# Optimizing Support Vector Machines with ISBA-A-g$_S$ Land Surface Variables as a Surrogate Model to Simulate ASCAT Derived Parameters

Master Thesis

Manish Kharagjitsing

Geoscience and Remote Sensing, 2020

# Optimizing Support Vector Machines with ISBA-A-$g_s$ Land Surface Variables as a Surrogate Model to Simulate ASCAT Derived Parameters

by

Manish Kharagjitsing

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday February 28, 2020 at 08:45 AM

Cover Page Source: https://www.ouest-france.fr/


| | | |
|---|---|---|
| Student number: | 4218884 | |
| Thesis committee: | Prof. dr. ir. S. Steele-Dunne, | TU Delft, Supervisor |
| | Dr. S. L. M. Lhermitte, | TU Delft, (Chair) |
| | Dr. ir. M. Coenders-Gerrits, | TU Delft |


*This thesis is confidential and cannot be made public until March 2020*

# Preface

This thesis is a product of hard work and persistence. I have learned so many invaluable lessons during this period and am very proud to make my contribution to science. Through this way, I would like to thank Susan Steele-Dunne who offered this project to me. Her patience with me, the faith that I can do this, ways how to manage a project like this and the expertise and insights into the subject were absolutely priceless. I also would like to thank Miriam Coenders-Gerrits and Stef Lhermitte for providing me with very useful and valuable feedback on the work which gave me a lot of clarity.

In addition to that, I would like to thank Clément Albergel and Sebastian Hahn from Météo-France for processing all ASCAT data and ISBA-A-$g_s$ on a quarter degree grid. Without their service, this work would not have been possible. Also, I would like to thank Barton Forman for the insightful Skype conversations in the beginning of the project who gave very relevant comments on the initiative.

My family, in particular my parents and girlfriend, have given me tremendous amount of support and patience during times I really struggled. Also, my fellow students of the department of Geoscience and Remote Sensing have been really inspiring and helping me during all times.

*Manish Kharagjitsing*
*Delft, February 2020*

# Summary

Soil moisture retrieval with the ASCAT scatterometer on top of the MetOp satellites requires proper corrections of the effects of vegetation on the backscatter signal. The TU-Wien developed a soil moisture retrieval algorithm that uses the incidence angle dependence of backscatter to obtain soil moisture estimates (Wagner et al., 1999). The core of this algorithm is a second order Taylor expansion with which the backscatter is normalized at a reference angle. Studies have shown that the first and second order derivative within this Taylor expansion, known as slope and curvature, are somehow related to the wet biomass and structure of vegetation. The general approach to forward model satellite observations with land surface variables in a data assimilation framework is through a radiative transfer model (Albergel et al., 2017). However, this requires plenty of assumptions about the vegetation canopy (such as stem height, shape, size, orientation etc.) and is therefore relatively inefficient for understanding the impact of soil moisture and vegetation dynamics on backscatter on a large scale.

This study investigates the possibility of using support vector machines as a surrogate model instead of a radiative transfer model to link the TU-Wien normalized backscatter and slope to land surface variables soil moisture and leaf area index. The land surface variables are simulations from the $CO_2$-responsive ISBA-A-$g_s$ land surface model. Support vector machines have the advantage of providing implicit kernel functions, which make them very useful for non-linear problems. The ISBA-A-$g_s$ data is provided by Météo-France and its spatial resolution is set to 25 kilometers, so that it matches with the spatial resolution of the ASCAT data. Each grid point consists of several land covers and the fractional cover is provided. The area that is investigated is France and only grid points that consist of more than 30% of a land cover type considered.

For the optimization of the support vector machines, a grid search is done with an internal cross validation function with the Scikit-Learn package in the Python programming language. The Radial Basis Function is chosen as kernel function. The parameters of the support vector machines that are optimized are $C$ and $\gamma$. For both parameters, 5 possible values with a difference on logarithmic scales were provided to the grid search, which makes it 25 combinations in total. Each combination is then tested with a k-fold cross validation, which means that the training data is split into k equal intervals with k=9. In this way, the parameter combinations with optimal results after cross validation are selected for the grid point. The performance metric that is used during the whole study is the coefficient of determination because of the ability to compare model performances on a map in a way that explains where the models are able to understand the variability of backscatter and slope and where not.

The performances of the normalized backscatter and normalized slope simulations are calculated over three different periods. The first period has a time span of one year. The other two periods have a time span of 6 months in which one period represents the vegetation growing season and the other period contains the winter months. By displaying the different performances on maps a clear indication was obtained that performances do indeed depend on the type of land covers within the grid point and the time span over which the performance is calculated.

For both simulations, the performance of different dominant land cover types is investigated. These are C3 crops, Temperate Grassland, Temperate Broadleaf Deciduous and Bare Land. For the $\sigma^0(40)$ simulations, an additional simulation is done by using the created models to simulate the climatology. This is done by taking the average day of year value for soil moisture and LAI training data and in this way a climatology estimate is obtained. Subtracting this estimate from the $\sigma^0(40)$ simulations gives an indication of what the created models are able to simulate besides annual patterns.

After displaying the parameters of the models for each grid point on a map, it is concluded that there is a spatial consistency with the optimization. Spatial consistency is also shown in terms of performance. Hence, land cover types and fractional covers determine the type of support vector machine and performance that comes with it. The focus of this study is not the in depth parameter justification, but rather exploring the feasibility of an optimization procedure for ASCAT parameter simulations and the performance for different

land cover types. As an initial approach to use support vector machines in forward simulation of ASCAT backscatter $\sigma^0(40)$ and slope $\sigma'(40)$, the conclusion is that this method is able to serve as a surrogate model. However, the optimization should be done for specific periods throughout the year, while in this study the same model is used for every period throughout the year.

# Contents

List of Figures ix

List of Tables xiii

Nomenclature xiv

1 Introduction 1
1.1 Microwave Remote Sensing for Soil Moisture . . . 1
1.2 Microwave Remote Sensing for Vegetation . . . 2
1.3 Advanced Scatterometer . . . 3
1.4 Problem Statement . . . 3
1.5 Area of Interest . . . 3
1.6 Research Objective & Questions. . . . 5

2 Background Theory 7
2.1 Microwave Remote Sensing . . . 7
2.2 ASCAT. . . . 7
2.2.1 System Characteristics . . . 7
2.3 TU-Wien Soil Moisture Retrieval . . . 9
2.3.1 Mathematical Approach . . . 9
2.3.2 ASCAT Soil Moisture Formulation . . . 10
2.3.3 ASCAT Data . . . 11
2.4 Land Surface Model . . . 12
2.4.1 ISBA . . . 12
2.4.2 A-$g_s$ . . . 12
2.4.3 Leaf Area Index . . . 12
2.4.4 Soil moisture . . . 13
2.4.5 Validation of Land Surface Model Variables . . . 13
2.4.6 ECOCLIMAP Database . . . 13
2.5 Support Vector Machines . . . 14

3 Methodology 17
3.1 General . . . 17
3.1.1 Requirements of machine learning model . . . 17
3.1.2 Workflow. . . . 17
3.2 Data Preparation . . . 19
3.2.1 Standardization . . . 19
3.2.2 Grid points. . . . 19
3.2.3 Data GPI for Day of Year . . . 19
3.3 Hyperparameter optimization . . . 21
3.3.1 Grid Search and Cross Validation . . . 21
3.3.2 Software . . . 21
3.3.3 Bias-Variance Trade-off . . . 21
3.3.4 SVM Parameters . . . 22
3.4 Performance Evaluation . . . 22

4 Results 25
4.1 Backscatter Simulations. . . . 25
4.1.1 Overall Performance . . . 25
4.1.2 Performance During Vegetation Growing Season . . . 28
4.1.3 Performances and Land Surface Variables . . . 29
4.1.4 Fractional Land Cover Good and Bad Performances . . . 34

4.2 Slope Simulations . . . 37
4.2.1 Overall Performance . . . 37
4.2.2 Performance During Vegetation Growing Season . . . 38
4.2.3 Performances and Land Surface Variables . . . 40
4.2.4 Fractional Land Cover Good and Bad Performances . . . 44
4.3 Spatial Consistency SVR Parameters . . . 46

5 Discussion & Conclusions 49
5.1 Discussion . . . 49
5.1.1 Support Vector Machine Optimization Overview . . . 49
5.1.2 Results Overview . . . 49
5.1.3 Sources of uncertainty . . . 51
5.2 Conclusions . . . 51
5.2.1 Research Questions . . . 51
5.2.2 Recommendations . . . 52

Bibliography 53

A Figures 57
A.1 Backscatter SVR parameter Combinations . . . 57
A.2 Slope SVR Parameter Combinations . . . 58
A.3 Histogram Parameter Occurrences . . . 59
A.4 R2 Performances Parameter Combinations . . . 60

B Tables 63
B.1 Slope Land Cover Fractions . . . 63
B.2 Parameter Occurrences Backscatter . . . 64
B.3 Parameter Occurrences Slope . . . 65

# List of Figures


1.1 Illustration of spatial and temporal resolution of both active and passive microwave systems. On the vertical axis the temporal resolution is displayed and the spatial resolution can be seen on the horizontal axis. (Wagner et al., 2007) . . . 2
1.2 Illustration of scattering mechanisms on a vegetated surface. (a) represents backscattering from the surface. (b) is referred as volume scattering and (c) is the effect of multi-bounce between the surface and the vegetation. [See Wagner et al. (2013)] . . . 2
1.3 High resolution land cover map. Obtained from Inglada et al. [11] . . . 4

2.1 The geometric view of the ASCAT on board of Metop-A. Image source: Wagner et al. [35]. . . . 8
2.2 Incidence angle dependence of the backscattering coefficient $\sigma^0$. Source: Wagner et al. [35] . . . 9
2.3 An illustration of the concepts within the TUW SMR method. The upper part of this figure shows ASCAT data from a grid point in Nebraska in a study of Steele-Dunne et al. (2019). The legend shows the relation of wet and dry reference with the backscatter. The bottom parts show the illustrate the incidence angle dependence for changes in soil moisture and vegetation. . . . 10
2.4 The mean (a) and standard deviation (b) of the ASCAT normalized backscatter parameter $\sigma^0(40)$ over grid points in Europe. . . . 11
2.5 The mean (a) and standard deviation (b) of the ASCAT normalized slope parameter $\sigma'(40)$ over grid points in Europe. . . . 12
2.6 $\varepsilon$ -intensive loss function. The filled circles represent the support vectors. It can be seen that some support vectors are outside of the $\varepsilon$-margin. Source: Rojas et al. (2017) . . . 14

3.1 Workflow of simulations for each individual grid points (GPI). . . . 18
3.2 On the left hand side of the figure (a) the dominant land cover types are displayed on the map. It can be seen that there are 5 main land cover types. On the right hand side of the figure, the fractional coverage of these land cover types is displayed. Only grid points that contain a land cover with a fraction equal to or larger than 30 % are displayed and processed. . . . 19
3.3 Climatology of Soil Moisture, LAI and slope $\sigma^0(40)$ for grid point 782653 . . . 20
3.4 Climatology of Soil Moisture, LAI and slope $\sigma'(40)$ for grid point 782653 . . . 20
3.5 Iteration scheme of the k-fold cross validation. The green boxes represent the test fold and the grey boxes are the train folds. The value of k for the cross validation of training data is set 9. This means that the 9 years of training data is equally split in 9 pieces and for each iteration 8 pieces are used for training and one for testing. . . . 21
3.6 The two models on the left hand side under-fit the training sample while the two models on the right hand side are clearly over fitting. Source image: Maimon and Rokach [17] . . . 22
3.7 Procedure for simulating climatology with standardization of the climatological fit. . . . 23

4.1 Map of backscatter $\sigma^0(40)$ performances. The left hand side of the figure (a) shows the R2 performance with the climatology included. The right hand side (b) shows the performance of the regression . . . 26
4.2 On the left hand side of the figure (a) the standard deviation of the backscatter $\sigma^0(40)$ is shown over a part of Europe. On the right hand side the RMSE performances are displayed for the selected grid points. . . . 26
4.3 Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages. Climatology is not removed. . . . 27
4.4 Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages. Climatology is removed. . . . 28

4.5 On the left hand side of the figure (a) the coefficient of determination R2 from $\sigma^0(40)$ simulations is displayed during the period from September to March which is during the Fall and Winter season. On the right hand side (b) the R2 is displayed from the period from March till September during the vegetation growing season. . . . . 29
4.6 Simulation for $\sigma^0(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value for the winter period. Climatology is simulated and removed from the total simulation. 30
4.7 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 30
4.8 Simulation for $\sigma^0(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value for the vegetation growing season. Climatology is simulated and removed from the total simulation. . . . . 31
4.9 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 31
4.10 Simulation for $\sigma^0(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months. Climatology is simulated and removed from the total simulation. . . . . 32
4.11 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 32
4.12 Simulation for $\sigma^0(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months. Climatology is simulated and removed from the total simulation. . . . . 33
4.13 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 33
4.14 Map of slope $\sigma'(40)$ performances. The left hand side of the figure (a) shows the R2 performance with the climatology included. The right hand side (b) shows the performance of the regression 37
4.15 Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages. . . . . 38
4.16 On the left hand side of the figure (a) the coefficient of determination R2 for $\sigma'(40)$ is displayed during the period from September to March which is during the Fall and Winter season. On the right hand side (b) the R2 is displayed from the period from March till September during the vegetation growing season. . . . . 39
4.17 Simulation for $\sigma'(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value during the winter months. . . . . 40
4.18 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 41
4.19 Simulation for $\sigma'(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value during the vegetation growing season. . . . . 41
4.20 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 42
4.21 Simulation for $\sigma'(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months. . . . . 42
4.22 Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown. . . . . 43
4.23 $\sigma^0(40)$ map of grid points of which the optimized SVR model has a certain value of C on the left hand side (a) and $\gamma$ on the right hand side (b) . . . . . 46
4.24 $\sigma'(40)$ map of grid points of which the optimized SVR model has a certain value of C on the left hand side (a) and $\gamma$ on the right hand side (b) . . . . . 46

A.1 Map of land cover grid points with SVR parameter combination $C = 10$ and $\gamma = 0.1$ for $\sigma^0(40)$ simulations. . . . . 57
A.2 Map of land cover grid points with SVR parameter combination $C = 1000$ and $\gamma = 0.001$ for $\sigma^0(40)$ simulations. . . . . 58
A.3 Map of land cover grid points with SVR parameter combination $C = 10$ and $\gamma = 0.1$ for $\sigma^0(40)$ simulations . . . . . 58
A.4 Map of land cover grid points with SVR parameter combination $C = 0.1$ and $\gamma = 1$ for $\sigma^0(40)$ simulations . . . . . 59
A.5 $\sigma^0(40)$ amount of occurrences of SVR parameter $C$ on the left hand side (a) and $\gamma$ on the right hand side (b) . . . . . 59
A.6 $\sigma'(40)$ amount of occurrences of SVR parameter $C$ on the left hand side (a) and $\gamma$ on the right hand side (b) . . . . . 59
A.7 $\sigma^0(40)$ R2 performances for simulations with removed climatology for the most occurring parameter combinations for the four most dominant land cover types in France. . . . . 60
A.8 $\sigma'(40)$ R2 performances for simulations for the most occurring parameter combinations for the four most dominant land cover types in France. . . . . 61

# List of Tables


2.1 Metop and ASCAT characteristics. Info retrieved from Imperatore and Riccio [10] . . . . . 8

4.1 $\mu$ is the mathematical mean. $\sigma$ is the mathematical standard deviation. R2 is the coefficient of determination. SS represents Spring + Summer and AW represents Autumn + Winter. . . . . 28
4.2 Fractional Cover Statistics for $\sigma^0(40)$ simulations with C3 Crops as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 34
4.3 Fractional Cover Statistics for $\sigma^0(40)$ simulations with Temperate Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 35
4.4 Fractional Cover Statistics for $\sigma^0(40)$ simulations with Temperate Broadleaf Deciduous as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. 35
4.5 Fractional Cover Statistics for $\sigma^0(40)$ simulations with Bare Land as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 36
4.6 $\mu$ is the mathematical mean. $\sigma$ is the mathematical standard deviation. R2 is the coefficient of determination. SS represents Spring + Summer and AW represents Autumn + Winter. . . . . 39
4.7 Fractional land cover for grid point 766805 with dominant land cover type Temperate Grassland. 43
4.8 Fractional Cover Statistics for $\sigma'(40)$ simulations with C3 Crops as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 44
4.9 Fractional Cover Statistics for $\sigma'(40)$ simulations with Temperate Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . 45
4.10 Fractional Cover Statistics for $\sigma'(40)$ simulations with Temperate Broadleaf Deciduous as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. 45

B.1 Fractional Cover Statistics for $\sigma'(40)$ simulations with Bare Land as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 63

B.2 Fractional Cover Statistics for $\sigma'(40)$ simulations with Boreal Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated. . . . . 64
B.3 Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the C3 crops land cover. . . . . 64
B.4 Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the temperate grassland land cover. . . . . 64
B.5 Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the Temperate Broadleaf Deciduous land cover. . . . . 64
B.6 Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the Bareland land cover. . . . . 65
B.7 Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the winter c3 crops land cover. . . . . 65
B.8 Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Temperate Grassland land cover. . . . . 65
B.9 Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Temperate Broadleaf Deciduous land cover. . . . . 65
B.10 Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Bareland land cover. . . . . 65

# Nomenclature

| | |
|---|---|
| $\lambda$ | Wavelength |
| $\mu$ | Mathematical mean |
| $\sigma$ | Mathematical standard deviation |
| $\sigma''(40)$ | Normalized curvature |
| $\sigma'(40)$ | Normalized slope |
| $\sigma^0(40)$ | Normalized backscatter |
| $\sigma^0_d$ | Dry reference soil moisture |
| $\sigma^0_w$ | Wet reference soil moisture |
| $\theta_a$ | Aftbeam incidence angle |
| $\theta_f$ | Forebeam incidence angle |
| $\theta_m$ | Midbeam incidence angle |
| $\theta_r$ | Reference incidence angle |
| $\Theta(t)$ | ASCAT soil moisture |
| $A - g_s$ | Net assimilation and stomatal conductance |
| ASCAT | Advanced Scatterometer |
| ISBA | Interactive Soil-Biosphere-Atmosphere |
| LAI | Leaf Area Index |
| MetOP | Meteorological OPerational |
| RT | Radiative Transfer |
| SVAT | Soil-Vegetation-Atmosphere Transfer |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| TUW-SMR | TU-Wien Soil Moisture Retrieval |
| WG2 | Soil moisture at depth of around 4 cm |

# 1 Introduction

This section serves as an introduction to the rest of the study. First, the concept of soil moisture retrieval using microwave remote sensing is introduced together with its limitations in the presence of densely vegetated areas. In this part it is justified why satellite remote sensing with microwave instruments are the chosen data source for soil moisture retrieval. After that, an introduction is made to the retrieval of vegetation dynamics through microwave remote sensing. This is followed by introducing the Advanced Scatterometer (ASCAT) on top of the Metop-A and Metop-B satellites and the justification why this source of data is used in this study. This is followed by the problem statement in which the objective of this study is presented, which is to create a forward model for simulation of ASCAT derived products.

## 1.1. Microwave Remote Sensing for Soil Moisture

Soil moisture is a crucial parameter for a number of fields such as agriculture, forest management, meteorology and climate related studies. It can be described as the amount of water in the empty spaces between soil particles in the unsaturated soil zone (Liang et al., 2012). The distinction can be made between surface soil moisture and root-zone soil moisture. The latter one describes the water availability to plants and is generally referred to the soil moisture in the upper 2 meters of soil. The surface soil moisture is known as the soil moisture in the upper 5 centimeters of soil.

The temporal and spatial resolution of soil moisture observations determine the usefulness for the specific fields that require these observations. In-situ soil moisture measurements can be valuable for agricultural activities on a local scale such as irrigation management for a farmer. The disadvantage of in-situ observations for regional or global scale applications is the inconsistent spatial coverage. With satellite remote sensing, the demand of spatial- and temporally consistent observations could be fulfilled (Liang et al., 2012). Figure 1.1 gives an illustration of the spatial and temporal resolutions of different microwave remote sensing technologies. The demand for this study is a high temporal resolution with the potential of a global scale and hence Synthetic Aperture Radar (SAR) and ScanSAR will not be considered.

Microwave remote sensing offers the advantage that it is capable of monitoring during day and night and under all weather conditions. The two modes in which microwave remote sensing observations could be obtained are the passive and active mode. With passive microwave remote sensing the microwave emission of a soil is measured. With active microwave remote sensing, microwaves are transmitted to the earth surface and the backscattered signal is measured. Both approaches are used for soil moisture retrieval.

Wagner et al. (2008) compared soil moisture products from the ERS scatterometer (active) with the AMSRE-E (passive) instrument. The AMSRE-E is a radiometer that operates at six frequencies with dual-polarization. The soil moisture dataset from the AMSRE-E were derived from C-band ($\lambda = 5.7cm$) passive microwave brightness temperatures. The ERS scatterometer operated at C-band and the method to estimate soil moisture was developed by the Vienna University of Technology (TU Wien) [see Wagner et al. (1999)]. A reasonable agreement between both products was obtained for areas where vegetation density was sparse to moderate. Densely vegetated areas showed poor similarity between both products. The explanation that was given for this result is the limitations of retrieving soil moisture in areas where vegetation cover is dense.

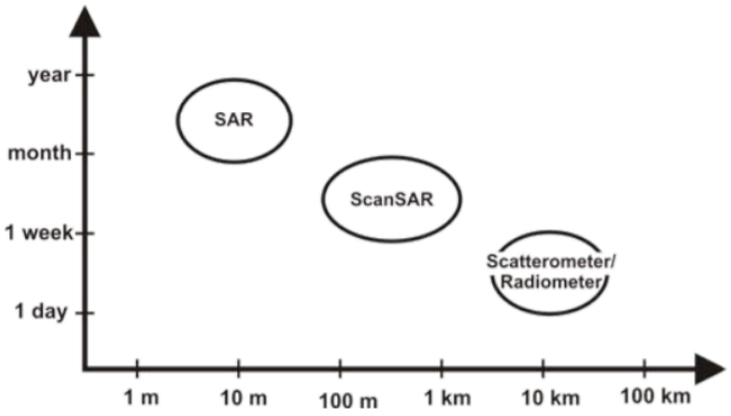



Figure 1.1: Illustration of spatial and temporal resolution of both active and passive microwave systems. On the vertical axis the temporal resolution is displayed and the spatial resolution can be seen on the horizontal axis. (Wagner et al., 2007)

## 1.2. Microwave Remote Sensing for Vegetation

Naeimi and Wagner (2010) describe the major components that attenuate the backscatter signal over (vegetated) land. Their focus is on the application of the C-band scatterometers on top of the ERS-1, ERS-2 and Metop-A satellites. The spatial resolution of the scatterometers on board of these satellites are equal or larger than 25 kilometers. The factors attenuating the backscatter are soil moisture, roughness, vegetation structure and vegetation water content. For this spatial resolution (≥ 25 kilometers) it is said that generally the surface roughness parameter does not vary much over time. The vegetation water content and canopy vary on a time scale of several days to weeks. Soil moisture has the highest frequency variations of the above mentioned factors with variations on a time scale of hours to days. The reason for the difficulty of modelling backscatter from vegetated surfaces is due to the impact of volume scattering. An illustration and description of the different scattering mechanisms is provided in the figure below.

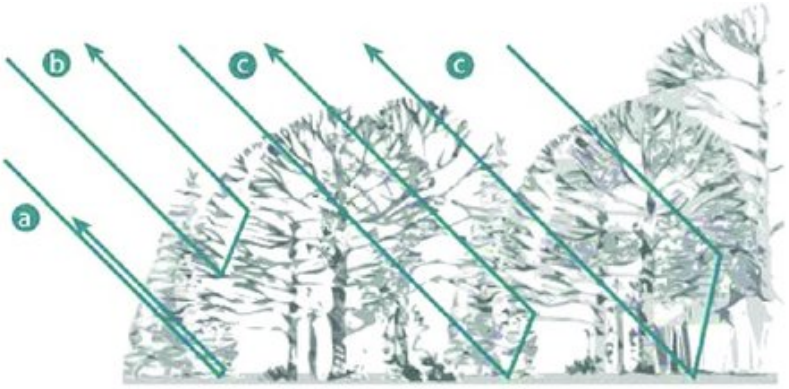



Figure 1.2: Illustration of scattering mechanisms on a vegetated surface. (a) represents backscattering from the surface. (b) is referred as volume scattering and (c) is the effect of multi-bounce between the surface and the vegetation. [See Wagner et al. (2013)]

The wavelength of the microwave determines the amount of possible penetration of vegetation. A large wavelength is more likely to penetrate vegetation than a low wavelength. Wagner et al. (2008) state that 3 bandwidths are typically used for soil moisture retrieval. These are the L-band ($f_c$ = 1.4 GHz, $\lambda$ = 24 cm), C-band ($f_c$= 6 GHz, $\lambda$ = 5.7 cm) and X-band ($f_c$ = 10 GHz, $\lambda$ = cm). The L-band is least sensitive to vegetation cover of all mentioned bands and the X-band is most sensitive and its backscatter will likely contain just information about the vegetation and not soil moisture as it does not necessarily reaches the soil surface. The C-band backscatter contains information regarding both soil moisture and vegetation.

## 1.3. Advanced Scatterometer

This study uses the Advanced Scatterometer (ASCAT) products on top of Meteological Operational satellites (MetOp) A and B. This satellite is an active instrument, hence it transmits energy and measures the received backscatter. The transmitted microwaves have a wavelength of 5.7 centimeters. Due to the dielectric properties of water, by transmitting energy with this wavelength on a soil that is not covered with vegetation, backscatter data is obtained that contains information regarding the water content on the top surface layer. If the water content in the soil is high, a high backscatter is obtained. For a low water content, its the opposite. The presence of vegetation causes that the backscatter contains information about the soil and vegetation. With the life span of the ASCAT missions, of which the operations started since 2007 and is still operating till date, the data is a very valuable source of information for understanding the effect of both components (soil and vegetation) on the backscatter ($\sigma^0$).

## 1.4. Problem Statement

The ability to accurately monitor and understand land surface processes such as fluxes of carbon, water and vegetation cycles depends on the quality of data provided to a land surface model. With data assimilation, an optimal combination of data sources is created to reduce the uncertainty difference between the simulation of a land surface model and (satellite) observations (Barbu et al.,2011). As mentioned earlier in this chapter, the retrieval of soil moisture through satellite observations cannot be directly done as different type of land covers impact the backscatter in different ways. Albergel et al. (2017) mention the need for different ways to simulate, because if data sources and methods of modelling are too similar, the errors could be more correlated and hence the data assimilation process will not be optimal. Consequently, the demand exists for new methods that are able to serve as forward operators for ASCAT backscatter.

The current state of art is that radiative transfer (RT) models are used for backscatter $\sigma^0$ simulations. Quast et al. (2019) used a first-order radiative transfer modelling approach to retrieve soil and vegetation related parameters from ASCAT $\sigma^0$ observations. The obtained model is able to forward simulate $\sigma^0$ using auxiliary leaf area index (LAI) and soil moisture (WG2) provided by the Interactions between Soil, Biosphere and Atmosphere land surface model (ISBA) or soil moisture inversion. The model has been successfully implemented for 158 locations in France and the model can be easily adjusted according to soil and vegetation types.

Even though RT models are widely used for simulations, the general problem with a RT model is that optimization requires a number of assumptions, such as specification of the vegetation geometry and dielectric properties. This can make model optimization very complex for large scale applications or even global models. An alternative and addition to a RT model is an assimilation framework could be a surrogate model. An advantage of such a model is that there is no need for a analytical function as basis for the simulation and hence the number of complexities of is reduced (Mahmoodian et al., 2018).

In Forman and Reichle (2015) a support vector machine (SVM) and an Artificial Neural Network combined with a land surface model are used to simulate passive microwave brightness temperatures. The motivation behind this attempt is to reduce the computational expense of brightness temperature estimation so that in can be used on a large scale without (only) using a snow emission model. Brightness temperature is relevant for retrieval of Snow Water Equivalent retrievals within a data assimilation framework. This concept serves as an inspiration for this thesis, in which the aim is to see if there is potential for using support vector machines as a surrogate model that is able to simulate ASCAT backscatter parameters $\sigma^0(40)$ and $\sigma'(40)$ using the ISBA-A-g$_s$ land surface model variables. Variations in $\sigma'(40)$ depend on vegetation dynamics and variations of $\sigma^0(40)$ on a bare soil are primarily caused by variations in soil moisture (Vreugdenhil et al., 2017).

## 1.5. Area of Interest

The chosen study region is France. The area suffices the condition stated in the problem statement that it should consist of several different vegetated land cover types. Figure 1.3 shows a map that is the result from a clustering and classification from Inglada et al. (2017). With the amount of very specific land cover types shown in the legend, it might be a hard task to specify the land cover type for grid points with a special resolution of 25 kilometers (ASCAT resolution). Therefore, grid points are be assigned to a particular land cover type if the fractional cover is larger than 30%. This will be explained in more depth in chapter 3.

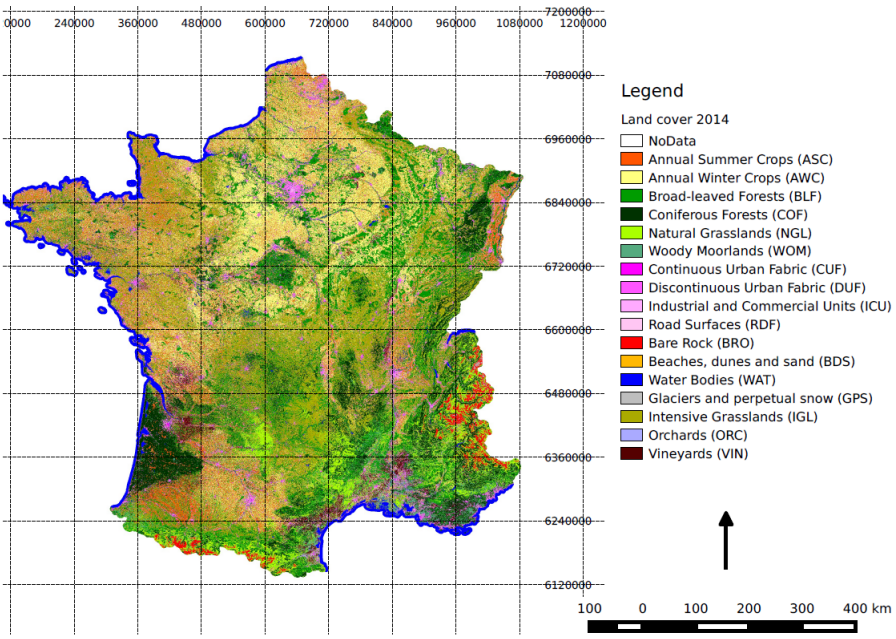



Figure 1.3: High resolution land cover map. Obtained from Inglada et al. [11] .

The land cover map used in this study is provided by Météo France. The data is from the ECOCLIMAP database. More information regarding this database, including land cover types, is provided in Section 2.

## 1.6. Research Objective & Questions

In the problem statement is mentioned that a surrogate model could solve many of the complexities of a radiative transfer model for large scale problems. The aim of this study is to develop such a surrogate model that is able to simulate ASCAT derived parameters $\sigma^0(40)$ (normalized backscatter) and $\sigma'(40)$ (normalized slope) using ISBA-A-$g_s$ land surface variables. These parameters are explained in Chapter 2. The land surface variables are from the ISBA-A-$g_s$ model and are provided by Météo France. The choice for the surrogate model is the support vector machine, as this machine learning method is able to deal with non-linear problems and has shown to be very useful by Forman and Reichle (2015).

**Research Question:**
*Is an optimized support vector machine with the ISBA-A-$g_s$ land surface model able to serve as a surrogate model to simulate $\sigma^0(40)$ and $\sigma'(40)$?*

In order to answer the main research question as complete as possible, the following sub-questions serve as focus questions during the research and are answered in the final chapter of this study.

**Sub-questions:**

- What are the limitations of the procedure of hyper-parameter optimization?
- What are the advantages and disadvantages of the support vector machines?
- What impact does land cover have on the performance of the simulation?

# 2

# Background Theory

This chapter gives all the background theory behind the relevant concepts for this study. First, a technical description of the ASCAT satellite is given. This includes the system characteristics and the processing of the radar signal. In addition to that, operational information is given about the ASCAT mission. At the core of this study is the TU-Wien Soil moisture retrieval algorithm. This algorithm describes the incidence angle dependence of the backscatter with a second order Taylor expansion. The first order derivative is called the slope and the second order is the curvature. These parameters depend on vegetation states and thus are important parameters to understand for better soil moisture retrieval on vegetative areas. The final part of this chapter is about the ISBA-A-g$_s$ land surface variables that will be used for simulation of the ASCAT parameters.

## 2.1. Microwave Remote Sensing

Microwaves are part of the electromagnetic spectrum and are within the frequency range of 300 MHz and 300 GHz. This translates into wavelengths of 1 meter and 1 millimeter respectively. The bands often used for applications regarding soil moisture and/or vegetation are the L-band ($\lambda = 24cm$) and C-band ($\lambda = 5.7cm$). Microwave remote sensing can be done an active or passive mode. Active remote sensing means that energy is transmitted and received while in passive mode energy is measured from the Earth's surface. This study focuses on backscatter obtained from an active remote sensing instrument known as the Advanced Scatterometer (ASCAT).

The measured backscatter from an active microwave is affected by a variety of factors. These factors can be distinguished in two categories: Radar parameters and target parameters (Ulaby, 1975). The factors that belong to the first category are the frequency, polarization and incidence angle. The factors that belong to the target parameters can be further separated into two groups: vegetation cover and underlying soil. For the vegetation cover, the factors are the geometry, roughness, density, pattern, height, dielectric properties and plant moisture content distribution. For the underlying soil, the factors are the surface roughness and dielectric properties.

The incidence angle is the angle between the radar beam and the normal of the target. Ulaby (1975) demonstrated that backscatter sensitivity of crops is higher with high incidence angles and high frequencies (thus low wavelengths) with VV polarization. In contrast, backscatter sensitiviy to soil moisture was found to be high with low incidence angles and low frequencies (hence large wavelengths) using HH polarization.

## 2.2. ASCAT

### 2.2.1. System Characteristics

The Advanced Scatterometer (ASCAT) is a real aperture radar on top of the MetOp-A and B sattelites. The design of the ASCAT is based on its predecessors which are the European Space Agency's Earth Remote Sensing (ERS) 1 and 2 satellites. Both the ASCAT and ERS satellites operate(d) at C-band (5.255 GHz) using vertically polarized waves. Vertical polarization means that the electromagnetic waves are perpendicular to the Earth's surface. The transmission of energy is done by a long pulse with linear frequency modulation (chirp). The radar contains three antennas on both the left and right side of the instrument. These antennas are the

fore-,mid- and aft-beam. Figure 2.1 illustrates the geometric view of ASCAT. It can be noticed that the range incidence angle is in between 25° and 64°.

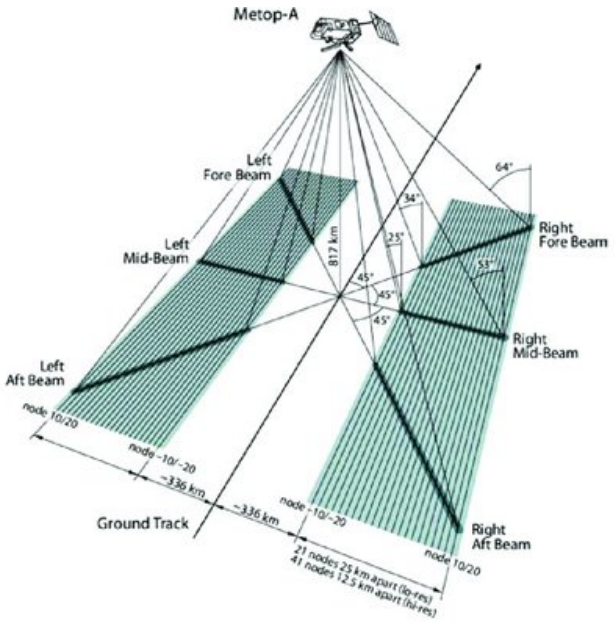



Figure 2.1: The geometric view of the ASCAT on board of Metop-A. Image source: Wagner et al. [35].

The relevant ASCAT and Metop-A and Metop-B properties are shown in the table below.

| Satellite | **Metop** | C-band Scatterometer | **ASCAT** |
|---|---|---|---|
| **Launch Year** | A: 2006<br>B: 2012 | **Wavelength** | $\lambda = 5.7cm$ |
| **Altitude Spacecraft** | 817 km | **Polarization** | VV |
| **Ground Track Repeat Cycle** | 29 Days | **Swath Width (Double)** | 550 km |
| **Orbit Period** | 101 minutes | **Spatial Resolution** | 50 km, 25 km |
| **Orbit** | Sun-synchronious | **Sampling Interval** | 25 km, 12.5 km |

Table 2.1: Metop and ASCAT characteristics. Info retrieved from Imperatore and Riccio [10]

As previously mentioned, the backscattered signal depends on the system characteristics and the target (or object). In case of a vegetated surface the factors attenuating the signal are the geometry, roughness, density, pattern, height, dielectric properties and plant moisture content distribution. The spatial resolution of the ASCAT is 25 kilometers. On this scale roughness is therefore considered to be temporally invariant. The most variant of all previously mentioned factors is the plant moisture content distribution. The time scale for this variation is a few days. The surface soil moisture however varies within hours. The density (or vegetation canopy) also varies within a time scale from several days to weeks.

The main factor from the system characteristics that impacts the strength of the backscattered signal is the incidence angle. This variable is different for every acquisition. One way to compare backscatter $\sigma^0$ variables is to average measurements over a certain period of time. The limit of this procedure is that vegetation cannot be distinguished from soil moisture. Instead, Wagner et al. (1999) proposed a different method to describe the backscatter based on incidence angle dependency. The first derivative of the backscatter $\sigma^0(\theta)$ with respect to the incidence angle $\theta$ is shown in equation 2.1.

$$\sigma'(\theta) = \frac{d\sigma^0(\theta)}{d\theta} \tag{2.1}$$

If this relationship between backscatter $\sigma^0$ and incidence angle $\theta$ is known, a $\sigma^0$ can be obtained at a particular reference angle $\theta_r$. Figure 2.2 gives an illustration of decreasing backscatter due to increasing in-

cidence angle under four different circumstances: dry soil without vegetation, dry soil with vegetation, wet soil without vegetation and finally wet soil with vegetation. Dry soil gives a lower $\sigma^0$ than wet soil at the same incidence angle, due to the lack of liquid water. The $\sigma^0$ increases for an increase in water content due to the dielectric properties of water.

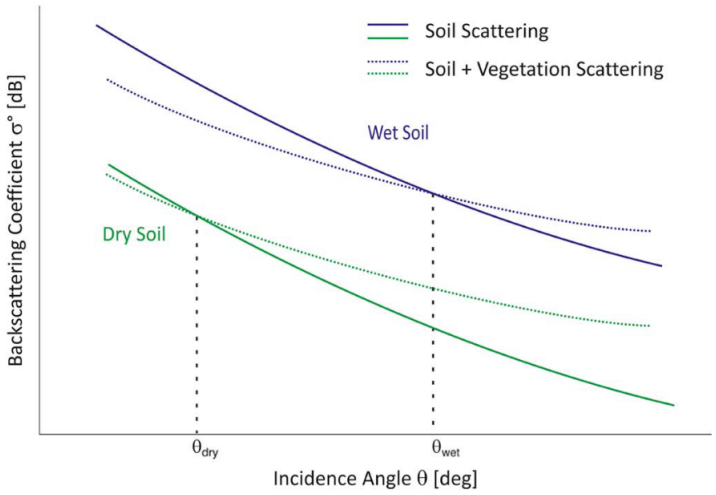



Figure 2.2: Incidence angle dependence of the backscattering coefficient $\sigma^0$. Source: Wagner et al. [35]

The second observation from Figure 2.2 is that the slopes depend upon the presence of vegetation. For the lines that represent soil scattering, the absolute value of the slope is larger than for the soil with vegetation scattering lines. Hence, the effect of vegetation on the backscattered signal can be found in the slope of the backscatter signal $\sigma^0$.

The last relevant observation from Figure 2.2 is that the relationship between $\sigma^0$ and incidence angle $\theta$ is non-linear [27]. These three observations are the principles used to describe the TU-Wien soil moisture retrieval algorithm. The approach described in [32] the slope $\sigma'$ is obtained per triplet, as the ASCAT measures $\sigma^0$ from the fore-, mid- and aft beam. The equation is shown below:

$$\sigma'(\frac{\theta_m - \theta_{a/f}}{2}) = \frac{\sigma^0_m(\theta_m) - \sigma^0_{a/f}(\theta_{a/f})}{\theta_m - \theta_{a/f}} \tag{2.2}$$

The $m, a$ and $f$ indices in the equation of the local slope $\sigma'$ stand for the mid-, fore- and aft beam. The amount of noise within the local slope values are determined by the amound of azimuthal effects, intrument noise and speckle in $\sigma_\theta$ [8]. By averaging both the local slope values (combination of forebeam with midbeam and aftbeam with midbeam) the noise can be reduced. This gives the following expression for the local slope:

$$\sigma'_l = \frac{\sigma'_{fm} + \sigma'_{am}}{2} \tag{2.3}$$

## 2.3. TU-Wien Soil Moisture Retrieval

### 2.3.1. Mathematical Approach

The TU-Wien soil moisture retrieval (TUW SMR) is a change detection method to obtain soil moisture from scatterometer observations. This method was initially developed by Wagner et al. (1999) for the ERS scatterometers. The incidence angle dependence of the backscatter forms the basis for this retrieval algorithm. The dependence is a second order Taylor expansion with the first (slope $\sigma'$) and second order derivative (curvature $\sigma''$) normalized at a reference angle. This reference angle is set to 40°. The relation between the $\sigma'$ and $\sigma''$ is linear. Hence, the equation for the slope at angle $\theta$ can be described as follows:

$$\sigma'(\theta) = \sigma'(40) + \sigma''(40)(\theta - 40) \tag{2.4}$$

The $\sigma'$ and $\sigma''$ are values at the reference angle. These values are obtained by a linear regression of the local slope values which are obtained from equation 2.2. Daily values are obtained for both $\sigma'(40)$ and $\sigma''(40)$ using a kernel smoother with a temporal window of 42 days. After obtaining the $\sigma'(40)$ and $\sigma''(40)$ the normalized backscatter $\sigma(40)$ can be obtained through extrapolation:

$$\sigma^0(40) = \sigma^0(\theta) - \sigma'(\theta - 40) - \frac{1}{2}\sigma''(\theta - 40) \tag{2.5}$$

### 2.3.2. ASCAT Soil Moisture Formulation

The historically highest (normalized) backscatter observation of a particular grid point is referred to as the wet reference and is denoted as $\sigma^0_{wet}$. Hence, the wet reference is independent from time. The dry reference ($\sigma^0_{dry}$) however, is the lowest value of the backscattering coefficient for a given date of a particular grid point. Soil moisture can be expressed as a function of time by scaling normalized backscatter observations using the wet and dry references. This gives the following expression for soil moisture:

$$\Theta(t) = \frac{\sigma^0(40,t) - \sigma^0_d(40,t)}{\sigma^0_w(40,t) - \sigma^0_d(40,t)} \tag{2.6}$$

The concepts of dry and wet reference are shown in the illustration in Figure 2.3.

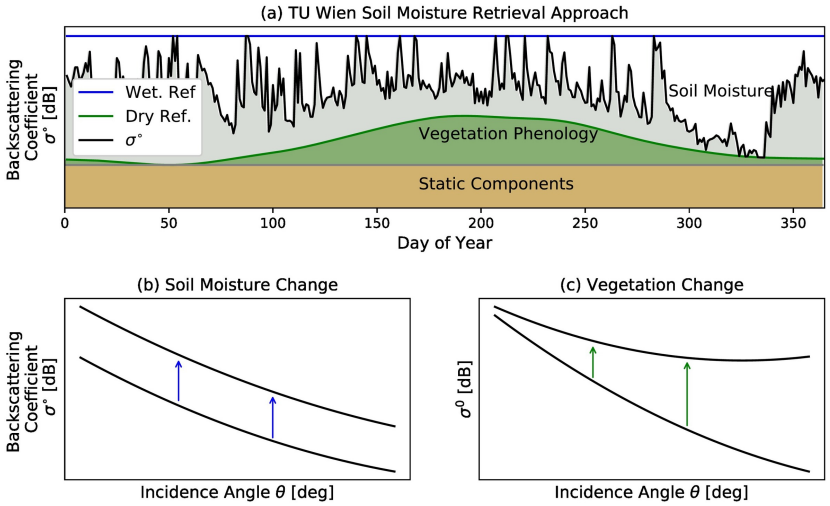


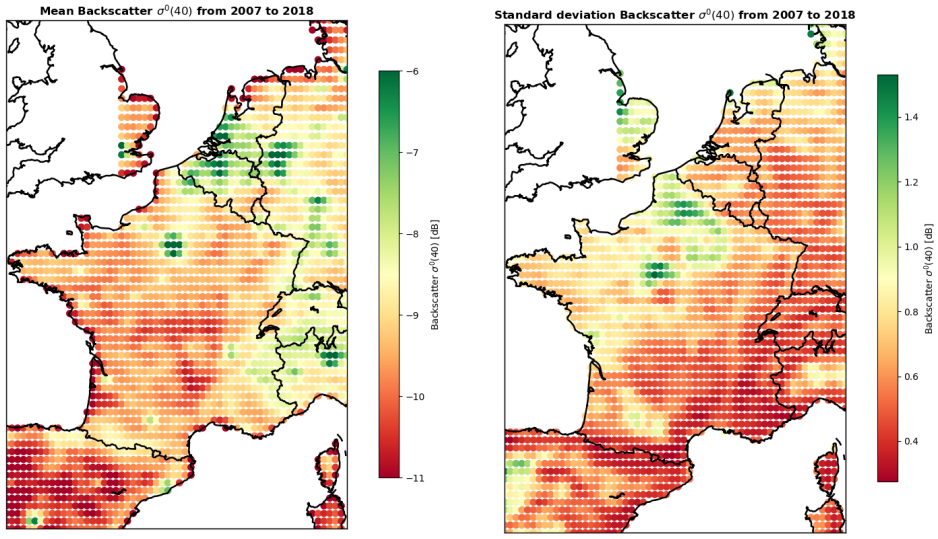

Figure 2.3: An illustration of the concepts within the TUW SMR method. The upper part of this figure shows ASCAT data from a grid point in Nebraska in a study of Steele-Dunne et al. (2019). The legend shows the relation of wet and dry reference with the backscatter. The bottom parts show the illustrate the incidence angle dependence for changes in soil moisture and vegetation.

### 2.3.3. ASCAT Data

The mean and standard deviation of normalized backscatter $\sigma^0(40)$ and normalized slope $\sigma'(40)$ time series for grid points spread over Europe are displayed in the maps in Figures 2.4 and 2.5. The grid points are computed on a 0.25° grid by Météo-France. The computed slope values have a temporal resolution of 1 day and the normalized backscatter has a temporal resolution of a few days. Only the morning measurements (descending orbit for Europe) from backscatter are used to keep a certain consistency over time.

Figure 2.4 (a) shows the mean normalized backscatter from 2007 to 2018. It can be seen that the backscatter values are high in urban areas such as Paris and the West-Coast of the Netherlands. This is because urban areas consist many smooth surfaces. If roughness increases, backscatter will be less because the energy is moving into different directions (keeping all other conditions the same). The Alps, which can be seen on the Southern-East side of France, consist of rocks and even layers of ice and hence there is less smoothness than in an urban area. Low backscatter values can be seen on the main land for grasslands, forests and agriculturally active areas. The two primary factors that influence the backscatter on areas that contain vegetation are surface and volume scattering (Wagner et al., 1999).

The standard deviations of the normalized backscatter are displayed in Figure 2.4 (b). It can be seen that in urban areas and the Alps are low compared to areas where there is agricultural activity (region around Paris) and grassland (Western part of France) (See Figure 1.3). This can be explained with the illustration in Figure 2.3 (a). During the growing season the amount of vegetation increases. Do to this increase in vegetation the backscatter behaviour for that specific area changes completely, as now the backscatter is a sum of scattering due to the surface and due to vegetation cover.

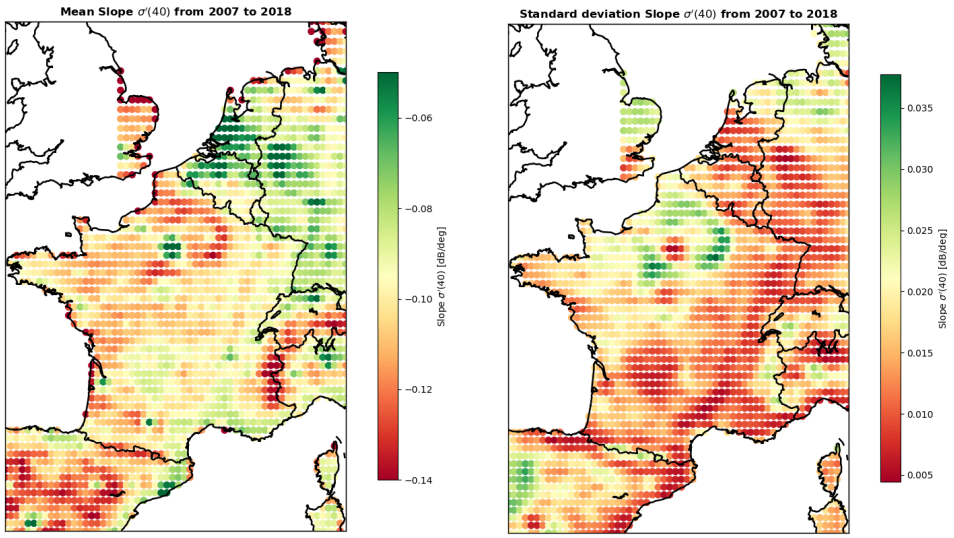



(a) Mean $\sigma^0(40)$

(b) Standard deviation $\sigma^0(40)$

Figure 2.4: The mean (a) and standard deviation (b) of the ASCAT normalized backscatter parameter $\sigma^0(40)$ over grid points in Europe.

Figure 2.5 (a) shows the mean slope $\sigma'(40)$ values in and around France for a period of 11 years. It can be seen that Paris and the urban areas in the west of the Netherlands and Belgium have a relatively high slope. This means that according to Figure 2.1, the line would be very steep. A land that contains vegetation, has a smaller slope which can be seen in the region around Paris for example. The standard deviation of the slope however (see Figure 2.5 (b), is very large in those areas for the simple reason that there is the presence of vegetation during winter time is much less than during summer.




(a) Mean $\sigma'(40)$ (b) Standard deviation $\sigma'(40)$

Figure 2.5: The mean (a) and standard deviation (b) of the ASCAT normalized slope parameter $\sigma'(40)$ over grid points in Europe.

## 2.4. Land Surface Model

This section serves as the theoretical background behind the ISBA-A-g$_s$ land surface model. This model is used for the simulation of the independent variables Soil Moisture and LAI, which are used as input for the Support Vector Regression.

### 2.4.1. ISBA

The Interactions between Soil Biosphere and Atmosphere (ISBA) model was developed at Meteo France with the aim of accurately describing surface processes for weather forecasting and climate models. The ISBA can be understood as a Soil Vegetation and Atmosphere Transfer (SVAT) model (Noilhan and Planton, 1989). The most important components of land surface processes are taken into consideration and several important coefficients are calibrated with sophisticated models and observational data.

The purpose of any SVAT scheme is to simulate energy and matter exchanges between land surface and atmosphere. These simulations could then be useful for improving the understanding of observations and hypothesis testing. ISBA is modelled within the SURFEX platform and each grid cell can be represented as a 1 dimensional layer consisting of three layers in total (Leroux et al., 2018). Transferring information from the top to bottom layer is done by force restore dynamics. Each grid cell consists of different vegetation tiles and the model is then run for each tile.

### 2.4.2. A-g$_s$

The ISBA-A-g$_s$ is the modified model of the ISBA and is developed to account for effect of the atmospheric carbon dioxide ($CO_2$) on the stomatal aperture (Calvet et al.,1998). The A-g$_s$ appended on the original ISBA is known as the Jacobs' model (Jacobs et al., 1996) and this model takes the relation between the stomatal conductance $g_s$ and the net assimilation $A_n$ of $CO_2$ into account.

### 2.4.3. Leaf Area Index

LAI can be defined as the ratio between the are of the leaf surface versus the area of the ground surface (Qiao et al.,2019). This variable is useful for a variety of reasons such as describing the geometry of the vegetation canopy, light interception within a certain area and seasonal dynamics of vegetation. Also, LAI can serve as a measure of vegetation density and is therefore a useful variable to relate to the ASCAT slope $\sigma'(40)$ parameter. There are several ways in which an estimate of LAI can be obtained. The most commonly known are:

- In situ measurements through destructive sampling.
- Model simulations.

- Retrievals with (optical) remote sensing.

The first given option requires a large network of in-situ measurements to get useful insights on a regional scale. LAI obtained with optical remote sensing are done with vegetation indices. These indices use the Near Infrared (NIR) and red band of the optical spectrum.

### 2.4.4. Soil moisture

Within this study, the soil moisture in the upper layers of soil (not root zone) is considered. This is important due to the fact that the wavelength of the ASCAT is only 5.7 centimeters, thus the signal is more related to surface soil moisture than to soil moisture at depths of 5 centimeters. The soil moisture used as input variable for the Support Vector Regressions is the simulation between 1-4 centimeters depth. In-situ observations of soil moisture are not dense enough for the region of this study. The importance of this variable is given in the chapter 1.

### 2.4.5. Validation of Land Surface Model Variables

Gibelin et al. (2006) showed that the simulated LAI by the ISBA-A-$g_s$ is able to capture the patterns of satellite derived LAI. The study area was not a specific place but rather a global study. They also found spatial consistency during the growing seasons of vegetation. Brut et al. [3] compared satellite derived LAI with the simulated LAI of ISBA-A-$g_s$ to determine if the ISBA-A-$g_s$ is able reproduce the seasonal and interannual variability of LAI on regional scale. The satellite derived LAI came from three sources: Estimation of LAI with MODIS Bidirectional Reflection Distribution Function product, LAI estimation through MODIS collection 4 dataset and the CYCLOPES V3 LAI. An absolute true value for LAI as both observations and simulations contain uncertainties. For the satellite products, these are the following:

- Saturation effect; Which occurs when there is a large amount of vegetation within a certain area. The energy that is reflected in the red and near infra red (NIR) does not increase beyond a certain threshold. Hence, the Normalized Vegetation Difference Index (NDVI), which is used as an indicator of LAI, is therefore limited in the estimation.
- All vegetation elements (including dead leaves) affect the derived LAI with the optical sensors.

For the LAI simulated by the ISBA-A-$g_s$, the following sources of uncertainties are mentioned in Brut et al. (2009):

- Highly uncertain parameters in the model are the maximum water capacity for plants and plant rooting depth. Consequently, there is an offset in the date at which the maximum simulated LAI is obtained.
- The spatial heterogeneity of the land cover, with the focus here on agriculturally active areas are quite challenging to account for. C3 crops for example can be wheat, rapeseed and barley within France. But C3 crops could also represent a Summer crop such as sunflower. These crops differ in growing cycles which makes the model uncertain.
- The other complexity regarding agriculturally active areas is that land management practices are not accounted for.
- The last potential source of uncertainty mentioned by Brut et al. (2009) is the response to temperature of photosynthesis.

### 2.4.6. ECOCLIMAP Database

The initialization of the used Soil, Vegetation and Atmosphere Transfer model (SVAT) requires proper allocation of land water masks and soil-vegetation characteristics [18]. The dataset that serves this purpose is the Ecoclimap. This is a global surface parameter dataset with a spatial resolution of 1 km. Each grid point consists of different vegetation patches and the spatial heterogeneity is therefore maintained. The possible land cover types within Ecoclimap are: Bare land, bare rock, permanent snow, boreal broadleaf deciduous, temperate broadleaf deciduous, tropical broadleaf deciduous, temperate broadleaf evergreen, boreal needleleaf evergreen, temperate needleleaf evergreen, boreal needleleaf deciduous, shrubs, boreal grassland, temperate grassland, tropical grassland, Winter C3 crops, Summer C3 crops, C4 crops, flooded trees, flooded grassland.

## 2.5. Support Vector Machines

The constraint for a support vector regressor is to minimize the error between the output of a function given a set of independent variables and the true value of the dependent variable (Smola and Schölkopf, 2004). The estimation function of a support vector regressor can take the form as is given in equation 2.7. $Y$ is the dependent variable which is estimated with the sum of the dot product of the weight vector $W$ and the Support Vector $X$ and the estimated bias $b$ .

$$Y = f(x) = W \cdot X + b \tag{2.7}$$

With

$$X, W \in \mathbb{R}^d$$

With support vector regression (SVR), feature engineering can be done within an equation itself. This implies that instead of transforming the (non-linear) data explicitly into a high dimensional feature space, the transformation is done though an inner dot product. This is known as the Kernel Function. This method allows for simulating non linear regression problems. A model $\hat{y} = f(x)$ is created based on trainingset $D_{train} = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)\}$ so that $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The linear support vector regression is given in equation 2.8.

$$f(x, w) = \sum_{i=1}^{N} \langle w_i, \phi(x_i) \rangle + b \tag{2.8}$$

- $w$ denotes the weight vector
- $b$ is a constant and referred to as the bias.
- $\phi(x)$ is the mapping function.

The weight vector $w$ and bias $b$ are estimated with the training data. $x$ is the set of independent variables or features vector. $\phi(x)$ is a function that can transform the feature space of the independent variables so that a linear fit in. The aim is to construct an estimation function $f(x, w)$ that is build on training data and is able to estimate data not used during training. To optimize the SVR, a cost- or loss function needs to be minimized. The optimum of the cost function is the hyperplane that is the best fit for the data.

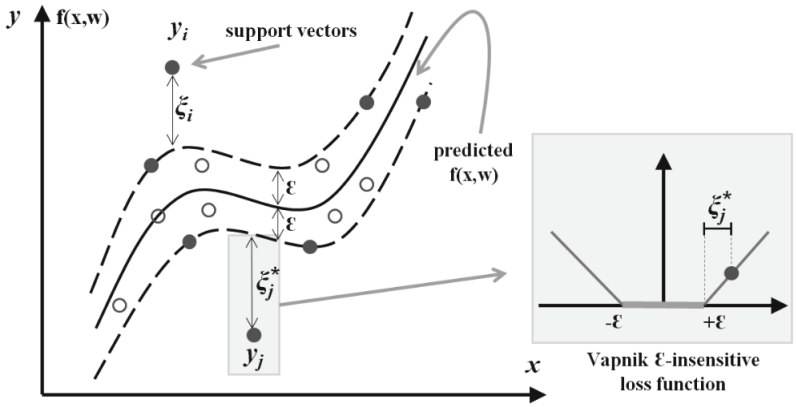



Figure 2.6: $\varepsilon$ -intensive loss function. The filled circles represent the support vectors. It can be seen that some support vectors are outside of the $\varepsilon$-margin. Source: Rojas et al. (2017)

Figure 2.6 introduces the Vapnik $\varepsilon$-intensive loss function. On the left hand side can be seen that a prediction model $f(x, w)$ is set with a tolerance margin $\varepsilon$. It is shown in the right part of the figure that only points which are outside of the $\varepsilon$ region are contributing to the loss function. The penalty for these points increase linearly with the deviations.

The aim is to get an optimized solution so that the error is minimized while prevent overfitting at the same time. To do this, a Langrange function needs to be constructed from the primal form using a set of conditions. The Karush-Kuhn-Tucker (KKT) conditions allow for obtaining an optimal solution by taking inequality constraints into account. The primal form of the constraint optimization problem of the support vector regression is given in equation 2.9

$$\text{minimize} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \quad w \in \mathbb{R}^d, \xi \ \& \ \xi^* \in [0, \infty]^n \tag{2.9}$$

Such that

- $y_i - w^T x_i - b \leq \varepsilon + \xi_i^* \quad \forall i$
- $w^T x_i + b - y_i \leq \varepsilon + \xi_i \quad \forall i$
- $\xi_i, \xi^* \geq 0 \quad \forall i$

The $C$ in equation 2.9 is a constant and determines to which amount deviations larger that $\varepsilon$ are tolerated. The 'slack' variables are necessary to obtain feasible constraints of the optimization problem, making it a soft margin regressor. The complete derivation can be seen in Smola and Schölkopf (2004). The result of the derivation is the optimization problem with constraints:

$$\text{maximize} -\frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*) \tag{2.10}$$

Such that:

- $\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0$
- $\alpha_i, \alpha_i^* \in [0, C]$

The Representer theorem states that the solution of weight vector $w$ can be described as a linear combination of the training data. Hence, $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ where N is the number of support vectors. Substituting this in the primal form $f(x) = w^T x + b$ gives equation.

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) x_i^T x + b \tag{2.11}$$

Equation 2.11 is the linear representation of the SVR.

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{2.12}$$

The linear kernel function is similar to a linear regression. From Figure 3.3 and Figure 3.4 can directly be observed that the problems are non-linear. Hence, a kernel must be chosen that will be able to linearly fits a model in the data in a higher dimensional space. Both Linear and Polynomial kernel have the advantage of being less time consuming then the Gaussian Radial Basis and the Sigmoid. The objective of using support vector machines for this is to investigate whether an optimized kernel is able to simulate the ASCAT parameters $\sigma^0(40)$ and $\sigma'(40)$ and this should be done for many grid points and thus different type of data distributions. Therefore, the Radial Basis kernel is chosen as initial approach towards a set of optimized models for grid points in and around France. This kernel method was also chosen by Forman and Reichle (2015). The expression of the radial basis function is given in 2.13.

$$K(x_i, x_j) = exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) = exp\left(-\gamma ||x_i - x_j||^2\right) \tag{2.13}$$

# 3

# Methodology

The objective is to create an optimization routine to simulate ASCAT $\sigma^0(40)$ and $\sigma'(40)$. This chapter briefly discusses every step in this process. First, the general approach of the methodology will be explained, which includes the requirements of the machine learning model and the choice for the machine learning model. After that, the workflow is explained which serves as an general overview of the whole procedure. This is followed by a more in depth understanding of the Support Vector Regression, primarily the role of hyper-parameters during the optimization process and the grid search optimization with cross validation is briefly discussed as well.

## 3.1. General

### 3.1.1. Requirements of machine learning model

The choice of the machine learning model is based on the requirements for the simulation of the normalized backscatter $\sigma^0(40)$ and slope $\sigma'(40)$. The key drivers of the backscatter signal and slope do not linearly depend on land surface variables. Hence, a non-linear model has to be obtained in order to validate this driving mechanisms behind the backscatter and its slope.

Generally, regression for non-linear data is done with either Neural Networks or Support Vector Machines. Other functions such as linear regression can be used as well, however the big disadvantage is that features need to be transformed explicitly. The model should allow for the optimization of parameters, as the aim is to understand if specific land cover types contain similar parameters and how the quality of simulations can be interpreted.

### 3.1.2. Workflow

Support vector regression is chosen as the estimation method. The three main steps taken to train the SVM kernel regressor are similar to the description of standard kernel classifiers in Huang and Lin (2016 and are enumerated below:

1. Scale the input features to the interval [-1,1]

2. Obtain the optimal model parameters $C$ and $\gamma$. This entails that for the chosen kernel function the highest accuracy is obtained with cross validation.

3. Create the model $w$ with the optimized hyperparameters $C$ and $\gamma$

The main principles are the same for regression and thus are used for this study. Figure 3.1 gives a complete overview from data to model validation. The period from Januari 2008 till September 2017 is time period of which the data is used for training and optimizing the support vector machines. The final year spans from September 2017 till September 2018 is kept for the final simulation and validation of the model. Each grid point (GPI) has its own optimized support vector machine.

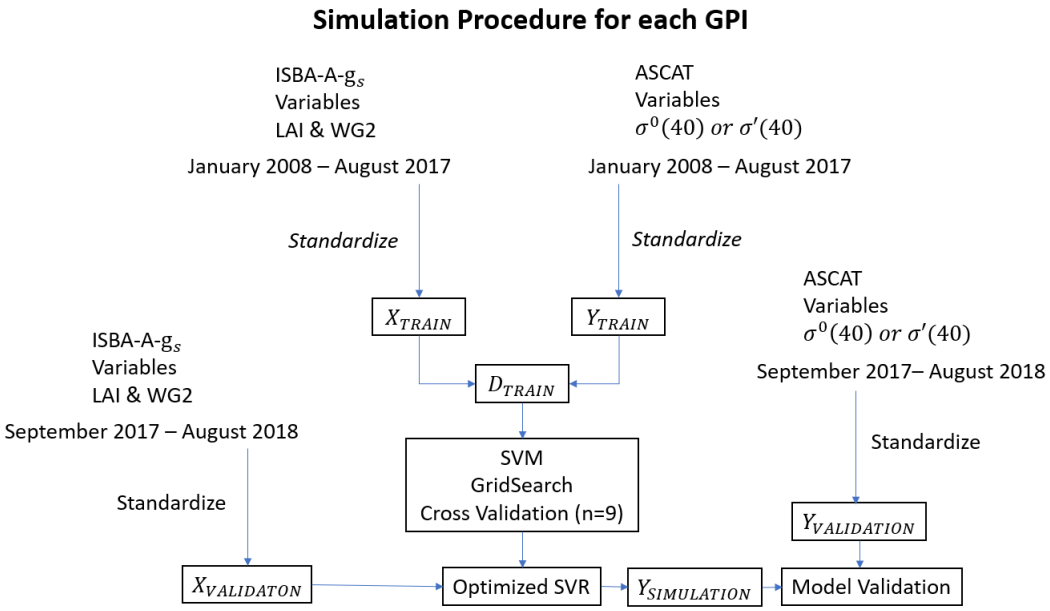



Figure 3.1: Workflow of simulations for each individual grid points (GPI).

Each step of the simulation procedure is briefly described in this chapter. In the next section, data preparation is described followed by the theory behind support vector machines in the case of regression.

## 3.2. Data Preparation

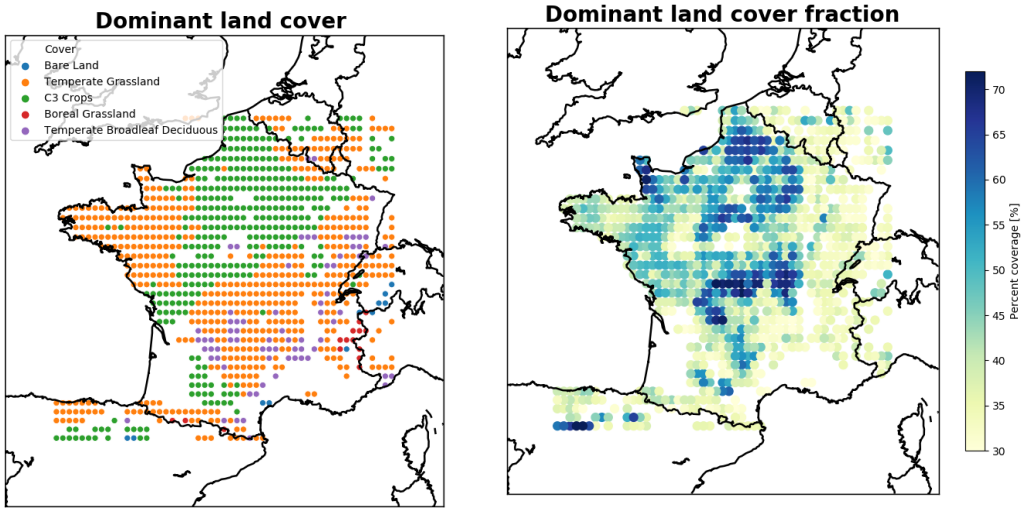
### 3.2.1. Standardization

The Support Vector Machine tries to optimize margins using input variables that initially have their own scale and units. As the distance between vectors becomes relevant, an optimized margin cannot be obtained. Therefore, prior to training the model the data should be rescaled so that there are no issues of different units and scales. Standardization means that the data is rescaled so that the data has the properties of a Gaussian distribution with the mean $\mu = 0$ and $\sigma = 1$. The equation for standardization is shown in 3.1.

$$z = \frac{x - \mu}{\sigma} \tag{3.1}$$

### 3.2.2. Grid points

The grid points are primarily located in France and only grid points are considered that have a dominant land cover type. For this study, this implies that a grid point should contain at least 30% of a particular land cover type. The map with the dominant land cover types is displayed in Figure 3.2 (a) and on the right hand side of the Figure (b) the dominant land cover fraction is displayed.




(a) Dominant land cover types

(b) Percentage of land cover in grid cell

Figure 3.2: On the left hand side of the figure (a) the dominant land cover types are displayed on the map. It can be seen that there are 5 main land cover types. On the right hand side of the figure, the fractional coverage of these land cover types is displayed. Only grid points that contain a land cover with a fraction equal to or larger than 30 % are displayed and processed.

### 3.2.3. Data GPI for Day of Year

Each gridpoint in the the ASCAT backscatter and ASCAT slope data is labeled with a unique point id. The backscatter is a dependent variable and its independent variables are the LAI and WG2. All data are within a time span from January 2008 till September 2018. Figure 3.3 displays the two feature variables WG2 and LAI and the ASCAT backscatter observations for a particular grid point, with a unique identity of 782653. This grid point consists of at least 30 % temperate grassland.

It can be observed om Figure 3.3 that that for the first 50 days of the year the backscatter $\sigma^0(40)$ shows similarities with the soil moisture. After that period, vegetation starts to come and the backscatter looks much less similar to the soil moisture values. In Figure 3.4 can be seen that the slope $\sigma'(40)$ seems to follow the main pattern of the leaf area index. However, the seasonal variations are not very explainable by just the leaf area index.

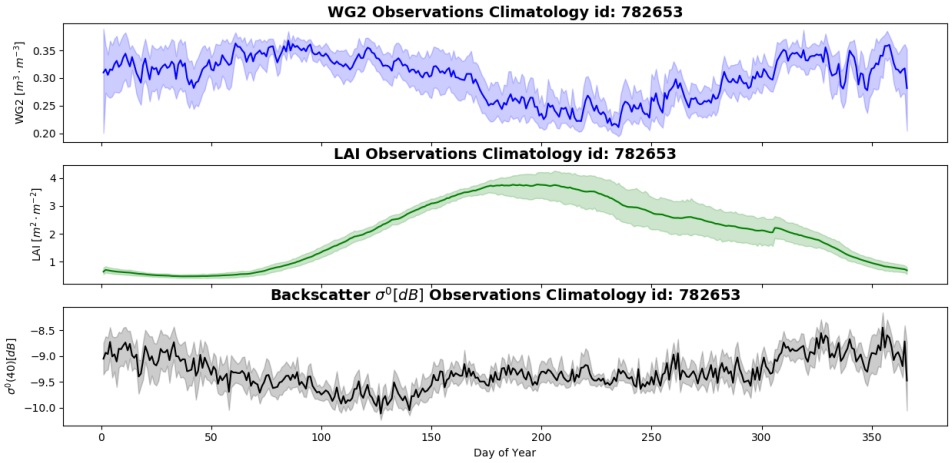



Figure 3.3: Climatology of Soil Moisture, LAI and slope $\sigma^0(40)$ for grid point 782653

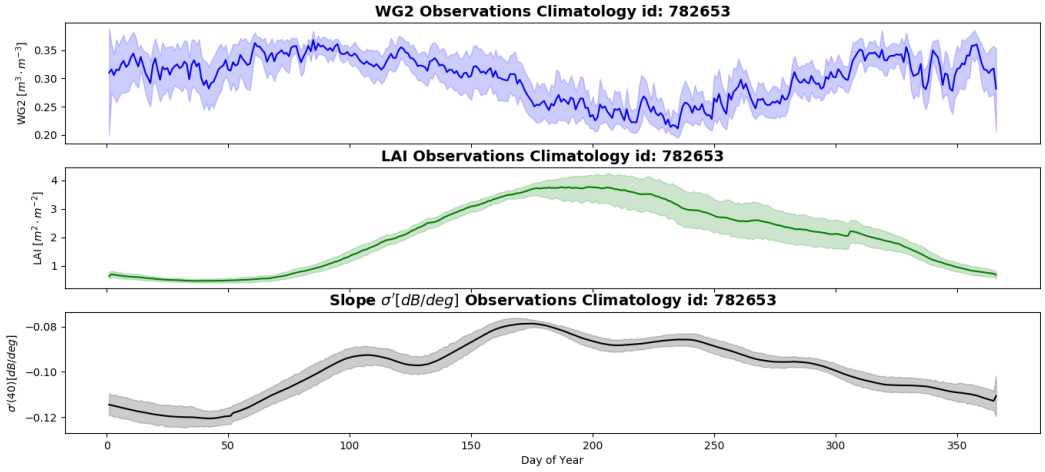



Figure 3.4: Climatology of Soil Moisture, LAI and slope $\sigma'(40)$ for grid point 782653

## 3.3. Hyperparameter optimization

### 3.3.1. Grid Search and Cross Validation

A set of possible parameter combination is given to the optimization model. This is the grid search space. Each combination in this space is tested and the performances of all combinations is compared. To make the parameter selection a reliable process, a k-fold cross validation is done to select the model parameters with on average the best regression score. An illustration of this process is given in Figure 3.5. The metric that is used for making the decision which parameter set is best is the R2 coefficient of determination. This metric is a measure for the proportion of variance of an estimated variable $y$ that is explained by the independent variables. Hence, it is a good indication for the goodness of fit. This metric is also chosen for the final validation of the model as it makes performances between different grid points comparable. The equation of the coefficient of determination is shown in 3.2.

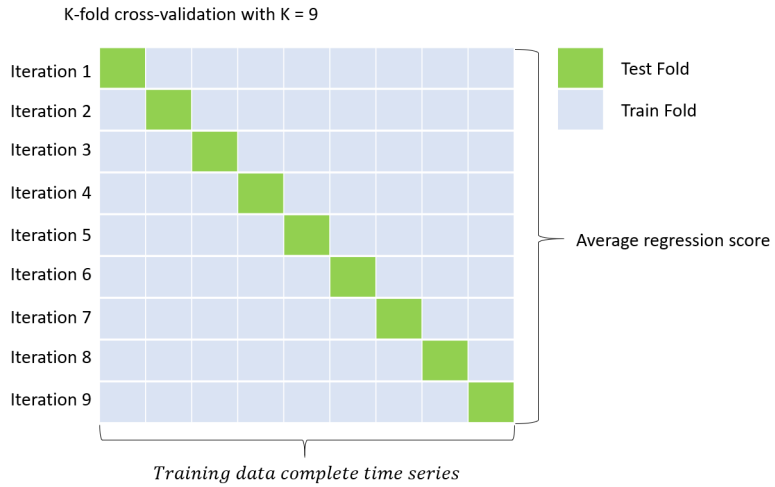



Figure 3.5: Iteration scheme of the k-fold cross validation. The green boxes represent the test fold and the grey boxes are the train folds. The value of k for the cross validation of training data is set 9. This means that the 9 years of training data is equally split in 9 pieces and for each iteration 8 pieces are used for training and one for testing.

### 3.3.2. Software

The package that is used to create and optimize the model is Scikit-Learn in Python.The Scikit-Learn packages offer many tools for machine learning (classification, regression and clustering). This includes the grid search with internal cross validation.

### 3.3.3. Bias-Variance Trade-off

The importance of a bias-variance trade of can be explained by the illustration in Figure 3.6. On the left hand side can be seen if the model under-fits and is too simpel to understand the problem. Hence, the bias of this model is high and the variance very low. On the right hand side its the opposite. The model overestimates and has a high variance while the bias is very low. If a model performs too well on a training set, it means that the model not only learns the training data, but the noise in the training data as well. Even if theoretically the training data does not have any noise, errors will occur due to the complexity of the model.

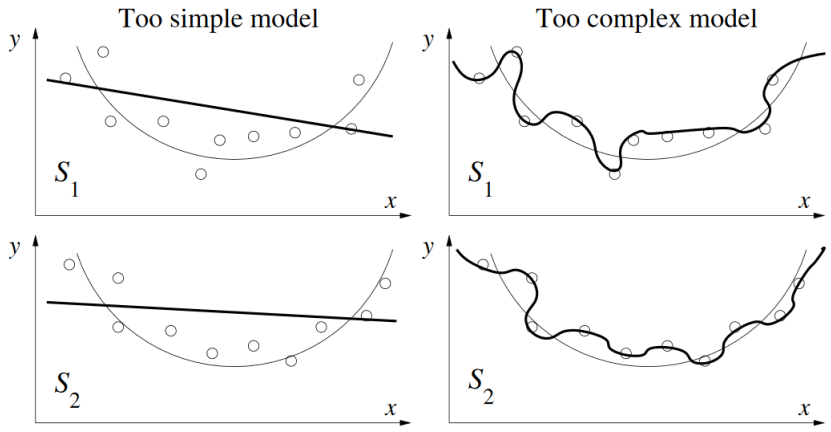



Figure 3.6: The two models on the left hand side under-fit the training sample while the two models on the right hand side are clearly over fitting. Source image: Maimon and Rokach [17]

The scikit-learn software package provides a cross validation tool in which optimization can be done with a given set of desired hyperparameters. This is called a grid search. Another way of tuning the hyperparameters is randomized searching. A big advantage of randomized searching over grid searching is that the number of iterations can be exactly specified by the user, which make it very applicable for applications with a high number of hyperparameter combinations. The disadvantage of using this way of tuning over grid search is that it is more difficult to check for spatial consistency as parameter values vary on a log scale.

The advantage of using this given set of desired hyperparameters is because 1324 regression models are created and this process of optimization is computationally less expensive than randomized searching. Different land cover types with different land cover characteristics require different orders of magnitude for the machine learning algorithm parameters.

### 3.3.4. SVM Parameters

The hyperparameters that are optimized are $C$ and $\gamma$. The selection of values for each of these variables is shown below. The way these values are obtained is by an iterative approach.

$$C \in \{0.1, 1.0, 10, 100, 1000\}$$

$$\gamma \in \{1.0 \cdot 10^{-4}, 1.0 \cdot 10^{-3}, 1.0 \cdot 10^{-2}, 0.1, 1.0\}$$

The aim of regularization parameter $C$ is to penalize the parameters in the model so that the model will much less vulnerable to noise and hence does better at generalization. A high $C$ means that the model likely estimates the training data more accurately and hence has a higher probability of overfitting. Consequently, a lower $C$ simplifies the SVR which means that the training accuracy is lower as well.

The $\gamma$ is the parameter within the Kernel function (see eq. 2.13). It serves as a similarity measure and determines to what extend Support Vectors are used in the function. A very low $\gamma$ overgeneralizes while a very high $\gamma$ causes the model to overfit. It is worth mentioning that the focus of this study is not to justify the obtained values for $C$ and $\gamma$, but rather explore the possibility of optimizing simulations and discover whether there is spatial consistency with parameters.

## 3.4. Performance Evaluation

The performance metric that is used is the coefficient of determination (R2). The expression is given by the SciKit-learn software package and is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3.2}$$

With $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$

A negative value of R2 means that the mathematical mean of the observations is a better estimate than the simulations. The intuitive reasoning for using the R2 over the root mean square error is because the aim is to see to what extend the created support vector machines are able to capture the variability of the $\sigma^0(40)$ or $\sigma'(40)$, rather than depending on the signal to noise ratio of different grid points. However, the RMSE is also calculated for each grid point and in Section 4.1 is shown that the RMSE displayed on a map is an indicator of the standard deviation. Therefore, this performance metric is less useful.

In order to assess the model performance for high frequency observations in the backscatter $\sigma^0(40)$, a climatology fit is created and subtracted from the initial backscatter regression for each grid point. This is done by averaging all LAI and soil moisture values for a particular day of the year for 10 years of data for each grid point, which are then used as input in the created SVRs and thus a climatology estimation is obtained. In this way, a fair comparison of high frequency signal performance can be made between different grid points. Figure 3.7 gives an illustration of this procedure.

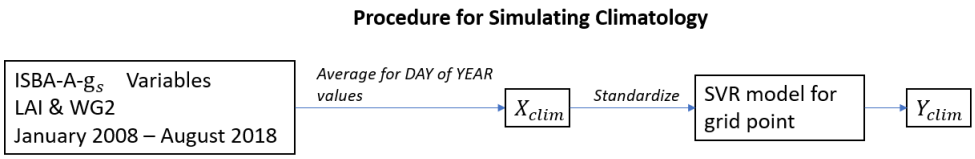



Figure 3.7: Procedure for simulating climatology with standardization of the climatological fit.

# 4

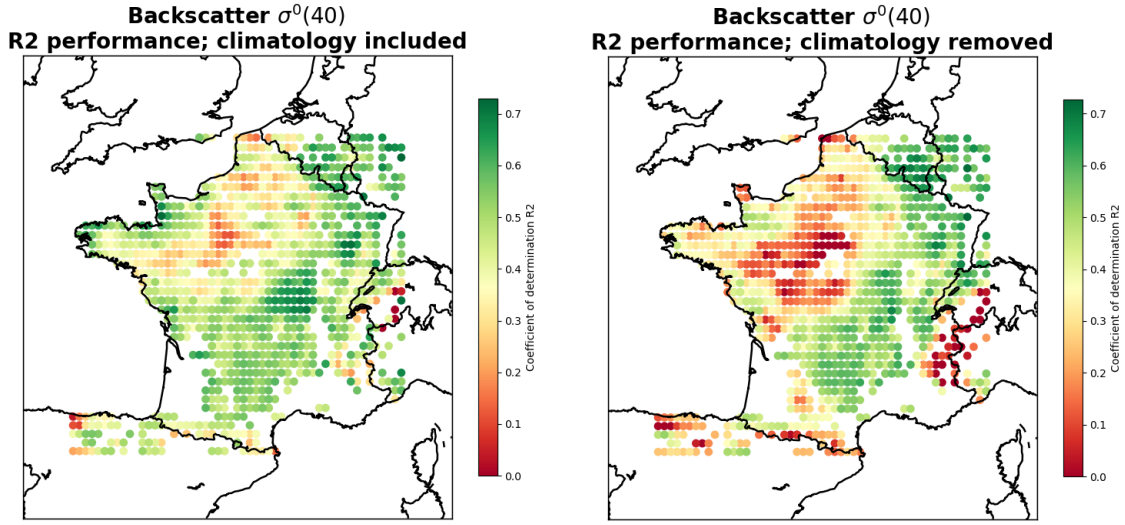
# Results

In this chapter, the performances of the simulations of grid points that contain at least 30% of a specific land cover type are shown and evaluated. For both backscatter and slope, the performances are displayed on maps. Performances are calculated over three different time periods: One full year and two separate periods of 6 months. The first period of 6 months is during the winter months and the second period of 6 months is during the growing season. The simulations, as described in section 3, are done with SVR models created with 9 years of training data and these models are then used to simulate backscatter $\sigma^0(40)$ or slope $\sigma'(40)$ on one year of unseen data, which is from September 2017 till September 2018. Simulations of individual grid points are shown together with the corresponding interannual variability of the land surface variables by which the model is trained. After that, statistics of land cover fractions of grid points with a certain dominant land cover type are shown and evaluated. In the end of this chapter, the spatial consistency of SVM parameters is shown.

## 4.1. Backscatter Simulations

### 4.1.1. Overall Performance

The overall performance of the simulation in this study is the value of the coefficient of determination (R2) for a time period of one year. This time period includes all seasons. The performances of the backscatter simulations are shown on the map in Figure 4.1. The left side of the figure (a) shows the R2 performance of the complete simulation and on the right side (b) the R2 performance is shown where climatology is simulated and subtracted from the complete simulation. It can be seen that in both maps a spatial consistency exists. Looking at the land cover type map and fractional cover map in Figure 3.2 (a) and (b) it can be suggested that the dominant land cover type and land cover fraction are significant factors for the $\sigma^0(40)$ performance. It can also be observed that the R2 performances of the grid points shown in (b) are generally lower than in (a). This suggests that for the points where this is the case that the model is able to capture the seasonal cycles, but not able to capture the high frequency variations not caused by the seasonal cycles.

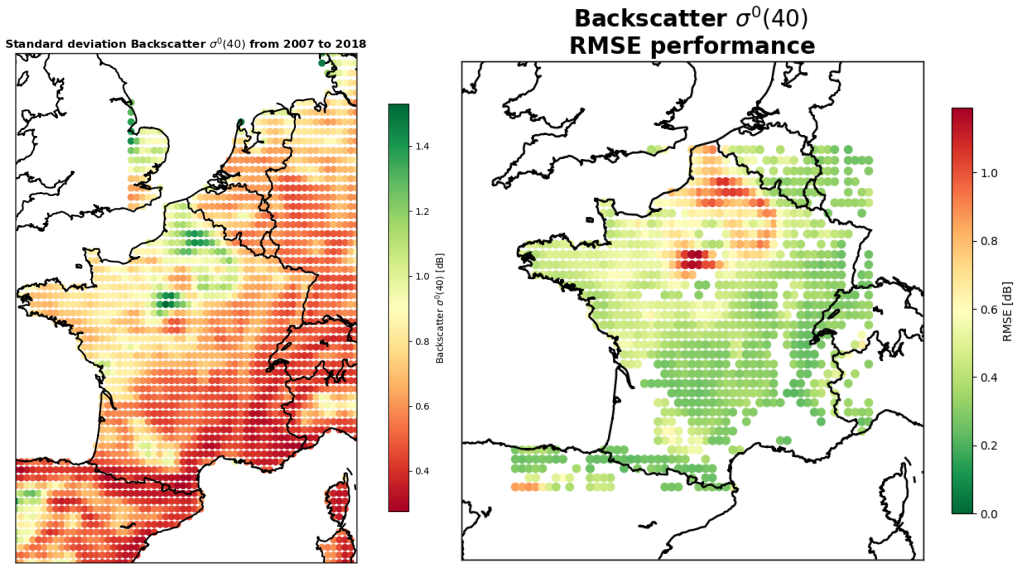



(a) Climatology included

(b) Climatology removed

Figure 4.1: Map of backscatter $\sigma^0(40)$ performances. The left hand side of the figure (a) shows the R2 performance with the climatology included. The right hand side (b) shows the performance of the regression

As mentioned in chapter 3, the coefficient of determination (denoted as R2 in this study) is the chosen performance metric for the evaluation of created support vector machines. The reason for this is to get a measure of how well variability of the ASCAT $\sigma^0(40)$ and $\sigma'(40)$ can be captured with the support vector machines and the land surface variables soil moisture and leaf area index from the ISBA-A-$g_s$. In addition to that, it makes the grid points displayed on a map more comparable, because the performances will otherwise (In case of using a RMSE or MSE as performance metric) only be linked to the standard deviation of the $\sigma^0(40)$ or $\sigma'(40)$. This is demonstrated in Figure 4.2. The left part (a) of the figure is a map with the standard deviation of the backscatter. The high standard deviations can be observed in the area left-beneath and right-above Paris. These are agriculturally active areas and the RMSE map (b) on the right hand side of the figure mirrors the standard deviation of the RMSE. It therefore lacks in providing information about which grid points have a model that works successfully and which ones not.




(a)

(b)

Figure 4.2: On the left hand side of the figure (a) the standard deviation of the backscatter $\sigma^0(40)$ is shown over a part of Europe. On the right hand side the RMSE performances are displayed for the selected grid points.

The spatial consistency that can be noticed in Figure 4.1 show that the type of land cover does indeed determine whether a simulation is possible or not. Each grid point displayed on the maps in Figure 4.1 (a) and (b) have a dominant land cover type. The requirement as stated in the methodology (see Section 3) is that a land cover is considered to be dominant if the fractional land cover is larger or equal to 30%. The dominant land cover types that at least ten times are Temperate Grassland (522), C3 Crops (351), Temperate Broadleaf Deciduous (73), Boreal Grassland (15) and Bare Land (11). Boreal Grassland is left out for further analysis as the locations of its grid points are near the Alps and during winter months the SVM will not be trained for the right reason, which is to simulate backscatter with soil moisture and vegetation density.

To have a general idea of how a dominant land cover type can be linked to the performance of the support vector machines, scatterplots have been created in which the coefficient of determination (R2) is plotted against the fractional coverage of the dominant land cover type. These can be seen in Figures 4.3 and 4.4 where the R2 performances of dominant land covers are shown for the total simulation and simulation with climatology removed. Both Figures show very similar results in terms of their distribution. As could be noticed in Figure 4.1 as well, a bunch of grid points have a high R2 for the total simulation but perform worse after climatology is removed. This implies that these grid points do reasonably well with low frequency variations such as seasonal variations, but besides these low frequency variations the support vector machine does not seem to understand much.

On the top left of Figures 4.3 and 4.4 can be seen that R2 values tend to be lower for a high fractional cover of C3 Crops than for a low fraction of C3 Crops. Grid points with C3 Crops as dominant land cover type are agriculturally active areas and land management practices are not understood well by the ISBA-A-$g_s$ land surface model (Canal et al.,2014). Therefore, it is suggested that an increase of presence of C3 Crops diminishes the performance of the simulation. However, the scatterplots show that this does not always have to be the case as individual grid points with a fractional cover of around 70% can still achieve an R2 of 0.4. On the top right scatterplot in Figures 4.3 and 4.4 can be observed that a high fractional cover of Temperate Grassland seems to give a minimum achievable performance. Fractional covers of Temperate Grassland of around 70% achieve a minimum R2 of 0.4 and a fractional cover of 50% ensures at least a R2 of 0.25 for the case of total simulation. The bottom scatter plots represent the performances of grid points with Temperate Broadleaf Deciduous and Bare Land as dominant land cover type. There is not enough data to indicate from these figures whether they enhance or diminish R2 performance.

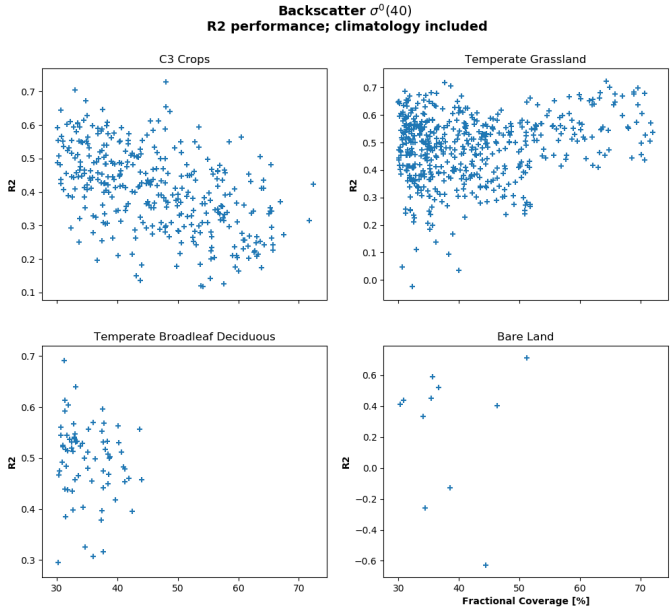



Figure 4.3: Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages. Climatology is not removed.

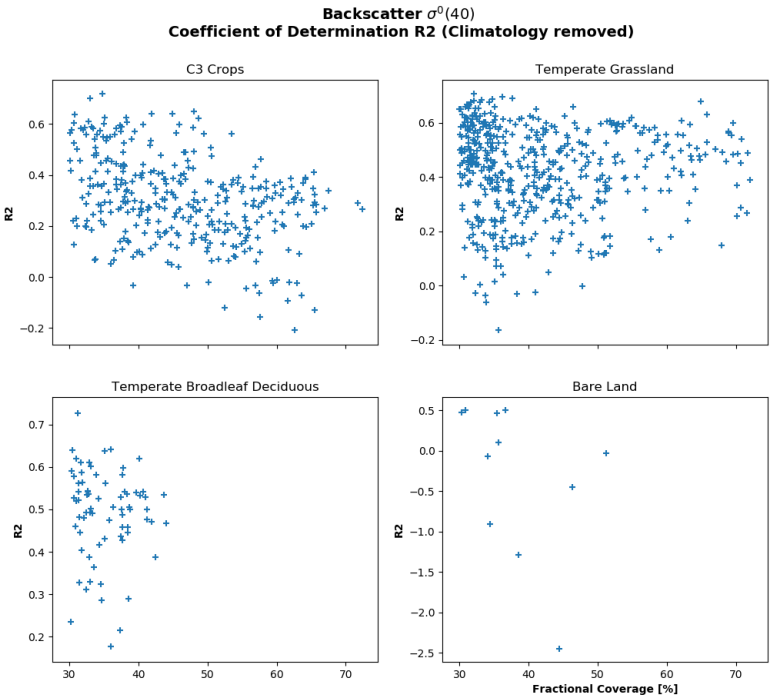



Figure 4.4: Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages. Climatology is removed.

The R2 values displayed in Figures 4.1, 4.3 and 4.4 represent the performance for a time period during the whole year. In the next section, the performance of the simulations during the vegetation growing season, which is from March till September and the 'winter' season from September to March, is evaluated.

### 4.1.2. Performance During Vegetation Growing Season

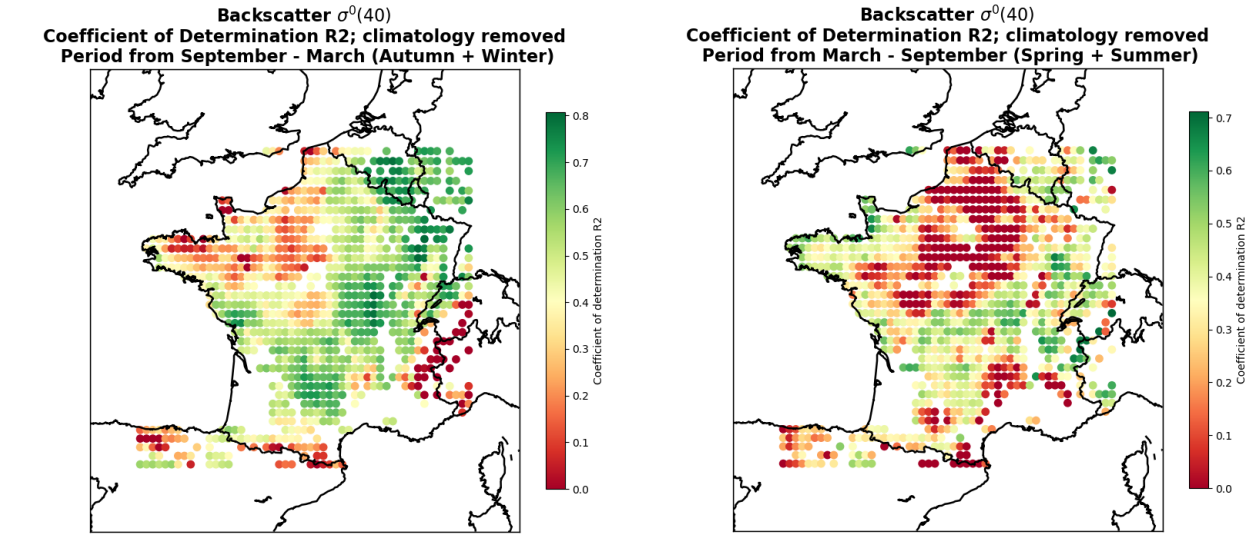
The presence of vegetation during summer might have a completely different impact on this performance than during winter. During winter, vegetation density is generally less than during summer and precipitation values are higher. In the absence of vegetation the backscatter is primarily driven by soil moisture. It is therefore expected that backscatter simulations are more successful during winter than during summer. It is shown in Figure 4.5 that this expectation is true for most grid points. On the left hand side of the figure (a) the R2 values for the autumn and winter season are shown on a map and on the right hand side (b) for the spring and summer season. It can be seen that from the centre North to centre Middle of France most grid points have a worse performance during the vegetation growing period than during winter months. However, it can also be seen that for the North-Western and South-Eastern part of France this is the opposite. So on these locations performance is worse in winter than during summer. The South-Eastern part of France contains grid points near or within the Alps which is covered with snow during winter and therefore the model is not trained for the reason to understand soil moisture and LAI impact on backscatter.

| | $\mu$ R2 | $\sigma$ R2 | $\mu$ R2 SS | $\sigma$ R2 SS | $\mu$ R2 AW | $\sigma$ R2 AW | GP |
|---|---|---|---|---|---|---|---|
| C3 Crops | 0.30 | 0.17 | -0.14 | 0.33 | 0.44 | 0.17 | 351 |
| Bare Land | -0.29 | 0.93 | -0.42 | 1.02 | -0.57 | 1.25 | 11 |
| Temperate Grassland | 0.42 | 0.16 | 0.24 | 0.21 | 0.46 | 0.22 | 522 |
| Temperate Broadleaf Deciduous | 0.49 | 0.11 | 0.32 | 0.17 | 0.50 | 0.17 | 73 |
| Boreal Grassland | 0.03 | 0.14 | -0.04 | 0.19 | -0.11 | 0.30 | 15 |

Table 4.1: $\mu$ is the mathematical mean. $\sigma$ is the mathematical standard deviation. R2 is the coefficient of determination. SS represents Spring + Summer and AW represents Autumn + Winter.

A statistical perspective on the performances of dominant land cover types during the vegetation growing season and winter months is shown in Table 4.1. The statistical properties that are considered in the table are the mean $\mu$ and standard deviation $\sigma$.

- Grid points with dominant land cover type C3 Crops on average perform bad during the vegetation growing season with an coefficient of determination below 0. This means that during this period the statistical mean of the observed $\sigma^0(40)$ is a better estimate than the SVM model simulated $\sigma^0(40)$.
- Grid points with Bare Land as dominant land cover type perform generally worse than other dominant land cover types. However, there is a huge variance in the performances. Presence of Bare Land does therefore not necessarily mean a high or low R2 value. There are insufficient amount of grid points with Bare Land as dominant land cover type to make an accurate statement.
- Grid points with Temperate Grassland as dominant land cover type show on average show poorer performance during summer than during winter. This is in line with the theoretical reasoning that during vegetation growing periods $\sigma^0(40)$ is harder to simulate than during periods where vegetation density is minimal. It can also be noted that compared to C3 crops, the performance during summer is very good.
- Grid points with Temperate Broadleaf Deciduous can arguably be considered most consistent in terms of the having the lowest standard deviations during all considered periods. The performance during winter months is higher than during the vegetation growing period. Hence, this example also supports the theoretical arguing that performance during vegetation growing season for backscatter simulations is worse than during winter months.




(a) (b)

Figure 4.5: On the left hand side of the figure (a) the coefficient of determination R2 from $\sigma^0(40)$ simulations is displayed during the period from September to March which is during the Fall and Winter season. On the right hand side (b) the R2 is displayed from the period from March till September during the vegetation growing season.

### 4.1.3. Performances and Land Surface Variables

In this part, the actual simulations of individual grid points are shown. In addition to that, the interannual variations of the land surface variables are displayed, to obtain insight in the driving factors behind the simulation and performance. Two types of simulations are shown for:

(I) Grid point with best performance during winter months

(II) Grid point with best performance during vegetation growing season

This is done for the C3 Crops and Temperate Grassland dominant land cover types as these land cover types are most present in the data. For the remaining land cover types, a few simulations can be seen in the Appendix. The reason for distinguishing these two particular cases is to obtain a better understanding of why performances are high or low during a specific time period. It was shown in the performance maps in Figure 4.5 (a) and (b) that for backscatter simulations performances are better during winter months than during summer months, due to the absence of vegetation. This subsection demonstrates the role of the ISBA-A-$g_s$ variables in the performance of the simulation.

**C3 Crops (I)**

The simulation of the grid point with dominant land cover type C3 crops that has the best performance during the winter period can be seen in Figure 4.6. It can be observed that from September 2017 till March 2018 the simulated $\sigma^0(40)$ is quite able to capture the high frequency variations of the $\sigma^0(40)$ observations. But after March 2018, the simulation is absolutely not able to understand the patterns.

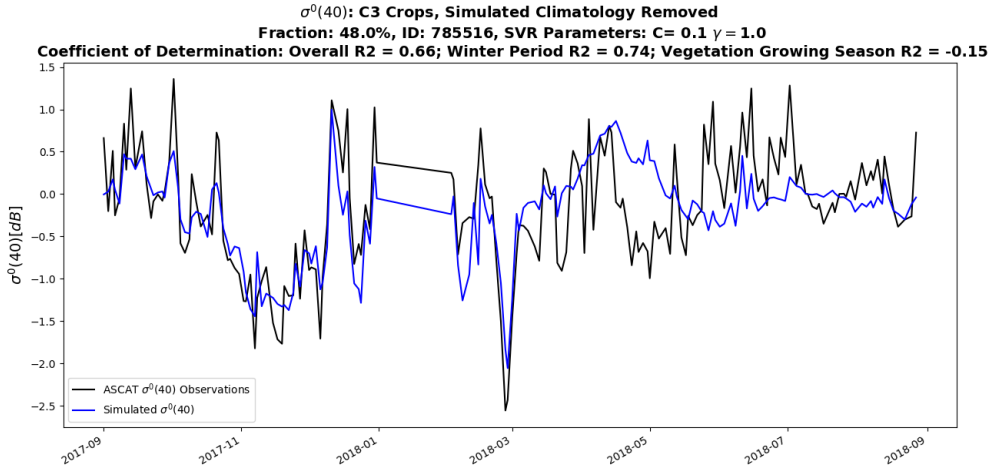



Figure 4.6: Simulation for $\sigma^0(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value for the winter period. Climatology is simulated and removed from the total simulation.

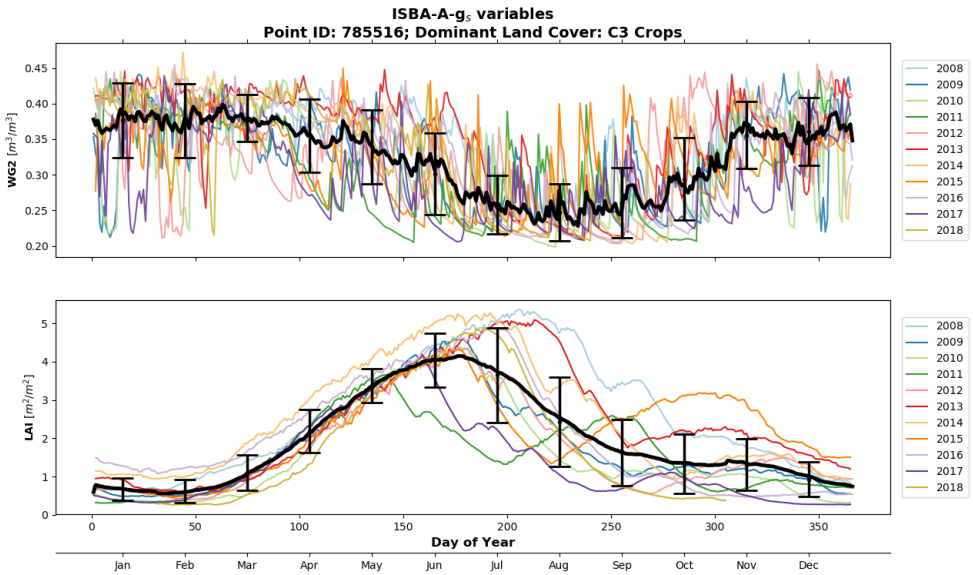



Figure 4.7: Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

An explanation for the low performance during the vegetation growing season and high performance during the winter period can be obtained from looking at the variability of the ISBA-A-$g_s$ variables soil moisture (WG2) and leaf area index (LAI) for each day of the year. This is given in Figure 4.7. It can be noticed that the variability of LAI during the months from May will September is very high and during this same period the WG2 is at its lowest. The basic assumption of the TUW SMR approach is that $\sigma^0$ is more or less linearly relatable to surface soil moisture (Vreugdenhil et al., 2016). It therefore makes sense that the presence of vegetation is a hindrance for the simulation. Also, there is a high deviation during the summer months and this is a difficulty for the support vector machine, because the model is trained by past data and this data is different from year to year.

**C3 Crops (II)**

The simulation for the second case (Best performance during vegetation growing season) for C3 Crops can be seen in Figure 4.8. It can be seen that the variability is captured relatively well during the whole year, even though the performance during the vegetation growing season is not as high as during the winter period.

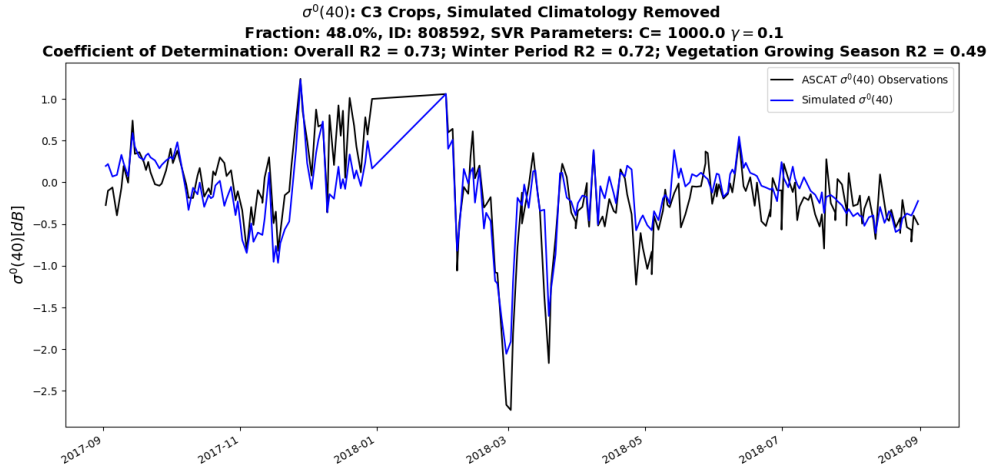



Figure 4.8: Simulation for $\sigma^0(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value for the vegetation growing season. Climatology is simulated and removed from the total simulation.

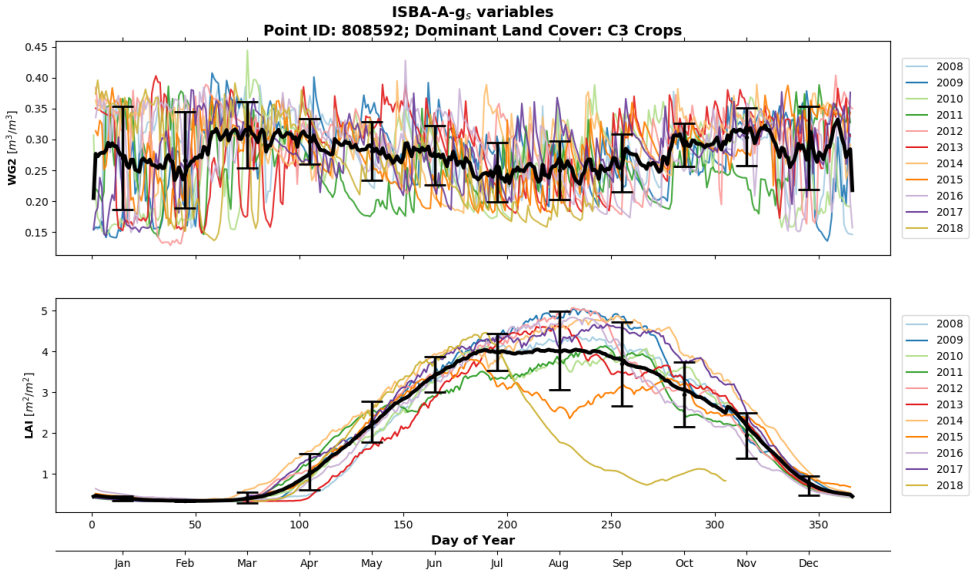



Figure 4.9: Interannual variability of the ISBA-A-g$_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

The reason for the reasonably good performance of this grid point can again be found in the variability of the ISBA-A-$g_s$ variables WG2 and LAI. These are shown in Figure 4.9. The WG2 fluctuates much less than the WG2 in the previous case (See Figure 4.7) and is consistently high compared to the grid point in the previous case. The LAI values are much more consistent for each day of the year during summer months than in the previous case, where there more irregularities can be observed. The SVM is able to simulate $\sigma^0(40)$ well if the amount of irregularities are minimal.

**Temperate Grassland (I)**

For case (I) with Temperate Grassland as dominant land cover type, the simulation with the best performance during winter months is shown in Figure 4.10. The performance during the winter period is relatively high and during the vegetation growing season it is lower but arguably not as bad as the case considered with C3 crops (See simulation in Figure 4.6).

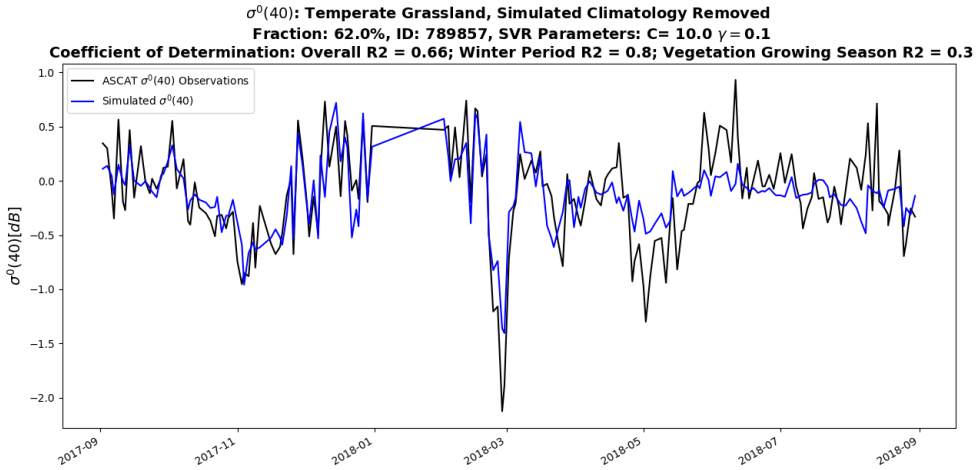



Figure 4.10: Simulation for $\sigma^0(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months. Climatology is simulated and removed from the total simulation.

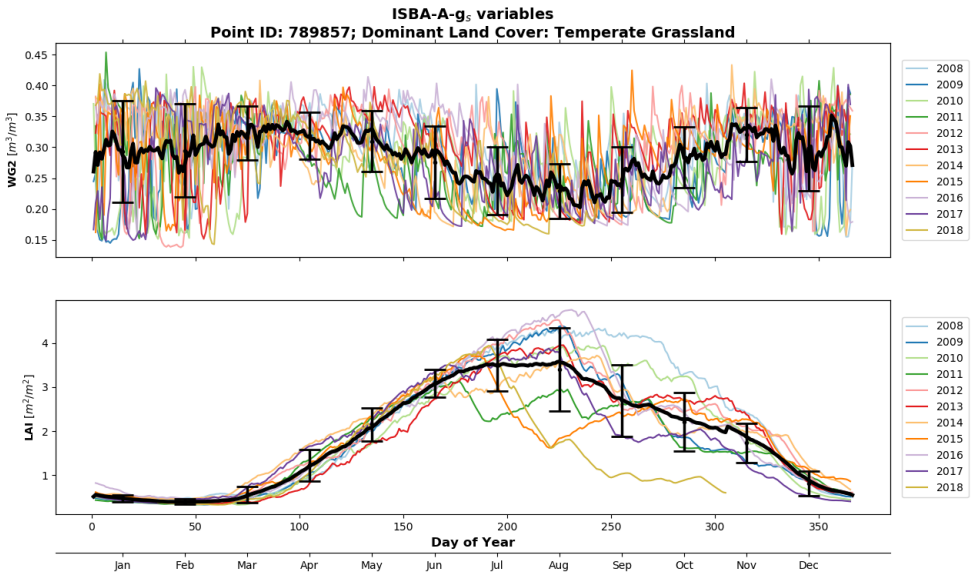



Figure 4.11: Interannual variability of the ISBA-A-$g_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

Figure 4.11 shows the interannual variations of the ISBA-A-$g_s$ and gives more insight into the possible reason for the poor performance during the vegetation growing season and good performance during the winter months. It can be said that the variability of the LAI and the low soil moisture values result in poorer performance during growing season than during winter months.

**Temperate Grassland (II)**

Figure 4.12 shows the simulation for the grid point with the highest performance during the vegetation growing season. Again, the performance during this period does not come near the performance during the winter months. The reason for this can be found in Figure 4.13 where the interannual variations for the ISBA-A-$g_s$ land surface variables soil moisture and LAI are shown. Again, soil moisture values decrease during summer months and LAI values are highly variable for each year, which consequently gives lower performances during growing season.

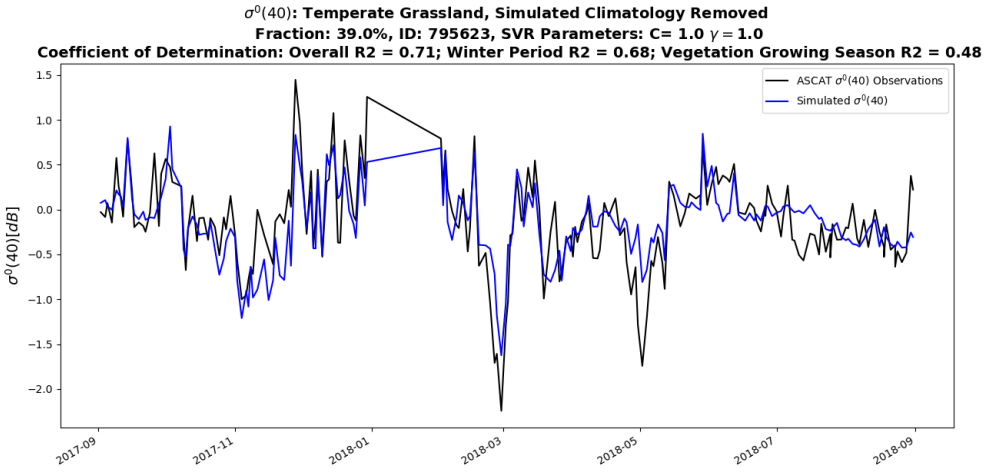



Figure 4.12: Simulation for $\sigma^0(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months. Climatology is simulated and removed from the total simulation.

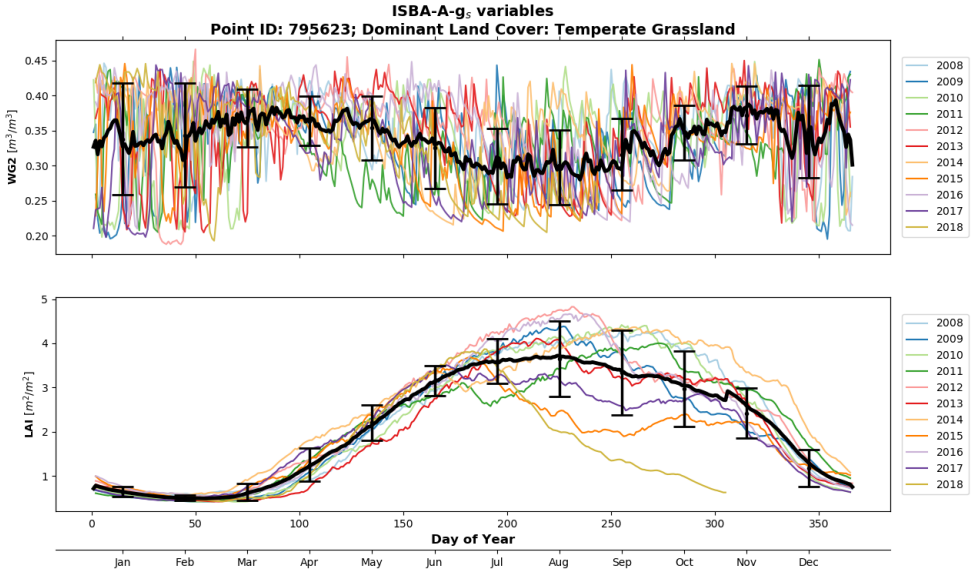



Figure 4.13: Interannual variability of the ISBA-A-g$_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

### 4.1.4. Fractional Land Cover Good and Bad Performances

It has been shown in the previous subsection that performances for backscatter simulations generally depend on the time period over which the R2 is calculated. The theoretical reasoning behind the backscatter simulations being higher during winter months than during the growing season is that the presence of vegetation still forms a hindrance for forward modelling of backscatter and soil moisture values are higher.

The purpose of this subsection is to look at the statistics of land cover fractions of grid points of which the $\sigma^0(40)$ simulation are successful and unsuccessful. In this study, a simulation of $\sigma^0(40)$ is considered successful (or good) if the coefficient of determination (R2) during winter months is larger than 0.6 (60%). The simulation is considered unsuccessful (or bad) if the R2 is smaller than 0.3 (30%). The reason for doing this is to see whether presence of certain land cover types significantly impact the performance. This should then be taken into account if future simulations are done. The statistical overview is presented in tables. Each table gives the amount of grid points that are considered either good or bad and the mean and standard deviations of the land cover fractions within these grid points. In total there are 5 tables, one for each dominant land cover type (C3 Crops, Temperate Grassland, Temperate Broadleaf Deciduous, Bare Land and Boreal Grassland).

The first table that is shown give the statistics of land cover fractions for grid points with C3 Crops as dominant land cover type and can be seen in Table 4.2. One of the major differences between grid points that are good versus those that are bad according to this table is the mean fraction of C3 crops itself. For grid points with a successful simulation, the fractional coverage of C3 Crops is almost 42% while for grid points with unsuccessful simulations, the fractional coverage is almost 52%. This difference of around 10% suggests that the presence of C3 Crops diminishes the performance of the simulation. This might be because the used land surface model is not fully optimized with land management practices.

Another difference that can be seen is the presence of Temperate Grassland which on average is almost 4% higher for grid points with a good simulation than for grid points with a bad simulation. Also, Temperate Needleleaf Evergreen has 2 % more presence on average in grid points with a good simulation. However, the deviation is high relative to the mean value, so the certainty for the statement that presence of Temperate Needleleaf Evergreen enhances performance is not high. The same can be said for the presence of Bare Land, Bare Rock and Shrubs.

| Fractional Cover Statistics $\sigma^0(40)$ | **Good: R2 (Winter Months) >0.6 \| Bad: R2 (Winter Months) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: C3 Crops** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.091 | 0.039 | 96 | 0.066 | 0.015 | 98 |
| Bare Rock | 0.032 | 0.026 | 96 | 0.024 | 0.012 | 98 |
| Permanent Snow | 0.000 | 0.000 | 96 | 0.000 | 0.000 | 98 |
| Temperate Broadleaf Deciduous | 0.141 | 0.080 | 96 | 0.147 | 0.070 | 98 |
| C3 Crops | 0.416 | 0.082 | 96 | 0.518 | 0.089 | 98 |
| C4 Crops | 0.031 | 0.018 | 96 | 0.034 | 0.009 | 98 |
| Summer C3 Crops | 0.009 | 0.013 | 96 | 0.001 | 0.005 | 98 |
| Temperate Grassland | 0.202 | 0.057 | 96 | 0.164 | 0.079 | 98 |
| Temperate Needleleaf Evergreen | 0.048 | 0.031 | 96 | 0.028 | 0.021 | 98 |
| Boreal Grassland | 0.000 | 0.000 | 96 | 0.000 | 0.000 | 98 |
| Boreal Broadleaf Deciduous | 0.000 | 0.000 | 96 | 0.000 | 0.000 | 98 |
| Shrubs | 0.022 | 0.025 | 96 | 0.015 | 0.022 | 98 |

Table 4.2: Fractional Cover Statistics for $\sigma^0(40)$ simulations with C3 Crops as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

For grid points with Temperate Grassland as the dominant land cover type, Table 4.3 gives the statistical overview of land cover fractions for successful and unsuccessful grid points. This time, the fraction of C3 Crops is higher on average for the grid points with good simulations than for those with bad simulations. Also, the deviation is higher so according to this table, it cannot be confessed that C3 Crops necessarily diminishes

performance. A big difference in land cover fraction is the fraction of Temperate Broadleaf Deciduous in the 'good' and 'bad' grid points. Grid points with successful simulations on average seem to have around 10% more presence of Temperate Broadleaf Deciduous and Temperate Needleleaf Evergreen than grid points with unsuccessful simulations. Note that the deviation is very high relative to the mean and hence there is less certainty for a conclusion. The rest of the land cover types do not seem do differ very significantly either due to the low fractional presence, or the high deviation in possible outcomes.

| Fractional Cover Statistics $\sigma^0(40)$ | **Good: R2 (Winter Months) >0.6 \| Bad: R2 (Winter Months) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Temperate Grassland** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.051 | 0.014 | 149 | 0.070 | 0.023 | 145 |
| Bare Rock | 0.016 | 0.008 | 149 | 0.027 | 0.019 | 145 |
| Permanent Snow | 0.000 | 0.000 | 149 | 0.000 | 0.000 | 145 |
| Temperate Broadleaf Deciduous | 0.207 | 0.072 | 149 | 0.098 | 0.105 | 145 |
| C3 Crops | 0.135 | 0.074 | 149 | 0.198 | 0.115 | 145 |
| C4 Crops | 0.021 | 0.015 | 149 | 0.032 | 0.017 | 145 |
| Summer C3 Crops | 0.000 | 0.002 | 149 | 0.001 | 0.003 | 145 |
| Temperate Grassland | 0.438 | 0.123 | 149 | 0.407 | 0.074 | 145 |
| Temperate Needleleaf Evergreen | 0.112 | 0.075 | 149 | 0.076 | 0.070 | 145 |
| Boreal Grassland | 0.000 | 0.000 | 149 | 0.006 | 0.018 | 145 |
| Boreal Broadleaf Deciduous | 0.000 | 0.000 | 149 | 0.002 | 0.007 | 145 |
| Shrubs | 0.007 | 0.010 | 149 | 0.032 | 0.031 | 145 |

Table 4.3: Fractional Cover Statistics for $\sigma^0(40)$ simulations with Temperate Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

Table 4.4 shows the land cover statistics for grid points with Temperate Broadleaf Deciduous as dominant land cover type. The amount of grid points labeled as 'good' and 'bad' performing grid points is much less then the previously considered cases. This gives lower uncertainty in the conclusions that can be drawn from the data. The only significant difference that is observed from this table, is the higher presence of C3 Crops in grid points with successful simulations. However, the high standard deviation adds up to the uncertainty for a statement on the presence of this land cover type.

| Fractional Cover Statistics $\sigma^0(40)$ | **Good: R2 (Winter Months) >0.6 \| Bad: R2 (Winter Months) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Temperate Broadleaf Deciduous** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.048 | 0.014 | 9 | 0.047 | 0.012 | 14 |
| Bare Rock | 0.020 | 0.011 | 9 | 0.016 | 0.008 | 14 |
| Permanent Snow | 0.000 | 0.000 | 9 | 0.000 | 0.000 | 14 |
| Temperate Broadleaf Deciduous | 0.337 | 0.025 | 9 | 0.334 | 0.025 | 14 |
| C3 Crops | 0.132 | 0.094 | 9 | 0.064 | 0.025 | 14 |
| C4 Crops | 0.023 | 0.011 | 9 | 0.017 | 0.008 | 14 |
| Summer C3 Crops | 0.001 | 0.001 | 9 | 0.000 | 0.001 | 14 |
| Temperate Grassland | 0.280 | 0.039 | 9 | 0.298 | 0.018 | 14 |
| Temperate Needleleaf Evergreen | 0.130 | 0.038 | 9 | 0.171 | 0.032 | 14 |
| Boreal Grassland | 0.000 | 0.000 | 9 | 0.000 | 0.000 | 14 |
| Boreal Broadleaf Deciduous | 0.000 | 0.000 | 9 | 0.000 | 0.000 | 14 |
| Shrubs | 0.000 | 0.001 | 9 | 0.005 | 0.011 | 14 |

Table 4.4: Fractional Cover Statistics for $\sigma^0(40)$ simulations with Temperate Broadleaf Deciduous as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

For grid points with Bare Land as dominant land cover type, the land cover fractional statistics are displayed in Table 4.5. No grid points were found that have a successful simulation according to the criteria that the R2 must be larger than 0.6 during winter months. This itself suggests that presence Bare Land is undesirable for backscatter simulations. Also the high presence of Bare Rock and Boreal Grassland and presence of Permanent Snow indicate that grid points with Bare Land as dominant land cover type should be avoided for simulations.

| Fractional Cover Statistics $\sigma^0(40)$ | **Good: R2 (Winter Months) >0.6 \| Bad: R2 (Winter Months) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Bare Land** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | - | - | 0 | 0.381 | 0.062 | 6 |
| Bare Rock | - | - | 0 | 0.121 | 0.053 | 6 |
| Permanent Snow | - | - | 0 | 0.004 | 0.005 | 6 |
| Temperate Broadleaf Deciduous | - | - | 0 | 0.035 | 0.040 | 6 |
| C3 Crops | - | - | 0 | 0.018 | 0.036 | 6 |
| C4 Crops | - | - | 0 | 0.006 | 0.013 | 6 |
| Summer C3 Crops | - | - | 0 | 0.005 | 0.011 | 6 |
| Temperate Grassland | - | - | 0 | 0.045 | 0.039 | 6 |
| Temperate Needleleaf Evergreen | - | - | 0 | 0.010 | 0.025 | 6 |
| Boreal Grassland | - | - | 0 | 0.196 | 0.112 | 6 |
| Boreal Broadleaf Deciduous | - | - | 0 | 0.019 | 0.014 | 6 |
| Shrubs | - | - | 0 | 0.048 | 0.075 | 6 |

Table 4.5: Fractional Cover Statistics for $\sigma^0(40)$ simulations with Bare Land as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

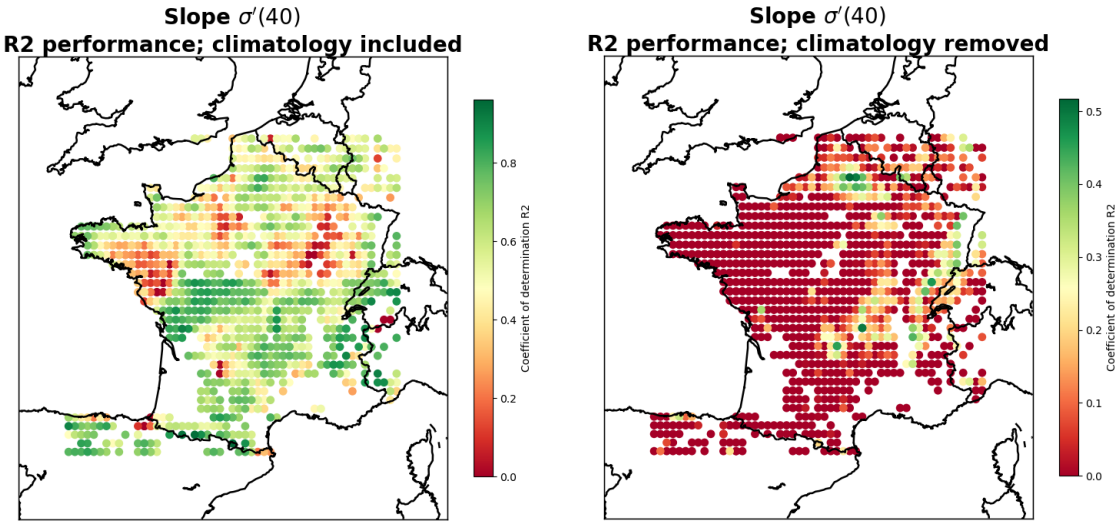
To summarize, this part of the Chapter 'Results' investigated the performances of ASCAT backscatter $\sigma^0(40)$ for grid points with a dominant land cover type. First, an overall performance was shown on a map and it could be seen that even though the climatology is simulated and removed from the total simulation, the performances were in general still reasonably well. After that, it was shown that performances differ depending on the time scale over which it is calculated. For $\sigma^0(40)$ simulations, vegetation growth forms a hindrance for linking soil moisture to $\sigma^0(40)$ observations through the support vector machines. In cases where LAI showed relatively stable results during different days of the year and the soil moisture is not too low during summer months, the $\sigma^0(40)$ performances were shown to be better as well during growing season.

## 4.2. Slope Simulations

In this part the chapter 'Results', the performances of the slope $\sigma'(40)$ simulations are evaluated in a similar manner to the first part, in which backscatter $\sigma^0(40)$ simulations were evaluated. The overall performances of the $\sigma'(40)$ simulations are displayed on a map first. Again, only grid points are considered that have a dominant land cover type (land cover fraction ≥ 30%). After that, the performances during the vegetation growing season and during the winter period are considered.

### 4.2.1. Overall Performance

The overall performances of the slope $\sigma'(40)$ simulations expressed as the coefficient of determination (R2) can be seen in Figure 4.14. On the left hand side of the figure (a) the performances are shown where simulated climatology is not removed from the total simulations. The map on the right hand side of the figure (b) shows the performance where simulated climatology is removed from the total simulation. It can be immediately concluded that the support vector machines are not able to simulate high frequency signal. What it is able to simulate are seasonal variations. In Chapter 2 was mentioned that variations in slope $\sigma'(40)$ depend on vegetation dynamics. The mean and standard deviation of the $\sigma'(40)$ shown in Figure 2.5 in Chapter 2 showed that variation is relatively high in agriculturally active areas and moderate to in areas with dominant presence of Temperate Grassland or Temperate Broadleaf Deciduous.




(a) Climatology included

(b) Climatology removed

Figure 4.14: Map of slope $\sigma'(40)$ performances. The left hand side of the figure (a) shows the R2 performance with the climatology included. The right hand side (b) shows the performance of the regression

A more complete perspective on the performance of the $\sigma'(40)$ for the main land cover types is given in Figure 4.15. On the top left the performance of the C3 Crops land cover type is shown against the fractional coverage. It can be seen that there is a wide variety of performance outcomes for fraction of land cover up to 55%. Fractional coverage of C3 Crops larger than 55% seems to enhance the overall performance. For the performances with Temperate Grassland as dominant land cover type, which can be seen on the top right of the figure, a similar pattern as the performance of C3 Crops can be observed. There is a wide spread of possible performance values for cases where the fractional coverage is less than approximately 50%. For a fractional coverage higher than 50%, a certain minimum can be ensured based on this scatter plot. Also, there seems to develop an upper boundary as well which narrows the range of possible R2 values for high fractions of Temperate Grassland. However, this cannot be ensures as the amount of grid points decreases for increasing fractional cover. For the Temperate Broadleaf Deciduous and Bar Land land cover type, the range of possible performance outcomes is large and there are insufficient amount of grid points to make a statement on the performance outcomes if the fraction of land cover increases.

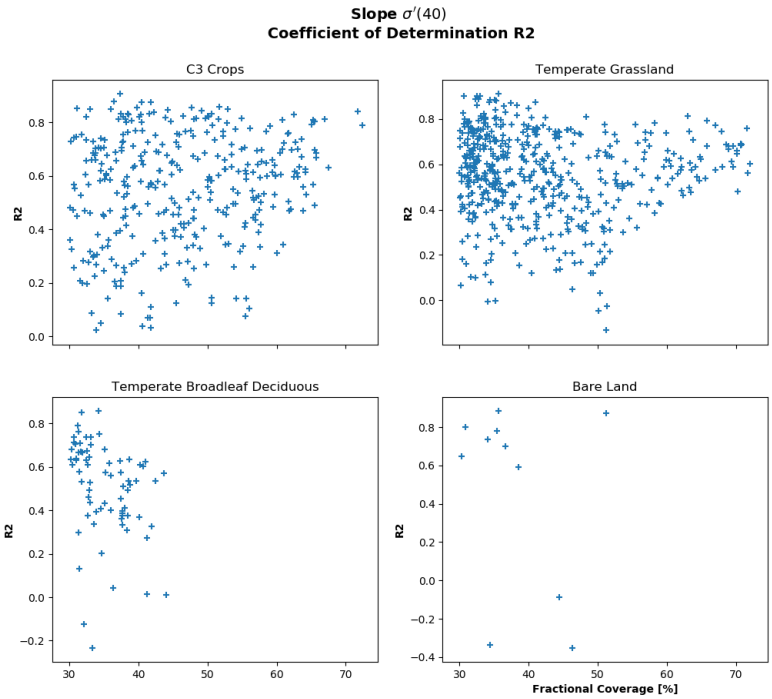


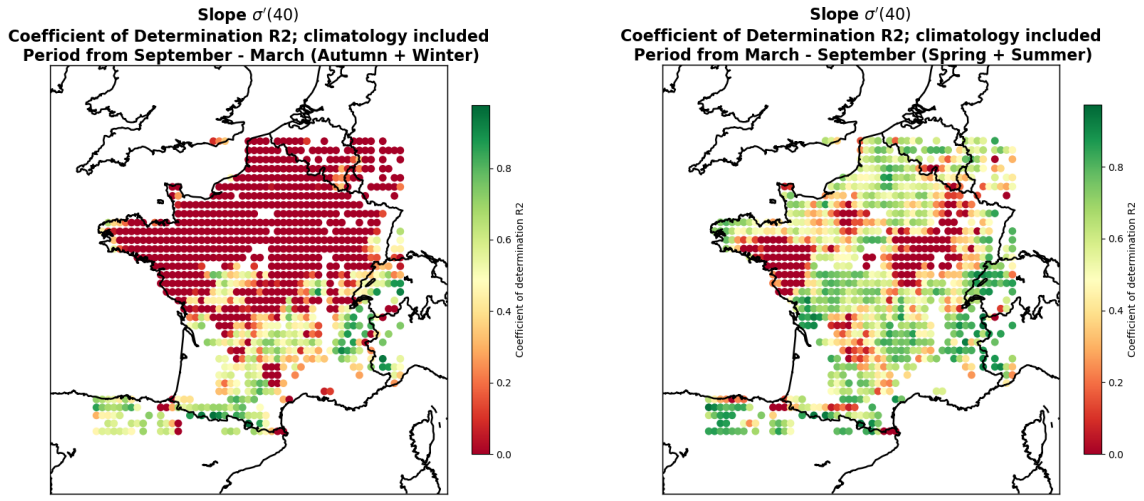

Figure 4.15: Scatterplots for the four main dominant land cover types. The performances are shown against the fractional coverages.

The performances are calculated over a time period of one full year. The performance within a shorter time span might or might not give a different result, as was shown in the previous section where backscatter simulations were evaluated. According to the hypothesis that variations of $\sigma'(40)$ is primarily driven by vegetation dynamics, it is expected that simulations during winter months will be worse compared to summer months. The results are shown and discussed in the next subsection.

### 4.2.2. Performance During Vegetation Growing Season

The maps in Figure 4.16 (a) and (b) display the performances of the slope $\sigma'(40)$ simulations. The left hand side of the figure (a), represent R2 values calculated for the winter months. R2 values seem to be very low compared to the R2 values calculated over a whole year. It can be seen at the locations near the Alps, the grid points have very high R2 performances (R2 $\geq$ 0.6). For these points it can be seen in Figures 2.5 (a) and (b) in Chapter 2 that the mean values of the slope is relatively very low and standard deviations are moderate. The main land cover types of these grid points are Bare Land, Temperate Grassland and Boreal Grassland, though the percentages of land cover fraction are very low as can be seen in Figures 3.2 (a) and (b). Thus, other land cover types within these grid points can determine the amount of predictability in the slope.

The right hand side (b) of Figure 4.16 shows the performances during the vegetation growing season. Compared to the performances on the left hand side (b) during winter months, simulations are much better. However, a number of grid points around the (imaginary) horizontal middle line through France still give a poor performance, while these grid points do not seem to perform as worse as is displayed in this map if it compared to the overall performance shown in Figure 4.14 (a). A possible reason for this might be the low value for the mean slope $\sigma'(40)$ and low standard deviation for the slope (See Figure 2.5) in combination with land cover type and the different behaviour in different time scales. A better understanding could be obtained if the performances of shorter time scales are studied. However, this is not within the scope of this study, as the performances are only evaluated during the vegetation growing season and winter months, each having a time period of 6 months.




(a) (b)

Figure 4.16: On the left hand side of the figure (a) the coefficient of determination R2 for $\sigma'(40)$ is displayed during the period from September to March which is during the Fall and Winter season. On the right hand side (b) the R2 is displayed from the period from March till September during the vegetation growing season.

A statistical overview of R2 performances can be seen in Table 4.6. In this table, the mean and standard deviation of performances during the the three time frames are considered for each dominant land cover class.

| | $\mu$ R2 | $\sigma$ R2 | $\mu$ R2 SS | $\sigma$ R2 SS | $\mu$ R2 AW | $\sigma$ R2 AW | GP |
|---|---|---|---|---|---|---|---|
| C3 Crops | 0.55 | 0.21 | 0.48 | 0.34 | -3.64 | 5.79 | 351 |
| Bare Land | 0.48 | 0.48 | 0.60 | 0.42 | 0.08 | 0.88 | 11 |
| Temperate Grassland | 0.55 | 0.19 | 0.36 | 0.56 | -1.13 | 2.51 | 522 |
| Temperate Broadleaf Deciduous | 0.50 | 0.22 | 0.40 | 0.42 | 0.05 | 0.89 | 73 |
| Boreal Grassland | 0.74 | 0.21 | 0.80 | 0.19 | 0.59 | 0.33 | 15 |

Table 4.6: $\mu$ is the mathematical mean. $\sigma$ is the mathematical standard deviation. R2 is the coefficient of determination. SS represents Spring + Summer and AW represents Autumn + Winter.

From the statistical overview given in Table 4.6 the following points can be noticed:

- Grid points with C3 Crops as dominant land cover type on average perform best for a time frame of a whole year. The performances during vegetation growing season are slightly worse than the overall performances, but not bad. During winter months, the performance of the simulation is extremely worse as the average R2 value is far below zero.
- Grid points with Bare Land as dominant land cover type have the highest standard deviation in the overall performance. Performance is better during winter than during summer, which suggests that the model does not work for the proper reason as variability in vegetation is more driven by the variability of the leaf area index, which very likely is not the case during this period of the year.
- Grid points with Temperate Grassland as dominant land cover type show an average overall performance that is higher than the performances for vegetation growing season and winter months. During winter months, the performance on average is very bad as the R2 is below 0 and the deviation is very high. This confirms the theory that simulations are more successful during vegetation growing season than during winter months.
- Grid points with Temperate Broadleaf Deciduous as dominant land cover type also on average have a high overall performance and low standard deviation compared to the the mean and standard deviation during the vegetation growing season and winter months. This again confirms the hypothesis that during winter months the slope is very unlikely to give a good performance.

- Grid points with Boreal Grassland as dominant land cover type show an mean overall R2 value which is highest of all dominant land cover types. As mentioned earlier in this section, the grid points are primarily located in the Alps region and it is therefore very unlikely that the results are obtained for the right reason.

### 4.2.3. Performances and Land Surface Variables

In this part, the simulation is shown of individual grid points to see whether the performances makes sense by looking at the variability of the land surface variables. The variations of the slope $\sigma'(40)$ are driven by dynamics of vegetation. During winter months, soil moisture (WG2) values are generally high and leaf area index (LAI) values low. During this period, it is expected that the variations of the $\sigma'(40)$ are harder to simulate due to the low LAI values. Consequently, it is expected that $\sigma'(40)$ simulations are more successful during the growing season. The following three cases are considered for grid points with dominant land cover types C3 Crops and Temperate Grassland:

(I) Grid point with best performance during winter months

(II) Grid point with best performance during vegetation growing season

(III) Grid point with good performance during winter months but bad performance during growing season.

#### C3 Crops (I)

The first considered case for $\sigma'(40)$ simulations is one with C3 Crops as dominant land cover type and a high performance during the winter months. The simulation can be seen in Figure 4.17. The black line in the figure represents the $\sigma'(40)$ observation and the blue line represents the simulation by the support vector machine. Even though the calculated coefficient of determination during the three different time spans are high, it can be seen that the simulation contain noticeable noise.

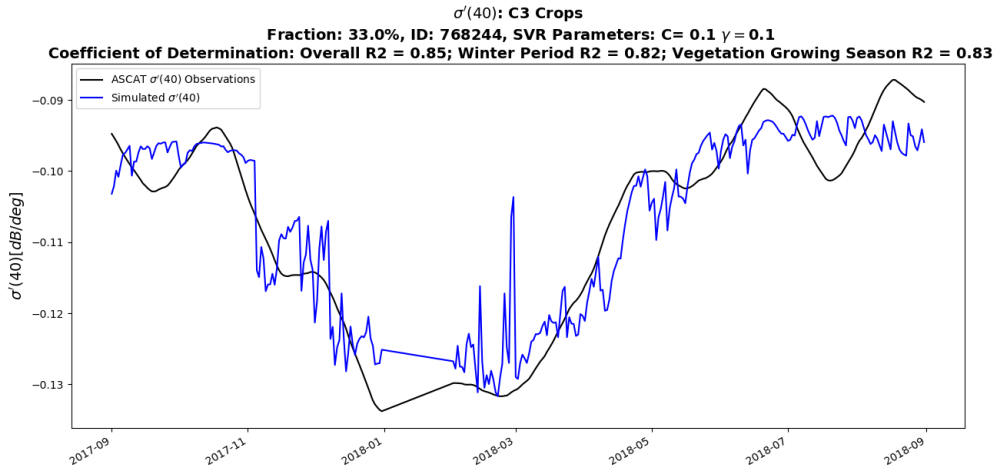



Figure 4.17: Simulation for $\sigma'(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value during the winter months.

Looking at the variability of the corresponding land surface variables, which can be seen in Figure 4.18, it can directly be seen that during July, August and September the deviation of LAI values is very high. With supervised machine learning in general, models are trained with the purpose to understand patterns from the past so that if an event occurs in the future, the model has some idea what the outcome should be. It therefore becomes difficult to train a model with data that has a high variability and this might be an important reason for the amount of noise in the simulation.

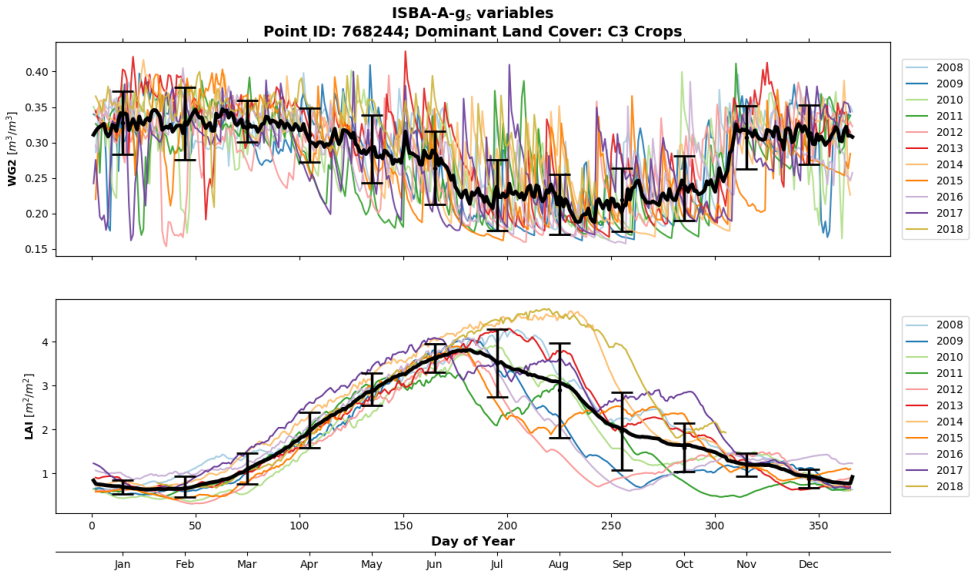



Figure 4.18: Interannual variability of the ISBA-A-g$_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

### C3 Crops (II)

The second case that is considered for grid points with C3 Crops as dominant land cover type is about the situation of the grid points with the best performance during the vegetation growing season. The simulation of this grid point is shown in Figure 4.19. The performance during the growing season is 93%, which can be considered as a very successful simulation period. During the winter, this performance is significantly worse with R2 = 63%. Looking at the winter months in the figure, it can be noticed that the large movements of the $\sigma'(40)$ are captured well, but the simulation looks very noisy compared to the simulation starting from March 2018.

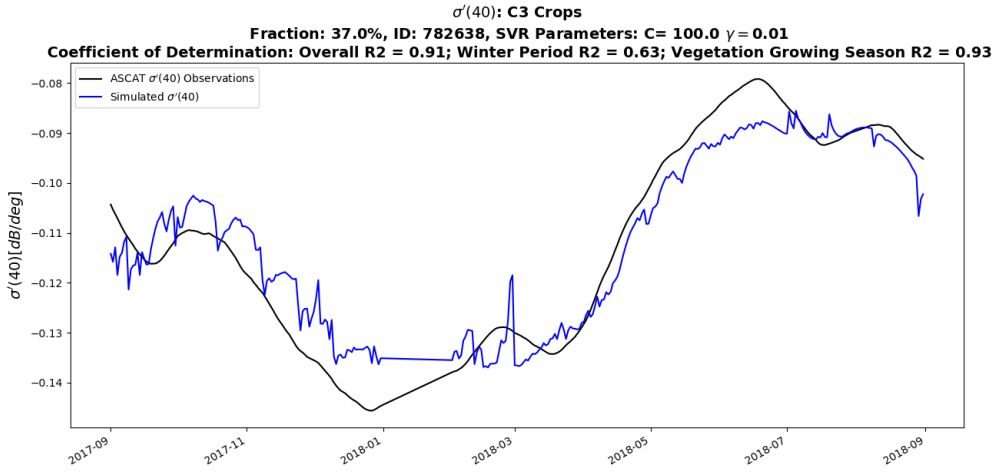



Figure 4.19: Simulation for $\sigma'(40)$ with C3 Crops as dominant land cover type. This grid point has the highest R2 value during the vegetation growing season.

The variability of the land surface variables by which it is trained and simulated can be seen in Figure 4.20. Again, during periods in which LAI values are high and soil moisture values low, the $\sigma'(40)$ is simulated relatively well. The low values of soil moisture during the growing season might contribute to the decrease of noise in the simulation, even though the variability of the LAI is high.

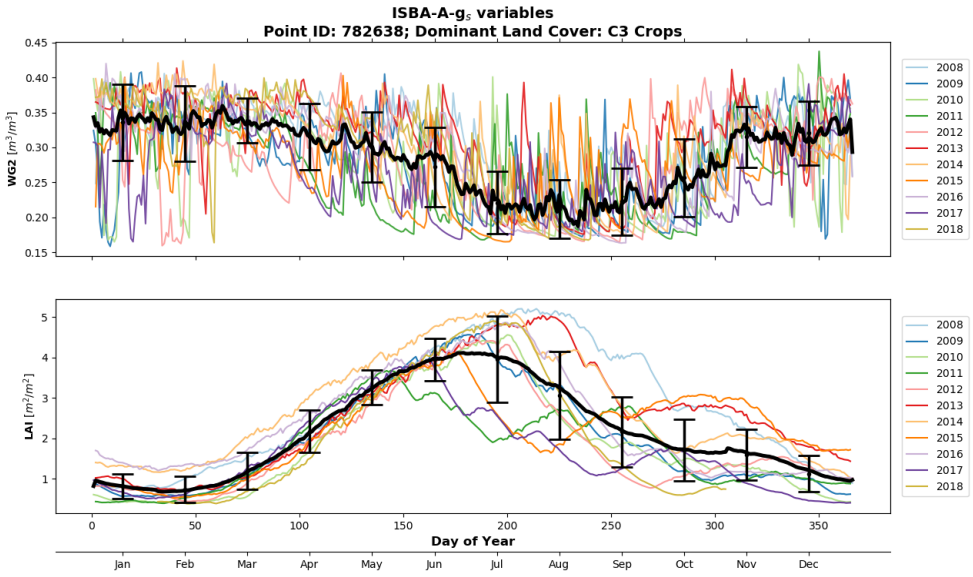



Figure 4.20: Interannual variability of the ISBA-A-g$_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

### Temperate Grassland (III)

For the third case a grid point is considered that has Temperate Grassland as dominant land cover type. The simulation is shown in Figure 4.21. It should be noted that the range of $\sigma'(40)$ values ranges from approximately -0.125 to -0.09 dB/deg while the range for the previously mentioned case in Figure 4.19 had a range of slope values from approximately -0.145 to -0.08 dB/deg. The point of mentioning this is that slope values seem to be smaller for this grid point. This suggests that there might be land covers present within this grid point that do not contain any vegetation. Therefore, for this particular grid point the land cover fractions are displayed in Table 4.7. It can be seen that Bare Land is present for about 7% and even Bare Rock for almost 2%. Also, there is presence of Temperate Needleleaf Evergreen for about 9%. For this land cover type, variations of vegetation dynamics are usually low.

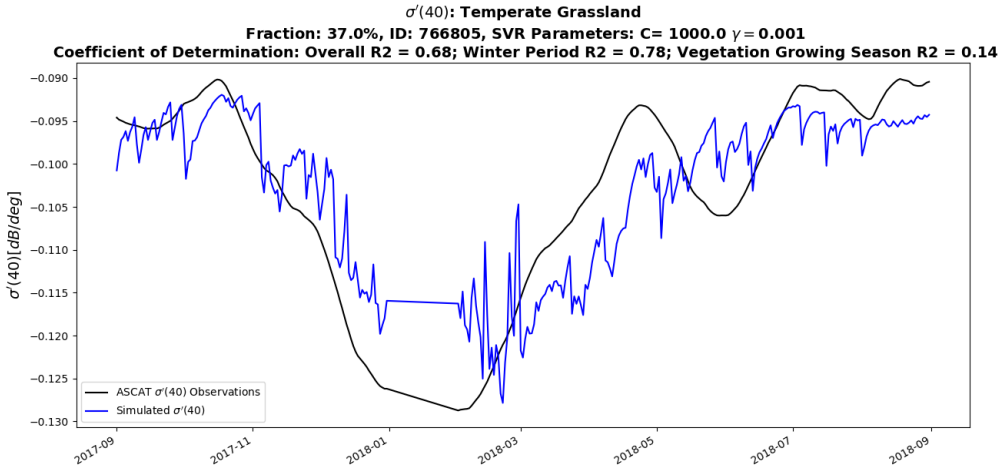



Figure 4.21: Simulation for $\sigma'(40)$ with Temperate Grassland as dominant land cover type. This grid point has the highest R2 value during the winter months.

| | Fraction [x 100 %] |
|---|---|
| Bare Land | 0.071 |
| Bare Rock | 0.019 |
| Permanent Snow | 0.000 |
| Temperate Broadleaf Deciduous | 0.237 |
| C3 Crops | 0.181 |
| C4 Crops | 0.017 |
| Summer C3 Crops | 0.004 |
| Temperate Grassland | 0.371 |
| Temperate Needleleaf Evergreen | 0.090 |
| Boreal Grassland | 0.000 |
| Boreal Broadleaf Deciduous | 0.000 |
| Shrubs | 0.008 |

Table 4.7: Fractional land cover for grid point 766805 with dominant land cover type Temperate Grassland.

The variability of the land surface variables is shown in Figure 4.22. The noise in the simulation seems again to be a consequence of the high frequency movements in both land surface variables.

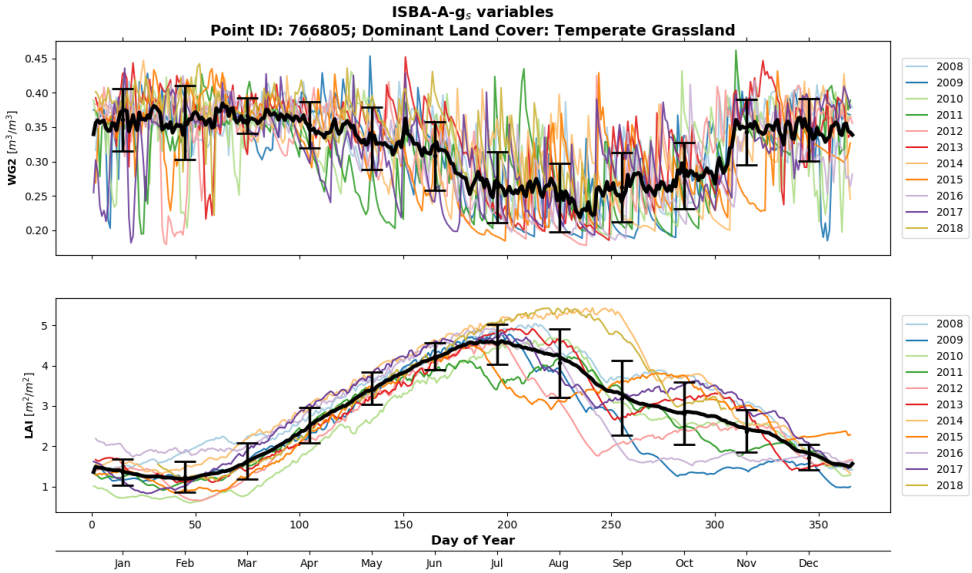



Figure 4.22: Interannual variability of the ISBA-A-g$_s$ variables from 2008 till 2019 plotted against the day of year. The black line represents the mean value for each day of the year. Also, the standard deviation for each month is shown.

In the next subsection, the statistics of fractional covers of grid points with a particular dominant land cover type is looked at. The purpose of this is to spot differences in land cover types for grid points with successful simulations are unsuccessful simulations.

### 4.2.4. Fractional Land Cover Good and Bad Performances

The purpose of this subsection is to see whether the presence of certain land cover types impact the performances for the slope $\sigma'(40)$ simulations. The dominant land cover types that are considered are C3 Crops, Temperate Grassland, Temperate Broadleaf Deciduous, Bare Land and Boreal Grassland. The criteria for a grid point to be called successful (or good) is that the coefficient of determination (R2) must be larger than 60% during the growing season. For a simulation to be considered unsuccessful the performance during the growing season must be below 30%.

For successful and unsuccessful grid points with dominant land cover type C3 Crops, the land cover fractional statistics can be seen in Table 4.8. It can be seen that successful grid points on average contain more Bare Land and Bare Rock. This does not necessarily mean that presence of these land cover types enhances or diminishes the performance, because the deviation values next to the mean values are in the same order of magnitude. Hence, it is very uncertain to conclude this. A major difference between good performing grid points and bad performing grid points is that on average Temperate Broadleaf Deciduous is more than 10% present in bad performing grid points than in good performing grid points. The deviations are low relative to the mean values so a suggestion can be made that presence of Temperate Broadleaf Deciduous seems to diminish performance. Presence of C3 Crops seems to be good for grid points with high performance. Though it should be noted that due to the high deviation of C3 Crops for good performing grid points, it cannot really be a suggestion yet. The rest of the land cover types are either to small in presence or the deviation is too high relative to the mean fraction to make useful suggestions.

| Fractional Cover Statistics $\sigma'(40)$ | **Good: R2 (Growing Season) >0.6 \| Bad: R2 (Growing Season) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: C3 Crops** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.094 | 0.040 | 158 | 0.056 | 0.013 | 81 |
| Bare Rock | 0.032 | 0.026 | 158 | 0.024 | 0.017 | 81 |
| Permanent Snow | 0.000 | 0.000 | 158 | 0.000 | 0.000 | 81 |
| Temperate Broadleaf Deciduous | 0.119 | 0.061 | 158 | 0.221 | 0.066 | 81 |
| C3 Crops | 0.486 | 0.110 | 158 | 0.422 | 0.078 | 81 |
| C4 Crops | 0.035 | 0.016 | 158 | 0.024 | 0.009 | 81 |
| Summer C3 Crops | 0.010 | 0.015 | 158 | 0.001 | 0.002 | 81 |
| Temperate Grassland | 0.166 | 0.066 | 158 | 0.188 | 0.067 | 81 |
| Temperate Needleleaf Evergreen | 0.031 | 0.024 | 158 | 0.054 | 0.026 | 81 |
| Boreal Grassland | 0.000 | 0.000 | 158 | 0.000 | 0.000 | 81 |
| Boreal Broadleaf Deciduous | 0.000 | 0.000 | 158 | 0.000 | 0.000 | 81 |
| Shrubs | 0.021 | 0.025 | 158 | 0.007 | 0.018 | 81 |

Table 4.8: Fractional Cover Statistics for $\sigma'(40)$ simulations with C3 Crops as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

Table 4.9 contains the land cover fractional statistics for grid points with Temperate Grassland as dominant land cover type with successful and unsuccessful simulations. In the previous case, the dominant land cover type was C3 Crops and it was said that it might be that presence of Bare Land and Bare Rock enhance performances of slope simulations. This table shows that this does not necessarily have to be the case, as the bad performing grid points have more presence of these two land cover types than the good performing grid points. Also, in the previously considered case with C3 Crops as dominant land cover type, it was mentioned that presence of C3 Crops might be good for the simulation. However, this table again says the opposite. What both tables have in common though is the higher presence of Temperate Broadleaf Deciduous in the good performing grid points.

| Fractional Cover Statistics $\sigma'(40)$ | **Good: R2 (Growing Season) >0.6 \| Bad: R2 (Growing Season) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Temperate Grassland** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.058 | 0.022 | 201 | 0.061 | 0.021 | 152 |
| Bare Rock | 0.022 | 0.015 | 201 | 0.025 | 0.020 | 152 |
| Permanent Snow | 0.000 | 0.000 | 201 | 0.000 | 0.000 | 152 |
| Temperate Broadleaf Deciduous | 0.178 | 0.087 | 201 | 0.143 | 0.119 | 152 |
| C3 Crops | 0.125 | 0.090 | 201 | 0.187 | 0.105 | 152 |
| C4 Crops | 0.024 | 0.018 | 201 | 0.025 | 0.016 | 152 |
| Summer C3 Crops | 0.001 | 0.004 | 201 | 0.000 | 0.001 | 152 |
| Temperate Grassland | 0.398 | 0.103 | 201 | 0.431 | 0.092 | 152 |
| Temperate Needleleaf Evergreen | 0.118 | 0.086 | 201 | 0.083 | 0.069 | 152 |
| Boreal Grassland | 0.006 | 0.018 | 201 | 0.000 | 0.000 | 152 |
| Boreal Broadleaf Deciduous | 0.003 | 0.009 | 201 | 0.000 | 0.000 | 152 |
| Shrubs | 0.014 | 0.023 | 201 | 0.017 | 0.023 | 152 |

Table 4.9: Fractional Cover Statistics for $\sigma'(40)$ simulations with Temperate Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

The fractional statistics for grid points with Temperate Broadleaf Deciduous as dominant land cover type can be seen in Table 4.10. It can be seen that the differences between the land cover fractions for good performing grid points and bad performing grid points do not differ very significantly. The number of grid points is much lower for this dominant land cover type than for the dominant land cover types C3 Crops and Temperate Grassland. Therefore, this is the last considered dominant land cover type. The tables for Bare Land and Boreal Grassland can be seen in the Appendix.

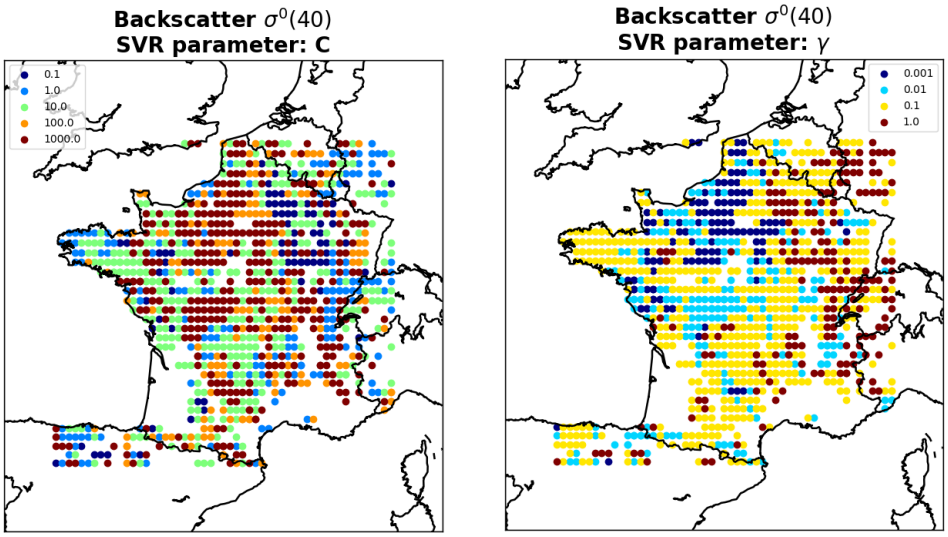
| Fractional Cover Statistics $\sigma'(40)$ | **Good: R2 (Growing Season) >0.6 \| Bad: R2 (Growing Season) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Temperate Broadleaf Deciduous** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.049 | 0.013 | 26 | 0.046 | 0.006 | 19 |
| Bare Rock | 0.022 | 0.013 | 26 | 0.013 | 0.003 | 19 |
| Permanent Snow | 0.000 | 0.000 | 26 | 0.000 | 0.000 | 19 |
| Temperate Broadleaf Deciduous | 0.324 | 0.023 | 26 | 0.368 | 0.037 | 19 |
| C3 Crops | 0.086 | 0.036 | 26 | 0.094 | 0.076 | 19 |
| C4 Crops | 0.025 | 0.016 | 26 | 0.027 | 0.013 | 19 |
| Summer C3 Crops | 0.001 | 0.002 | 26 | 0.000 | 0.000 | 19 |
| Temperate Grassland | 0.295 | 0.019 | 26 | 0.295 | 0.026 | 19 |
| Temperate Needleleaf Evergreen | 0.155 | 0.047 | 26 | 0.141 | 0.028 | 19 |
| Boreal Grassland | 0.003 | 0.013 | 26 | 0.000 | 0.000 | 19 |
| Boreal Broadleaf Deciduous | 0.003 | 0.014 | 26 | 0.000 | 0.000 | 19 |
| Shrubs | 0.002 | 0.006 | 26 | 0.001 | 0.003 | 19 |

Table 4.10: Fractional Cover Statistics for $\sigma'(40)$ simulations with Temperate Broadleaf Deciduous as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

To summarize, this section shows that ASCAT $\sigma'(40)$ simulations are generally far more successful during the vegetation growing season than during the winter months. Even though the scatter plots in Figure 4.15 indicates that high presence of C3 Crops and Temperate Grassland is beneficial for the overall performance, it could not be justified with the tables of land cover fraction statistics. The simulations of $\sigma'(40)$ were shown to contain much noise, which can be explained by the high frequency variations of the land surface variables by which the model is trained.

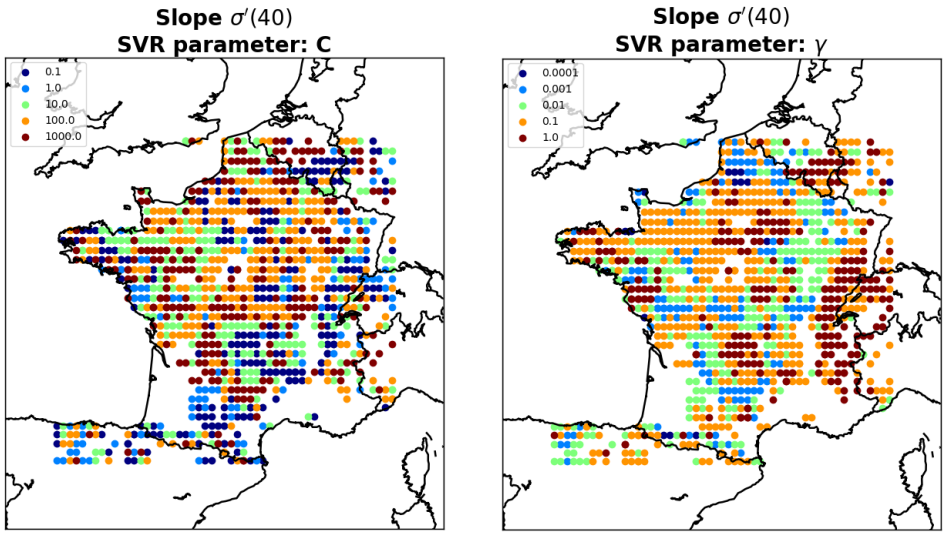
## 4.3. Spatial Consistency SVR Parameters

This section shows that the created Support Vector Machines for each grid point have spatial consistency. This study excludes the link of parameters and parameter combinations to land cover types and performances. However, this is definitely a recommendation for future studies of the application of Support Vector Machines to simulate ASCAT backscatter $\sigma^0(40)$ or slope $\sigma'(40)$. Tables of parameter combinations for dominant land cover types can be seen in the Appendix. The map of grid points with their corresponding SVR parameters for backscatter $\sigma^0(40)$ simulations and slope $\sigma'(40)$ simulations can be seen in Figures 4.23 and 4.24 respectively.




(a) (b)

Figure 4.23: $\sigma^0(40)$ map of grid points of which the optimized SVR model has a certain value of C on the left hand side (a) and $\gamma$ on the right hand side (b)




(a) (b)

Figure 4.24: $\sigma'(40)$ map of grid points of which the optimized SVR model has a certain value of C on the left hand side (a) and $\gamma$ on the right hand side (b)

The map of grid points with their corresponding SVR parameters for $\sigma'(40)$ simulations can be seen in Figure 4.24. With the optimization procedure, the aim was to prevent over- and under fitting. The fact that

spatial consistency can be observed suggests that land cover types and therefore also performances, means that land covers and/or a combination of land cover types can be linked to the hyper parameters of the support vector machines.

Remember from the Methodology chapter that the models were optimized in a way so that it can simulate $\sigma^0(40)$ or $\sigma'(40)$ during each period of a year. The main reason for using a grid search during optimization was to see whether spatial consistency exists. Now that it is the case, the following possibilities can be explored.

- Optimize in the same way, but for each grid point make two models. One for the winter months (autumn + winter season) and one for the summer months. In this way, the support vector machines can be optimized for completely different conditions during the year and will probably give better simulations during each period of the year.
- Once the general pattern between land cover type and hyper parameters is found, a more accurate combination of hyper parameters can be used to optimize. Suppose that for a particular grid point the $C$ parameter equals to 100, but ideally it should be 50. Just because the options were 1, 100 or 1000 the optimization was not allowed to take a value of 50. Thus, this ensures that optimization can be done in a more precise way.
- The obtained hyper parameters can be seen as a first approach in minimizing the error of the objective function (The optimization problem). Improved optimization can be done by using the upper and lower limits of the hyper parameters for randomized or Bayesian optimization.

# 5

# Discussion & Conclusions

This chapter discusses several outcomes of the results and ends with the conclusion in which the research question(s) are answered and recommendations are given.

## 5.1. Discussion

### 5.1.1. Support Vector Machine Optimization Overview

The objective of this work was to find out if support vector machines (SVMs) could serve as a surrogate model to simulate ASCAT parameters $\sigma^0(40)$ and $\sigma'(40)$ with the ISBA-A-$g_s$ land surface variables soil moisture and leaf area index. The advantage of SVMs is its ability to internally do the feature engineering, by making use of the Kernel function. The Gaussian Kernel function has been successfully implemented in studies that simulate brightness temperatures derived from passive microwave observations with land surface variables (See Forman and Reichle (2015)) and is therefore chosen as the kernel function.

Prior to this study, it was not known what domain space is sufficient for the SVM hyper parameters $C$ and $\gamma$ for all grid points in France. Therefore, a set of parameter combinations with values logarithmically spaced are used as an initial approach. This parameter combinations grid that will be used for minimizing the objective function. To ensure that models are created that are able to understand and act upon unseen data, a k-fold cross validation is applied with k=9 for 9 years of training data. In this way, over-fitting and/or under-fitting is taken into account and the bias for choosing a certain set of parameters is reduced.

The performance metric that was used to decide whether the best results are obtained for a parameter set is the coefficient of determination (R2). This metric serves as an indicator for how much of the variability the model is able to understand. The final evaluation of the models is done on one year of unseen data. The R2 is also used as the main performance metric of the final simulation. The main reason for this is to compare the success of simulations of different grid points on a map according to the models ability to capture the variability of the $\sigma^0(40)$ and $\sigma'(40)$, rather than comparing the mean squared errors or root mean squared errors, which are more related to the standard deviation of the $\sigma^0(40)$ and $\sigma'(40)$ if displayed on a map.

For 1324 grid points in and around France, support vector machines have been created. This study only investigates grid points with a dominant land cover fraction larger than 30%. Spatial consistency was found for the hyper parameters and this is useful for further optimization studies.

### 5.1.2. Results Overview

This study gives four perspectives on the obtained $\sigma^0(40)$ and $\sigma'(40)$ simulations:

(I) Maps with performances. The aim of this is to show that the performance of the simulations depend on land cover types. Scatter plots are created for the main dominant land cover types to see if an increase in fractional cover generally enhances or diminishes performance.

(II) Performance is calculated over three spans of time. The period is a full year. The other two periods

are the Autumn + Winter seasons and Spring + Summer seasons. In the Chapter 4 (Results), they are referred to as winter months and (vegetation) growing season respectively.

(III) Individual simulations are displayed together with the ISBA-A-$g_s$ land surface variables soil moisture and leaf area index.

(IV) Grid points are distinguished between performances above 60%, which are considered as good performing grid points, and below 30%, which are considered as bad performing grid points, to look at the land cover fractional statistics of grid points with successful and unsuccessful simulations.

For the simulations of the backscatter and slope, these three cases are separately discussed.

**Backscatter** $\sigma^0(40)$

(I) The backscatter $\sigma^0(40)$ simulations generally gave reasonable results as for many points even the high frequency fluctuations could be understood by the SVR model. This is demonstrated by simulating the climatology with the same model and subtracting this from the total backscatter simulation. The main dominant land cover types are C3 Crops, Temperate Grassland, Temperate Broadleaf Deciduous and Bare Land. According to the scatter plots where overall performance is shown against land cover fraction, it could be seen that in the presence of C3 Crops increases, the performance tends to decrease. For Temperate Grassland, it could be seen that an increase in land cover fraction seems to enhance performance. The amount of grid points for the remaining land cover types is very low and hence the scatter plots do not reveal much about the performance due to an increase or decrease of their presence.

(II) Performances are calculated during the growing season and winter season separately as well. It was expected that backscatter simulation would be more successful during winter months than during the growing season, due to the hindrance that vegetation dynamics can form and the low values of soil moisture during dry months, because backscatter variations are primarily driven by soil moisture. Table 4.1 shows that the performance during winter months is indeed better than during the growing season.

(III) Backscatter simulations of individual grid points show that the high interannual variability of leaf area index values in combination with the low soil moisture values seem to significantly impact the performance during the growing season. This is relevant information as this suggests that models in general should be constructed with more focus on the specific periods during a year.

(IV) The statistics of the fraction of different land cover types for successful and unsuccessful grid points revealed that bad performing grid points seem to have more C3 crops than good performing grid points. However, this is not the case for every situation so this cannot be concluded. For presence of Temperate Needleleaf Evergreen and Temperate Broadleaf Deciduous the presence seems to enhance performance. But again, there are cases where this is not the case and so this can again not be concluded. The grid points is considered successful if the coefficient of determination (R2) during the winter months is larger than 60% and unsuccessful if R2 is smaller than 30%.

**Slope** $\sigma'(40)$

(I) From the maps of the performances of slope $\sigma'(40)$ simulations (See Figure 4.14) was demonstrated that the SVMs are generally able to simulate the seasonal effects very well, but if climatology is removed like was done with the backscatter, the performances are very bad. This is because the slope is very smooth compared to the backscatter. The performance scatter plots for the main land cover types show that increase of the fraction of C3 Crops within a grid point seems to improve performance. The same can be said for an increase of the fractional cover of Temperate Grassland. For Temperate Broadleaf Deciduous and Bare Land, there are not enough data points to support any statement on their presence.

(II) The performances were expected to be good during the growing season, as variations in slope are primarily driven by vegetation dynamics (Vreugdenhil et al., 2017). This assumption was indeed confirmed with the maps that showed the performances calculated during the winter months and growing season (See Figure 4.16).

(III) The simulations of individual grid points revealed that the simulations in general contain much more noise than the smooth slope $\sigma'(40)$. The possible reason for this are the high annually varying soil moisture

and leaf area index.

(IV) No suggestions or conclusions were obtained from the fractional cover statistics for the considered dominant land cover types. For both grid points with good performances and points with bad performances, no real reason is found why presence of certain land cover types is beneficial or not for slope simulations.

### 5.1.3. Sources of uncertainty

- The ISBA-A-g$_s$ land surface model is able to simulate global land surface variable time series. However, further improvements of the model were recommended by Gibelin et al. (2006) for the treatment of vegetation in cultivated areas.
- Seasonal biases in the performance maps. It was shown that performances calculated over different time spans give different interpretations of the level of success of the created models. These differences were found to be very significant. It is therefore expected that high differences can be expected as well for performance calculated over a specific season.
- Inaccurate assignment of land cover types. The land cover type classifications are based on the ECOCLIMAP classes with a spatial resolution of 1 kilometer. Whether the grid points give an accurate representation of reality or not, with the time span of 10 years land cover succession might occur.
- Importance of input parameters soil moisture and leaf area during the optimization is not known. Hence, it is not known if the results are obtained for the right reasons.

## 5.2. Conclusions

### 5.2.1. Research Questions

- *What are the limitations of the optimization procedure?*

  A limitation of the methodology is that actually the grid points are not completely optimized by the Grid Search. If during cross validation there is just a bit more probability that C will be 10 instead of 100, it could mean that actually the performance would be much higher if for example C is optimal at a value of 50. The limited set of grid search parameters was based on efficiency as optimizing SVMs generally is computationally expensive.

  The methodology itself could be used to obtain SVR parameters worldwide for simulations of backscatter $\sigma^0$. For the slope $\sigma'(40)$ however, obtaining the right set of parameter combinations is more challenging because besides seasonal effects, higher frequency fluctuations are not captured by the SVR model. In both cases the seasonal cycles of the $\sigma^0(40)$ and $\sigma'(40)$ were simulated well.

- *What are the advantages and disadvantages of the used method?*

  The foremost advantage is the ability to create a bunch of models with the SVR and select the best set of parameter combinations with the best cross validation. It was shown in chapter 4 that there is a spatial consistency with SVR parameters and hence a new relationships between the method and land cover types can be established. No real disadvantage of the method is experienced, besides the time and computational expenses it takes to train a model.

- *What impact does land cover have on the performance of the simulation?*

  Land cover has shown to be a very relevant factor for the performance of the simulations. For the backscatter $\sigma^0(40)$, it was found that presence of agriculturally active areas within a grid cell of 25 kilometers results in a poor performance while forested areas are understood relatively well by the SVR. Even with the $\sigma'(40)$ simulations land cover plays a role, though it is less clear than for $\sigma^0(40)$. Besides seasonal fluctuations, higher frequency fluctuations are not captured with the SVRs.

**Main research question**:

*Is an optimized support vector machine with the ISBA-A-$g_s$ land surface model able to serve as a surrogate model to simulate $\sigma^0(40)$ and $\sigma'(40)$?*

This study shows that the method of optimizing support vector machines is very promising for simulating backscatter $\sigma^0(40)$ and slope $\sigma'(40)$. The models are optimized giving it a set of parameter combinations without priory taking land covers into account. The performance (coefficient of determination) maps show that reasonable results can be obtained for models that are not even optimized per season but for a time period of a whole year.

### 5.2.2. Recommendations

The following recommendations are given based on the results of this study:

- Shamambo et al. (2019) simulated normalized ASCAT backscatter with a semi-empirical water cloud model and they focused more on performances during spring and summer. In this study both seasons are taken together as the growing season. But both seasons have different characteristics in terms of vegetation growth and soil/vegetation wetness.
- Optimize models separately for each specific season. This is especially important for agriculturally active areas as these areas change a lot per season compared to a forest for example. In this study, SVMs are optimized for a whole year and hence the best possible parameter combination is taken for a whole year, which is not optimal for seasonal differences.
- A limit of this study is the relation between the soil moisture and leaf area index (which are the variables used to forward model $\sigma^0(40)$ and $\sigma'(40)$) with performances of simulations. The advantage of including the statistics of this relation is to correlate wet and dry soil conditions and/or sparse and dense vegetation conditions with performances of the simulations during different periods throughout the year.
- Test the stability of the created SVRs by performing a sensitivity analysis. With sensitivity analysis, the importance of the model input variables (soil moisture and leaf area index) can be determined for the simulation (Iooss and Saltelli, 2015). In total, 1324 support vector machines have been optimized. It is not known whether the right results are obtained for the right reasons.
- Understand parameter combinations in relation to land cover types and performances to enhance the optimization procedure. This could be useful after optimizing per season or optimizing for a period of two seasons.
- The reason for providing a parameter grid for optimization, was to find if initial values of SVM parameters are spatially consistent. After optimization, a spatial consistency was found and hence the assigning of parameters to grid points is not random. A disadvantage of doing this in the future is that all parameter combinations are tested, which is computationally expensive and time consuming. An alternative strategy is (Sequential Model based) Bayesian optimization (See Shahriari et al., 2016). This would be a good improvement of the methodology because results can be improved as not only hyperparameter spaces are predicted but also the uncertainty of the prediction is given. So this method improves itself as it improves previous results

# Bibliography


[1] Clément Albergel, Simon Munier, Delphine Jennifer Leroux, Hélène Dewaele, David Fairbairn, Alina Lavinia Barbu, Emiliano Gelati, Wouter Dorigo, Stéphanie Faroux, Catherine Meurey, Patrick Le Moigne, Bertrand Decharme, Jean-Francois Mahfouf, and Jean-Christophe Calvet. Sequential assimilation of satellite-derived vegetation and soil moisture products using SURFEX_v8.0: LDAS-monde assessment over the euro-mediterranean area. *Geoscientific Model Development*, 10(10):3889–3912, October 2017. doi: 10.5194/gmd-10-3889-2017. URL https://doi.org/10.5194/gmd-10-3889-2017.

[2] A. L. Barbu, J.-C. Calvet, J.-F. Mahfouf, C. Albergel, and S. Lafont. Assimilation of soil wetness index and leaf area index into the ISBA-a-gs land surface model: grassland case study. *Biogeosciences*, 8(7):1971–1986, July 2011. doi: 10.5194/bg-8-1971-2011. URL https://doi.org/10.5194/bg-8-1971-2011.

[3] A. Brut, C. Rüdiger, S. Lafont, J.-L. Roujean, J.-C. Calvet, L. Jarlan, A.-L. Gibelin, C. Albergel, P. Le Moigne, J.-F. Soussana, K. Klumpp, D. Guyon, J.-P. Wigneron, and E. Ceschia. Modelling LAI at a regional scale with ISBA-a-gs: comparison with satellite-derived LAI over southwestern france. *Biogeosciences*, 6(8):1389–1404, August 2009. doi: 10.5194/bg-6-1389-2009. URL https://doi.org/10.5194/bg-6-1389-2009.

[4] Jean-Christophe Calvet, Joël Noilhan, Jean-Louis Roujean, Pierre Bessemoulin, Maurice Cabelguenne, Albert Olioso, and Jean-Pierre Wigneron. An interactive vegetation SVAT model tested against data from six contrasting sites. *Agricultural and Forest Meteorology*, 92(2):73–95, July 1998. doi: 10.1016/s0168-1923(98)00091-4. URL https://doi.org/10.1016/s0168-1923(98)00091-4.

[5] N. Canal, J.-C. Calvet, B. Decharme, D. Carrer, S. Lafont, and G. Pigeon. Evaluation of root water uptake in the ISBA-a-gs land surface model using agricultural yield statistics over france. *Hydrology and Earth System Sciences*, 18(12):4979–4999, December 2014. doi: 10.5194/hess-18-4979-2014. URL https://doi.org/10.5194/hess-18-4979-2014.

[6] Barton A. Forman and Rolf H. Reichle. Using a support vector machine and a land surface model to estimate large-scale passive microwave brightness temperatures over snow-covered land in north america. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(9):4431–4441, September 2015. doi: 10.1109/jstars.2014.2325780. URL https://doi.org/10.1109/jstars.2014.2325780.

[7] Anne-Laure Gibelin, Jean-Christophe Calvet, Jean-Louis Roujean, Lionel Jarlan, and Sietse O. Los. Ability of the land surface model ISBA-a-gs to simulate leaf area index at the global scale: Comparison with satellites products. *Journal of Geophysical Research*, 111(D18), 2006. doi: 10.1029/2005jd006691. URL https://doi.org/10.1029/2005jd006691.

[8] Sebastian Hahn, Christoph Reimer, Mariette Vreugdenhil, Thomas Melzer, and Wolfgang Wagner. Dynamic characterization of the incidence angle dependence of backscatter using metop ASCAT. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):2348–2359, May 2017. doi: 10.1109/jstars.2016.2628523. URL https://doi.org/10.1109/jstars.2016.2628523.

[9] Hsin-Yuan Huang and Chih-Jen Lin. Linear and kernel classification: When to use which? In *Proceedings of the 2016 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, June 2016. doi: 10.1137/1.9781611974348.25. URL https://doi.org/10.1137/1.9781611974348.25.

[10] Pasquale Imperatore and Daniele Riccio, editors. *Geoscience and Remote Sensing New Achievements*. InTech, February 2010. doi: 10.5772/214. URL https://doi.org/10.5772/214.

[11] Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95, January 2017. doi: 10.3390/rs9010095. URL https://doi.org/10.3390/rs9010095.

[12] Bertrand Iooss and Andrea Saltelli. Introduction to sensitivity analysis. In *Handbook of Uncertainty Quantification*, pages 1–20. Springer International Publishing, 2015. doi: 10.1007/978-3-319-11259-6_31-1. URL https://doi.org/10.1007/978-3-319-11259-6_31-1.

[13] C.M.J. Jacobs, B.M.M. van den Hurk, and H.A.R. de Bruin. Stomatal behaviour and photosynthetic rate of unstressed grapevines in semi-arid conditions. *Agricultural and Forest Meteorology*, 80(2-4):111–134, July 1996. doi: 10.1016/0168-1923(95)02295-3. URL https://doi.org/10.1016/0168-1923(95)02295-3.

[14] Delphine Leroux, Jean-Christophe Calvet, Simon Munier, and Clément Albergel. Using satellite-derived vegetation products to evaluate LDAS-monde over the euro-mediterranean area. *Remote Sensing*, 10(8):1199, July 2018. doi: 10.3390/rs10081199. URL https://doi.org/10.3390/rs10081199.

[15] Shunlin Liang, Xiaowen Li, and Jindi Wang. Chapter 19 - soil moisture content. In *Advanced Remote Sensing*, pages 589 – 614. Academic Press, Boston, 2012. ISBN 978-0-12-385954-9. doi: https://doi.org/10.1016/B978-0-12-385954-9.00019-8. URL http://www.sciencedirect.com/science/article/pii/B9780123859549000198.

[16] Mahmood Mahmoodian, Jairo Arturo Torres-Matallana, Ulrich Leopold, Georges Schutz, and Francois H. L. R. Clemens. A data-driven surrogate modelling approach for acceleration of short-term simulations of a dynamic urban drainage simulator. *Water*, 10(12):1849, December 2018. doi: 10.3390/w10121849. URL https://doi.org/10.3390/w10121849.

[17] Oded Maimon and Lior Rokach, editors. *Data Mining and Knowledge Discovery Handbook*. Springer US, 2010. doi: 10.1007/978-0-387-09823-4. URL https://doi.org/10.1007/978-0-387-09823-4.

[18] Valéry Masson, Jean-Louis Champeaux, Fabrice Chauvin, Christelle Meriguet, and Roselyne Lacaze. A global database of land surface parameters at 1-km resolution in meteorological and climate models. *Journal of Climate*, 16(9):1261–1282, May 2003. doi: 10.1175/1520-0442-16.9.1261. URL https://doi.org/10.1175/1520-0442-16.9.1261.

[19] Vahid Naeimi and Wolfgang Wagner. C-band scatterometers and their applications. In *Geoscience and Remote Sensing New Achievements*. InTech, February 2010. doi: 10.5772/9102. URL https://doi.org/10.5772/9102.

[20] J. Noilhan and S. Planton. A simple parameterization of land surface processes for meteorological models. *Monthly Weather Review*, 117(3):536–549, March 1989. doi: 10.1175/1520-0493(1989)117<0536:aspols>2.0.co;2. URL https://doi.org/10.1175/1520-0493(1989)117<0536:aspols>2.0.co;2.

[21] Kun Qiao, Wenquan Zhu, Zhiying Xie, and Peixian Li. Estimating the seasonal dynamics of the leaf area index using piecewise LAI-VI relationships based on phenophases. *Remote Sensing*, 11(6):689, March 2019. doi: 10.3390/rs11060689. URL https://doi.org/10.3390/rs11060689.

[22] Raphael Quast, Clément Albergel, Jean-Christophe Calvet, and Wolfgang Wagner. A generic first-order radiative transfer modelling approach for the inversion of soil and vegetation parameters from scatterometer observations. *Remote Sensing*, 11(3):285, February 2019. doi: 10.3390/rs11030285. URL https://doi.org/10.3390/rs11030285.

[23] Ignacio Rojas, Gonzalo Joya, and Andreu Catala, editors. *Advances in Computational Intelligence*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-59153-7. URL https://doi.org/10.1007/978-3-319-59153-7.

[24] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, January 2016. doi: 10.1109/jproc.2015.2494218. URL https://doi.org/10.1109/jproc.2015.2494218.

[25] Daniel Chiyeka Shamambo, Bertrand Bonan, Jean-Christophe Calvet, Clément Albergel, and Sebastian Hahn. Interpretation of ASCAT radar scatterometer observations over land: A case study over southwestern france. *Remote Sensing*, 11(23):2842, November 2019. doi: 10.3390/rs11232842. URL https://doi.org/10.3390/rs11232842.

[26] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004. doi: 10.1023/b:stco.0000035301.49549.88. URL https://doi.org/10.1023/b:stco.0000035301.49549.88.

[27] Susan C. Steele-Dunne, Sebastian Hahn, Wolfgang Wagner, and Mariette Vreugdenhil. Investigating vegetation water dynamics and drought using metop ASCAT over the north american grasslands. *Remote Sensing of Environment*, 224:219–235, April 2019. doi: 10.1016/j.rse.2019.01.004. URL https://doi.org/10.1016/j.rse.2019.01.004.

[28] F. Ulaby. Radar response to vegetation. *IEEE Transactions on Antennas and Propagation*, 23(1):36–45, January 1975. doi: 10.1109/tap.1975.1140999. URL https://doi.org/10.1109/tap.1975.1140999.

[29] Mariette Vreugdenhil, Wouter A. Dorigo, Wolfgang Wagner, Richard A. M. de Jeu, Sebastian Hahn, and Margreet J. E. van Marle. Analyzing the vegetation parameterization in the TU-wien ASCAT soil moisture retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3513–3531, June 2016. doi: 10.1109/tgrs.2016.2519842. URL https://doi.org/10.1109/tgrs.2016.2519842.

[30] Mariette Vreugdenhil, Sebastian Hahn, Thomas Melzer, Bernhard BauerMarschallinger, Christoph Reimer, Wouter Arnoud Dorigo, and Wolfgang Wagner. Assessing vegetation dynamics over mainland australia with metop ASCAT. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(5):2240–2248, May 2017. doi: 10.1109/jstars.2016.2618838. URL https://doi.org/10.1109/jstars.2016.2618838.

[31] W. Wagner, G. Lemoine, M. Borgeaud, and H. Rott. A study of vegetation cover effects on ERS scatterometer data. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):938–948, March 1999. doi: 10.1109/36.752212. URL https://doi.org/10.1109/36.752212.

[32] W. Wagner, J. Noll, M. Borgeaud, and H. Rott. Monitoring soil moisture over the canadian prairies with the ERS scatterometer. *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):206–216, 1999. doi: 10.1109/36.739155. URL https://doi.org/10.1109/36.739155.

[33] W. Wagner, R. A. M. De Jeu, T. R. H. Holmes, A. J. Dolman, N. C. van de Giesen, and J. Friesen. Global soil moisture patterns observed by space borne microwave radiometers and scatterometers. *Surveys in Geophysics*, 29(4-5):399–420, October 2008. doi: 10.1007/s10712-008-9044-0. URL https://doi.org/10.1007/s10712-008-9044-0.

[34] Wolfgang Wagner, Günter Blöschl, Paolo Pampaloni, Jean-Christophe Calvet, Bizzarro Bizzarri, Jean-Pierre Wigneron, and Yann Kerr. Operational readiness of microwave remote sensing of soil moisture for hydrologic applications. *Hydrology Research*, 38(1):1–20, February 2007. doi: 10.2166/nh.2007.029. URL https://doi.org/10.2166/nh.2007.029.

[35] Wolfgang Wagner, Sebastian Hahn, Richard Kidd, Thomas Melzer, Zoltan Bartalis, Stefan Hasenauer, Julia Figa-Saldaña, Patricia de Rosnay, Alexander Jann, Stefan Schneider, Jürgen Komma, Gerhard Kubu, Katharina Brugger, Christoph Aubrecht, Johann Züger, Ute Gangkofner, Stefan Kienberger, Luca Brocca, Yong Wang, Günter Blöschl, Josef Eitzinger, and Kla Steinnocher. The ASCAT soil moisture product: A review of its specifications, validation results, and emerging applications. *Meteorologische Zeitschrift*, 22(1):5–33, February 2013. doi: 10.1127/0941-2948/2013/0399. URL https://doi.org/10.1127/0941-2948/2013/0399.

# A

# Figures

## A.1. Backscatter SVR parameter Combinations

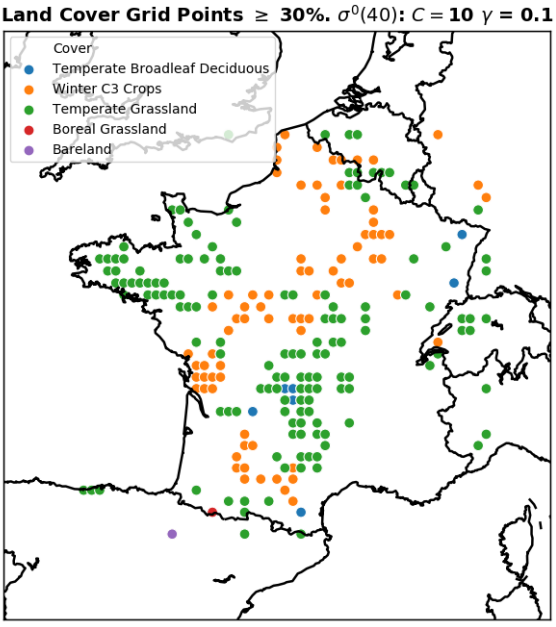



Figure A.1: Map of land cover grid points with SVR parameter combination $C = 10$ and $\gamma = 0.1$ for $\sigma^0(40)$ simulations.

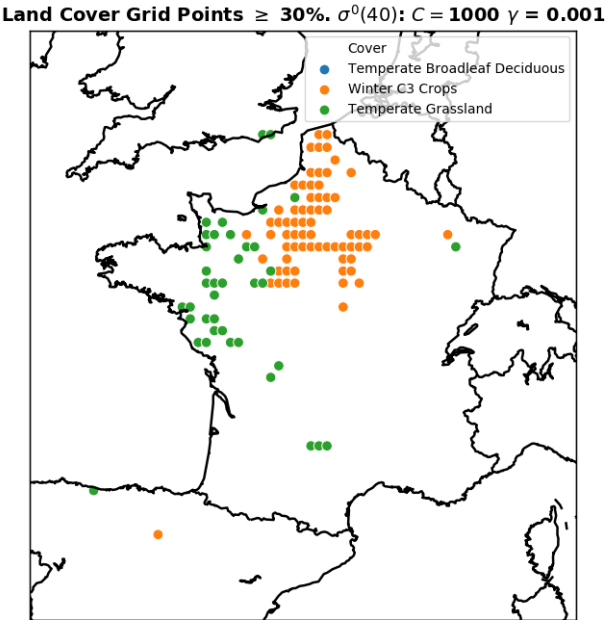



Figure A.2: Map of land cover grid points with SVR parameter combination $C = 1000$ and $\gamma = 0.001$ for $\sigma^0(40)$ simulations.

## A.2. Slope SVR Parameter Combinations

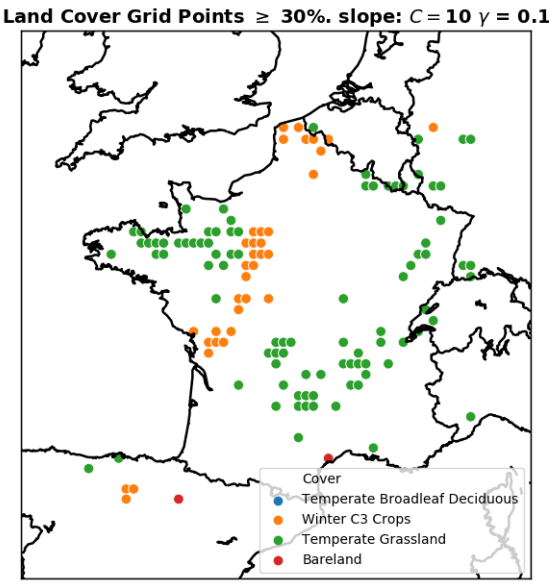



Figure A.3: Map of land cover grid points with SVR parameter combination $C = 10$ and $\gamma = 0.1$ for $\sigma^0(40)$ simulations

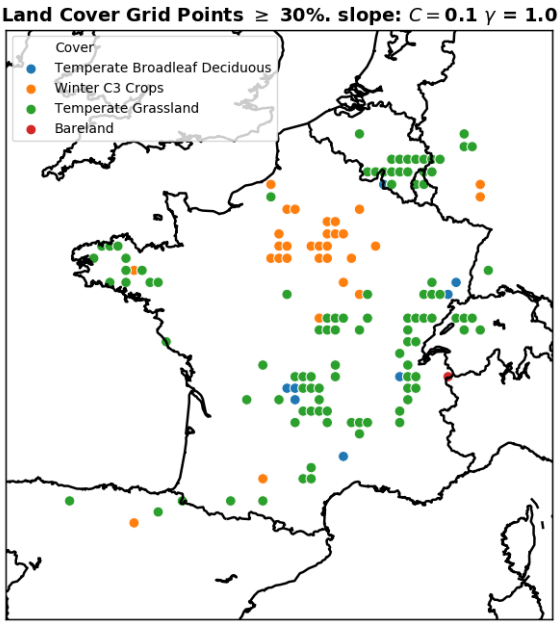
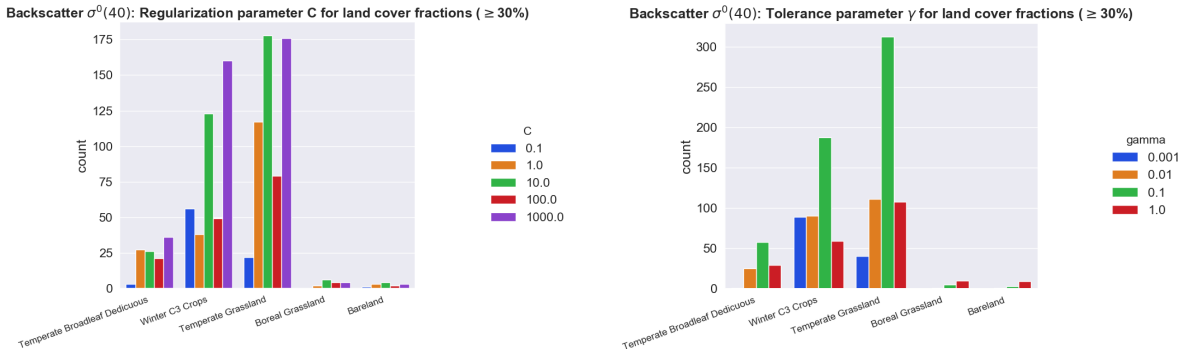


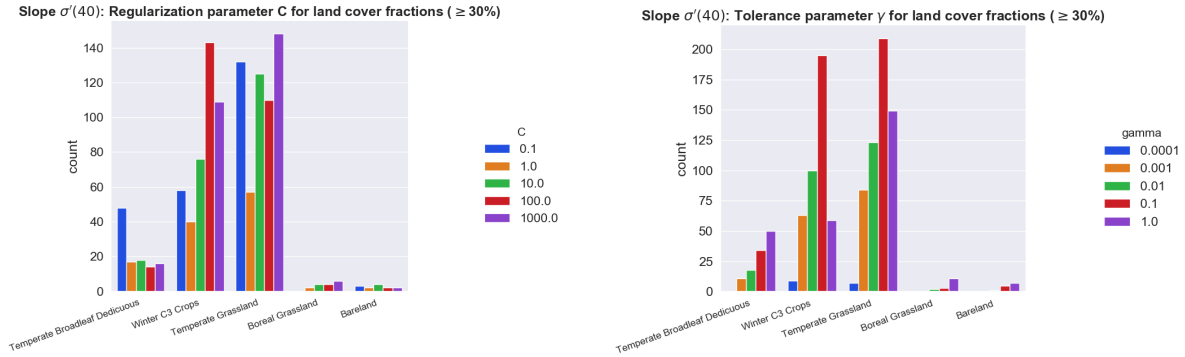

Figure A.4: Map of land cover grid points with SVR parameter combination $C = 0.1$ and $\gamma = 1$ for $\sigma^0(40)$ simulations

## A.3. Histogram Parameter Occurrences




(a) label 1 (b) label 2

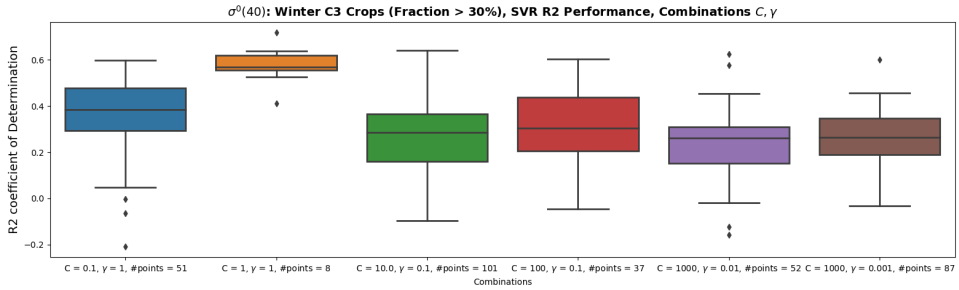
Figure A.5: $\sigma^0(40)$ amount of occurrences of SVR parameter $C$ on the left hand side (a) and $\gamma$ on the right hand side (b)



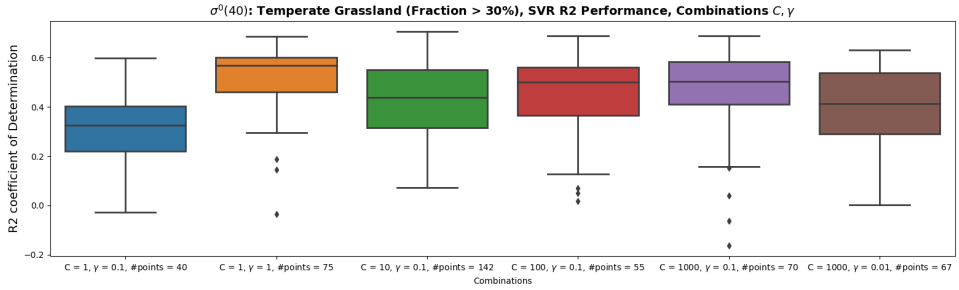

(a) label 1 (b) label 2

Figure A.6: $\sigma'(40)$ amount of occurrences of SVR parameter $C$ on the left hand side (a) and $\gamma$ on the right hand side (b)

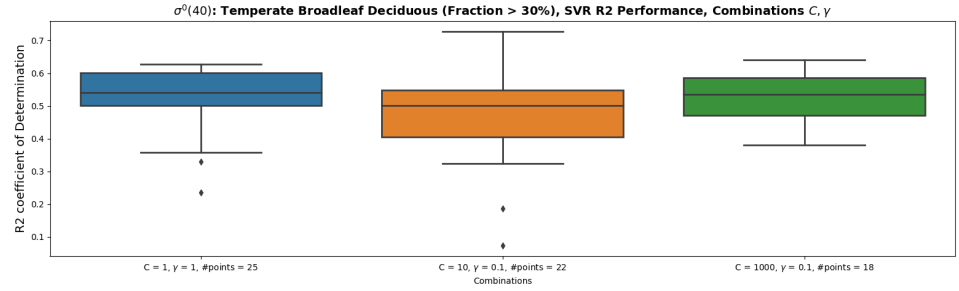
## A.4. R2 Performances Parameter Combinations



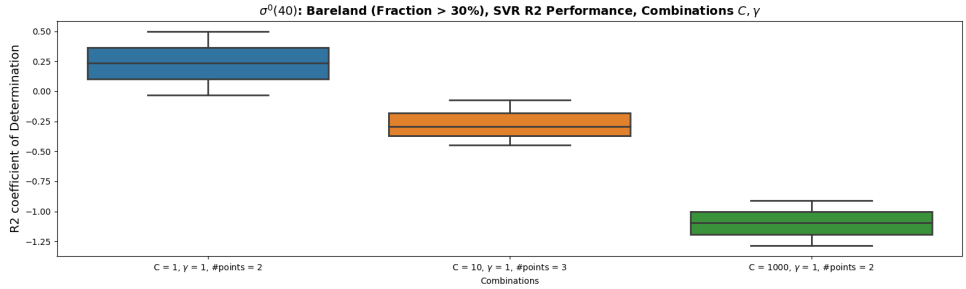

(a)




(b)

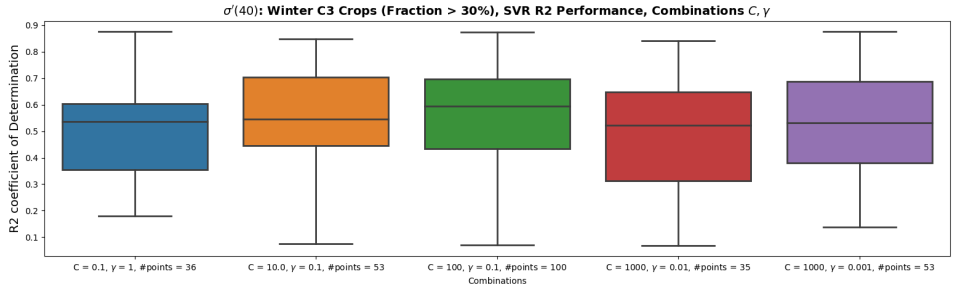



(c)

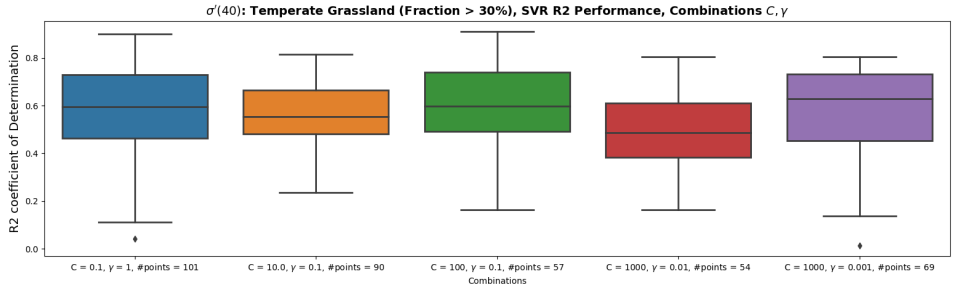



(d)

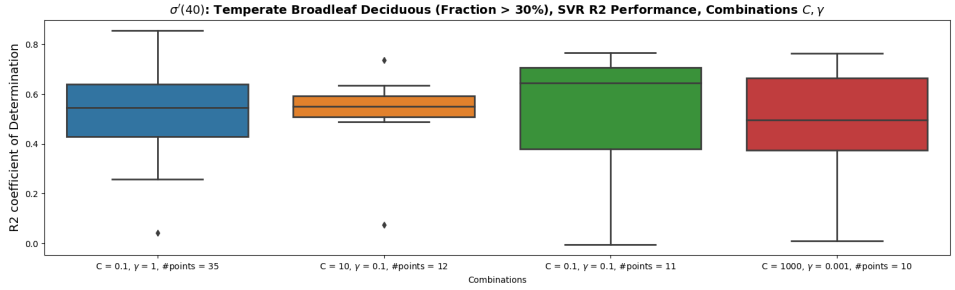
Figure A.7: $\sigma^0(40)$ R2 performances for simulations with removed climatology for the most occurring parameter combinations for the four most dominant land cover types in France.



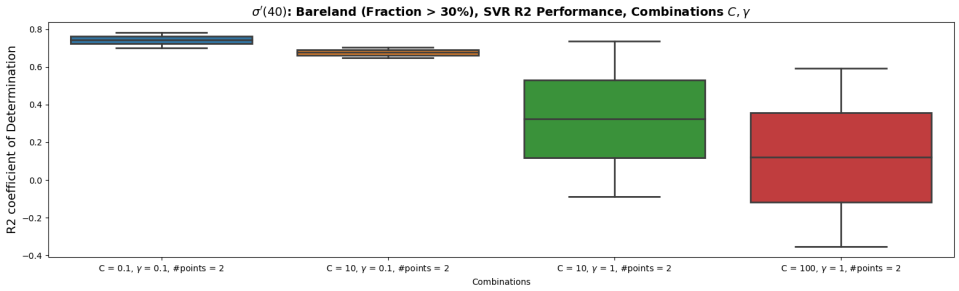

(a)




(b)




(c)




(d)

Figure A.8: $\sigma'(40)$ R2 performances for simulations for the most occurring parameter combinations for the four most dominant land cover types in France.

# B

# Tables

## B.1. Slope Land Cover Fractions

| Fractional Cover Statistics $\sigma'(40)$ | **Good: R2 (Growing Season) >0.6 \| Bad: R2 (Growing Season) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Bare Land** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.363 | 0.070 | 7 | 0.404 | 0.084 | 2 |
| Bare Rock | 0.072 | 0.061 | 7 | 0.126 | 0.018 | 2 |
| Permanent Snow | 0.008 | 0.016 | 7 | 0.004 | 0.002 | 2 |
| Temperate Broadleaf Deciduous | 0.044 | 0.042 | 7 | 0.029 | 0.032 | 2 |
| C3 Crops | 0.109 | 0.121 | 7 | 0.004 | 0.002 | 2 |
| C4 Crops | 0.014 | 0.015 | 7 | 0.001 | 0.002 | 2 |
| Summer C3 Crops | 0.019 | 0.020 | 7 | 0.000 | 0.000 | 2 |
| Temperate Grassland | 0.099 | 0.070 | 7 | 0.033 | 0.044 | 2 |
| Temperate Needleleaf Evergreen | 0.025 | 0.026 | 7 | 0.000 | 0.000 | 2 |
| Boreal Grassland | 0.107 | 0.143 | 7 | 0.231 | 0.031 | 2 |
| Boreal Broadleaf Deciduous | 0.008 | 0.010 | 7 | 0.027 | 0.017 | 2 |
| Shrubs | 0.072 | 0.084 | 7 | 0.018 | 0.001 | 2 |

Table B.1: Fractional Cover Statistics for $\sigma'(40)$ simulations with Bare Land as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

| Fractional Cover Statistics $\sigma'(40)$ | **Good: R2 (Growing Season) >0.6 \| Bad: R2 (Growing Season) <0.3** | | | | | |
|---|---|---|---|---|---|---|
| **Dominant Land Cover Type: Boreal Grassland** | $\mu$ Good [x 100%] | $\sigma$ Good [x 100%] | # Points | $\mu$ Bad [x 100%] | $\sigma$ Bad [x 100%] | # Points |
| Bare Land | 0.173 | 0.059 | 13 | - | - | 0 |
| Bare Rock | 0.127 | 0.037 | 13 | - | - | 0 |
| Permanent Snow | 0.003 | 0.003 | 13 | - | - | 0 |
| Temperate Broadleaf Deciduous | 0.025 | 0.021 | 13 | - | - | 0 |
| C3 Crops | 0.005 | 0.002 | 13 | - | - | 0 |
| C4 Crops | 0.001 | 0.001 | 13 | - | - | 0 |
| Summer C3 Crops | 0.000 | 0.000 | 13 | - | - | 0 |
| Temperate Grassland | 0.063 | 0.058 | 13 | - | - | 0 |
| Temperate Needleleaf Evergreen | 0.003 | 0.005 | 13 | - | - | 0 |
| Boreal Grassland | 0.346 | 0.030 | 13 | - | - | 0 |
| Boreal Broadleaf Deciduous | 0.043 | 0.018 | 13 | - | - | 0 |
| Shrubs | 0.028 | 0.006 | 13 | - | - | 0 |

Table B.2: Fractional Cover Statistics for $\sigma'(40)$ simulations with Boreal Grassland as dominant land cover type. Grid points are distinguished in good and bad performing grid points. $\mu$ is the mathematical mean of the land cover fraction of the corresponding grid points. $\sigma$ is the mathematical standard deviation of the land cover fraction. # Points means the amount of grid points that are considered as either good or bad grid points over which the $\mu$ and $\sigma$ are calculated.

## B.2. Parameter Occurrences Backscatter

| **Winter C3 Crops** | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 0 | 5 | 51 |
| 1.0 | 0 | 0 | 6 | 24 | 8 |
| 10.0 | 0 | 0 | 22 | 101 | 0 |
| 100.0 | 0 | 2 | 10 | 37 | 0 |
| 1000.0 | 0 | 87 | 52 | 21 | 0 |

Table B.3: Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the C3 crops land cover.

| **Temperate Grassland** | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 0 | 6 | 16 |
| 1.0 | 0 | 0 | 2 | 40 | 75 |
| 10.0 | 0 | 0 | 19 | 142 | 17 |
| 100.0 | 0 | 1 | 23 | 55 | 0 |
| 1000.0 | 0 | 39 | 67 | 70 | 0 |

Table B.4: Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the temperate grassland land cover.

| **Temperate Broadleaf Deciduous** | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 0 | 0 | 3 |
| 1.0 | 0 | 0 | 0 | 2 | 25 |
| 10.0 | 0 | 0 | 3 | 22 | 1 |
| 100.0 | 0 | 0 | 5 | 16 | 0 |
| 1000.0 | 0 | 1 | 17 | 18 | 0 |

Table B.5: Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the Temperate Broadleaf Deciduous land cover.

| Bare Land | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 0 | 0 | 1 |
| 1.0 | 0 | 0 | 0 | 1 | 2 |
| 10.0 | 0 | 0 | 0 | 1 | 3 |
| 100.0 | 0 | 0 | 0 | 1 | 1 |
| 1000.0 | 0 | 0 | 1 | 0 | 2 |

Table B.6: Amount of occurrence for SVR parameter combinations for $\sigma^0(40)$ regressions for the Bareland land cover.

## B.3. Parameter Occurrences Slope

| Winter C3 Crops | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 8 | 14 | 36 |
| 1.0 | 0 | 0 | 21 | 8 | 11 |
| 10.0 | 5 | 0 | 7 | 53 | 11 |
| 100.0 | 3 | 10 | 29 | 100 | 1 |
| 1000.0 | 1 | 53 | 35 | 20 | 0 |

Table B.7: Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the winter c3 crops land cover.

| Temperate Grassland | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 11 | 20 | 101 |
| 1.0 | 0 | 3 | 10 | 18 | 26 |
| 10.0 | 7 | 1 | 14 | 90 | 13 |
| 100.0 | 0 | 11 | 34 | 57 | 8 |
| 1000.0 | 0 | 69 | 54 | 24 | 1 |

Table B.8: Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Temperate Grassland land cover.

| Temperate Broadleaf Deciduous | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 2 | 11 | 35 |
| 1.0 | 0 | 0 | 2 | 6 | 9 |
| 10.0 | 0 | 0 | 2 | 12 | 4 |
| 100.0 | 0 | 1 | 8 | 3 | 2 |
| 1000.0 | 0 | 10 | 4 | 2 | 0 |

Table B.9: Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Temperate Broadleaf Deciduous land cover.

| Bare Land | | | | | |
|---|---|---|---|---|---|
| **C \| $\gamma$** | 0.0001 | 0.0010 | 0.0100 | 0.1000 | 1.0000 |
| 0.1 | 0 | 0 | 0 | 2 | 1 |
| 1.0 | 0 | 0 | 0 | 1 | 1 |
| 10.0 | 0 | 0 | 0 | 2 | 2 |
| 100.0 | 0 | 0 | 0 | 0 | 2 |
| 1000.0 | 0 | 0 | 1 | 0 | 1 |

Table B.10: Amount of occurrence for SVR parameter combinations for $\sigma'(40)$ regressions for the Bareland land cover.