described in the state of the art chapter. Moreover, the innovation of this work as previously explained, resides among others in the logic and the chronology adopted to solve the problem, i.e. to locally train a method based on data, to enrich it by physics and to complement it by a spatial component in order to solve a physical problem. From a model improvement perspective, other model tuning options can be explored such as using a custom cost function to optimize the model. A specific cost function can be elaborated to take into account the specifics of soil moisture time series such as moisture peaks that sometimes could not be identified with a standard cost function. In some cases, comparison of the in-situ RZSM time series with the predicted RZSM time series showed that the model missed peaks or on the contrary, generated false peaks. A physics-guided cost function can also be interesting if it assigns a specific weight for dry and wet cycles for instance.

Another avenue of improvement can be explored with respect to the choice of the process-related variables that were included in the model. The impact of adding additional temporal

information, such as precipitation or irrigation events, can be investigated in future studies. Spatial information, namely soil texture information, can be considered as well. However, the inaccuracies affecting these variables make them a source of bias for the model. In this context, machine learning or deep learning methods can be used as calibration tools for low precision data or as gap filling tools for missing data. The choice of the type of ANN can be also questioned in order to include these static variables in addition to temporal dynamic variables. Actually, MLPs are not suitable for a joint use of static and dynamic variables since all temporal interdependencies may be lost. Hybrid ANNs which combine an MLP for static variables and Long Short Term Memory (LSTM) for dynamic variables can be a promising alternative.

Another perspective as far this second axis is concerned, consists in the separation of the data into clusters of the same climate class. Specific trainings can be done based on the climate type. This aims at potentially clarifying the relations governing the variables used and the climate class.

When it comes to the last research axis, higher quality remote sensing SSM products can be used to further emphasize the impact of input information on RZSM predictions. The launch of additional satellites in the Sentinel-1 constellation is promising. The unavailability of Sentinel-1B due to a technical anomaly since December 2021 makes the launch of the new Sentinel-1C satellite a highly anticipated event. Actually, Sentinel-1C is the third satellite of the Sentinel-1 radar constellation and will be launched in 2023. It is equipped with C-band radar and will ensure global data continuity as part of the Copernicus program.

The launch of other L-band missions is also promising. Radar Observing System for Europe L (ROSE-L) is a future mission developed under the European Union's Copernicus program. The improved penetration capability of the L-band SAR flown on board of this satellite will improve the capability of soil moisture retrieval. This mission will contribute to the continuity of Copernicus observations on a global scale, for example by improving their accuracy, product quality, temporal and spatial resolution of the collected data. Also, the NASA ISRO Synthetic Aperture Radar (NISAR) mission that will be launched in 2023 is quite promising for high resolution soil moisture observations. This mission will provide data on the dynamics of the Earth's surface on a global scale. These data are essential for various Earth science disciplines such as carbon and water cycle observations. One of the objectives of this mission is to provide global soil moisture products with a spatial resolution of 200m and a temporal repetitiveness of 6 days.

Always in the context of remote sensing, the future potential of extracting RZSM via P-band sensors, such as the BIOMASS mission scheduled for launch in 2024, could provide a better understanding and quantification of soil moisture in deep layers.

Conclusion générale et perspectives (français)

Les représentations classiques du cycle de l'eau font généralement table rase des effets anthropiques. En effet, seuls 15 % des diagrammes du cycle de l'eau décrivent l'interaction humaine avec l'eau (Abbott et al., 2019). L'utilisation de l'humidité du sol pour les cultures est l'une des formes de l'impact humain sur le cycle de l'eau. Dans un contexte agricole, l'humidité du sol est une variable clé pour suivre la santé des cultures et la prédire les rendements par exemple. La composante de la zone racinaire présente un grand intérêt puisque les plantes puisent l'eau dans le profil du sol.

Cette thèse est principalement centrée sur la prédiction de la RZSM à grande échelle et à résolution kilométrique. Une première étape a consisté à prédire la RZSM tout en la reliant à l'humidité de surface aux RZSM en se basant sur une importante base de données provenant du réseau mondial ISMN. . Ce raisonnement repose sur le fait que les deux composantes de l'humidité du sol sont interconnectées par des processus de diffusion par exemple. À ce stade, un modèle ANN a été développé et entraîné sur les données in-situ de la SSM fournies par l'ISMN sur des stations d'humidité du sol de différentes caractéristiques. Différentes configurations ont été appliquées pour obtenir le modèle ANN le plus performant. La précision et la transférabilité de l'approche ont été évaluées, ainsi que la contribution de chaque réseau d'humidité du sol. Les résultats ont démontré une forte fiabilité de la méthode sous certaines conditions. Des valeurs médiane, maximale et minimale de corrélation égales à 0.77, 0.96 et 0.65 ont été enregistrées après avoir appliqué une méthode de filtrage de données qui visait à éliminer les stations de mauvaise qualité des opérations d'apprentissage et de test. Dans d'autres contextes, plus spécifiquement là où il y a des phénomènes de découplage entre surface et zone racinaire tel qu'en zones arides, l'approche s'est avérée insuffisante. Dans l'optique de mieux cerner ces cas particuliers et d'une meilleure prise en compte des cas plus complexes, nous avons considéré l'option d'enrichir le modèle avec des variables supplémentaires représentant les processus physiques ayant un fort impact sur l'humidité du sol améliorerait les prédictions RZSM, là où le modèle échoue.

En se basant sur les résultats précédents, différents modèles ANN ont été développés de telle sorte que leurs variables d'entrée incluent la SSM et une variable liée à un processus physique. Le modèle ANN le plus complexe est formé de données SSM et une combinaison de variables liées aux processus. Les variables considérées sont le SWI calculé à l'aide d'un filtre exponentiel récursif, l'efficacité d'évaporation calculée à l'aide d'un produit PET issu de la télédétection et de données de NDVI. Les différents modèles ont été entraînés sur les stations identifiées de bonne qualité dans la partie précédente. De même, ces stations couvrent des contextes très larges en termes de caractéristiques de sol et de type de climat. Chaque modèle a été testé sur un ensemble de stations ISMN non utilisées lors du processus d'apprentissage du modèle. L'analyse par classe climatique a montré que l'utilisation d'autres variables en plus de la SSM pouvait améliorer les performances de la méthode. Plus précisément, la prise en compte de l'efficacité de l'évaporation dans les zones à fort taux d'évaporation s'est avérée bénéfique. Sur les champs agricoles, le NDVI s'est révélé être la variable liée aux processus la plus pertinente pour la prédiction de la RZSM. Le modèle le plus fiable s'est avéré être le modèle le plus complexe dans lequel la RZSM est prédite en se basant sur la SSM et les trois variables liées aux processus à savoir le NDVI, l'efficacité

d'évaporation et le SWI. La robustesse de la méthodologie a été également évaluée à travers des tests supplémentaires appliqués sur stations externes à la base de données ISMN. Des stations situées au centre de la Tunisie, en Inde et en Italie ont été considérées. La qualité des prédictions a augmenté de manière significative dans le cas de la Tunisie lorsque le modèle ANN le plus complexe a été utilisé. Cependant, les performances ont légèrement évolué en Inde et en Italie, potentiellement à cause de la présence de nuages et de l'hétérogénéité des cultures qui affectent les produits MODIS.

Les prévisions RZSM à l'échelle locale n'étant pas pertinentes pour des applications agricoles, nous avons étendu l'application à large échelle. Le modèle ANN le plus complexe, précédemment décrit, a été utilisé en mode prédiction afin de produire des cartes à grande échelle de RZSM à une résolution spatiale de 1 km. Le modèle a été testé sur l'Europe continentale en utilisant en entrée des variables issues de la télédétection. Différents produits SSM issus de la télédétection ont été considérés. Des cartes de RZSM à une résolution spatiale de 1 km ont été produites à l'aide d'un produit SSM en bande C (Sentinel-1) de 1 km de résolution spatiale et d'une répétitivité temporelle de 2 à 4 jours, selon la localisation, depuis le lancement de Sentinel-1B en Octobre 2016 et jusqu'à son arrêt en décembre 2021 (SSM1km). Ceci constitue l'objectif ultime de ce travail. Le modèle de réseaux de neurones localement entraîner a pu suivre les variations de l'humidité en zone racinaire avec des statistiques raisonnables sur plusieurs zones de l'Europe. Les prédictions ont été validées suite à une comparaison avec le produit de réanalyse de ERA5-Land et avec une comparaison par rapport aux données in-situ. De plus, le choix d'utiliser des moyennes glissantes d'humidité de surface en entrée du modèle nous a permis de s'affranchir du problème des trous de données du produit en entrée et de produire un des cartes journalières d'humidité de zone racinaire.

Néanmoins, le produit SSM1km a présenté certaines limitations qui ont eu des répercussions sur la précision des prédictions RZSM et nous a donc conduit à explorer d'autres produits SSM. Ainsi, des cartes RZSM supplémentaires ont été produites par le modèle ANN à des résolutions spatiales qui dépendent de la source de SSM. Bien que la plupart des limitations soulevées par l'utilisation du produit SSM1km aient été traitées de manière satisfaisante avec le produit SMAP en bande L de résolution spatiale égale à 36km et de résolution temporelle égale à 3 jours, les prédictions n'étaient pas aussi précises lorsqu'elles étaient comparées aux données RZSM in-situ. Le modèle a également été appliqué sur des jeux de données SSM de réanalyse ERA5-Land et a marché mieux que les modèles ANN basés sur les jeux de données SSM issus de la télédétection. Sa précision a été également supérieure à celle du produit RZSM de réanalyse ERA5-Land de résolution temporelle native horaire et dont les moyennes journalières ont été utilisées dans cette étude.

Tout au long de cette thèse, nous avons démontré la faisabilité d'un modèle ANN pour cartographier la RZSM à grande échelle et à une résolution kilométrique lorsque certaines considérations sont prises en compte. Cette étape constitue une base pour confirmer la faisabilité de la prédiction du RZSM à l'échelle globale et à une résolution kilométrique à l'aide d'un modèle ANN.

Dans la suite, nous présentons des pistes d'amélioration de l'approche. Les perspectives peuvent viser le choix de la méthode, le traitement de données, les produits utilisés, etc. Durant cette thèse, on a fait le choix de travailler avec une méthode de machine learning et plus particulièrement les réseaux de neurones artificiels vu leur fiabilité par rapport aux méthodes physiques et d'assimilation de données comme décrit dans le chapitre état de l'art. De plus, l'innovation de ce travail, comme précédemment expliqué, réside entre autres dans la logique et la chronologie adoptées pour résoudre le problème à savoir entraîner localement une méthode basée sur les données, l'enrichir par la physique et la complémenter par une composante spatiale pour enfin résoudre un problème physique. Dans une optique d'amélioration de modèle, on peut explorer d'autres options de réglage du modèle comme par exemple l'utilisation d'une fonction de coût personnalisée pour optimiser le modèle en fonction des spécificités des séries temporelles d'humidité du sol comme les pics d'humidité qui n'ont pas pu parfois être identifiés avec une fonction de coût standard. Dans certains cas, la comparaison entre les séries RZSM in-situ et les séries prédites de RZSM par le modèle ANN basé uniquement sur la SSM, a montré que le modèle a loupé des pics ou, au contraire, a généré des faux pics. Une fonction de coût guidée par la physique peut également être intéressante si elle attribue un poids spécifique aux cycles secs et humides par exemple.

Une autre piste de recherche peut être explorée par rapport au choix des variables reliées aux processus physiques qui ont été incluses pour enrichir le modèle. L'impact de l'ajout d'informations temporelles supplémentaires, comme la précipitation ou les événements d'irrigation, peut être investigué dans des études futures. Des informations spatiales, à savoir des informations sur la texture du sol, peuvent être considérées aussi. Cependant, les contraintes de précision affectant ces variables en font une source de biais pour le modèle. Dans ce contexte, des méthodes de machine learning ou de deep learning peuvent être utilisés comme outils de calibration des données peu précises ou des outils de « gap filling » pour les données manquantes ou discontinues. Le choix du type d'ANN peut également être remis en question afin d'inclure ces variables statiques en plus des variables dynamiques temporelles. En fait, les MLPs ne conviennent pas à une utilisation conjointe de variables statiques et dynamiques, car toutes les interdépendances temporelles peuvent être perdues. Les ANN hybrides qui combinent un MLP pour les variables statiques et une mémoire à long terme (LSTM) pour les variables dynamiques peuvent être une alternative prometteuse.

Une autre perspective par rapport à ce deuxième axe est la séparation des différents ensembles de données en clusters d'une même classe climatique. Des entraînements spécifiques peuvent être effectués par type de climat pour clarifier potentiellement les relations régissant les variables utilisées et la classe climatique.

En ce qui concerne le dernier axe de recherche, des produits SSM issus de la télédétection de meilleure qualité peuvent être utilisés pour souligner davantage l'impact de la qualité des informations d'entrée sur les prédictions RZSM. Le lancement d'autres satellites de la constellation Sentinel-1 est également prometteur. L'indisponibilité de Sentinel-1B suite à une anomalie technique depuis décembre 2021 fait que le lancement et la mise en orbite du nouveau satellite Sentinel-1C est un évènement très attendu. Sentinel-1C est le troisième satellite de la constellation radar Sentinel-1 et sera lancé en 2023. Il est équipé d'un radar en

bande C et assurera la continuité des données à l'échelle globale dans le cadre du programme Copernicus.

Le lancement d'autres missions en bande L est aussi prometteur. Radar Observing System for Europe L (ROSE-L) est une future mission développée dans le cadre du programme Copernicus de l'Union Européenne. Le satellite disposera d'un SAR en bande L. La meilleure capacité de pénétration de la bande L améliorera la capacité d'acquisition de l'humidité de sol. Cette mission contribuera à la continuité des observations Copernicus à l'échelle globale, par exemple en améliorant leur précision, la qualité des produits, la résolution temporelle et spatiale des données collectées. Aussi, la mission NASA ISRO Synthetic Aperture Radar (NISAR) qui sera lancée en 2023 est prometteuse en matière d'observation d'humidité du sol à haute résolution. Cette mission fournira des données sur la dynamique de la surface terrestre à l'échelle globale. Ces données sont essentielles pour différentes disciplines des sciences de la Terre telles que les observations des cycles du carbone et de l'eau. Un des objectifs de cette mission est de fournir des produits globaux d'humidité de sol à une résolution spatiale de 200m et une répétitivité temporelle de 6 jours.

Toujours dans un contexte de télédétection, le potentiel futur d'extraction de la RZSM via des capteurs en bande P, telle que la mission BIOMASS dont le lancement est prévu 2024, pourrait assurer une meilleure compréhension et quantification de l'humidité du sol dans les couches profondes.

Acronyms

AI: Artificial Intelligence

AMSR-E: Advanced Microwave Scanning Radiometer–Earth Observing System

ANN: Artificial Neural Network

ANR: Agence Nationale de la Recherche

ASCAT: Advanced Scatterometer

ASCE: American Society of Civil Engineers

ATI: Apparent Thermal Inertia

CDF: Cumulative Density Function

CESBIO: Centre d'Etudes Spatiales de la BIOsphère

CLM: Community Land Model

CNES: Centre National d'Etudes Spatiales

CNN: Convolutional Neural Network

CNRS: Centre National de la Recherche Scientifique

CLSM: Catchment Land Surface Model

DL: Deep Learning

ECV: Essential Climate Variable

EKF: Extended Kalman Filter

EnKF: Ensemble Kalman Filter

EnPF: Ensemble Particle filter

ERS: European Remote-Sensing

ESA: European Space Agency

ET: EvapoTranspiration

FAO: Food and Agriculture Organization

FDR: Frequency Domain Reflectometry

FF: FeedForward

GLDAS: Global Land Data Assimilation System

GNSS: Global Navigation Satellite Systems

GPS: Global Positioning System

GRU: Gated Recurrent Unit

HESS: Hydrology and Earth System Sciences

IEM: Integral Equation Model

ISBA: Interaction Sol-Biosphère-Atmosphère

ISMN: International Soil Moisture Network

JULES: Joint UK Land Environment Simulator

KF: Kalman Filter

LDAS: Land Data Assimilation System

LRN: Local Response Normalization

LSM: Land Surface Model

LST: Land Surface Temperature

LSTM: Long Short-Term Memory

MARMIT: MultilAyer Radiative transfer Model of soIl reflectance

MetOp: Meteorological operational satellite

ML: Machine Learning

MMSCA: MinMax SCAling

MODIS: Moderate-Resolution Imaging Spectroradiometer

NASA: National Aeronautics and Space Administration

NDVI: Normalized Difference Vegetation Index

NSMI: Normalized Soil Moisture Index

PET: Potential EvapoTranspiration

RF: Random Forest

RMSE: Root Mean Square Error

RNN: Recurrent Neural Network

RS: Remote Sensing

RZSM: Root-Zone Soil Moisture

RZWQM: Root Zone Water Quality Model

S1: Sentinel-1

SAR: Synthetic Aperture Radar

SASI: Shortwave Angle Slope Index

SDG: Sustainable Development Goal

SM: Soil moisture

SMAP: Soil Moisture Active Passive

SMAR: Soil Moisture Analytical Relationship

SMOS: Soil Moisture and Ocean Salinity

SMOSREX: Surface Monitoring Of the Soil Reservoir Experiment

SNR: Signal-to-noise ratio

SSCA: Standard SCAling

SSM: Surface Soil Moisture

SURFEX: SURface Externalisée

SVAT: Soil–Vegetation–Atmosphere Transfer

SVM: Support Vector Machine

SWI: Soil Water Index

SWIR: Short-Wave Infrared

TB: Brightness Temperature

TDR: Time Domain Reflectometry

TI: Thermal Inertia

TIR: Thermal Infrared

UbRMSE: Unbiased Root Mean Square Error

WASAG: Global Framework *on* Water Scarcity in Agriculture

WCM: Water Cloud Model

References

Abdullah, N. H. H., Kuan, N. W., Ibrahim, A., Ismail, B. N., Majid, M. R. A., Ramli, R., and Mansor, N. S.: Determination of soil water content using time domain reflectometer (TDR) for clayey soil, ADVANCES IN CIVIL ENGINEERING AND SCIENCE TECHNOLOGY, Penang, Malaysia, 020016, https://doi.org/10.1063/1.5062642, 2018.

Abbott, B. W., Bishop, K., Zarnetske, J. P., Minaudo, C., Chapin, F. S., Krause, S., Hannah, D. M., Conner, L., Ellison, D., Godsey, S. E., Plont, S., Marçais, J., Kolbe, T., Huebner, A., Frei, R. J., Hampton, T., Gu, S., Buhman, M., Sara Sayedi, S., Ursache, O., Chapin, M., Henderson, K. D., and Pinay, G.: Human domination of the global water cycle absent from depictions and perceptions, Nat. Geosci., 12, 533–540, https://doi.org/10.1038/s41561-019-0374-y, 2019.

Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., and Martin, E.: From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations, Hydrology and Earth System Sciences, 12, 1323–1337, https://doi.org/10.5194/hess-12-1323-2008, 2008.

Albers, M. A., Dobos, R. R., and Robotham, M. P.: User Guide for the National Commodity Crop Productivity Index (NCCPI), Version 3.0, Available online: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/16/nrcs143_020559.pdf (accessed on 28 July 2022)

Al-Mukhtar, M.: Modelling the root zone soil moisture using artificial neural networks, a case study, Environ Earth Sci, 75, 1124, https://doi.org/10.1007/s12665-016-5929-2, 2016.

Ångström, A. : The Albedo of Various Surfaces of Ground, Geografiska Annaler, 7, 323, https://doi.org/10.2307/519495, 1925.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology: Artificial Neural Networks in Hydrology. II: Hydrologic Applications, J. Hydrol. Eng., 5, 124–137, https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(124), 2000.

Awad, M. and Khanna, R.: Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4302-5990-9, 2015.

A, Y., Wang, G., Hu, P., Lai, X., Xue, B., and Fang, Q.: Root-zone soil moisture estimation based on remote sensing data and deep learning, Environmental Research, 212, 113278, https://doi.org/10.1016/j.envres.2022.113278, 2022.

Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H., and Tuller, M.: Ground, Proximal, and Satellite Remote Sensing of Soil Moisture, Rev. Geophys., 57, 530–616, https://doi.org/10.1029/2018RG000618, 2019.

Bablet, A., Viallefont-Robinet, F., Jacquemoud, S., Fabre, S., and Briottet, X.: High-resolution mapping of in-depth soil moisture content through a laboratory experiment coupling a spectroradiometer and two hyperspectral cameras, Remote Sensing of Environment, 236, 111533, https://doi.org/10.1016/j.rse.2019.111533, 2020.

Bablet, A., Vu, P. V. H., Jacquemoud, S., Viallefont-Robinet, F., Fabre, S., Briottet, X., Sadeghi, M., Whiting, M. L., Baret, F., and Tian, J.: MARMIT: A multilayer radiative transfer model of soil reflectance to estimate surface soil moisture content in the solar domain (400–2500 nm), Remote Sensing of Environment, 217, 1–17, https://doi.org/10.1016/j.rse.2018.07.031, 2018.

Bach, H. and Mauser, W.: Modelling and model verification of the spectral reflectance of soils under varying moisture conditions, in: Proceedings of IGARSS '94 - 1994 IEEE International Geoscience and Remote Sensing Symposium, IGARSS '94 - 1994 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 2354–2356, https://doi.org/10.1109/IGARSS.1994.399735, 1994.

Baghdadi, N., Aubert, M., and Zribi, M.: Use of TerraSAR-X Data to Retrieve Soil Moisture Over Bare Soil Agricultural Fields, IEEE Geosci. Remote Sensing Lett., 9, 512–516, https://doi.org/10.1109/LGRS.2011.2173155, 2012.

Barrett, B., Dwyer, E., and Whelan, P.: Soil Moisture Retrieval from Active Spaceborne Microwave Observations: An Evaluation of Current Techniques, Remote Sensing, 1, 210–242, https://doi.org/10.3390/rs1030210, 2009.

Belcher, D. J., Cuykendall, T. R., and Sack, H. S.: Measurement of Soil Moisture and Density By Neutron and Gamma-ray Scattering. 1950.

Ben-Dor, E. and Banin, A.: Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties, Soil Science Soc of Amer J, 59, 364–372, https://doi.org/10.2136/sssaj1995.03615995005900020014x, 1995.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, Geosci. Model Dev., 4, 677–699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Bindlish, R., Jackson, T., Cosh, M., Tianjie Zhao, and O'Neill, P.: Global Soil Moisture From the Aquarius/SAC-D Satellite: Description and Initial Assessment, IEEE Geosci. Remote Sensing Lett., 12, 923–927, https://doi.org/10.1109/LGRS.2014.2364151, 2015.

Bishop, C. M.: Pattern Recognition and Machine Learning, Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009)., Springer New York, New York, NY, 738 pp., 2016.

Bollegala, D.: Dynamic feature scaling for online learning of binary classifiers, Knowledge-Based Systems, 129, 97–105, https://doi.org/10.1016/j.knosys.2017.05.010, 2017.

Bordoni, M., Bittelli, M., Valentino, R., Chersich, S., Persichillo, M. G., and Meisina, C.: Soil Water Content Estimated by Support Vector Machine for the Assessment of Shallow Landslides Triggering: the Role of Antecedent Meteorological Conditions, Environ Model Assess, 23, 333–352, https://doi.org/10.1007/s10666-017-9586-y, 2018.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Bryant, R., Thoma, D., Moran, S., Holifield, C., Goodrich, D., Keefer, T., Paige, G., Williams, D., and Skirvin, S.: Evaluation of Hyperspectral, Infrared Temperature and Radar Measurements for Monitoring Surface Soil Moisture, First Interagency Conference on Research in the Watersheds : October 27-30, 2003.

Burnash, R.: The NWS River Forecast System-Catchment Modeling. In: Singh, V., Ed., Computer Models of Watershed Hydrology, Water Resources Publication, Colorado, 311-366, 1995.

Carlson, T.: An Overview of the "Triangle Method" for Estimating Surface Evapotranspiration and Soil Moisture from Satellite Imagery, Sensors, 7, 1612–1629, https://doi.org/10.3390/s7081612, 2007.

Carranza, C., Nolet, C., Pezij, M., and van der Ploeg, M.: Root zone soil moisture estimation with Random Forest, Journal of Hydrology, 593, 125840, https://doi.org/10.1016/j.jhydrol.2020.125840, 2021.

Ceballos, A., Scipal, K., Wagner, W., and Martínez-Fernández, J.: Validation of ERS scatterometer-derived soil moisture data in the central part of the Duero Basin, Spain, Hydrol. Process., 19, 1549–1566, https://doi.org/10.1002/hyp.5585, 2005.

Chanasyk, D. S. and Naeth, M. A.: Field measurement of soil moisture using neutron probes, Can. J. Soil. Sci., 76, 317–323, https://doi.org/10.4141/cjss96-038, 1996.

Chang, C.-W., Laird, D. A., Mausbach, M. J., and Hurburgh, C. R.: Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties, Soil Sci. Soc. Am. J., 65, 480–490, https://doi.org/10.2136/sssaj2001.652480x, 2001.

Cihlar, J. and Ulaby, F. T.: Dielectric properties of soils as a function of moisture content, 1974.

Claps, P. and Laguardia, G.: Assessing spatial variability of soil water content through thermal inertia and NDVI, Remote Sensing, Barcelona, Spain, 378, https://doi.org/10.1117/12.510984, 2004.

Crow, W. T., Kumar, S. V., and Bolten, J. D.: On the utility of land surface models for agricultural drought monitoring, Hydrol. Earth Syst. Sci., 16, 3451–3460, https://doi.org/10.5194/hess-16-3451-2012, 2012.

Crow, W. T. and Wood, E. F.: The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using Ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97, Advances in Water Resources, 26, 137–149, https://doi.org/10.1016/S0309-1708(02)00088-X, 2003.

Curran PJ. Principles of Remote Sensing. Longman Scientific and Tech., UK, 1985, 282.

da Silva, I. N., Andrade Flauzino, R., dos Reis Alves, S. F., Hernane Spatti, D., and Liboni, L. H. B.: Artificial Neural Networks: A Practical Course, 1st ed. 2017., Springer International Publishing : Imprint: Springer, Cham, 1 pp., 2017.

Datta, S., Taghvaeian, S., and Stivers, J. W.: Understanding Soil Water Content and Thresholds For Irrigation Management, https://doi.org/10.13140/RG.2.2.35535.89765, 2017.

De Jeu, R. A. M., Wagner, W., Holmes, T. R. H., Dolman, A. J., van de Giesen, N. C., and Friesen, J.: Global Soil Moisture Patterns Observed by Space Borne Microwave Radiometers and Scatterometers, Surv Geophys, 29, 399–420, https://doi.org/10.1007/s10712-008-9044-0, 2008.

De Lannoy, G. J. M. and Reichle, R. H.: Global Assimilation of Multiangle and Multipolarization SMOS Brightness Temperature Observations into the GEOS-5 Catchment Land Surface Model for Soil Moisture Estimation, Journal of Hydrometeorology, 17, 669–691, https://doi.org/10.1175/JHM-D-15-0037.1, 2016.

De Lannoy, G. J. M., de Rosnay, P., and Reichle, R. H.: Soil Moisture Data Assimilation, in: Handbook of Hydrometeorological Ensemble Forecasting, edited by: Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., and Schaake, J. C., Springer, Berlin, Heidelberg, 701–743, https://doi.org/10.1007/978-3-642-39925-1_32, 2019.

Dobriyal, P., Qureshi, A., Badola, R., and Hussain, S. A.: A review of the methods available for estimating soil moisture and its implications for water resource management, Journal of Hydrology, 458–459, 110–117, https://doi.org/10.1016/j.jhydrol.2012.06.021, 2012.

Dorigo, W. A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., and Jackson, T.: The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements, Hydrol. Earth Syst. Sci., 15, 1675–1698, https://doi.org/10.5194/hess-15-1675-2011, 2011.

Elshorbagy, A. and Parasuraman, K.: On the relevance of using artificial neural networks for estimating soil moisture content, Journal of Hydrology, 362, 1–18, https://doi.org/10.1016/j.jhydrol.2008.08.012, 2008.

Entekhabi, D., Nakamura, H., and Njoku, E. G.: Solving the inverse problem for soil moisture and temperature profiles by sequential assimilation of multifrequency remotely sensed observations, IEEE Trans. Geosci. Remote Sensing, 32, 438–448, https://doi.org/10.1109/36.295058, 1994.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., Spencer, M. W., Thurman, S. W., Tsang, L., and Van Zyl, J.: The Soil Moisture Active Passive (SMAP) Mission, IEEE, 2010.

Faridani, F., Farid, A., Ansari, H., and Manfreda, S.: A modified version of the SMAR model for estimating root-zone soil moisture from time-series of surface soil moisture, WSA, 43, 492, https://doi.org/10.4314/wsa.v43i3.14, 2017.

Farthing, M. W. and Ogden, F. L.: Numerical Solution of Richards' Equation: A Review of Advances and Challenges, Soil Sci. Soc. Am. j., 81, 1257–1269, https://doi.org/10.2136/sssaj2017.02.0058, 2017.

Feki, M., Ravazzani, G., Ceppi, A., Milleo, G., and Mancini, M.: Impact of Infiltration Process Modeling on Soil Water Content Simulations for Irrigation Management, Water, 10, 850, https://doi.org/10.3390/w10070850, 2018.

Ford, T. W., Harris, E., and Quiring, S. M.: Estimating root zone soil moisture using near-surface observations from SMOS, Hydrol. Earth Syst. Sci., 18, 139–154, https://doi.org/10.5194/hess-18-139-2014, 2014.

Francesca, V., Osvaldo, F., Stefano, P., and Paola, R. P.: Soil Moisture Measurements: Comparison of Instrumentation Performances, Journal of Irrigation and Drainage Engineering, 136, 81–89, https://doi.org/10.1061/(ASCE)0733-9437(2010)136:2(81), 2010.

Francois, C., Quesney, A., and Ottlé, C.: Sequential Assimilation of *ERS-1* SAR Data into a Coupled Land Surface–Hydrological Model Using an Extended Kalman Filter, J. Hydrometeor, 4, 473–487, https://doi.org/10.1175/1525-7541(2003)4<473:SAOESD>2.0.CO;2, 2003.

Gao, Z., Xu, X., Wang, J., Yang, H., Huang, W., and Feng, H.: A method of estimating soil moisture based on the linear decomposition of mixture pixels, Mathematical and Computer Modelling, 58, 606–613, https://doi.org/10.1016/j.mcm.2011.10.054, 2013.

GARDNER, W., and KIRKHAM, D.: Determination of soil moisture by neutron scattering. Soil Sci. 73, 391-401, 1952.

Grillakis, M. G., Koutroulis, A. G., Alexakis, D. D., Polykretis, C., and Daliakopoulos, I. N.: Regionalizing Root-Zone Soil Moisture Estimates From ESA CCI Soil Water Index Using Machine Learning and Information on Soil, Vegetation, and Climate, Water Res, 57, https://doi.org/10.1029/2020WR029249, 2021.

Han, E., Merwade, V., and Heathman, G. C.: Application of data assimilation with the Root Zone Water Quality Model for soil moisture profile estimation in the upper Cedar Creek, Indiana: RZWQM DATA ASSIMILATION WITH ENKF, Hydrol. Process., 26, 1707–1719, https://doi.org/10.1002/hyp.8292, 2012.

Hassan-Esfahani, L., Torres-Rua, A., Jensen, A., and McKee, M.: Assessment of Surface Soil Moisture Using High-Resolution Multi-Spectral Imagery and Artificial Neural Networks, Remote Sensing, 7, 2627–2646, https://doi.org/10.3390/rs70302627, 2015.

Haubrock, S. -N., Chabrillat, S., Lemmnitz, C., and Kaufmann, H.: Surface soil moisture quantification models from reflectance data under field conditions, International Journal of Remote Sensing, 29, 3–29, https://doi.org/10.1080/01431160701294695, 2008.

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

Hirschi, M., Davin, E. L., Schwingshackl, C., Wartenburger, R., Meier, R., Gudmundsson, L., and Seneviratne, S. I.: Soil moisture and evapotranspiration, ETH Zurich, https://doi.org/10.3929/ETHZ-B-000389455, 2020.

Hoekstra, P. and Delaney, A.: Dielectric properties of soils at UHF and microwave frequencies, J. Geophys. Res., 79, 1699–1708, https://doi.org/10.1029/JB079i011p01699, 1974.

Howell, T.A. Lysimetry. p. 379– 386. D. Hillel (ed.) Encyclopedia of soils in the environment. Elsevier, New York. 2004

IPCC, 2022: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 3056 pp., doi:10.1017/9781009325844.

Islam, S. and Engman, T.: Why bother for 0.0001% of Earth's water? Challenges for soil moisture research, Eos Trans. AGU, 77, 420, https://doi.org/10.1029/96EO00290, 1996.

Jackson, R. B., Canadell, J., Ehleringer, J. R., Mooney, H. A.,Sala, O. E., and Schulze, E. D.: A global analysis of root distributions for terrestrial biomes, Oecologia, 108, 389–411, doi:10.1007/BF00333714, 1996.

Jackson, T. J., Cosh, M. H., Bindlish, R., Starks, P. J., Bosch, D. D., Seyfried, M., Goodrich, D. C., Moran, M. S., and Du, J.: Validation of Advanced Microwave Scanning Radiometer Soil Moisture Products, IEEE Trans. Geosci. Remote Sensing, 48, 4256–4272, https://doi.org/10.1109/TGRS.2010.2051035, 2010.

Jackson, T. J., Le Vine, D. M., Swift, C. T., Schmugge, T. J., and Schiebe, F. R.: Large area mapping of soil moisture using the ESTAR passive microwave radiometer in Washita'92, Remote Sensing of Environment, 54, 27–37, https://doi.org/10.1016/0034-4257(95)00084-E, 1995.

Jackson, T. J. and Schmugge, T. J.: Passive microwave remote sensing system for soil moisture: some supporting research, IEEE Trans. Geosci. Remote Sensing, 27, 225–235, https://doi.org/10.1109/36.20301, 1989.

Janiesch, C., Zschech, P., and Heinrich, K.: Machine learning and deep learning, Electron Markets, 31, 685–695, https://doi.org/10.1007/s12525-021-00475-2, 2021.

Jarvis, N. J.: Simple physics-based models of compensatory plant water uptake: concepts and eco-hydrological consequences, Hydrol. Earth Syst. Sci., 15, 3431–3446, https://doi.org/10.5194/hess-15-3431-2011, 2011.

Jayalakshmi, T. and Santhakumaran, A.: Statistical Normalization and Back Propagationfor Classification, IJCTE, 89–93, https://doi.org/10.7763/IJCTE.2011.V3.288, 2011.

Kerr, Y. H., Al-Yaari, A., Rodriguez-Fernandez, N., Parrens, M., Molero, B., Leroux, D., Bircher, S., Mahmoodi, A., Mialon, A., Richaume, P., Delwart, S., Al Bitar, A., Pellarin, T., Bindlish, R., Jackson, T. J., Rüdiger, C., Waldteufel, P., Mecklenburg, S., and Wigneron, J.-P.: Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation, Remote Sensing of Environment, 180, 40–63, https://doi.org/10.1016/j.rse.2016.02.042, 2016.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.-J., Font, J., Reul, N., Gruhier, C., Juglea, S. E., Drinkwater, M. R., Hahne, A., Martín-Neira, M., and Mecklenburg, S.: The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle, Proc. IEEE, 98, 666–687, https://doi.org/10.1109/JPROC.2010.2043032, 2010.

KGS Pub. Inf. Circ. 22--Part 2 of 4: https://www.kgs.ku.edu/Publications/pic22/pic22_2.html, last access: 5 September 2022.

Khanna, S., Palacios-Orueta, A., Whiting, M. L., Ustin, S. L., Riaño, D., and Litago, J.: Development of angle indexes for soil moisture estimation, dry matter detection and land-cover discrimination, Remote Sensing of Environment, 109, 154–165, https://doi.org/10.1016/j.rse.2006.12.018, 2007.

Kolassa, J., Reichle, R. H., and Draper, C. S.: Merging active and passive microwave observations in soil moisture data assimilation, Remote Sensing of Environment, 191, 117–130, https://doi.org/10.1016/j.rse.2017.01.015, 2017.

Kornelsen, K. C. and Coulibaly, P.: Root-zone soil moisture estimation using data-driven methods, Water Resour. Res., 50, 2946–2962, https://doi.org/10.1002/2013WR014127, 2014.

Koster, R. D., Guo, Z., Yang, R., Dirmeyer, P. A., Mitchell, K., and Puma, M. J.: On the Nature of Soil Moisture in Land Surface Models, Journal of Climate, 22, 4322–4335, https://doi.org/10.1175/2009JCLI2832.1, 2009.

Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M., and Kumar, P.: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure, J. Geophys. Res., 105, 24809–24822, https://doi.org/10.1029/2000JD900327, 2000.

Krinner, G., et al. "A Dynamic Global Vegetation Model for Studies of the Coupled Atmosphere-Biosphere System: DVGM FOR COUPLED CLIMATE STUDIES." *Global Biogeochemical Cycles*, vol. 19, no. 1, 2005, https://doi.org/10.1029/2003GB002199.

Kubelka, P., & Munk, F. : Ein Beitrag zur Optik der Farbanstriche. Zeitschrift für Technische Physik, 12, 593– 601, 1931.

Kumar, S. V., Reichle, R. H., Koster, R. D., Crow, W. T., and Peters-Lidard, C. D.: Role of Subsurface Physics in the Assimilation of Surface Soil Moisture Observations, Journal of Hydrometeorology, 10, 1534–1547, https://doi.org/10.1175/2009JHM1134.1, 2009.

Lakshmi, V.: Remote Sensing of Soil Moisture, ISRN Soil Science, 2013, 1–33, https://doi.org/10.1155/2013/424178, 2013.

Larson, K. M., Small, E. E., Gutmann, E., Bilich, A., Axelrad, P., and Braun, J.: Using GPS multipath to measure soil moisture fluctuations: initial results, GPS Solut, 12, 173–177, https://doi.org/10.1007/s10291-007-0076-6, 2008.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Lecun, Y.: Generalization and network design strategies. Connectionism in perspective, 1989, vol. 19, no 143-155, p. 18.

Leenaars, J. G. B., Claessens, L., Heuvelink, G. B. M., Hengl, T., Ruiperez González, M., van Bussel, L. G. J., Guilpart, N., Yang, H., and Cassman, K. G.: Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa, Geoderma, 324, 18–36, https://doi.org/10.1016/j.geoderma.2018.02.046, 2018.

Le Moigne, P., Boone, A., Calvet, J.-C., Decharme, B., Faroux, S., Gibelin, A.-L., Lebeaupin, C., Mahfouf, J.-F., Martin, E., Masson, V., Mironov, D., Noilhan, J., Tulet, P., and Van Den Hurk, B.: SURFEX scientific documentation, Groupe de météorologie à moyenne échelle, note de centre, 87, 211 p., 2009.

Le Vine, D. M., Lagerloef, G. S. E., and Torrusio, S. E.: Aquarius and Remote Sensing of Sea Surface Salinity from Space, Proc. IEEE, 98, 688–703, https://doi.org/10.1109/JPROC.2010.2040550, 2010.

Lievens, H., Reichle, R. H., Liu, Q., De Lannoy, G. J. M., Dunbar, R. S., Kim, S. B., Das, N. N., Cosh, M., Walker, J. P., and Wagner, W.: Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates, Geophys. Res. Lett., 44, 6145–6153, https://doi.org/10.1002/2017GL073904, 2017.

Lievens, H. and Verhoest, N. E. C.: Spatial and temporal soil moisture estimation from RADARSAT-2 imagery over Flevoland, The Netherlands, Journal of Hydrology, 456–457, 44–56, https://doi.org/10.1016/j.jhydrol.2012.06.013, 2012.

Liu, H.-J., Zhang, Y.-Z., Zhang, X.-L., Zhang, B., Song, K.-S., Wang, Z.-M., and Tang, N.: Quantitative Analysis of Moisture Effect on Black Soil Reflectance, Pedosphere, 19, 532–540, https://doi.org/10.1016/S1002-0160(09)60146-6, 2009.

Liu, Z., Zhao, L., Peng, Y., Wang, G., and Hu, Y.: Improving Estimation of Soil Moisture Content Using a Modified Soil Thermal Inertia Model, Remote Sensing, 12, 1719, https://doi.org/10.3390/rs12111719, 2020.

Lü, H., Li, X., Yu, Z., Horton, R., Zhu, Y., Hao, Z., and Xiang, L.: Using a H∞ filter assimilation procedure to estimate root zone soil water content, Hydrol. Process., 24, 3648–3660, https://doi.org/10.1002/hyp.7778, 2010.

Makkink, G. F.: Limitations and perspectives of lysimeter research, Colloque de Hannoversch-Muenden 8 - 14 September 1959 = Symposium of Hannoversch-Muenden 8 - 14 September 1959, 13–25, 1959.

Manfreda, S., Brocca, L., Moramarco, T., Melone, F., and Sheffield, J.: A physically based approach for the estimation of root-zone soil moisture from surface measurements, Hydrol. Earth Syst. Sci., 18, 1199–1212, https://doi.org/10.5194/hess-18-1199-2014, 2014.

Mishra, S. K., Tyagi, J. V., and Singh, V. P.: Comparison of infiltration models, Hydrol. Process., 17, 2629–2652, https://doi.org/10.1002/hyp.1257, 2003.

Mitchell, T. M.: Machine Learning, McGraw-Hill, New York, 414 pp., 1997.

Moran, M. S., Clarke, T. R., Inoue, Y., and Vidal, A.: Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index, Remote Sensing of Environment, 49, 246–263, https://doi.org/10.1016/0034-4257(94)90020-5, 1994.

Musters, P. A. D. and Bouten, W.: Assessing rooting depths of an austrian pine stand by inverse modeling soil water content maps, Water Resour. Res., 35, 3041–3048, https://doi.org/10.1029/1999WR900173, 1999.

Njoku, E. G. and Entekhabi, D.: Passive microwave remote sensing of soil moisture, Journal of Hydrology, 184, 101–129, https://doi.org/10.1016/0022-1694(95)02970-2, 1996.

Njoku, E. G. and Kong, J.-A.: Theory for passive microwave remote sensing of near-surface soil moisture, J. Geophys. Res., 82, 3108–3118, https://doi.org/10.1029/JB082i020p03108, 1977.

Ottlé, C. and Vidal-Madjar, D.: Assimilation of soil moisture inferred from infrared remote sensing in a hydrological model over the HAPEX-MOBILHY region, Journal of Hydrology, 158, 241–264, https://doi.org/10.1016/0022-1694(94)90056-6, 1994.

Paloscia, S., Pettinato, S., Santi, E., Notarnicola, C., Pasolli, L., and Reppucci, A.: Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation, Remote Sensing of Environment, 134, 234–248, https://doi.org/10.1016/j.rse.2013.02.027, 2013.

Pan, X., Kornelsen, K. C., and Coulibaly, P.: Estimating Root Zone Soil Moisture at Continental Scale Using Neural Networks, J Am Water Resour Assoc, 53, 220–237, https://doi.org/10.1111/1752-1688.12491, 2017.

Petropoulos, G. P., Ireland, G., and Barrett, B.: Surface soil moisture retrievals from remote sensing: Current status, products & future trends, Physics and Chemistry of the Earth, Parts A/B/C, 83–84, 36–56, https://doi.org/10.1016/j.pce.2015.02.009, 2015.

Philpot, W.: Spectral Reflectance of Wetted Soils, https://doi.org/10.13140/2.1.2306.0169, 2010.

Pratt, D. A. and Ellyett, C. D.: The thermal inertia approach to mapping of soil moisture and geology, Remote Sensing of Environment, 8, 151–168, https://doi.org/10.1016/0034-4257(79)90014-2, 1979.

Priddy, K. L. and Keller, P. E.: Artificial Neural Networks: An Introduction, SPIE Press, 184 pp., 2005.

Rahimzadeh-Bajgiran, P., Berg, A. A., Champagne, C., and Omasa, K.: Estimation of soil moisture using optical/thermal infrared remote sensing in the Canadian Prairies, ISPRS Journal of Photogrammetry and Remote Sensing, 83, 94–103, https://doi.org/10.1016/j.isprsjprs.2013.06.004, 2013.

Reichle, R. H.: Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model, Geophys. Res. Lett., 32, L02404, https://doi.org/10.1029/2004GL021700, 2005.

Reichle, R. H., Lannoy, G. J. M. D., Liu, Q., Ardizzone, J. V., Colliander, A., Conaty, A., Crow, W., Jackson, T. J., Jones, L. A., Kimball, J. S., Koster, R. D., Mahanama, S. P., Smith, E. B., Berg, A., Bircher, S., Bosch, D., Caldwell, T. G., Cosh, M., González-Zamora, Á., Collins, C. D. H., Jensen, K. H., Livingston, S., Lopez-Baeza, E., Martínez-Fernández, J., McNairn, H., Moghaddam, M., Pacheco, A., Pellarin, T., Prueger, J., Rowlandson, T., Seyfried, M., Starks, P., Su, Z., Thibeault, M., Velde, R. van der, Walker, J., Wu, X., and Zeng, Y.: Assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture Product Using In Situ Measurements, Journal of Hydrometeorology, 18, 2621–2645, https://doi.org/10.1175/JHM-D-17-0063.1, 2017.

Reichle, R. H., Walker, J. P., Koster, R. D., and Houser, P. R.: Extended versus Ensemble Kalman Filtering for Land Data Assimilation, J. Hydrometeor, 3, 728–740, https://doi.org/10.1175/1525-7541(2002)003<0728:EVEKFF>2.0.CO;2, 2002.

Reynolds, S. G.: The gravimetric method of soil moisture determination Part I A study of equipment, and methodological problems, Journal of Hydrology, 11, 258–273, https://doi.org/10.1016/0022-1694(70)90066-1, 1970.

Richards, L. A.: CAPILLARY CONDUCTION OF LIQUIDS THROUGH POROUS MEDIUMS, Physics, 1, 318–333, https://doi.org/10.1063/1.1745010, 1931.

Rivieccio, R., Di Bene, C., Paolanti, M., Marchetti, M., and Napoli, R.: Soil rooting depth of Italy, Journal of Maps, 16, 36–42, https://doi.org/10.1080/17445647.2019.1690595, 2020.

Sabater, J. M., Jarlan, L., Calvet, J.-C., Bouyssel, F., and De Rosnay, P.: From Near-Surface to Root-Zone Soil Moisture Using Different Assimilation Techniques, Journal of Hydrometeorology, 8, 194–206, https://doi.org/10.1175/JHM571.1, 2007.

Sadeghi, M., Jones, S. B., and Philpot, W. D.: A linear physically-based model for remote sensing of soil moisture using short wave infrared bands, Remote Sensing of Environment, 164, 66–76, https://doi.org/10.1016/j.rse.2015.04.007, 2015.

Sanchez-Azofeifa, A., Antonio Guzmán, J., Campos, C. A., Castro, S., Garcia-Millan, V., Nightingale, J., and Rankine, C.: Twenty-first century remote sensing technologies are revolutionizing the study of tropical forests, Biotropica, 49, 604–619, https://doi.org/10.1111/btp.12454, 2017.

Schenk, H. J. and Jackson, R. B.: THE GLOBAL BIOGEOGRAPHY OF ROOTS, Ecological Monographs, 72, 311–328, https://doi.org/10.1890/0012-9615(2002)072[0311:TGBOR]2.0.CO;2, 2002.

Schmugge, T. J., Kustas, W. P., Ritchie, J. C., Jackson, T. J., and Rango, A.: Remote sensing in hydrology, Advances in Water Resources, 25, 1367–1385, https://doi.org/10.1016/S0309-1708(02)00065-9, 2002.

Selig, E., Wobschall, D., Mansukhani, S., and Motiwala, A.: Capacitance sensor for soil moisture measurement. Transportation Research Record, 532:64–76, 1975.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture–climate interactions in a changing climate: A review, Earth-Science Reviews, 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Shen, X., Walker, J. P., Ye, N., Wu, X., Boopathi, N., Yeo, I.-Y., Zhang, L., and Zhu, L.: Soil Moisture Retrieval Depth of P- and L-Band Radiometry: Predictions and Observations, IEEE Trans. Geosci. Remote Sensing, 59, 6814–6822, https://doi.org/10.1109/TGRS.2020.3026384, 2021.

Shewalkar, A. N.: Comparison of RNN, LSTM and GRU on Speech Recognition Data, 2018.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., and Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, https://doi.org/10.48550/ARXIV.1506.04214, 2015.

Souissi, R., Al Bitar, A., and Zribi, M.: Accuracy and Transferability of Artificial Neural Networks in Predicting in Situ Root-Zone Soil Moisture for Various Regions across the Globe, Water, 12, 3109, https://doi.org/10.3390/w12113109, 2020.

Souissi, R., Zribi, M., Corbari, C., Mancini, M., Muddu, S., Tomer, S. K., Upadhyaya, D. B., and Al Bitar, A.: Integrating process-related information into an artificial neural network for root-zone soil moisture prediction, Hydrol. Earth Syst. Sci., 26, 3263–3297, https://doi.org/10.5194/hess-26-3263-2022, 2022.

Stefan, V.-G., Indrio, G., Escorihuela, M.-J., Quintana-Seguí, P., and Villar, J. M.: High-Resolution SMAP-Derived Root-Zone Soil Moisture Using an Exponential Filter Model Calibrated per Land Cover Type, Remote Sensing, 13, 1112, https://doi.org/10.3390/rs13061112, 2021.

Tarboton, D.: Rainfall Runoff Processes, A Workbook to Accompany the Online Module Prepared for the National Weather Service Comet Outreach Program, 2003.

Thomas, A. M.: In situ measurement of moisture in soil and similar substances by `fringe' capacitance, J. Sci. Instrum., 43, 21–27, https://doi.org/10.1088/0950-7671/43/1/306, 1966.

Tinet, A.-J., Chanzy, A., Braud, I., Crevoisier, D., and Lafolie, F.: Development and evaluation of an efficient soil-atmosphere model (FHAVeT) based on the Ross fast solution of the Richards equation for bare soil conditions, Hydrology and Earth System Sciences, 19, 969–980, https://doi.org/10.5194/hess-19-969-2015, 2015.

Topp, G. C., Davis, J. L., and Annan, A. P.: Electromagnetic determination of soil water content: Measurements in coaxial transmission lines, Water Resour. Res., 16, 574–582, https://doi.org/10.1029/WR016i003p00574, 1980.

Tuller, M. and Or, D.: WATER RETENTION AND CHARACTERISTIC CURVE, in: Encyclopedia of Soils in the Environment, Elsevier, 278–289, https://doi.org/10.1016/B0-12-348530-4/00376-3, 2005.

Upadhyaya, D. B., Evans, J., Muddu, S., Tomer, S. K., Al Bitar, A., Yeggina, S., S, T., Morrison, R., Fry, M., Tripathi, S. N., Mujumdar, M., Goswami, M., Ganeshi, N., Nema, M. K., Jain, S. K., Angadi, S. S., and Yenagi, B. S.: The Indian COSMOS Network (ICON): Validating L-Band Remote Sensing and Modelled Soil Moisture Data Products, Remote Sensing, 13, 537, https://doi.org/10.3390/rs13030537, 2021.

Van doninck, J., Peters, J., De Baets, B., De Clercq, E. M., Ducheyne, E., and Verhoest, N. E. C.: The potential of multitemporal Aqua and Terra MODIS apparent thermal inertia as a soil moisture indicator, International Journal of Applied Earth Observation and Geoinformation, 13, 934–941, https://doi.org/10.1016/j.jag.2011.07.003, 2011.

Vaz, C. M. P., Jones, S., Meding, M., and Tuller, M.: Evaluation of Standard Calibration Functions for Eight Electromagnetic Soil Moisture Sensors, Vadose Zone Journal, 12, vzj2012.0160, https://doi.org/10.2136/vzj2012.0160, 2013.

Verhoest, N., Lievens, H., Wagner, W., Álvarez-Mozos, J., Moran, M., and Mattia, F.: On the Soil Roughness Parameterization Problem in Soil Moisture Retrieval of Bare Surfaces from Synthetic Aperture Radar, Sensors, 8, 4213–4248, https://doi.org/10.3390/s8074213, 2008.

Verstraeten, W., Veroustraete, F., and Feyen, J.: Assessment of Evapotranspiration and Soil Moisture Content Across Different Scales of Observation, Sensors, 8, 70–117, https://doi.org/10.3390/s8010070, 2008.

Verstraeten, W. W., Veroustraete, F., van der Sande, C. J., Grootaers, I., and Feyen, J.: Soil moisture retrieval using thermal inertia, determined with visible and thermal spaceborne data, validated for European forests, Remote Sensing of Environment, 101, 299–314, https://doi.org/10.1016/j.rse.2005.12.016, 2006.

Wagner, W., Blöschl, G., Pampaloni, P., Calvet, J.-C., Bizzarri, B., Wigneron, J.-P., and Kerr, Y.: Operational readiness of microwave remote sensing of soil moisture for hydrologic applications, Hydrology Research, 38, 1–20, https://doi.org/10.2166/nh.2007.029, 2007.

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., de Rosnay, P., Jann, A., Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang, Y., Blöschl, G., Eitzinger, J., and Steinnocher, K.: The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications, metz, 22, 5–33, https://doi.org/10.1127/0941-2948/2013/0399, 2013.

Wagner, W., Lemoine, G., and Rott, H.: A Method for Estimating Soil Moisture from ERS Scatterometer and Soil Data, Remote Sensing of Environment, 70, 191–207, https://doi.org/10.1016/S0034-4257(99)00036-X, 1999.

Walker, J. P., Willgoose, G. R., and Kalma, J. D.: One-Dimensional Soil Moisture Profile Retrieval by Assimilation of Near-Surface Measurements: A Simplified Soil Moisture Model and Field Application, J. Hydrometeor, 2, 356–373, https://doi.org/10.1175/1525-7541(2001)002<0356:ODSMPR>2.0.CO;2, 2001.

Walker, J. P., Willgoose, G. R., and Kalma, J. D.: In situ measurement of soil moisture: a comparison of techniques, Journal of Hydrology, 293, 85–99, https://doi.org/10.1016/j.jhydrol.2004.01.008, 2004.

Walker, J. P., Willgoose, G. R., and Kalma, J. D.: Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: Simplified Kalman filter covariance forecasting and field application: THREE-DIMENSIONAL SOIL MOISTURE ASSIMILATION, Water Resour. Res., 38, 37-1-37–13, https://doi.org/10.1029/2002WR001545, 2002.

Wang, L. (Ed.)Kacprzyk, J.: Support Vector Machines: Theory and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/b95439, 2005.

Wang, L. and Qu, J. J.: Satellite remote sensing applications for surface soil moisture monitoring: A review, Front. Earth Sci. China, 3, 237–247, https://doi.org/10.1007/s11707-009-0023-7, 2009.

Weidong, L., Baret, F., Xingfa, G., Qingxi, T., Lanfen, Z., and Bing, Z.: Relating soil surface moisture to reflectance, Remote Sensing of Environment, 81, 238–246, https://doi.org/10.1016/S0034-4257(01)00347-9, 2002.

Wigneron, J.-P., Calvet, J.-C., deRosnay, P., Kerr, Y., Waldteufel, P., Saleh, K., Escorihuela, M. J., and Kruszewski, A.: Soil Moisture Retrievals From Biangular L-Band Passive Microwave Observations, IEEE Geosci. Remote Sensing Lett., 1, 277–281, https://doi.org/10.1109/LGRS.2004.834594, 2004.

Wigneron, J.-P., Schmugge, T., Chanzy, A., Calvet, J. C., and Kerr, Y. H.: Use of passive microwave remote sensing to monitor soil moisture, Agronomie, 18, 27–43, 1998.

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., and McMahon, T.: Experimental design and initial results from the Mahurangi River variability experiment: MARVEX, in: Water Science and Application, vol. 3, edited by: Lakshmi, V., Albertson, J., and Schaake, J., American Geophysical Union, Washington, D. C., 201–213, https://doi.org/10.1029/WS003p0201, 2001.

Xu, J., Ma, X., Logsdon, S. D., and Horton, R.: Short, Multineedle Frequency Domain Reflectometry Sensor Suitable for Measuring Soil Water Content, Soil Sci. Soc. Am. j., 76, 1929–1937, https://doi.org/10.2136/sssaj2011.0361, 2012.

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

Yu, J., Zhang, X., Xu, L., Dong, J., and Zhangzhong, L.: A hybrid CNN-GRU model for predicting soil moisture in maize root zone, Agricultural Water Management, 245, 106649, https://doi.org/10.1016/j.agwat.2020.106649, 2021.

Zazueta, F. S. and Xin, J.: Soil moisture sensors, Bulletin (Florida Cooperative Extension Service) (USA), 1994.

Zeng, X.: Global Vegetation Root Distribution for Land Modeling, J. Hydrometeor, 2, 525–530, https://doi.org/10.1175/1525-7541(2001)002<0525:GVRDFL>2.0.CO;2, 2001.

Zhang, Z. and Moore, J. C.: Data Assimilation, in: Mathematical and Physical Fundamentals of Climate Change, Elsevier, 291–311, https://doi.org/10.1016/B978-0-12-800066-3.00009-7, 2015

Zhan, Z., Qin, Q., Ghulan, A., and Wang, D.: NIR-red spectral space based new method for soil moisture monitoring, SCI CHINA SER D, 50, 283–289, https://doi.org/10.1007/s11430-007-2004-6, 2007.