TU WIEN Informatics

Maschinelles Lernen in der Fernerkundung

Bewertung der Auswirkungen langfristiger Landbedeckungsänderungen auf Metop ASCAT's Rückstreuung mittels Drift-Erkennung

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Jana Elena Huber, BSc
Matrikelnummer 01551001

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner
Mitwirkung: Dipl.-Ing. Sebastian Hahn

Wien, 23. August 2025

Jana Elena Huber	Wolfgang Wagner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Machine Learning in Remote Sensing

Assessing the Impact of Long-Term Land Cover Changes on Metop ASCAT Backscatter using Drift Detection

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Jana Elena Huber, BSc
Registration Number 01551001

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner
Assistance: Dipl.-Ing. Sebastian Hahn

Vienna, August 23, 2025

Jana Elena Huber	Wolfgang Wagner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Jana Elena Huber, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 23. August 2025

Jana Elena Huber

Danksagung

Zuallererst möchte ich Univ.-Prof. Dr. Wolfgang Wagner meinen aufrichtigen Dank aussprechen. Durch seine fachliche Unterstützung, die regelmäßigen Besprechungen und sein konstruktives Feedback hat er maßgeblich zum Gelingen dieser Arbeit beigetragen.

Ebenso danke ich Dipl.-Ing. Sebastian Hahn, der mir stets mit Rat und Tat zur Seite stand. Seine Impulse haben mich immer wieder auf den richtigen Weg geführt und mir geholfen, den Fokus nicht zu verlieren.

Ein besonderer Dank gilt meinem kleinen Bruder Gideon. Gidi, du warst mein „inoffizieller dritter Betreuer“. Mit deinem Wissen, deiner Erfahrung in der Betreuung von Masterarbeiten und vor allem deiner entspannten und stets positiven Art hast du mir viele Sorgen genommen und mich enorm gestärkt.

Von Herzen danke ich meiner Familie und meinen Freund:innen, allen voran meinen Eltern. Ohne euch wäre dieser Weg nicht möglich gewesen. Ihr habt mir Vertrauen in mich selbst gegeben und mich von Anfang an in jeder Hinsicht unterstützt – Danke Mama! Danke Papa!

Meinem Ehemann bin ich unendlich dankbar. Seit zwölf Jahren stehst du an meiner Seite und gibst jeden Tag dein Bestes für unsere kleine Familie mit unserem Sohn Leonidas. Deine Liebe und dein Rückhalt bedeuten mir sehr viel.

Zum Schluss danke ich Annabel – wir haben uns in diesem Prozess gegenseitig motiviert und durch so manches Tief getragen. Ebenso möchte ich Timi und Bella danken, die meine Studienzeit auf wunderbare Art und Weise mitgeprägt haben.

Allen, die mich auf diesem Weg begleitet haben, gilt mein aufrichtiger Dank.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to Univ.-Prof. Dr. Wolfgang Wagner. His academic support, regular meetings, and constructive feedback contributed significantly to the success of this thesis.

I am also very grateful to Dipl.-Ing. Sebastian Hahn, who continuously supported me with advice and guidance. His input helped me stay on track and maintain focus throughout the process.

A special thanks goes to my younger brother Gideon. Gidi, you were my "unofficial third supervisor." Your knowledge, your experience in supervising master's theses, and above all your calm and positive attitude helped me worry less and gave me great strength.

I wholeheartedly thank my family and friends, especially my parents. Without you, this journey would not have been possible. You have given me confidence in myself and supported me in every way from the very beginning – thank you, Mama! Thank you, Papa!

My deepest thanks go to my husband. You have been by my side for twelve years and do your best every day for our little family with our son Leonidas. Your love and encouragement mean a lot to me.

Finally, I would like to thank Annabel - we motivated each other during this process and carried each other through quite a few lows. I would also like to thank Timi and Bella, who contributed to my time at university in a wonderful way.

To everyone who has supported me on this path – thank you from the bottom of my heart.

Kurzfassung

Diese Arbeit widmet sich der Analyse langfristiger Trends im ASCAT-Rückstreusignal (Backscatter) und deren potenziellen Ursachen. Angesichts der engen Kopplung von Radar-Backscatter an oberflächennahe Bodenfeuchte liefert ein verbessertes Verständnis seiner zeitlichen Dynamiken wertvolle Beiträge zur Umweltüberwachung – insbesondere in den Bereichen Landwirtschaft, Wasserressourcenmanagement und Ökosystemforschung.

Zur Modellierung des Backscatter-Signals wurden drei Regressionsansätze – Elastic Net, Random Forest und Support Vector Regression – systematisch verglichen. Das Elastic Net-Modell zeigte die überzeugendste Kombination aus Prognosegüte und Effizienz und wurde anschließend auf ERA5-Land-Bodenfeuchte und MODIS-LAI für Südostaustralien im Zeitraum 2007–2022 angewendet. Darauf aufbauend erfolgte eine Driftanalyse mittels Mann-Kendall-Tests, die auf Zielvariable, Modellvorhersagen, Residuen und Eingangsdaten angewendet wurden. Die identifizierten Trends wurden kategorisiert und hinsichtlich ihrer Stärke und Verlaufsmuster segmentiert. Eine umfassende visuelle Analyse durch Zeitreihen und räumliche Karten ergänzte die Ergebnisse.

Die Resultate belegen einen signifikanten Einfluss der Landbedeckung auf das Rückstreuverhalten und die Modellgüte. Während an vielen Orten statistisch signifikante Trends festgestellt wurden, traten starke und kontinuierliche Entwicklungen vergleichsweise selten auf. Einzelne abrupte Veränderungen lassen sich häufig auf Extremereignisse zurückführen. Insgesamt zeigt die Studie, dass die kombinierte Anwendung datengetriebener Modellierung und statistischer Trendanalyse ein wirkungsvolles Instrument zur Identifikation und Interpretation längerfristiger Veränderungen in radarfernerkundeten Umweltgrößen darstellt.

Abstract

This thesis investigates long-term trends in ASCAT radar backscatter and their potential causes. Given the strong relationship between backscatter and surface soil moisture, understanding the temporal dynamics of the radar signal provides valuable insights for environmental monitoring, with relevance to agriculture, water resource management, and ecosystem analysis.

Three regression models — Elastic Net, Random Forest, and Support Vector Regression — were systematically evaluated for their ability to estimate backscatter based on ERA5-Land soil moisture and MODIS LAI. The Elastic Net model offered the best balance of predictive accuracy and computational efficiency and was applied to time series data for southeastern Australia from 2007 to 2022. A subsequent drift analysis employed Mann-Kendall trend tests on the target variable, model predictions, residuals, and input features. The resulting trends were categorized and segmented according to magnitude and pattern, supported by a visual exploration through time series plots and spatial heatmaps.

The results reveal a substantial influence of land cover on both backscatter behavior and model performance. While many grid points exhibited statistically significant trends, strong and continuous developments were relatively rare. In contrast, abrupt changes could often be attributed to extreme events. Overall, the study demonstrates that the combined use of data-driven modeling and trend analysis represents a robust framework for detecting and interpreting long-term changes in radar-derived environmental signals.

Contents

CHAPTER 1

Introduction

The first chapter of the thesis introduces the motivation behind the topic and defines the problem to be addressed. It states the purpose of the thesis and explains the research questions, followed by a summary of the thesis outline.

1.1 Motivation

Climate change is leading to increasingly extreme and unpredictable weather conditions. Events such as droughts and floods occur more frequently and pose serious threats to human life, health, and agricultural productivity. Crop failures caused by such events endanger global food security in the long term. To effectively respond to these challenges, the early prediction of extreme events is essential. Reliable information on soil moisture is crucial in this context, as it enables targeted measures such as irrigation and supports informed agricultural and environmental decision-making. Therefore, the accurate analysis and use of vegetation and soil moisture data are key components in anticipating and managing the impacts of a changing climate.

Soil moisture is a key hydrological variable with far-reaching implications for agriculture, ecology, and public health. On global and regional scales, it plays a critical role in regulating surface-atmosphere interactions and influencing weather patterns and climate systems. Accurate and timely soil moisture information is increasingly important to meet global demands for food, water, and fiber, while protecting groundwater resources and ensuring sustainable agricultural productivity [DMK+24].

Monitoring soil moisture at large scales requires robust and consistent observational tools. Satellite-based microwave remote sensing offers a powerful solution in this context. Unlike optical or infrared techniques, microwave sensors can penetrate cloud cover, vegetation, and even the upper soil layers, providing valuable data regardless of atmospheric conditions [WLM+22]. These methods can capture changes in plant water content and soil moisture

using active and passive microwave signals, making them particularly useful for monitoring land surface dynamics over time and across large areas [SSDH+22].

Soil moisture also serves as a critical regulator in the exchange of water, energy, and carbon between the land surface and the atmosphere. It influences processes such as evapotranspiration, surface albedo, and soil respiration — factors central to hydrological cycles, energy balance, and carbon fluxes. Understanding these dynamics is key to improving predictive models and managing natural resources sustainably in the face of ongoing climate change. While satellite-based microwave remote sensing primarily captures moisture in the upper few centimeters of the soil, numerous studies have shown strong correlations with deeper soil moisture levels, confirming its relevance for larger-scale hydrological applications [DDJC+12].

1.2 Problem Statement & Aim of the Thesis

Recent analyses of ASCAT backscatter time series have revealed unexpected trend behavior in various regions. In some areas, the signal shows a clear increase or decrease over time — either gradually or abruptly — that cannot be adequately explained by known environmental processes. These trends are problematic from a physical standpoint, as such directional changes in backscatter are not expected under stable surface conditions.

This discrepancy becomes even more apparent when comparing ASCAT data with independent sources such as soil moisture and vegetation indices. While backscatter is assumed to be mainly influenced by vegetation water content and surface soil moisture, other surface characteristics — such as land cover, structure, or roughness — also affect the radar signal. If a predictive model is constructed using vegetation and soil moisture as input variables, it can explain the backscatter signal well in many cases. However, in some locations the model either fails to produce accurate estimates or fails to replicate the trend direction observed in the backscatter time series. This suggests the presence of additional unaccounted for influences on the radar signal that go beyond the input variables typically considered.

The study of long-term changes in backscatter and their potential links to land cover change is still underexplored. The ASCAT research community remains relatively small, and only limited work has addressed signal drift or unexplained trends in a spatially and temporally resolved manner. This thesis aims to contribute to filling this gap by systematically analyzing trends in backscatter and comparing them to trends in vegetation, soil moisture, and model predictions.

The core objective is to identify gradual or sudden changes in the backscatter signal and assess whether these can be attributed to changes in the known input variables. Grid points are analyzed individually and grouped based on their trend behavior, enabling both spatial and temporal interpretation. A secondary goal is to evaluate how well vegetation and soil moisture can explain backscatter variations across space and time. It is expected that most locations will be well-explained by the selected input variables

and show little or no trend. However, for locations exhibiting clear trends or prediction mismatches, additional influences — possibly related to long-term land cover changes — may be hypothesized.

1.3 Research Questions

This work can be thematically divided into two main areas. Firstly, it involves identifying a machine learning model that is well-suited to the given research framework and delivers strong performance based on various evaluation metrics. Specifically, the ASCAT backscatter signal is estimated using different regression models, and their performance is assessed based on relative root mean squared error (rRMSE), relative mean absolute error (rMAE), Pearson correlation coefficient, and computational runtime. The goal is to determine to what extent the signal can be reliably predicted from selected input variables and to identify a model that balances accuracy and efficiency.

Secondly, the focus shifts to the analysis of long-term trends in both the observed and modeled data. Using the Mann-Kendall test, monotonic trends are detected in the ASCAT backscatter signal, model predictions, and residuals. These trends are then used to classify spatial patterns and identify potential drift effects. The underlying assumption is that systematic trends in the residuals or inconsistencies between observations and predictions may indicate shifts in environmental conditions, such as land cover change, that are not fully captured by the input variables. This part of the analysis aims to explore whether such trend patterns can provide meaningful insight into spatial and temporal drift phenomena. The following research questions are addressed:

RQ 1: To what extent can ASCAT backscatter signal be estimated with a ML approach, evaluated based on Relative Root Mean Squared Error (rRMSE), Relative Mean Absolute Error (rMAE), Pearson Correlation Coefficient and Runtime?

RQ 2: To what extent can Mann-Kendall Test results for ASCAT backscatter, model predictions, and residuals be used to identify spatial and temporal drift patterns that may reflect long-term land cover change?

1.4 Thesis Outline

This thesis is structured into seven chapters. Chapter 1 provides an introduction to the topic, including the motivation behind the study, the problem statement, and the main research questions. It also outlines the structure of the thesis. Chapter 2 presents the relevant background and theoretical foundations, covering key concepts related to soil moisture, remote sensing, and an overview of related work in the field.

Chapter 3 describes the data used in the study. It introduces the ASCAT backscatter measurements, ERA5-Land soil moisture data, and MODIS LAI products, and specifies the spatial and temporal scope of the analysis. Chapter 4 details the methodology

employed in this work. It includes the preprocessing steps, the process of model selection and evaluation, the description of the machine learning models used, and the approach for detecting drift patterns using the Mann-Kendall test.

Chapter 5 presents the results of the analysis. It begins with a comparison of model performance, followed by an in-depth investigation of temporal trends in both the observed and predicted data. A trend-based classification and segmentation of the grid points is carried out to better understand model behavior and the nature of changes in the backscatter signal.

Chapter 6 provides a summary and discussion of the main findings. It reflects on the implications of the results, outlines limitations of the current study, and suggests directions for future work. Finally, Chapter 7 contains the appendix, which includes additional figures, maps, and extended tables that support the analyses presented in the main chapters.

CHAPTER 2

Background

Currently, ground-based, airborne, and spaceborne remote sensing (RS) methods offer highly effective tools for assessing and tracking large-scale near-surface soil properties and processes with practical temporal and spatial resolutions [BSJ+19].

Numerous physical, chemical, and biological processes occurring at the land surface are heavily impacted by the amount of water stored in the upper soil layers. As a result, various scientific fields rely on soil moisture data to develop, test, and refine their models. Meteorology is one such field, as soil moisture plays a key role in regulating the exchange of heat and moisture between the ground and the lower atmosphere. Accurate soil moisture measurements can, therefore, enhance forecasts of air temperature, humidity, and precipitation [WHK+13].

2.1 Soil Moisture

Soil moisture (SM) is a crucial hydrologic variable with major significance for various Earth and environmental science applications that influence both the global environment and human well-being. Its applications include weather and climate forecasting, drought prediction and monitoring, water resource management, enhancing agricultural productivity and mitigating famine, preventing natural disasters like wildfires, landslides, floods, and dust storms, and tracking ecosystems' responses to climate change [BSJ+19].

Effective monitoring of large-scale soil moisture (SM) is crucial for managing agricultural water resources, as it enhances understanding of plant water availability, water use efficiency, and overall water productivity while helping to mitigate the risks of harmful human-induced environmental impacts. SM also serves as a key indicator for drought and flood risks — two significant meteorological hazards — by contributing to their early prediction. For instance, during droughts, anomalies in soil moisture often precede changes in vegetation indices [BSJ+19].

Soil moisture (SM) has been recognized as one of the 50 most critical climate variables [DGD+15] to assist international organizations in evaluating the impacts of climate change. In 2010, the Global Climate Observing System initiative officially designated SM as a fundamental climate variable [Sec10]. As a result, the European Space Agency (ESA) has integrated soil moisture (SM) into its climate change initiative [HMS+13].

Measuring soil moisture at scales relevant to hydrological studies is challenging. Ground-based measurements are typically taken from individual soil profiles at several locations within a catchment. However, logistical limitations make it difficult to deploy sensors over large areas, and each measurement generally represents only a few centimeters of soil. As a result, deriving accurate spatial averages for entire catchments using in-situ measurements is problematic. Remote sensing offers an effective alternative to overcome these limitations and monitor soil moisture over broader areas [WHK+13].

2.2 Remote Sensing

In recent decades, remote sensing has advanced significantly, with geophysical products expanding beyond traditional land cover maps. Hyperspectral sensors on aircraft and satellites now capture large areas with high spectral, spatial, and temporal detail, enabling the extraction of detailed information from Earth's surface [BDPCV+13].

The term "remote sensing" was introduced in the early 1960s to refer to tools used for observing Earth from space, initially involving traditional aerial photography, with cameras serving as the first sensors. Nowadays, there are thousands of satellites that have been launched into orbit around Earth, although a large part of them are no longer active. Satellites mainly follow two types of orbits: communication and Earth observation. Communication satellites enable global connections for navigation, TV, GPS, radio, and more. Observatory satellites are designed to monitor Earth's surface by capturing observations and images of both marine and terrestrial environments. This process relies on the interaction of energy between the surface objects and the sensor. Remote sensing involves two main types of sensors: passive and active. Passive sensors, such as cameras and electro-optical devices, detect natural electromagnetic radiation emitted by Earth's surface. Active sensors, like radar and lasers, emit energy toward a target and capture the radiation reflected back. Satellite remote sensors do not capture all electromagnetic radiation traveling from Earth. Mainly due to technological and economic constraints, they focus on specific bands of the spectrum [Sha22].

As a result, the potential applications of a remote sensor depend on the number of bands it can detect and their location within the electromagnetic spectrum. Remote sensing supports a wide range of applications, including assessment, monitoring, and exploration. Its uses extend to areas such as water resources, agriculture, natural hazards, marine environments, pollution monitoring, and change detection, among others [Sha22].

Spaceborne sensors offer the advantage of providing integrated soil moisture measurements over large areas, with data often available on a global scale. Unlike ground-based

measurements, they capture broader spatial averages rather than point-specific values [WHK+13].

Microwave Remote Sensing (MRS) offers key advantages over other signal types, such as optical sensors. Microwave signals in the lower microwave frequency range, particularly the L-band (1–2 GHz) and C-band (4–8 GHz), are less affected by precipitation or clouds because their longer wavelengths are less prone to scattering and absorption, allowing sensors to "see through" adverse weather conditions. This capability is especially beneficial for monitoring variables closely linked to rainfall, like soil moisture and floods. Unlike optical sensors, microwave sensors do not rely on sunlight, enabling observations both day and night. They can also penetrate vegetation canopies to a certain extent, further enhancing their usefulness.

However, they come with certain limitations, including relatively large pixel sizes compared to the scale of hydrological processes, low revisit frequencies (typically every few days), and limited penetration depths of microwave signals, usually only a few centimeters or less [WHK+13].

Vegetation significantly affects the radar backscatter coefficient, not only by attenuating the signal but also through various scattering mechanisms. As illustrated in Figure 2.1, the total backscatter from vegetated areas is composed of multiple contributions: (a) surface backscatter from the ground (subject to attenuation), (b) volume scattering within the canopy and (c) interactions between canopy and ground [WHK+13].

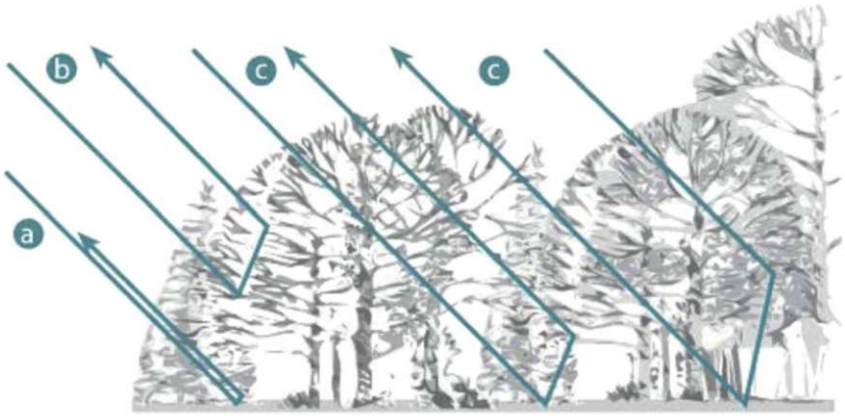


Figure 2.1: Schematic illustration of radar signal interaction with vegetation-covered land surfaces. Source: Wagner, 2013 [WHK+13]

The strength of the radar backscatter signal (σ^0) is strongly influenced by whether the microwave energy is reflected specularly (like a mirror) or diffusely (scattered in many directions) at the Earth's surface. This scattering behavior is primarily determined by the surface roughness relative to the radar wavelength. A commonly used approximation for this relationship is the Rayleigh criterion, which defines a surface as "smooth" if its root-mean-square (RMS) height deviation is smaller than a fraction of the wavelength,

adjusted for the incidence angle:

$$h < \frac{\lambda}{8 \cos \theta} \tag{2.1}$$

Here, h is the RMS surface roughness, λ is the wavelength of the radar signal (e.g., approximately 5.6 cm for C-band systems like ASCAT), and θ is the incidence angle, i.e., the angle between the incoming radar beam and the surface normal.

If this condition is met, the radar signal is reflected specularly, mostly away from the sensor. As a result, the backscatter signal received by the satellite is very low (σ^0 close to zero). This behavior is typical for calm water bodies or homogeneous, smooth surfaces such as dry crusts or ice sheets. In contrast, when the surface has roughness on the order of or greater than the radar wavelength — for example, due to vegetation, soil structure, or urban infrastructure — the scattering becomes diffuse (Figure 2.2). In such cases, a larger portion of the energy is returned to the sensor, resulting in higher backscatter values [LKC04].

Understanding this physical relationship is crucial for interpreting active microwave data correctly. A low backscatter value does not necessarily indicate dry conditions, but may instead result from specular reflection on a smooth surface. Conversely, high backscatter values are typically associated with rougher or more complex surfaces, which may also correlate with higher soil moisture content or vegetation density.

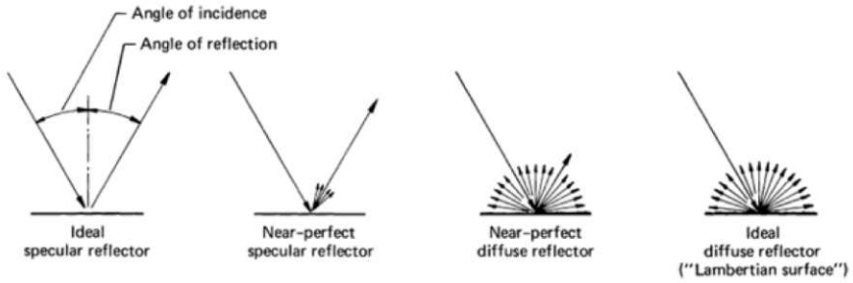


Figure 2.2: Specular versus diffuse reflectance. Source: Lillesand, Kiefer, Chipman, 2004 [LKC04]

2.3 State of the Art / Related Work

The ASCAT research community is not very big and especially the field of analyzing long-term land cover changes effects on ASCAT backscatter observations is relatively new.

ASCAT backscatter has already been investigated with multiple machine leaning algorithms. In 2022, Xu Shan et al. [SSDH+22] investigated the relationships between ASCAT soil moisture retrievals and the land surface variables (LSVs) using a Deep Neural Network. A specified "Normalised Senstivity Coefficient" indicates the sensitivity

of ASCAT observables to changes in LSVs. In this work it was found that there is a sensitivity between ASCAT backscatter/slope and specific LSVs respectively. These sensitivities vary strongly with time and are massively dependent on the land cover type.

Based on only semi-arid regions, Xulong Duan et al. [DMK+24] applied several machine leaning algorithms to estimate ASCAT backscatter: Random Forest, Support Vector Machine, Artificial Neural Network and Elastic Net Regression. Among other findings, land-use was found to be a highly influential variable and essential to explain regional variations in soil moisture content. The estimated forecasts and the measured soil moisture values show a high degree of agreement in all 4 models, with RF delivering the best results in the testing scenario. Nevertheless, it is explicitly mentioned that surface soil moisture is influenced by other spatial factors and that these are essential for a complete understanding of surface soil moisture.

One spatial factor with a high influence is subsurface scattering. Until now, it was assumed that the soil backscatter increases monotonically with increasing soil moisture content. In fact, however, it has now been found that under certain soil conditions (dry soil and near-surface rocks/stones) the exact opposite happens. Wagner et al. [WLM+22] have investigated this phenomenon. In addition to other experiments, a subsurface scattering term was integrated into the model and then both models (with and without the additional term) were compared. This made it possible to determine which spatial conditions favour the occurrence of subsurface scattering and in which regions of the study area these anomalous signals occur.

Wagner et al. [WLH+23] extended the mapping of subsurface scattering effects to a global scale. Furthermore, the influence on ASCAT soil moisture retrievals was investigated and the correlation with land surface characteristics was examined. The regional differences in the occurrence of subsurface scattering and the most driving factors were identified. Subsurface scattering represents the most significant of the unconsidered errors in the existing version of the ASCAT soil moisture data and a lack of knowledge in the past led to far-reaching misinterpretations.

CHAPTER 3

Data

This chapter provides an overview of the datasets used in the analysis. It introduces the data sources for the target and input variables, including ASCAT backscatter, ERA5-Land soil moisture, and MODIS LAI. In addition, the spatial and temporal scope of the study is defined by describing the selected study area and the observation period.

3.1 ASCAT backscatter

The Advanced Scatterometer (ASCAT) is a real-aperture radar system that operates in the C-band with VV polarization, offering both day and night measurement capabilities while remaining largely unaffected by cloud cover. It measures the Normalized Radar Cross Section (NRCS), commonly known as the backscatter coefficient σ^0 , expressed in either m^2m^{-2} or decibels (dB), from the Earth's surface. The instrument's design and performance specifications are based on the experience gained from the scatterometers aboard the ERS-1 (1991-2000) and ERS-2 (1995-2011) satellites [GSW00].

ASCAT is hosted on a series of three polar-orbiting Meteorological Operational (MetOp) satellites operated by EUMETSAT. The first satellite, MetOp-A, was launched in October 2006 and retired in November 2021 after more than 15 years of service. MetOp-B, launched in September 2012, continues to operate as a backup platform. MetOp-C, launched in November 2018, currently serves as the primary operational satellite, ensuring the continuous delivery of ASCAT backscatter data. Together, these satellites have provided long-term, consistent measurements that are critical for numerical weather prediction, ocean monitoring, and climate research [WLH+23]. METOP operates in a near-polar, sun-synchronous orbit at an altitude of approximately 817 km, with a 29-day repeat cycle. In this orbit, the satellite circles the Earth in about 100 minutes, completing 14 orbits per day. The equator crossing occurs at 9:30 during the descending pass and at 21:30 during the ascending pass [WHK+13].

ASCAT's advanced measurement geometry provides twice the coverage compared to its predecessors on ERS-1 and ERS-2, increasing daily global coverage from 41% to 82%. Additionally, the spatial resolution was enhanced from 50 km to 25-34 km, all while maintaining the same level of radiometric accuracy as the ERS-1/2 scatterometers. ASCAT achieves improved global coverage with two sets of three fan-beam antennas, compared to one set on the ERS satellites. Positioned broadside and at ±45 degrees, the antennas provide observations in three azimuthal directions within two 550 km swaths, separated by about 360 km. The multi-angle viewing capability, with incidence angles ranging from 25° to 64.5°, is crucial for accurately retrieving geophysical variables like ocean wind speed and land vegetation conditions [FSWA+02].

The Advanced Scatterometer (ASCAT) was originally designed to measure wind speed and direction over oceans, with no initial intention of monitoring soil moisture on land. However, its key features — such as its 5.7 cm wavelength, high radiometric accuracy, and multi-angle observation capabilities — make it a valuable sensor for soil moisture measurement [WHK+13].

In 2008, soil moisture data from Metop ASCAT became the first operational soil moisture product. The European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) manages the ASCAT soil moisture data service. EUMETSAT provides a range of soil moisture products, all based on the same physical principles but varying in aspects such as spatio-temporal resolution, data latency (the time between data acquisition and public availability), data format, and consistency [WHK+13] [ESDS21]. Many of these products are developed and distributed through the Satellite Application Facility on Support to Operational Hydrology and Water Management (H SAF), a specialized program within EUMETSAT focused on satellite-derived hydrological parameters such as precipitation, snow cover, and soil moisture [noa].

Challenges faced by the ASCAT soil moisture data service include quality flagging, instrument calibration, algorithmic improvements and application development.

The irregular spatial coverage of the ASCAT constellation results in highly variable temporal coverage. In some cases, two or more acquisitions may be recorded over a region in a single day, especially near the poles, while on other days, no data may be collected. This poses a significant challenge for using ASCAT soil moisture data, as applications must either handle the irregular coverage or rely on interpolated measurements, which come with increased uncertainty [WHK+13].

The strong link between backscatter intensity and soil moisture allows ASCAT σ^0 measurements to serve as a reliable indicator of soil moisture in bare soil areas. However, when vegetation is present, this relationship weakens, as vegetation reduces the sensitivity of σ^0 to moisture changes. Properly accounting for the impact of vegetation is essential for accurate soil moisture estimation [WHK+13].

In this study, the ASCAT backscatter at 40 ° incidence angle was used, as this is the standard reference angle employed in the EUMETSAT H SAF processing chain to retrieve surface soil moisture from ASCAT data [WHK+13]. Throughout this work, the term

backscatter refers exclusively to the backscatter at 40 °, which is labeled as backscatter40 in the original dataset.

Details on the temporal coverage of the dataset are provided in the Study Area & Period chapter. Data preprocessing procedures are described in the Methodology chapter, preceding their use in the subsequent analysis.

3.2 ERA5-Land soil moisture

ERA5-Land is a global dataset focused on the land component of the fifth-generation European ReAnalysis (ERA5), produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5-Land consistently tracks the evolution of water and energy cycles over land throughout its production period, making it useful for analyzing trends and anomalies [MSDAP+21].

It integrates model data with observations to track the evolution of water and energy cycles over land, using meteorological inputs from the ERA5 climate reanalysis and the Carbon Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land (CHTESSEL) model. ERA5-Land offers a significant advantage with its enhanced global horizontal resolution of 9 km, compared to 31 km in ERA5 and 80 km in ERA-Interim, while keeping the same hourly temporal resolution as ERA5 [WLH+23].

Another important distinction is that, unlike ERA5, ERA5-Land does not assimilate ASCAT soil moisture data or other land surface observations. This ensures that ERA5-Land soil moisture data and ASCAT backscatter measurements can be considered independent variables [MSDAP+21].

The modeled soil moisture data from ERA5-Land's volumetric soil water layer 1 (0–7 cm) were resampled to match ASCAT's 12.5 km global grid and temporally aligned with ASCAT measurement timestamps, using hourly data from 2007 to 2021.

Throughout this work, soil moisture is frequently abbreviated as SM. This refers specifically to the previously described modeled surface soil moisture variable provided by the ERA5-Land dataset, where it is labeled as swvl1_era5land in the original data. Unless stated otherwise, all mentions of SM in the following sections refer to this variable.

3.3 Modis LAI

Large parts of the world are covered by vegetation, such as forests, grasslands, shrublands, and agricultural areas, with plant growth and decay cycles. All of that is influenced by factors like temperature, precipitation, wildfires, and human activity. The Leaf Area Index (LAI) is a key dataset used to monitor vegetation status globally. Alongside related datasets — such as vegetation indices, seasonal growth and productivity parameters — the vegetation property index LAI helps assess plant responses to disturbances like droughts, storms, and wildfires, supporting applications ranging from agricultural policy reporting to climate change mitigation.

This thesis utilizes the MOD15A2H Version 6.1 Moderate Resolution Imaging Spectroradiometer (MODIS) which is combining Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation (FPAR) product. It is an 8-day composite dataset with 500 meter (m) pixel size. Several adjustments of calibration leads to the improved Version 6.1 Level-1B product. In concrete terms, LAI can be described as the green leaf area per unit ground area in broadleaf canopies, whereas only one side of the leaf is counted [MKP21].

MODIS delivers key data for vegetation monitoring and analyzing climate change. However, global LAI/FPAR products have drawbacks, such as spatial and temporal inconsistencies and accuracy issues [PYR+24].

3.4 Study Area and Period

The study was conducted in southeastern Australia, covering parts of the states of Victoria, South Australia, and New South Wales. The spatial domain is defined by the latitude range from 39 ° S to 30 ° S and the longitude range from 138 ° E to 155 ° E. This region is characterized by diverse land surface conditions, including forests, croplands, grasslands, urban areas, and both dry and humid zones. Such heterogeneity makes it particularly suitable for analyzing backscatter behavior under varying environmental conditions.

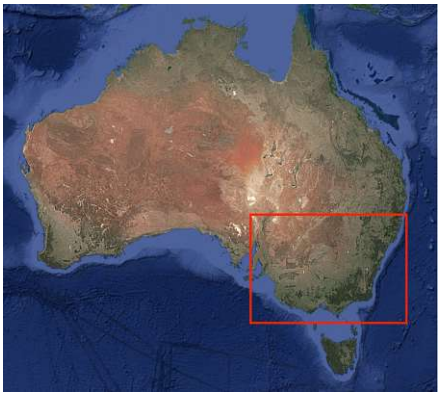
Figure 3.1 illustrates both the geographic extent of the study area and the spatial sampling. The left panel shows the location of the study region within southeastern Australia, while the right panel visualizes the randomly selected grid points used for analysis. From all available grid points within the defined boundaries, a random sample of 33% was selected to ensure computational feasibility, resulting in a total of 2283 grid points.

The selection of this area was further motivated by the availability of high-quality auxiliary data from national and regional authorities. Publicly accessible datasets provide detailed information on surface properties, weather events, and land use, which support a more comprehensive interpretation of remote sensing signals.

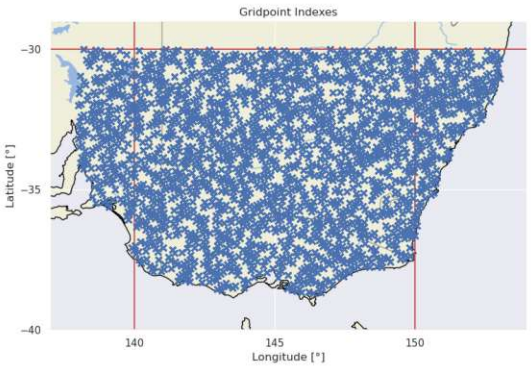
The observation period extends from January 2007 to December 2022, allowing for the analysis of long-term developments and potential drift effects over a span of 16 years.

Figure 3.2 shows the land cover types in the study area, and Figure 3.3 provides the full classification legend. Not all categories are present in the selected region. The area includes a wide variety of surface conditions, such as forested, agricultural, and urban zones, which are relevant for interpreting spatial variation in backscatter behavior.

The study area primarily consists of shrubland, cropland, regions with tree cover, and areas of sparse vegetation. Additionally, there are some urban areas in and around major cities such as Sydney, Melbourne, Canberra, Newcastle, and Adelaide. The dataset also includes regions classified as grassland, irrigated cropland or post-flooded areas, and



(a) Location of the study area in south-eastern Australia.



(b) Selected grid points for analysis.

Figure 3.1: Overview of the study area and spatial sampling.

areas where scrub or herbaceous land has been subject to flooding. These are the main types observed, although there may be additional land cover types present in the study area.

Land cover is known to strongly influence microwave backscatter through its impact on surface roughness, vegetation structure, and soil moisture retention. Therefore, the spatial distribution of land cover classes provides valuable context for later analyses of model performance and trend behavior.

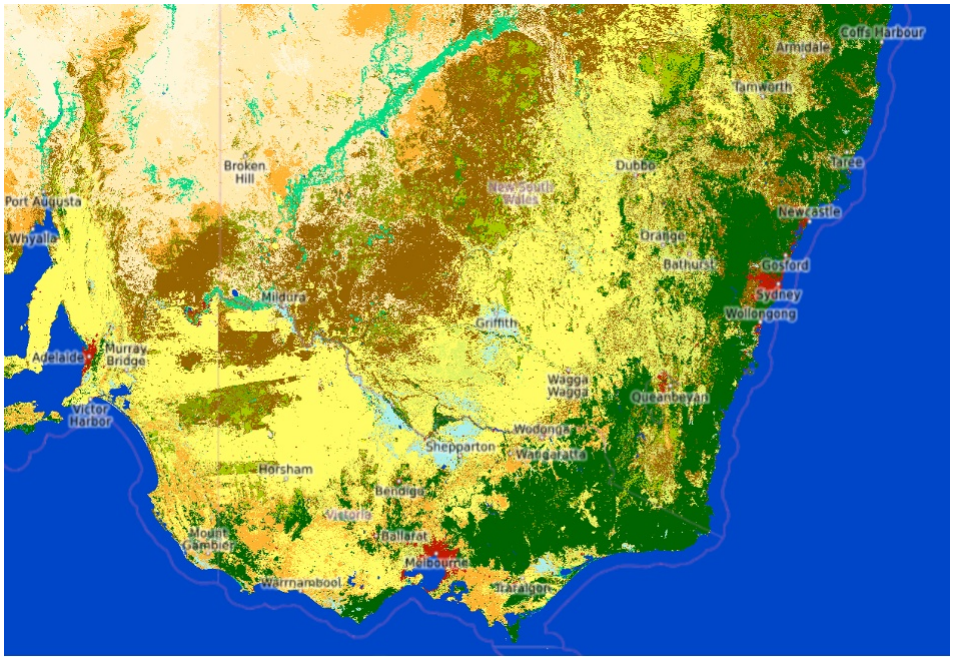


Figure 3.2: Land cover types in the study area.

Legend of the global CCI-LC maps, based on LCCS

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

Figure 3.3: Legend for the land cover classification shown in Figure 3.2.

CHAPTER 4

Methodology

This chapter presents the methodological framework of the study, which follows a mixed research approach that combines model development with the investigation of temporal drift in model performance and data. The methodology is structured into two main components: the first focuses on model selection and evaluation, while the second addresses drift detection. The overall approach combines quantitative and qualitative methods. In the Model Selection part, machine learning models are trained and evaluated using established performance metrics. In the Drift Detection part, statistical tests are used to identify temporal trends, while a wide range of visual analyses supports the qualitative interpretation of spatial and temporal model behavior. This combination allows for a comprehensive understanding of how model performance varies across locations and changes over time.

4.1 Model Selection

This chapter outlines the methodological approach taken to develop and evaluate machine learning models to predict ASCAT backscatter. It consists of three main sections: The first section (Workflow) provides an overview of the entire modeling pipeline, including data gathering, preprocessing, model selection, hyperparameter tuning, train-test splitting, evaluation procedures, and the tools and libraries used. The second section (ML Models) presents a detailed description of the individual machine learning models employed in this study. The third section (Evaluation Metrics) explains the performance metrics used to assess model quality. Together, these sections form the basis for the spatially explicit modeling approach applied in this work.

4.1.1 Workflow

Data gathering The analysis is based on three primary datasets: ASCAT backscatter, surface soil moisture (SM) from ERA5-Land, and Leaf Area Index (LAI) from MODIS.

All datasets were provided by the Department of Geodesy and Geoinformation (GEO) at TU Wien.

Data preprocessing Entries containing missing values were entirely removed from the dataset, as the data coverage was sufficiently high to justify this approach. Additionally, all observations over snow-covered areas or where soil temperatures were below freezing were excluded, since soil moisture values are considered unreliable under such conditions. To apply this filtering, snow depth and soil temperature at level 1 (0–7 cm) from ERA5 were used. Snow depth represents the depth of water that would result if the snow in a grid cell were melted and uniformly distributed; it is expressed in meters of water equivalent.

ML model The candidate models considered for this study include Random Forest (RF), Elastic Net (EN), and Support Vector Regression (SVR). A detailed description of these machine learning models is provided in chapter 4.1.2 "Machine Learning Models". A separate model was trained for each grid point rather than applying a single model across the entire study area. The primary reason for this approach lies in the spatially explicit nature of the analysis: the goal is not only to generate accurate estimates of backscatter but also to assess where reliable predictions are possible and where they are not. Capturing this spatial variability requires localized models. Additionally, substantial heterogeneity exists between grid points with respect to the statistical properties of the target variable. Although ASCAT backscatter typically ranges between approximately –22 and –4 dB, the mean backscatter varies notably across locations, ranging from –18 to –7 dB, depending on land cover, topography, and other local conditions. Furthermore, the relationship between the input variables and the target variable differs significantly across space. The correlation with Leaf Area Index (LAI) is generally moderate, while soil moisture (SM) tends to show medium to high correlation with backscatter. However, the strength of these correlations varies markedly from one grid point to another, as indicated by spatially heterogeneous Pearson correlation coefficients. Taken together, these factors justify the use of individual models per grid point, both to account for local differences in data relationships and to fulfill the overarching goal of spatially explicit model evaluation.

Hyperparameter tuning Hyperparameter tuning and selection were performed on a representative subset of 20 grid points. For each model, relevant hyperparameters (e.g., alpha, l1_ratio, n_estimators, etc.) were optimized individually using RandomizedSearchCV. A detailed description of the hyperparameters is provided in the ML Models chapter. To ensure consistency and reduce overfitting to individual locations, the final hyperparameter values used for all grid points were set to the median of the best-performing values across the 20 selected locations.

Table 4.1 summarizes the final hyperparameter values used in the evaluation. The specific meaning and function of each parameter are explained in the corresponding sections of the ML Models chapter.

Table 4.1: Final hyperparameter values used for each regression model.

Model	Hyperparameter	Value
Random Forest	`n_estimators`	129
	`max_depth`	12
	`min_samples_split`	4
	`min_samples_leaf`	2
Elastic Net	`alpha`	0.069621
	`l1_ratio`	0.510747
Support Vector Regression	`C`	1.793158
	`epsilon`	0.231559
	`kernel`	linear
	`gamma`	scale

Train-test split For model training and evaluation, the dataset was split chronologically. Data from the years 2007 to 2014 were used for training, while the period from 2015 to 2022 was reserved for testing. This temporal split was chosen to reflect a realistic forecasting scenario, ensuring that model performance is assessed on future, previously unseen data. It also allows for evaluating the models' ability to generalize over longer time periods and under potentially changing environmental conditions.

Evaluation Model evaluation was conducted based on a subset of 100 grid points, selected to represent the spatial and temporal variability of the dataset. For each model, performance was assessed using several metrics: relative root mean squared error (rRMSE), relative mean absolute error (rMAE), Pearson correlation coefficient (PCC), and runtime. Relative metrics were used in particular due to the substantial differences in scale across locations, ensuring a fair comparison of model performance. The final model selection was based on the average performance across all grid points, considering all metrics. A detailed description and interpretation of each metric is provided in the dedicated Evaluation Metrics chapter.

Tools and Libraries All analyses were conducted in Python 3.11.10 [VRD09] using JupyterLab version 4.2.5 [jup] as the working environment. A wide range of packages supported data handling, visualization, statistical analysis, and machine learning. Data processing and manipulation were primarily performed using Pandas [M+10] and NumPy [HMvdW+20]. Visualization and mapping relied on Matplotlib [Hun07], Seaborn [WBO+17], eomaps [Qua25], and cartopy [noa25]. Statistical computations were conducted using SciPy [VGO+20] and statsmodels [SP10]. The implemented machine learning models included Elastic Net, Random Forest, and Support Vector Regression (SVR), all built using scikit-learn [PVG+11]. Associated tools from scikit-learn were used for preprocessing (e.g., scaling), performance evaluation, and hyperparameter tuning (e.g., RandomizedSearchCV). The overall workflow was supported by standard Python libraries such as os [pyt], datetime [pyt], and math [VR20].

4.1.2 Machine Learning Models

This section introduces the three machine learning models used in this study: Random Forest, Elastic Net, and Support Vector Regression. Each of these models offers specific advantages and challenges, making them valuable for analyzing the ASCAT backscatter data in different ways.

Random Forest (RF)

If you were to ask a complex question to a large, randomly selected group of people and then combine their responses, the overall answer would often be more accurate than that of a single expert [Gé17]. Similarly, in machine learning, combining the predictions of multiple models — such as classifiers or regressors — typically leads to improved accuracy compared to relying on just one model. A collection of multiple predictors is referred to as an ensemble, and this approach is known as Ensemble Learning. For instance, one can train multiple Decision Tree classifiers on different randomly selected subsets of the training data. This collection of Decision Trees forms what is known as a Random Forest. Despite its straightforward design, it is considered one of the most effective machine learning algorithms available today [Gé17].

The core concept of Random Forests is to enhance the variance reduction achieved by bagging while minimizing the correlation between individual trees, without significantly increasing variance. This is accomplished during the tree-building process by randomly selecting a subset of input variables at each split [HTF09a].

In recent years, "ensemble learning" has gained significant attention. This approach involves generating multiple classifiers and combining their predictions to improve overall performance. Two widely recognized ensemble methods are boosting ([BFLS98]) and bagging ([Bre96]), both applied to classification and regression trees. Boosting adjusts the weighting of data points, giving more emphasis to those that were misclassified by previous models, and final predictions are made through a weighted voting process. In contrast, bagging constructs trees independently, without relying on the performance of earlier models [LW01].

Bagging, or bootstrap aggregation, is a method used to reduce the variance of a predictive model. This technique is particularly effective for high-variance, low-bias models, such as decision trees. In regression tasks, multiple regression trees are trained on different bootstrap samples of the training data, and their predictions are averaged. For classification, multiple trees are trained in the same manner, and the final prediction is determined by majority voting. The key idea behind bagging is to combine several noisy but mostly unbiased models, thereby lowering overall variance. Decision trees are well-suited for this approach because they can capture complex interactions within the data. When grown deep enough, they tend to have low bias but high variance, making them highly responsive to averaging, which significantly improves stability and performance [HTF09a].

Breimann (2001) [Bre01] introduced Random Forests, which build upon the bagging

technique by introducing an additional layer of randomness. While bagging constructs each tree using a different bootstrap sample of the data, Random Forests also modify the way trees are built. Unlike traditional decision trees, where each node is split using the best possible feature across all variables, Random Forests select a random subset of predictors at each split and choose the best split from that subset. Although this approach may seem counterintuitive, it has been shown to enhance performance and often outperforms other classifiers, such as discriminant analysis and neural networks [Bre01] [LW01].

Algorithm 4.1: Random Forest algorithm for regression or classification (adapted from Hastie et al. (2009))

Data: Training data with N observations and p variables
Result: Ensemble of random forest trees $\{T_b\}_1^B$

for $b = 1$ **to** B **do**

1. Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
2. Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node, until a minimum node size $n_{\min}$ is reached:
   a) Select m variables at random from the p variables.
   b) Pick the best variable/split-point among the m.
   c) Split the node into two daughter nodes.

end

Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

- **Regression:**
$$\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$
- **Classification:** Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then
$$\hat{C}_{\text{rf}}^B(x) = \text{majority vote}\left\{\hat{C}_b(x)\right\}_1^B.$$

The algorithm for Random Forest for regression and classification in Algorithm 4.1 is defined and written by Hastie et al. ([HTF09a]).

For many tasks, Random Forests achieve performance comparable to boosting while being easier to train and optimize. This simplicity makes them highly popular, and they are widely available in numerous machine learning libraries [HTF09a].

In this thesis, I am using Python and the scikit-learn package [PVG+11] to implement

and analyze Random Forest models.

Random Forests in Python rely on three key user-defined parameters: the number of trees in the forest (n_estimators), the minimum number of data points required in each terminal node (min_samples_leaf), and the number of features randomly selected for consideration at each split (max_features). In scikit-learn, the default values for these parameters are n_estimators=100, min_samples_leaf=1, and max_features="n_features". Additionally, another important parameter is min_samples_split, which controls the minimum number of samples required to split an internal node and has a default value of 2.

[HTF09a] state that, for regression, the default value for max_features is p/3 (where p is the total number of features), and the default min_samples_leaf is 5. In practice, the optimal values for these parameters depend on the specific problem and should be considered as hyperparameter to be tuned.

The importance of a variable can be assessed by calculating the difference in misclassification rates between the modified and original out-of-bag data. This difference is then divided by the standard error, providing a standardized measure of the variable's contribution to the model's performance [CEB+07].

Random Forests are highly effective for prediction and naturally resist overfitting due to the Law of Large Numbers. Each tree is trained on a different bootstrap sample of the data ([GBME08]). Introducing controlled randomness improves their accuracy as both classifiers and regressors. Moreover, the balance between the strength of individual predictors and their correlation helps explain the model's predictive performance. The use of out-of-bag (OOB) estimation provides a practical way to measure these theoretical properties [Bre01].

The ability to model high-dimensional non-linear relationships, the inclusion of an unbiased error rate estimation, built-in measures of variable importance, and the need for only a few user-defined parameters make Random Forests a robust and practical machine learning method [GBME08].

Random Forests offer several advantages over other statistical modeling approaches [Bre01] [LW01]. Unlike models such as linear regression or support vector machines, they can handle both continuous and categorical variables. Additionally, they are highly robust to noise in predictor variables, eliminating the need for extensive pre-selection of features [DUADA06]. Since only a randomly selected subset of predictors is considered at each split, the diversity among trees increases (reducing correlation) while also lowering computational cost.

Pruning is unnecessary, as deep trees naturally result in low bias and high variance while maintaining computational efficiency. By averaging a large number of trees, Random Forests achieve both low bias and low variance [DUADA06]. The algorithm's resistant to overfitting is particularly important in this study case since the model is applied to all individual grid points separately. Additionally, Random Forests require only a few user-defined parameters to achieve strong performance and are well-suited for large datasets.

They are also effective when the number of variables exceeds the number of observations and can naturally capture interactions among predictor variables [DUADA06] [Gé17].

Random Forest provides reliable error estimates through Out-Of-Bag (OOB) data, which consists of the portion of the dataset not included in the bootstrap samples — typically about one-third of the data. Since these excluded samples serve as an internal validation set, Random Forest eliminates the need for a separate validation dataset.

A key drawback of Random Forest and ensemble methods, in general, is their limited interpretability. Unlike linear models or single decision trees, where relationships between predictors and the response variable are easy to analyze, Random Forest is often considered a "black box" approach [Gé17]. However, compared to deep learning models, it is not entirely intransparent, as individual trees and feature importance measures can still be examined to gain insights into the model's behavior.

Elastic Net (EN)

The problem of overfitting is quite common when working with machine learning models. The model performs very well, sometimes surprisingly well on the training data but behaves poorly when it comes to build actual predictions using unseen data. A good way to reduce overfitting and therefore improve the prediction result is to use regularize the model. Regularization is a commonly used approach to improve the generalization ability of a machine learning model by adding a penalty term to the loss function [Gé17]. Improved generalization is not the only reason for the development of different regularization methods such as Elastic net which works as a combination of Ridge or Lasso regression. In order to understand Elastic Net and its development, it is important to first look at Ridge and Lasso.

Ridge regression Ridge regression, introduced by Hoerl and Kennard (1970) [Ha70] [HK70], is a regularization method that improves prediction accuracy by applying continuous shrinkage to the regression coefficients. This is achieved through an L2 penalty term added to the loss function, which constrains the overall magnitude of the coefficients and reduces their variance. The degree of penalization is controlled by a tuning parameter, which balances model fit and complexity. Although ridge regression leads to improved generalization via the bias-variance trade-off, it does not yield a sparse or parsimonious model, as it retains all predictors [Bre96].

$$\hat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad \text{, subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq t. \tag{4.1}$$

The ridge uses the ℓ_2 norm which is defined as $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ [JWHT13].

Ridge regression was created to deal with multicollinearity, which happens when predictor variables are highly correlated. In such cases, ordinary least squares (OLS) can give

unstable and overly large coefficient estimates, which can hurt prediction accuracy [BG09]. Ridge regression solves this by shrinking the coefficients. It also has a "grouping effect," meaning that strongly related predictors get similar coefficients. However, it does not remove any variables from the model — all predictors stay in [ZH05].

Lasso Tibshirani (1996) [Tib96] introduced a method known as the lasso (Least absolute shrinkage and selection operator) which enhances least squares regression by applying an L1 penalty to the coefficients. This penalty not only shrinks the coefficients but also performs variable selection by setting some of them exactly to zero. Later studies by Tibshirani (1996) [Tib96] and Fu (1998) [Fu98] compared the predictive abilities of the lasso and ridge regression, showing that neither method consistently outperforms the other. Nonetheless, because it produces simpler models by excluding irrelevant variables, the lasso has become particularly attractive in modern data analysis where variable selection plays a key role [ZH05].

The lasso minimizes the residual sum of squares under the constraint of ℓ_1 penalization. The ℓ_1 norm is defined as $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j$ [JWHT13].

$$\hat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \quad \text{, subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t. \tag{4.2}$$

If the penalization parameter is well chosen, the predictive accuracy of the estimator improves in comparison to the least squares estimator, just as with ridge regression.

Like ridge regression, the lasso reduces the size of the coefficient estimates, pulling them toward zero. However, due to the ℓ_1 penalty, the lasso can shrink some coefficients all the way to zero when the regularization parameter λ is large enough. This means it effectively selects variables, similar to best subset selection. As a result, lasso models are typically easier to interpret than those from ridge regression, since they include only a portion of the original predictors. Such models are referred to as sparse, meaning they rely on a limited number of variables [JWHT13].

Although the lasso offers many benefits, its limitations should also be acknowledged. When the number of predictors exceeds the number of observations (p > n), it can select at most n variables, which restricts its ability to perform variable selection. Additionally, when predictors are grouped or highly correlated, the lasso often picks just one variable from the group and ignores the rest [ZH05]. Unlike ridge regression, it does not exhibit a grouping effect. While this can lead to sparse models — as previously noted — it may also result in arbitrary choices among correlated predictors.

Naive Elastic Net A combination of the ridge and lasso method is the Naive Elastic Net, proposed by Zou and Hastie (2005) [ZH05]. It combines the advantages of ridge and lasso. This estimation method accomplishes unrestricted variable selection and has the grouping effect. Simulation studies and real data examples show that the elastic net often outperforms the lasso in terms of prediction accuracy.

The Naive Elastic Net uses both the ℓ_1 penalization of the lasso and the ℓ_2 penalization of the ridge. Zou & Hastie [ZH05] describe the Naive Elastic Net as follows:

Let the response be $y = (y_1, ..., y_n)^T$, $X = [x_1]...[x_p]$ is the model matrix with $x_j = (x_{1j}, ..., x_{nj})^T$ and $j = 1, ..., p$ are the predictions.

For any fixed non-negative λ_1 and λ_2, we define the naive elastic net criterion

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{4.3}$$

where

$$\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2 \quad \text{and} \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|. \tag{4.4}$$

The naive elastic net estimator $\hat{\beta}$ is the minimizer of the naive elastic net criterion (4.1):

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2, \quad \text{subject to } (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t \text{ for some } t. \tag{4.5}$$

The elastic net penalty can be expressed as:

$$(1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \tag{4.6}$$

where $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$

When $\alpha = 1$, the naive elastic net estimator is equivalent to ridge estimator and when $\alpha = 0$, its equivalent to the lasso estimator.

Elastic Net Although this naive elastic net criterion (4.3) is a natural combination of lasso and ridge regression, it has certain disadvantages: in particular, double shrinkage.

The naive elastic net estimator is a two-stage procedure: for each fixed λ_2 we first find the ridge regression coefficients, and then we do the lasso type shrinkage along the lasso coefficient solution paths.

By simultaneously penalizing with L1 and L2 norms, the coefficients are reduced more than necessary, which can lead to a loss of prediction accuracy. By addressing this double shrinkage, the predictive performance of the naive elastic net can be enhanced.

Zou and Hastie [ZH05] show that the naive elastic net problem can be transformed to an equivalent lasso problem on augmented data. However, the details of this transformation are not further discussed in this thesis, as the main focus lies on the final formulation of the elastic net estimator.

There is one important property: The reformulations of the naive elastic net can potentially select all p predictors in all situations (because of $dim(X^*) = (n + p)xp$), in contrast to lasso. It can also perform an automatic variable selection in a fashion similar

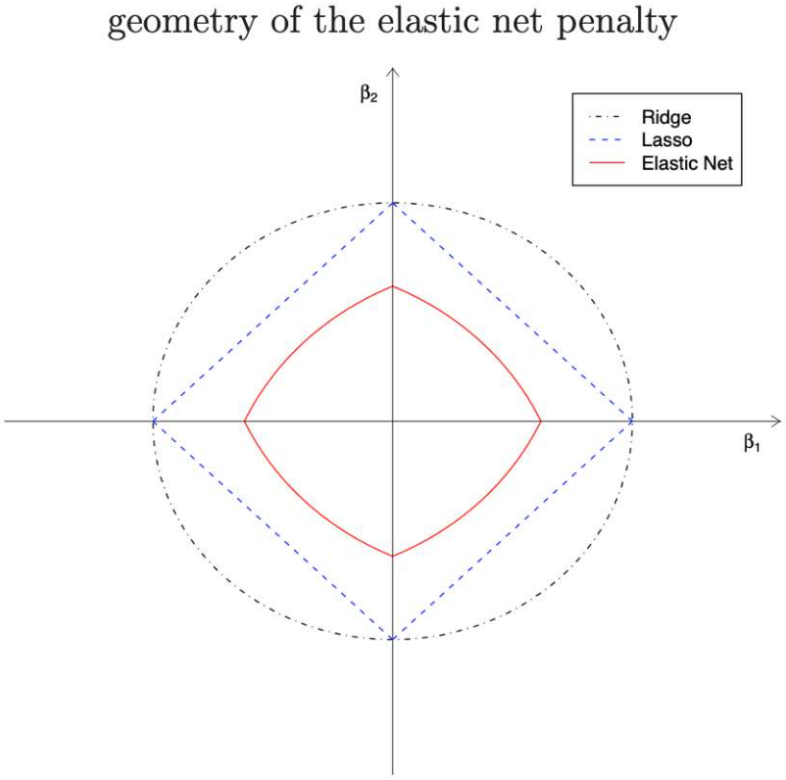


Figure 4.1: Contour plot of ridge, lasso and elastic net penalty where $\alpha = 0.5$ (Zou und Hastie, 2005) [ZH05]

to the lasso. Furthermore the naive elastic net has the ability of selecting "grouped" variables, a property not shared by the lasso.

The elastic net estimator can be expressed in terms of the naive elastic net estimator as:

$$\hat{\beta}_{\text{elastic net}} = (1 + \lambda_2)\hat{\beta}_{\text{naive elastic net}} \tag{4.7}$$

The elastic net penalty introducted by Zou and Hastie (2005) [ZH05]:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha)|\beta_j| \right), \tag{4.8}$$

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda \left((1 - \alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \right) \right\} \tag{4.9}$$

This method combines the advantages of Lasso (sparsity) and Ridge (group effect with correlated predictors) without completely adopting their respective disadvantages.

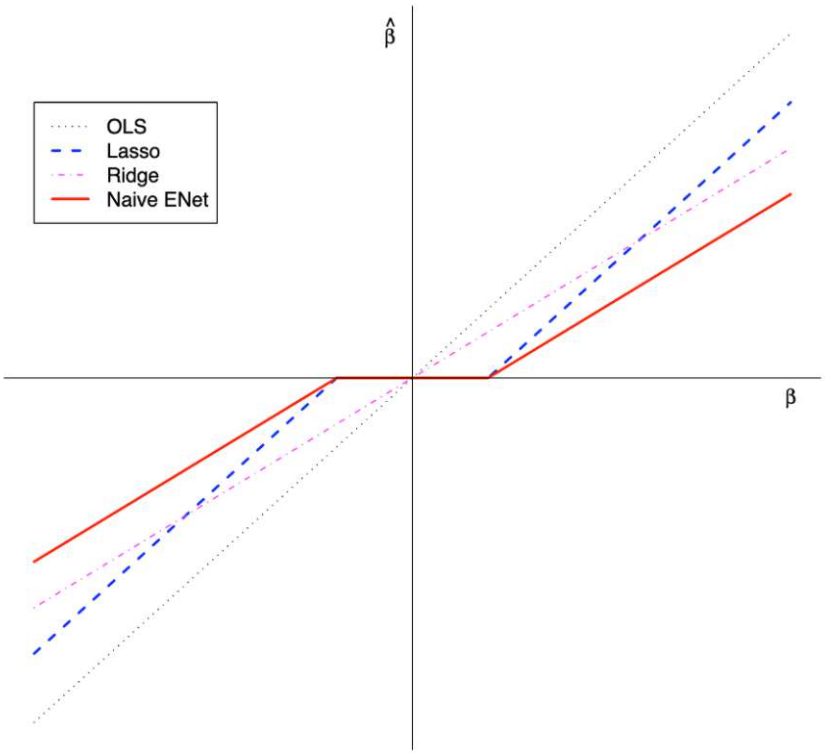


Figure 4.2: Ridge, lasso und elastic net solution with $\lambda_1 = 2$, $\lambda_2 = 1$ (Zou und Hastie, 2005) [ZH05]

According to empirical results by Zou and Hastie [ZH05], the elastic net performs strongly in comparison to both the lasso and ridge regression.

Figure 4.1 shows the Ridge's penalization function is a circle, strictly convex and smooth, whereas Lasso's is only convex and non-differentiable at the intersections of the axes. The naive elastic net lies between the two: It is strictly convex, but also non-differentiable at the axis intersections.

The strict convexity of ridge and elastic net leads to the grouping effect, i.e. similar coefficients for strongly correlated variables. The non-differentiability of Lasso and Elastic Net enables variable selection.

Figure 4.2 illustrates how the three penalization methods behave in an orthogonal design. In this setting, the two-stage approach of the naive elastic net is visible.

The Elastic Net model includes two important tuning parameters: λ and α. The parameter λ determines the overall strength of the regularization, controlling how strongly the regression coefficients are penalized. Larger values of λ produce more significant shrinkage, while smaller values allow for more flexibility in the model. The parameter α

determines the balance between the L1 (lasso) and L2 (ridge) penalties. When $\alpha = 1$, Elastic Net reduces to ridge regression; when $\alpha = 0$, it reduces to lasso regression. Intermediate values of α combine the strengths of both methods, producing sparse models with some grouping effect. In this work, the implementation of the Elastic Net uses the `ElasticNet()` function from scikit-learn [PVG+11]. In this implementation, the parameter `alpha` corresponds to the overall regularization strength λ, while `l1_ratio` represents the balance between L1 and L2 penalties (equivalent to α in the original theoretical formulation). These tuning parameters were optimized using cross-validation on a representative subset of grid points to ensure consistent model behavior and robust trend detection across the study area.

Support Vector Regression (SVR)

Support Vector Machines (SVM) were originally introduced for classification tasks [Vap95], but have since been extended to regression problems under the framework of Support Vector Regression (SVR). Instead of minimizing the squared error as in ordinary least squares regression, SVR focuses on finding a function that deviates from the true target values by no more than a specified margin, denoted by ε. The goal is to balance model complexity and predictive accuracy while maintaining robustness to outliers and overfitting [SS04].

The core idea behind SVR is the so-called ε-insensitive loss function. It defines a margin of tolerance around the predicted values. Errors within this ε-tube are not penalized, and only deviations beyond this margin contribute to the optimization objective. The model searches for a function $f(x) = \langle w, x \rangle + b$ that is as flat as possible, meaning that the norm of the weight vector $\|w\|$ is minimized.

This leads to the following convex optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to:

$$\begin{aligned} y_i - \langle w, x_i \rangle - b &\leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

Here, C is a regularization parameter, which controls the trade-off between the flatness of the function and the extent to which deviations larger than ε are tolerated. (Slack variables ξ_i and ξ_i^* allow the model to handle data points outside the ε-margin.) A large value of C implies strong penalization of errors outside the margin, potentially leading to a model that closely fits the training data, whereas a smaller value allows more flexibility and smoother predictions [HTF09b].

While linear SVR can capture linear relationships between input and output, real-world data often exhibit non-linear behavior. To address this, SVR can be extended through the use of kernel functions that implicitly map the input features into a higher-dimensional feature space where linear relationships may exist. Common kernels include:

- **Linear kernel:** $K(x_i, x_j) = x_i^\top x_j$
- **Polynomial kernel:** $K(x_i, x_j) = (\gamma x_i^\top x_j + r)^d$
- **Radial Basis Function (RBF):** $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

The RBF kernel is particularly popular due to its ability to model complex non-linear relationships with relatively few hyperparameters. It is often the default choice in practice, especially when no prior knowledge of the data's structure is available [SS02].

In these formulations, γ is a parameter that controls the influence of individual training samples: low values of γ lead to smoother models with broader generalization, while high values allow the model to fit the training data more tightly, potentially at the cost of overfitting. In the polynomial kernel, the parameter r serves as a constant term that shifts the inner product before applying the polynomial transformation, and d defines the degree of the polynomial, which directly impacts the complexity of the resulting model. These kernel-specific parameters, particularly γ and d, are typically tuned as part of the model selection process, as they play a crucial role in determining the flexibility and behavior of the SVR model.

In this thesis, SVR was implemented using the `scikit-learn` library in Python [PVG+11]. Prior to model training, all features were standardized to zero mean and unit variance, as SVR is sensitive to feature scales.

To determine the final model configuration, a randomized hyperparameter search was conducted. The parameters considered included the regularization strength C, sampled from a uniform distribution between 0.1 and 10, and the tolerance parameter ε, drawn from a uniform range between 0.01 and 0.5. Additionally, two kernel types were tested — `linear` and `rbf` — while the `gamma` parameter was fixed to `scale` throughout (not relevant for the linear kernel selected in the final model).

Support Vector Regression offers several advantages. It is robust to outliers (due to the ε-insensitive loss), requires relatively few hyperparameters, and performs well even with smaller datasets. Its kernel-based formulation allows it to flexibly model both linear and non-linear patterns.

However, SVR also comes with certain drawbacks. It lacks interpretability compared to models like linear regression or decision trees and does not provide native measures of feature importance. Additionally, model training can be computationally expensive for large datasets, especially when using non-linear kernels.

4.2 Evaluation Metrics

To evaluate the performance of the predictive models in this study, four key metrics were used: the relative Root Mean Squared Error (rRMSE), the relative Mean Absolute Error (rMAE), the Pearson correlation coefficient, and the computational runtime. Each metric captures different aspects of model accuracy and reliability, providing a comprehensive assessment of the results.

RMSE and MAE are both frequently used and widely discussed, as each brings its own advantages and disadvantages to the evaluation of model performance.

The RMSE have historically been widely used, in large part because of their close connection to statistical theory and their desirable mathematical properties in many modeling contexts. However, one notable limitation is their higher sensitivity to outliers, as squaring the errors gives disproportionate weight to large deviations. In contrast, the MAE is less affected by such extreme values and therefore often provides a more balanced view of typical prediction errors [HK06].

It cannot be said that one of these metrics is inherently better than the other. It is suggest to use a combination of different metrics, including but not limited to RMSEs and MAEs, in order to comprehensively assess model performance [CD14].

In this study, I consistently used a normalized version of the error metrics, which I refer to as "relative." Specifically, the RMSE and MAE were each divided by the mean of the observed data and multiplied by 100. This approach addresses the substantial variability in scale across different grid points. By expressing the errors relative to the mean, the resulting measure becomes independent of the absolute scale of the grid point, allowing for a more meaningful comparison of prediction accuracy across regions.

The relative Root Mean Squared Error (rRMSE) measures the square root of the average squared differences between the predicted and observed values, normalized by the mean of the observed data. It is given by:

$$\text{rRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100\% \tag{4.10}$$

where y_i and $\hat{y}_i$ are the observed and predicted values, and $\bar{y}$ is the mean of the observed values. The relative Root Mean Squared Error (rRMSE) is particularly sensitive to larger errors and highlights how well the model captures the overall variability of the data in relation to its mean.

The relative Mean Absolute Error (rMAE) quantifies the average absolute difference between observed and predicted values, again relative to the mean of the observations:

$$\text{rMAE} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\bar{y}} \times 100\% \tag{4.11}$$

In contrast to rRMSE, the relative Mean Absolute Error (rMAE) is less influenced by outliers and provides a more balanced view of the typical prediction error across different scales [HK06].

The Pearson correlation coefficient measures the linear relationship between predicted and observed values:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{4.12}$$

This coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship. In this work, it is especially useful for assessing how well the models reproduce temporal patterns and trends, even when absolute error values might vary due to differences in data scales.

Finally, the computational runtime provides a measure of the practical feasibility of the models, particularly in large-scale Earth observation applications where thousands of grid points need to be analyzed. In this study, runtime was an important factor for comparing model efficiency and generalizability across the entire study area, despite differences in data scale or resolution.

Together, these four metrics ensure a balanced evaluation of both model accuracy and practical considerations, such as the ability to reproduce long-term trends and to scale the analysis to different regions.

4.3 Drift Detection

This section introduces the approach taken to detect and analyze drift in the study area. It first explains the conceptual background and different types of drift, followed by a description of the specific detection method applied in this work. The Mann-Kendall test, a widely used non-parametric trend detection tool, is then presented as the main analytical approach for identifying temporal changes in backscatter and model residuals.

4.3.1 Understanding Drift

Concept drift describes the phenomenon where the statistical properties of data change over time, causing predictive models to become less accurate if they do not adapt. Such shifts can reduce the accuracy and reliability of models trained on earlier data, as the assumptions on which they were built no longer fully hold [LLD+18].

Gama et al. ([GZB+14]) typically distinguish between four main types of drift:

- **Sudden drift** involves an abrupt change in data distribution within a very short period of time, such as when a sensor is replaced and suddenly produces different readings.

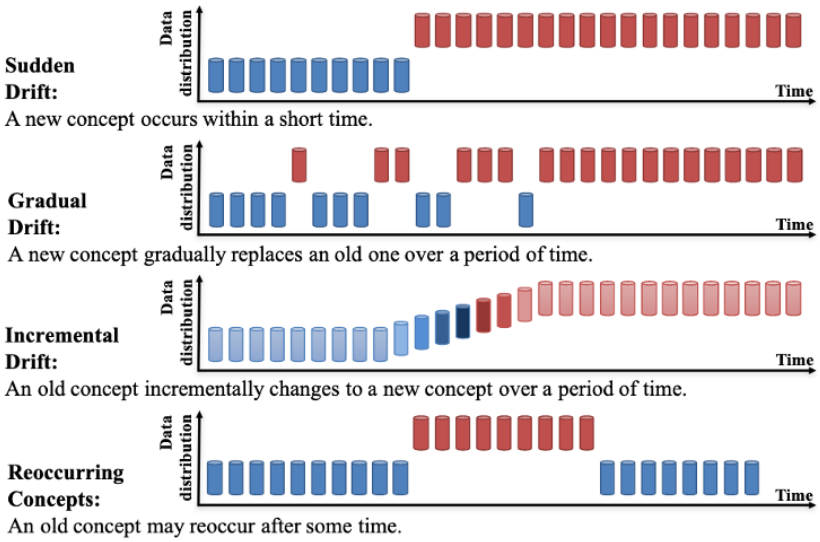


Figure 4.3: Types of Concept drift [LLD+18]

- **Gradual drift** occurs when a new concept slowly replaces the old one, with a period of overlap where both data patterns are present.
- **Incremental drift** describes a continuous and steady transition from one concept to another, often with intermediate states that blend characteristics of both.
- **Reoccurring concepts** refer to situations where a previously seen concept returns after a period of absence, creating cycles in the data that can challenge model stability.

These four types of concept drift are summarized in Figure 4.3.

Understanding drift requires answering questions about when it occurs, how severe it is, and where in the data domain it manifests. This information guides the development of algorithms to detect and adapt to drift, ensuring that models remain accurate and relevant over time. However, one of the challenges in drift detection is to distinguish genuine concept drift from random noise or isolated outliers, which do not require adaptive changes [GZB+14].

In this work, the focus is on detecting such drift effects in the ASCAT backscatter time series and associated residuals. Identifying gradual or incremental changes can help reveal long-term environmental shifts not fully captured by the input variables, providing a deeper understanding of surface dynamics in the study region.

4.3.2 Drift Detection Approach

To identify and interpret temporal changes in the target variable, model predictions, and residuals, a combined approach involving statistical testing, trend-based classification, and exploratory visualization was applied. The objective was to systematically detect different types of drift — specifically target drift, prediction drift, and residual drift — and to assess their spatial and temporal characteristics.

The first step involved applying Mann-Kendall trend tests to all grid points, separately for the target variable (backscatter), the model predictions, the residuals, and the input variables (leaf area index, soil moisture). Tau and slope values derived from these tests were used to identify spatial patterns and quantify the strength and direction of trends.

Based on the presence or absence of significant trends in the target, prediction, and residuals, all grid points were assigned to one of eight groups in a trend-based classification. These were visualized on a spatial map to highlight regional patterns and identify representative points for deeper investigation.

In addition, a trend strength segmentation was performed, grouping grid points according to the magnitude of their Tau and slope values derived from the Mann-Kendall test on the target variable. Four segments were defined: (i) high slope & high Tau, (ii) high slope & low Tau, (iii) low slope & high Tau, and (iv) low slope & low Tau. This segmentation helps distinguish between continuous trends and abrupt, event-driven changes.

From each category and segment, one to two representative grid points were selected for detailed visual analysis. The following visualizations were produced:

- Time series of the target variable, model predictions, and residuals
- Absolute residuals and rolling mean of residuals
- Rolling signal plots of the backscatter variable
- Histogram comparison of observed vs. predicted values
- Yearly change maps for selected points

To complement the trend analysis, the *autocorrelation of residuals* was also assessed, with the goal of identifying systematic temporal error structures or periodic effects in the model behavior.

Technical implementation All analyses were conducted in Python 3.11.10 [VRD09] using JupyterLab version 4.2.5 [jup] environment. Data processing relied primarily on `Pandas` [M+10] and `NumPy` [HMvdW+20]. Visualization was performed using `Matplotlib` [Hun07], `Seaborn` [WBO+17], `eomaps` [Qua25], and `cartopy` [noa25]. Statistical analyses were implemented with `SciPy` [VGO+20] and `pymannkendall` [Sho25]. Random elements (e.g., point sampling) were controlled via the standard `random` [VR20] module.

4.3.3 Mann-Kendall Test

The Mann-Kendall test is a non-parametric method widely used to detect monotonic trends in time series data. Originally developed by Mann (1945) and refined by Kendall (1975), the test has become a standard approach in environmental and hydrological studies due to its robustness and simplicity [YPC02].

The hypotheses tested by the Mann-Kendall method are:

- **Null hypothesis ($\mathbf{H_0}$):** No monotonic trend exists in the data (the data are randomly ordered in time).
- **Alternative hypothesis ($\mathbf{H_1}$):** A monotonic trend (increasing or decreasing) exists in the data over time.

The Mann-Kendall statistic **S** is computed by evaluating the sign of the difference between all possible pairs of data values x_i and x_j (with $j > i$):

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_j - x_i)$$

where the sign function is defined as:

$$\operatorname{sgn}(\theta) = \begin{cases} 1 & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ -1 & \text{if } \theta < 0 \end{cases}$$

Here, x_i and x_j represent the data values in the time series, while θ in the sign function denotes the pairwise differences.

For large sample sizes ($n > 10$), the test statistic S can be approximated by a normal distribution, which allows the calculation of a Z-score to determine statistical significance.

A key aspect of the Mann-Kendall test is the calculation of the Kendall's Tau (τ) coefficient, which quantifies the strength and direction of the monotonic trend. Positive values of Tau indicate an increasing trend, negative values indicate a decreasing trend, and values near zero suggest no significant trend. Additionally, the Sen's slope estimator is often used to quantify the magnitude of the trend by calculating the median of all pairwise slopes in the data series, providing a robust estimate of the overall rate of change [UAY+20].

The Mann-Kendall test has several advantages. It is non-parametric, meaning it does not rely on assumptions about the data's distribution. It is robust to outliers and non-linear trends, making it particularly suitable for environmental data with high variability, and it requires only ordinal data while handling missing values effectively. However, the test

assumes that the data points are independent and identically distributed. If significant autocorrelation is present in the time series, it can inflate the test's significance and lead to incorrect inferences [YPC02].

In this study, the Mann-Kendall test is applied to ASCAT backscatter time series, model predictions, and residuals to identify and classify temporal trends. This approach is well-suited for long-term trend analysis in environmental datasets, as it does not require the assumption of linearity or stationarity.

CHAPTER 5

Results and Discussion

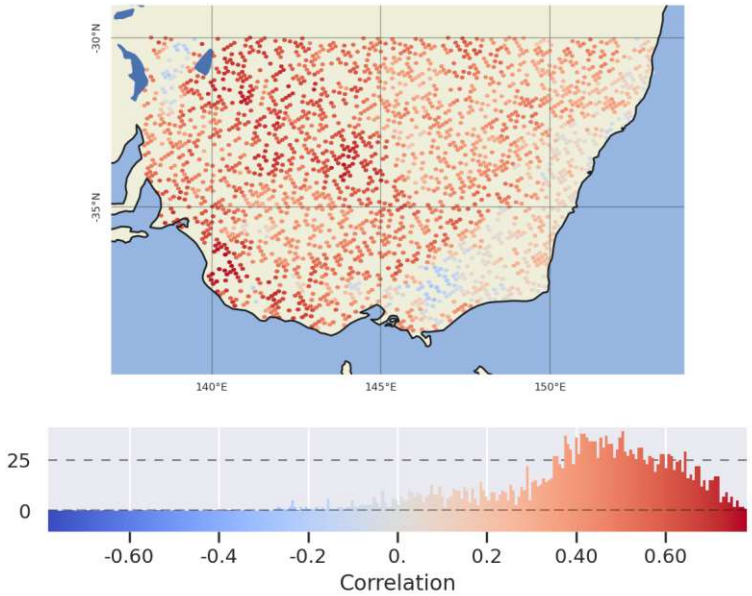
This chapter presents the results of the modeling and trend analysis introduced in the previous chapter as well as a comprehensive discussion including limitations and critical reflection. First, the performance of the selected regression models is evaluated, both in terms of accuracy and computational efficiency. Based on this comparison, the Elastic Net model is applied across the full study area. Subsequently, the focus shifts to the detection of temporal drift patterns, including a spatial analysis of trends, a categorization based on the alignment between predictions and observations, and a segmentation according to the strength and consistency of observed trends.

Figure 5.1 shows the spatial distribution of Pearson correlation coefficients between ASCAT backscatter and the two input variables: LAI and soil moisture. Moderate to high positive correlations are observed in areas dominated by cropland, grassland, shrubland, and sparse vegetation. In contrast, low or even negative correlations appear in arid regions, particularly between Lake Torrens and Lake Frome. Regions with evergreen tree cover also exhibit weak correlations. These spatial patterns differ between the two input variables, reflecting the varying influence of vegetation and soil moisture on backscatter across land cover types. Notable regional differences in the correlation are discussed in more detail in Chapter 6.1.

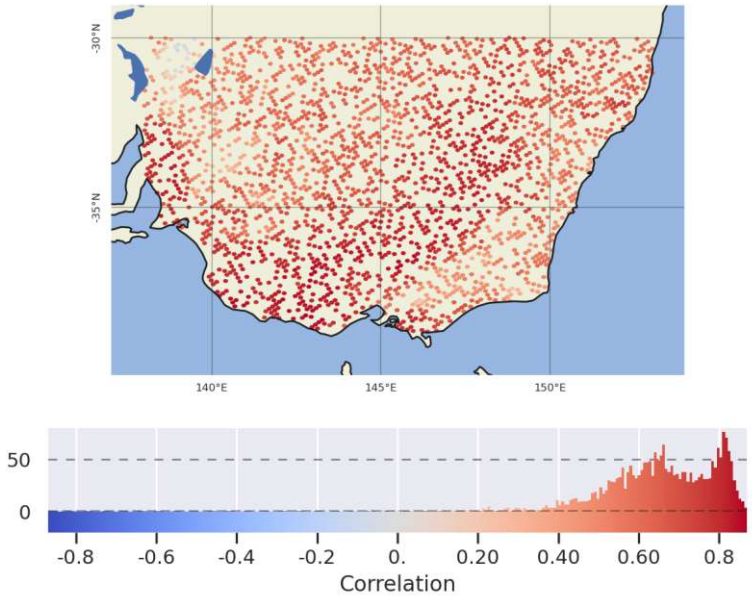
5.1 Model Selection

To evaluate the performance of the three selected regression models (Random Forest, Elastic Net, Support Vector Regression), a fixed set of hyperparameters was defined for each model based on the median values derived from 20 representative grid points.

Subsequently, all three models were applied to a sample of 100 randomly selected grid points using the fixed parameter settings. The evaluation was based on the metrics relative Root Mean Squared Error (rRMSE), relative Mean Absolute Error (rMAE),



(a) Correlation of backscatter and LAI



(b) Correlation of backscatter and soil moisture

Figure 5.1: Spatial distribution of Pearson correlation coefficients between ASCAT backscatter and the two input variables: LAI (left) and soil moisture (right).

Pearson Correlation Coefficient (PCC) and runtime per grid point. The averaged results across all grid points are summarized in the following table:

Model	rRMSE	rMAE	PCC	Runtime
Random Forest	4.895	3.658	0.771	0.529
Elastic Net	5.188	3.951	0.752	**0.001**
SVR	**4.805**	**3.487**	**0.780**	0.833

Table 5.1: Comparison of average model performance over 100 grid points.

Although Support Vector Regression achieved the best scores in rRMSE, rMAE and PCC, it showed the highest computational cost. Random Forest achieved consistent results across all metrics but was still significantly slower than Elastic Net. Elastic Net showed the lowest runtime by far, while maintaining reasonable prediction accuracy. The differences between the models in terms of rRMSE, rMAE, and PCC are relatively small, indicating that all three models perform similarly in terms of predictive accuracy. However, the differences in runtime are substantial. While prediction quality is comparable, the runtime varies by orders of magnitude, making it a decisive factor in model selection. Consequently, Elastic Net was selected as the final model for application across all grid points.

Spatial Evaluation

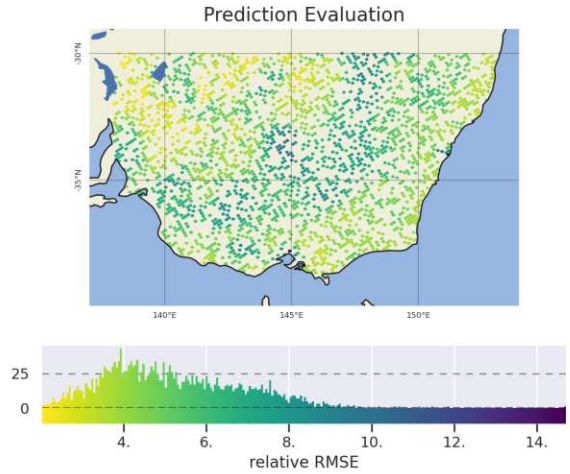
To further assess the spatial performance of the selected model (Elastic Net), evaluation metrics were computed and visualized for the entire study area (Figure 5.2). The spatial distribution and histograms of rRMSE, rMAE and PCC reveal substantial spatial variability in model performance.

The Pearson Correlation shows moderate to high values in most regions, with a few isolated areas of low or even negative correlation. The rRMSE and rMAE values are generally low but vary depending on location, indicating that model accuracy is strongly grid-point dependent.

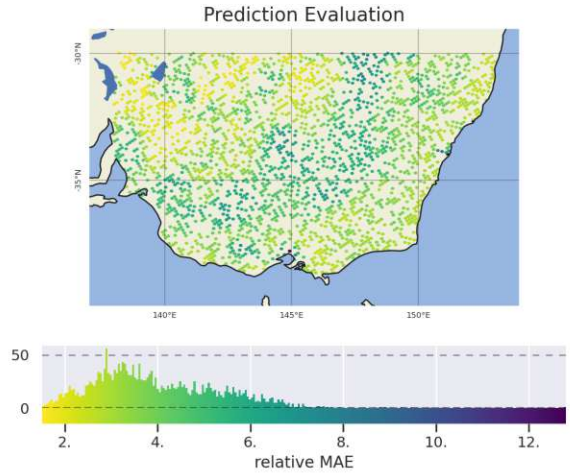
5.2 Drift Detection

5.2.1 Overview of Trend Patterns

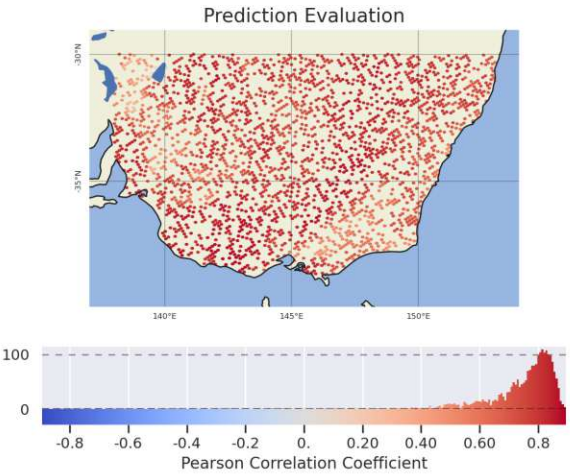
To gain an initial understanding of the magnitude and spatial distribution of temporal trends, Mann-Kendall tests were conducted at all available grid points. The analysis was applied to (i) the observed backscatter values, (ii) the model predictions, (iii) the residuals, and (iv) the input variables. The resulting Kendall's Tau values indicate both the direction (positive/negative) and strength of monotonic trends over the observation period. Figure 5.3 and Figure 5.4 presents the spatial distribution of Kendall's Tau values, allowing a direct comparison of temporal trend strength and direction across the



(a) rRMSE



(b) rMAE

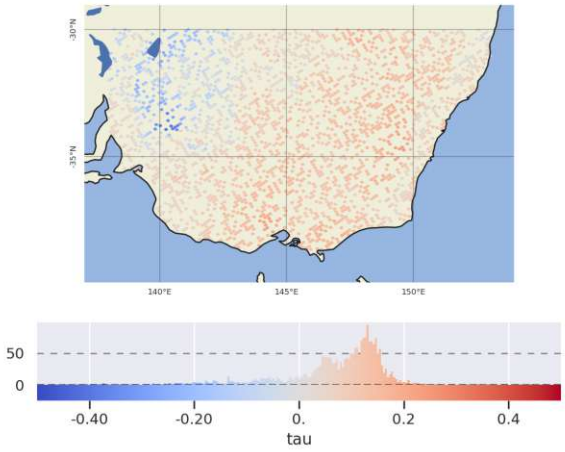


(c) PCC

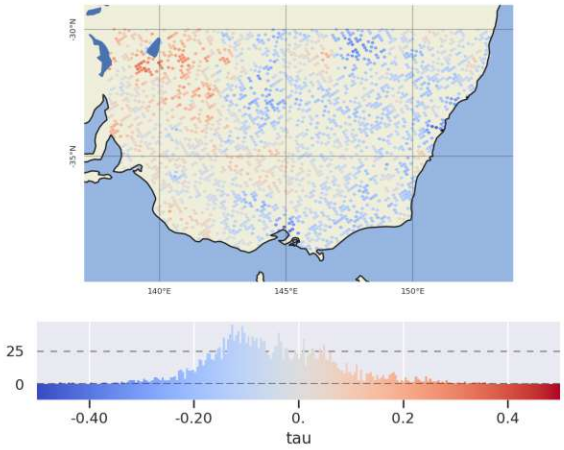
Figure 5.2: Model evaluation of rRMSE, rMAE and PCC.

study area. Clear regional patterns can be visually identified, highlighting areas with particularly strong or weak trends.

In addition, Sen's slope estimates were calculated to quantify the absolute magnitude of change. While slope values are presented in the appendix (7.1), key findings are briefly described in the following sections.



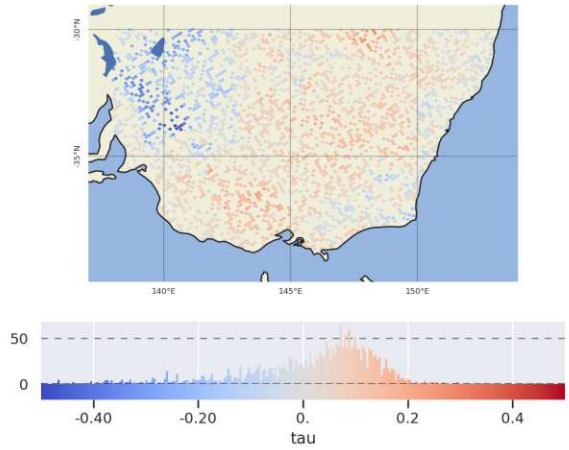
(a) Tau – Prediction



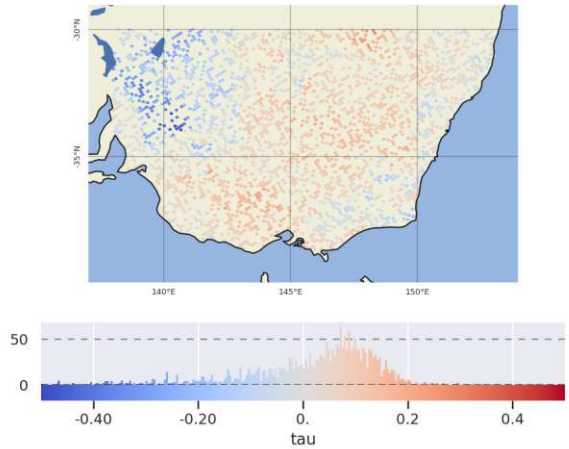
(b) Tau – Residuals

Figure 5.3: Kendall's Tau for each prediction and residuals across the study area.

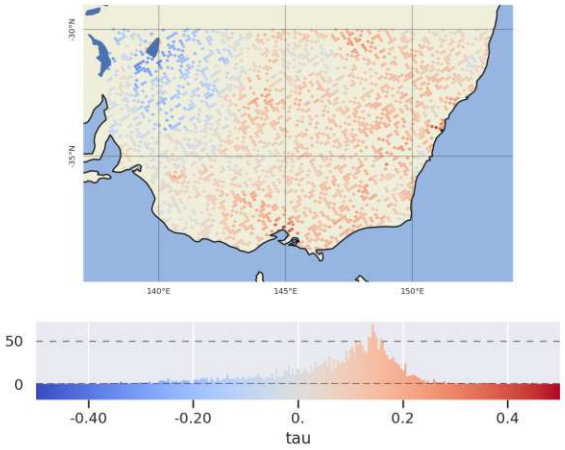
Target variable The Kendall's Tau values for observed backscatter reveal a clear spatial pattern. While positive trends dominate across large parts of the study area, negative trends are concentrated in the northwest and western regions. Overall, 74% of all grid points show positive trends, 18% negative trends, and 8% no significant trend.



(a) Tau – LAI



(b) Tau – SM



(c) Tau – backscatter

Figure 5.4: Kendall's Tau for each input variable across the study area.

Very strong positive trends are found in areas surrounding major cities such as Sydney and Melbourne. Small clusters of weakly negative trends also appear along the east coast and in the northern central part of the region. Despite the generally negative trend in the northwest, some grid points show positive trends as well.

An examination of the Sen's slope values highlights two distinct clusters with strong trends: one negative cluster in the west with decreases ranging from 0.164 to 0.274 dB per year, and one positive cluster in the northeast with increases of approximately 0.329 dB per year for the period 2015–2022. Across the entire study area, the average trend amounts to +0.054 dB per year.

Predictions The Kendall's Tau values for the model predictions show a similar overall pattern to the observed backscatter trends. However, the positive trends are generally less pronounced, particularly in urban regions where strong trends were observed in the backscatter. Across all grid points, 80% show positive trends, 14% negative trends, and 7% no significant trend.

In contrast to the backscatter results, the Sen's slope values for the predictions exhibit a narrower range of variation. The model outputs display no sharply localized clusters of strong increase or decrease. Instead, the slope values are more evenly distributed across space. The estimated slopes range between –0.149 and +0.201 dB per year, with a mean value of +0.041 dB per year.

Residuals The residuals show a distinct pattern compared to the observed and predicted values. The Kendall's Tau values indicate a dominance of negative trends across the study area: 63% of grid points exhibit negative trends, 25% positive trends, and 12% show no significant trend. Urban areas, which already stood out in the backscatter results, are again clearly visible.

The corresponding Sen's slope values (Appendix) reveal several localized clusters with stronger trend magnitudes — similar in location to those found in the backscatter data, but not captured to the same extent by the model predictions. These areas stand out more clearly in the residual slope map than in the prediction slopes.

In most regions, residual trends show an opposite direction compared to the trends in observed values, resulting in markedly contrasting spatial patterns.

Input Variables The Kendall's Tau values for LAI indicate a mixed but spatially structured trend pattern. 60% of grid points show positive trends, 28% negative, and 12% no significant trend. Strong negative trends are particularly concentrated in the western part of the study area, with additional areas of weak negative trends along the east coast.

The corresponding MK test results Sen's slope confirm these patterns, showing clearly distinguishable regions with both increasing and decreasing LAI over time. The mean slope is 0.00876 per year, with especially strong negative slopes along the east cost and

also in the west, while the area of southeastern Murray-Darling Basin exhibit consistent positive changes throughout the period 2015–2022.

The Kendall's Tau results for SM show a clear dominance of increasing trends: 90% of grid points exhibit positive trends, 2% negative, and 8% no significant trend. No strong negative clusters are visible.

The Sen's slope values support this observation, indicating a general increase in soil moisture across most of the study area. The area of southeastern Murray-Darling Basin with elevated slope values suggests locally stronger daily increases over the test period.

5.2.2 Trend-Based Classification

In order to better understand the interaction of trends in the target variable, the model predictions and the residuals, the grid points were classified. Points were assigned to specific categories depending on the presence of significant trends in (1) the observed backscatter, (2) the model predictions and (3) the residuals. The aim of this categorization is to identify patterns in which the model either "drifts" with the target variable, remains stable or develops systematic errors.

A total of 8 categories were defined, which represent different combinations of trend constellations (e.g. positive trend in target and prediction, but stable residuals).

Category	Backscatter	Prediction	Residuals
A	No trend	No trend	No trend
B	Trend	Trend	No trend
C	Trend	No trend	Trend
D	Trend	Trend	Trend
E	No trend	Trend	Trend
F	No trend	Trend	No trend
G	No trend	No trend	Trend
H	Trend	No trend	No trend

Table 5.2: Definition of trend-based categories based on the presence of significant trends in the target variable (Backscatter), model predictions, and residuals.

The interpretation of each category helps to better understand how the model responds to different trend constellations. Category A is of little relevance in this context, as no significant trend is observed in any of the three time series (backscatter, prediction, residuals). In contrast, category B shows matching trends in the target variable and model predictions, while the residuals remain stable. This suggests that the input variables explain the trend well and the model captures it correctly.

Category C shows a trend in the backscatter that is not reflected in the predictions, while the residuals exhibit a significant trend. In this case, the model fails to represent the observed development, indicating that relevant information is either missing from the

input data or not captured by the model. The cause of the trend likely lies outside the scope of the model.

The majority of grid points fall into category D. Here, both the target variable and the model predictions show a trend, but the residuals also exhibit a systematic pattern. This may indicate that the model captures the general direction of change but not its full dynamics – for example, it may respond too slowly or with the wrong magnitude. The model appears to "drift along" with the target, but not in perfect alignment.

Category E includes points with no trend in the backscatter, yet both predictions and residuals show trends. This suggests that the model systematically forecasts a development that does not occur in reality, possibly due to misleading patterns in the input variables. This may point to overfitting to irrelevant input trends.

In category F, predictions show a trend despite no trend in the target variable, but the residuals remain stable. This could indicate a slight model bias that does not substantially affect the prediction error. It might reflect a minor misinterpretation of input patterns by the model.

Category G exhibits no trend in the target or predictions, yet the residuals show a systematic development. This might indicate that subtle shifts in the target variable are not fully captured by the model, even though the predictions appear stable. It may reflect non-linear effects or slow changes that accumulate in the residuals.

Finally, category H shows a trend in the backscatter, while both predictions and residuals remain constant. This may imply that the model does not react to the ongoing change, either because the trend is too weak to be detected or because the model lacks sensitivity to gradual developments. It is also possible that small errors cancel each other out over time, masking a systematic drift.

Spatial Distribution of Categories Figure 5.5 illustrates the spatial distribution of the trend-based categories across the study area. Category D is by far the most prevalent, occurring throughout the entire region and dominating the classification. It is widespread and appears in all parts of the study area.

Category C is primarily concentrated in the western part of the region, with additional occurrences along the northeastern coastline. Similarly, category A is mostly found in the west but also appears sporadically along the eastern coast.

Category B is scattered across the study area without clear spatial clustering. While points often occur in small groups, there are no larger accumulations or dominant zones associated with this category.

Detailed Analysis of Categories In order to better understand the model behavior within the defined categories, exemplary grid points were selected and their time series analyzed. The visualizations include the target variable (backscatter), the model predictions and the residuals.

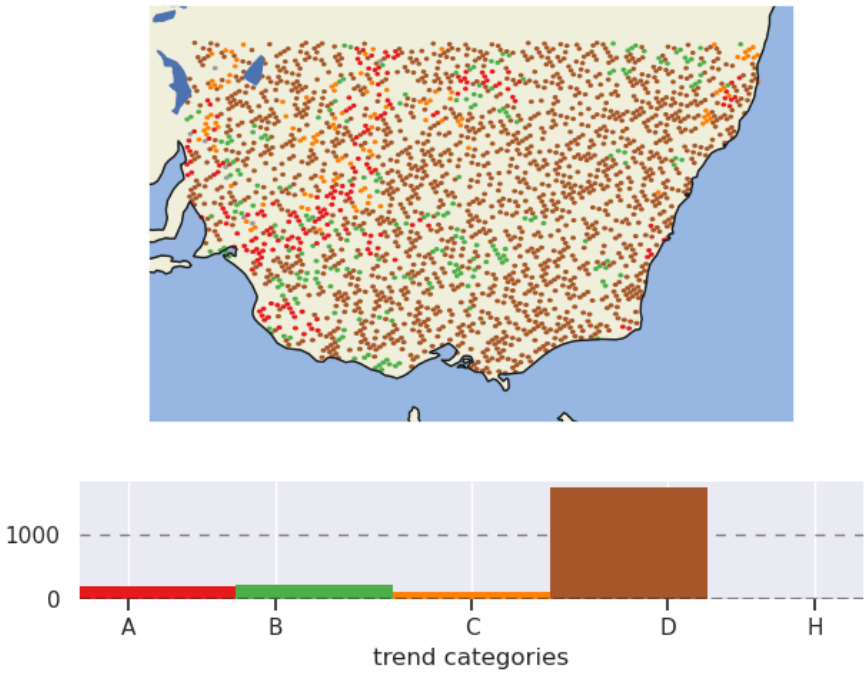


Figure 5.5: Trend categories in study area

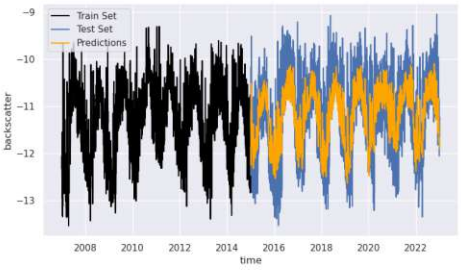
Category	Number of gridpoints
D	1751
B	219
A	193
C	112
H	8

Table 5.3: Number of grid points assigned to each trend category.

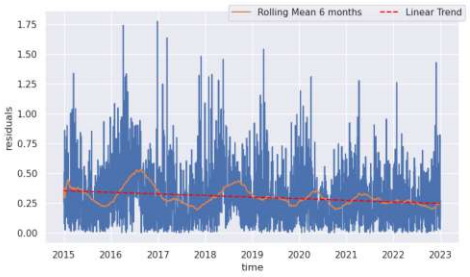
The initial focus is on categories B and C, as a direct comparison can be made here between a well-functioning model (category B) and clearly faulty model performance (category C).

The time series presented for category B (Figure 5.6a) demonstrate that the model captures the observed patterns very well. The trends in the target variable – even if only weakly pronounced in some cases – are reliably reflected in the model predictions. There is strong agreement between backscatter and prediction, both in terms of overall trend and seasonal dynamics.

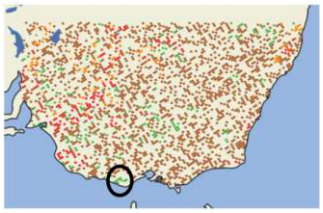
The rolling mean of the residuals (Figure 5.6b) supports this observation: the residuals remain stable throughout the entire period and do not show any systematic drift over time. This indicates robust model performance without evidence of temporal bias.



(a) Timeseries of backscatter-signal



(b) Rolling Mean of residuals



(c) Location of gridpoint

Figure 5.6: Timeseries (a), rolling mean (b) and location (c) of analyzed gridpoint 2278197 at 38.45 ° S, 143.34 ° E from category B.

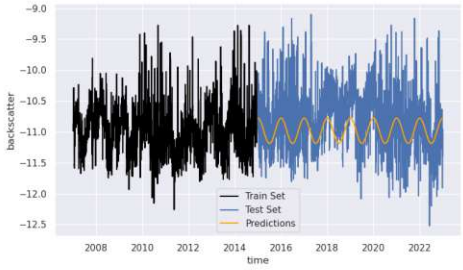
The time series for category C (Figure 5.7a) show that the model is unable to produce reliable predictions based on the available input data. While the backscatter exhibits a change over a specific time period, the model predictions remain largely constant and do not capture this development.

This poor model performance can be explained by the weak correlation between the input variables and the target. At this grid point, the correlation between backscatter and soil moisture is –0.016, and with LAI it is –0.028. These values indicate that neither variable is suitable for explaining backscatter in this part of the study area. As a result, the model relies entirely on the day-of-year as input – a variable that, on its own, is insufficient to represent the observed changes.

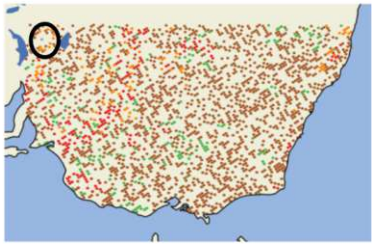
The time series for category D (Figure 5.8a) show that both the target variable (backscatter) and the model predictions exhibit a clear trend. Overall, the model depicts the overarching development well and reacts appropriately to the observed changes. Nevertheless, the residuals show a systematic pattern, which indicates minor deviations in the strength of the trend or over time - for example, when the backscatter increases faster than the prediction.

Such deviations can be explained by differences in the slope or a temporal shift between input and target variable. Overall, however, the model demonstrates a solid ability to capture the underlying processes, even if individual aspects of the trend are not fully represented.

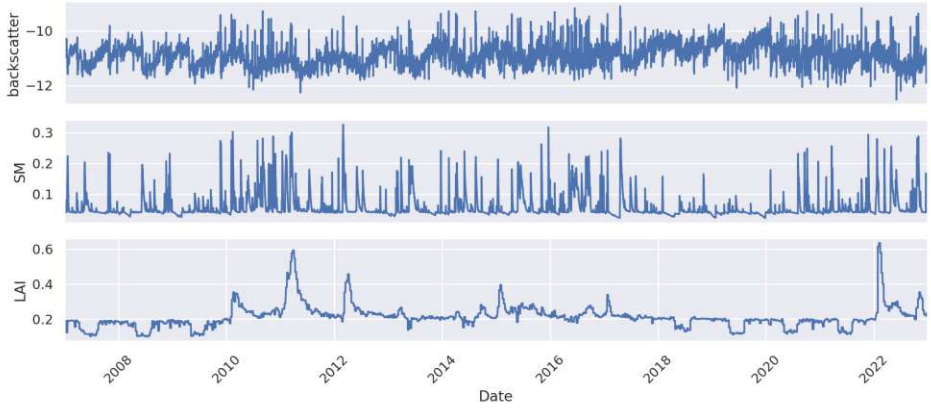
In one example from category D (Figure 5.9), the model's behavior varies notably over



(a) Timeseries of backscatter-signal



(b) Location of gridpoint



(c) Timeseries of input variables

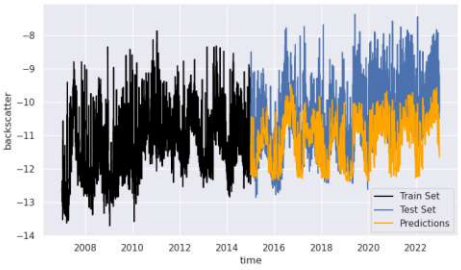
Figure 5.7: Timeseries (a,c) and location (b) of analyzed gridpoint 2463868 at 30.62 ° S, 139.28 ° E from category C.

time. During certain periods, the prediction closely follows the observed deviations in the backscatter, indicating a strong alignment between input-driven dynamics and target response. However, in other time intervals, the model fails to capture emerging trends, resulting in a growing mismatch.

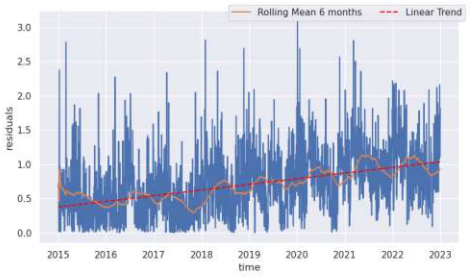
To illustrate these patterns more clearly, Figure 5.10 shows the time series of soil moisture, LAI, and backscatter for the representative grid point.

A closer look at the period from 2010 to 2012 in Figure 5.10 reveals that soil moisture and the Leaf Area Index (LAI) increase almost simultaneously, albeit with a slight time lag. The rise in soil moisture promotes plant growth, which is reflected in the increasing LAI. In parallel, the backscatter signal also increases during the first part of this period, capturing the rising water content in the soil.

Notably, the backscatter signal decreases much more slowly than both soil moisture and LAI: while soil moisture drops significantly and LAI declines as well, the backscatter remains elevated for a longer time and only gradually decreases. A similar pattern can be observed during the 2016–2018 period.



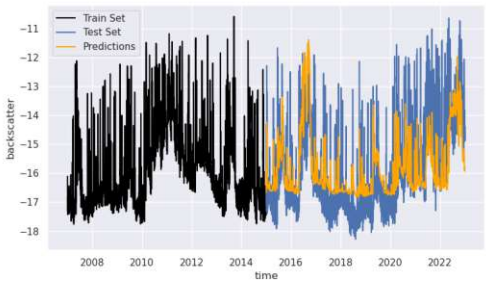
(a) Timeseries of backscatter-signal



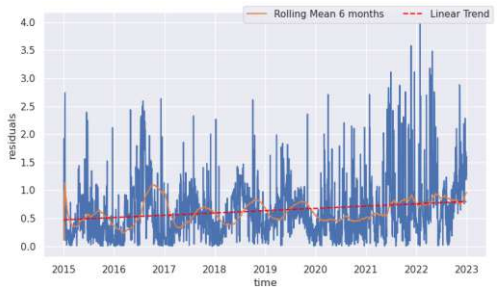
(b) Rolling Mean of residuals

(c) Location of gridpoint

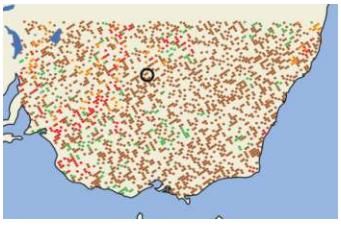
Figure 5.8: Timeseries (a), rolling mean (b) and location (c) of analyzed gridpoint 2294777 at 37.72 ° S, 144.61 ° E from category D.



(a) Timeseries of backscatter-signal



(b) Rolling mean of residuals



(c) Location of gridpoint

Figure 5.9: Timeseries (a), rolling mean (b) and location (c) of analyzed gridpoint 2420728 at 32.37 ° S, 144.22 ° E from category D.

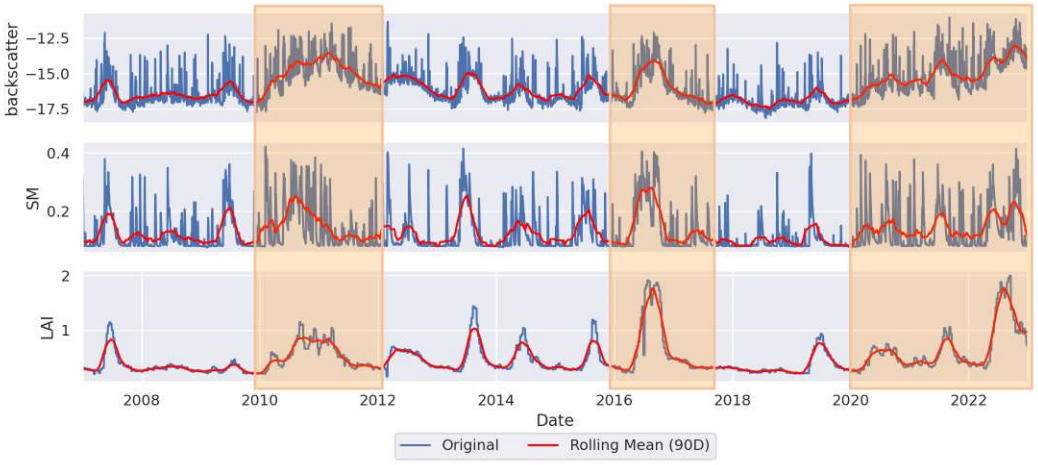


Figure 5.10: Time Series of input variables for gridpoint 2420728 at 32.37 ° S, 144.22 ° E.

The period from 2020 to 2023 is particularly interesting. Unlike the previous intervals, soil moisture does not exhibit longer phases of steadily increasing and consistently high values, but rather a series of short-term peaks. This pattern suggests a general trend towards increasing soil saturation, albeit without sustained plateaus. The backscatter signal captures this increase in soil moisture accordingly. In contrast, the LAI follows a clear seasonal pattern, with annual peaks during the middle of the year.

This example is particularly interesting as it does not show a continuous long-term trend but rather distinct episodes of sharp change, likely related to extreme events. These abrupt shifts appear to drive the overall trend signal detected in the Mann-Kendall test. Such cases highlight the limitations of the current trend-based classification, which does not distinguish between continuous developments and event-driven changes – a gap that is addressed in the following segmentation based on trend strength.

5.2.3 Trend Strength Segmentation

While the category-based classification provides valuable insights into the relationship between target variable, prediction, and residuals, it does not specify the nature of the trend in the target variable itself. In particular, it does not distinguish whether the observed trend is continuous or confined to specific time periods, such as during extreme events. Moreover, the strength of the trend remains unaddressed within the categorical framework.

The following section therefore introduces a segmentation based on the strength of the trend in the target variable. This segmentation is derived from the absolute value of Sen's slope and the Kendall's Tau coefficient. The aim is to identify and characterize different types of trend behavior in the backscatter time series – whether gradual or

abrupt, weak or strong – and to better understand their role in overall trend formation.

To operationalize this differentiation, four distinct trend segments were defined based on the combination of Sen's slope and Kendall's Tau. As described in the chapter 4.2.3 Mann-Kendall Test: While Sen's slope captures the magnitude of change, Tau reflects the consistency and directionality of the trend over time. This dual criterion allows for a more nuanced description of trend dynamics.

The segmentation distinguishes between low and high slope values, as well as weak and strong monotonicity, resulting in the following groups:

- **Segment 1 – No or weak trend:** Tau close to 0, low slope
- **Segment 2 – Continuous but weak trend:** High Tau, but low slope
- **Segment 3 – Strong but irregular trend:** Low Tau, but high slope
- **Segment 4 – Strong and continuous trend:** High Tau and high slope

This classification makes it possible to take into account both the intensity and the consistency of the trend - for example, to distinguish whether a strong increase is caused by a few "jumps" (e.g. extreme events) or by a consistent development.

The segmentation is based on threshold values of 0.15 for Kendall's Tau and 0.0002 for the absolute value of Sen's slope. The same thresholds are applied symmetrically for negative trends, allowing for a consistent classification across all directions of change.

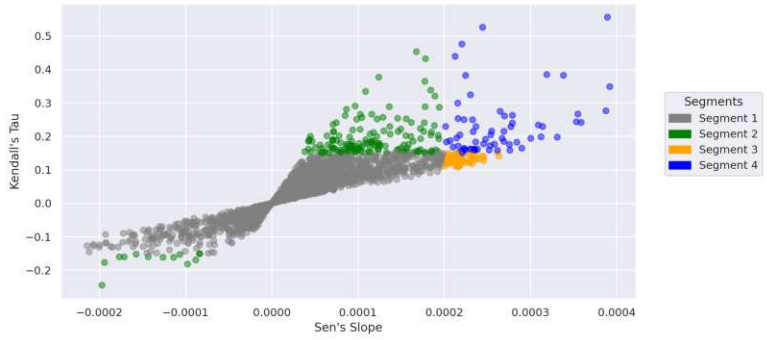
Figure 5.11b shows the spatial distribution of the four trend segments (S1–S4), with each grid point color-coded according to its assigned group.

The majority of points (87 %) fall into Segment 1, while the remaining segments are distributed as follows: S2 – 7 %, S3 – 3 %, S4 – 3 %."

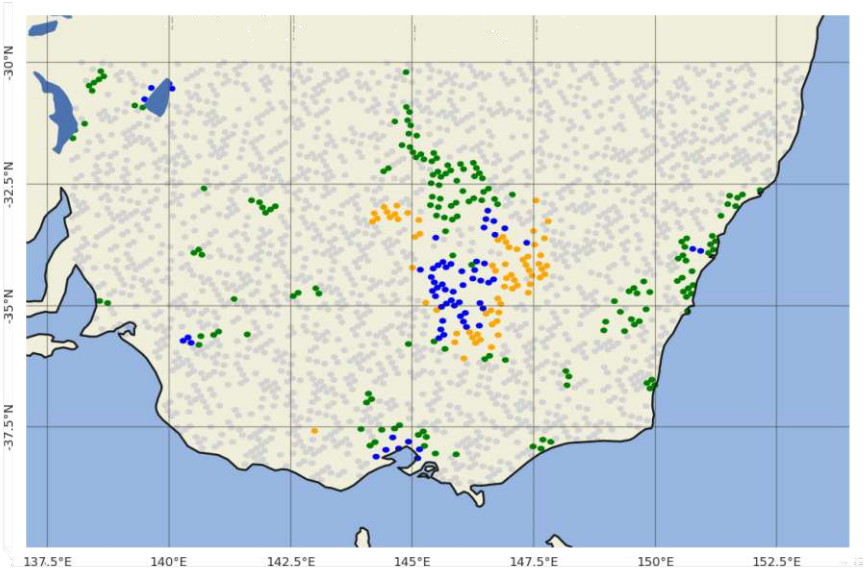
This indicates little to no discernible trend in the backscatter time series across most of the study area. Notably, this is the case despite the relatively low threshold chosen for Kendall's Tau (±0.15), which was already intended to capture weak monotonic signals. This suggests that large parts of the region exhibit highly stable or noisy backscatter dynamics. Adjusting this threshold could be an avenue for future sensitivity analysis.

The remaining segments (S2–S4), which reflect stronger or more consistent trends, are primarily located in the interior of the study area. Segment 4, representing strong and continuous trends, is mostly concentrated in central inland regions, with a few points near urban centers such as Melbourne and Sydney. Segment 3, capturing strong but irregular changes, is also predominantly found inland. Segment 2, which corresponds to weak but consistent trends, is more widely dispersed, including along the eastern coastline and at several scattered locations throughout the study area.

To better illustrate the characteristics of the defined trend segments, representative grid points from segments S2 to S4 were selected for visual inspection. For each point, time



(a) Sen's Slope vs. Kendall's Tau with segment-based color coding



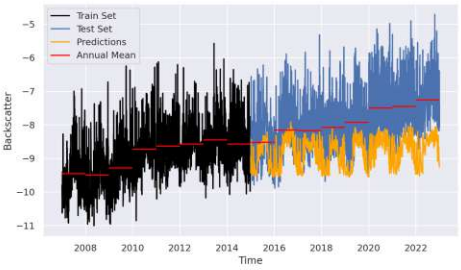
(b) Spatial distribution of the classified grid points

Figure 5.11: Trend segmentation and spatial distribution of grid points

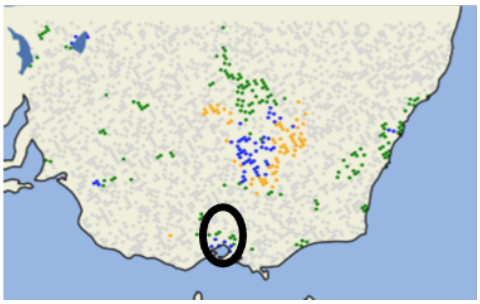
series of the target variable (backscatter), model predictions, and residuals are shown, along with the input variables. The examples highlight typical dynamics within each segment and allow for a qualitative interpretation of trend type and model behavior.

A representative grid point from Segment 4 is located in the urban region of Melbourne. The time series (Figure 5.12a) reveals a clear and continuous upward trend in backscatter across the entire period, which is reflected in the consistently increasing annual means. This point shows the highest Kendall's Tau (0.556) and Sen's slope (0.000389) of the entire dataset, making it the upper-right outlier in the Tau–Slope scatterplot. In the previous trend-based classification, this grid point was assigned to Category D.

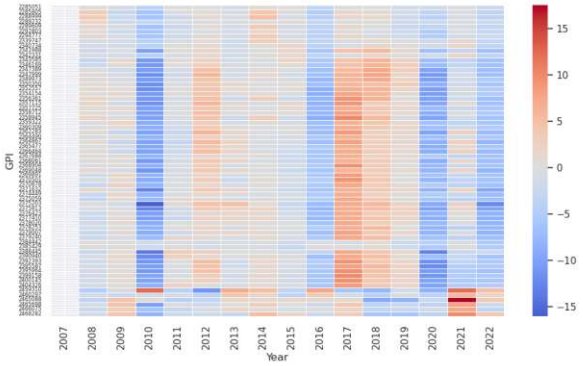
While the model predictions generally follow the seasonal dynamics, they fail to keep up with the steadily rising backscatter values. As a result, the residuals increase



(a) Timeseries of backscatter-signal



(b) Location of gridpoints



(c) Annual percentage change of backscatter in Segment 4

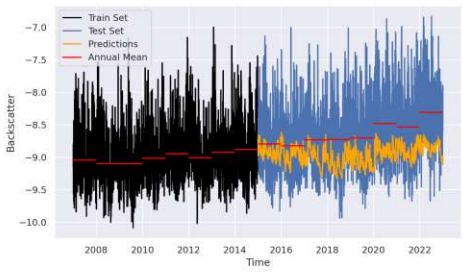
Figure 5.12: (a) Time series and (b) location of a representative grid point from Segment 4. (c) Annual percentage changes in backscatter for all grid points in Segment 4.

systematically over time – a sign of underestimation in the presence of a strong trend.

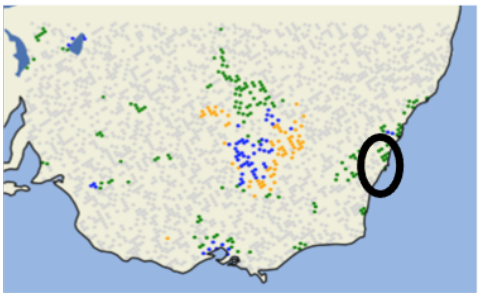
Figure 5.12b provides spatial context by showing the location of the selected grid point within the study area. Figure 5.12c presents a heatmap of the annual percentage change in backscatter for all grid points in Segment 4. Most points, particularly those in the upper part of the heatmap (including Melbourne), show small but consistent year-to-year increases. In contrast, a few points in the lower section of the plot display opposing trend behavior. These will be revisited in the discussion. Overall, the heatmap suggests that some years are drier (positive annual changes), while others are wetter (negative changes), even within a segment defined by overall strong and continuous trends.

The selected point for segment 2 is in the region around Sydney (Figure 5.13c). The time series of the backscatter signal (Figure 5.13c) shows a positive, steady trend. With a tau value of 0.376891, this point is one of the stronger trend points within the entire study region, whereas Sen's Slope only has a moderate slope of 0.000124.

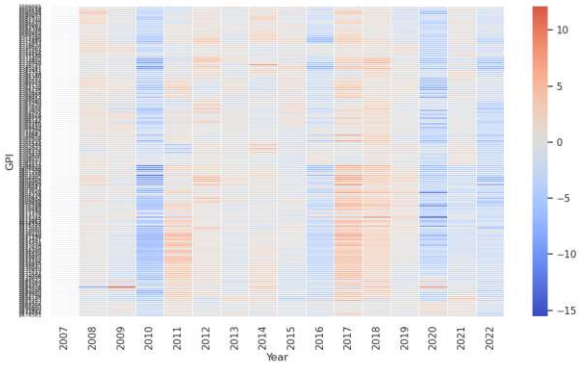
In the classification from section 5.2.2, this point falls into category D, which means that both the target and the residuals show a positive trend.



(a) Timeseries of backscatter-signal



(b) Location of gridpoints



(c) Annual percentage change of backscatter in Segment 2

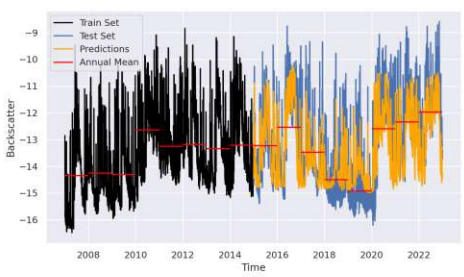
Figure 5.13: (a) Time series and (b) location of a representative grid point from Segment 2. (c) Annual percentage changes in backscatter for all grid points in Segment 2.

The associated heat map (Figure 5.13c) of the annual percentage changes confirms the impression of continuous change: extreme events such as abrupt jumps or breaks do not occur. Nevertheless, certain large-scale trends can be identified, such as years with drier or wetter conditions overall.

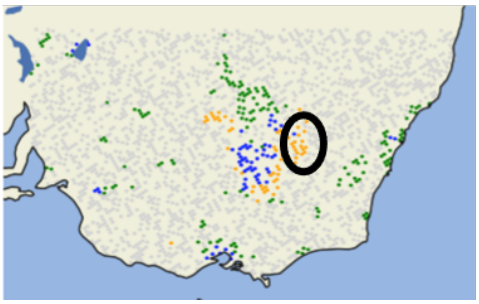
The point selected for segment 3 is located inland, near Hillston in New South Wales (Figure 5.14b). The time series (Figure 5.14a) shows large differences between years, with phases in which trends intensify over several years before reversing. The point belongs to category D from the classification described in section 5.2.2.

The tau value is 0.148599, the Sen's slope is 0.000205. This means that the trend is weaker than in segment 2, but still clearly pronounced. It is striking that this segment also includes some points where large seasonal fluctuations - without the occurrence of extreme events - provide a classification.

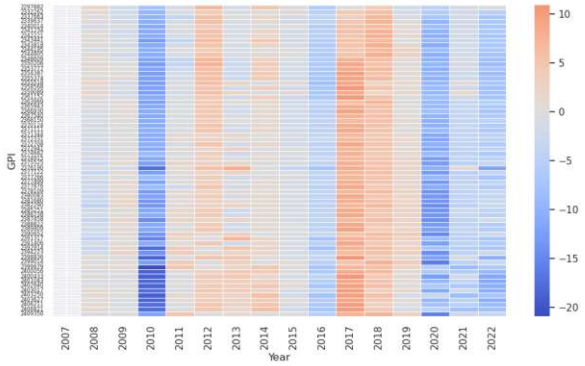
The associated heat map (Figure 5.14c) clearly shows the patterns: strongly positive and strongly negative years are clearly recognizable. All points in this segment follow these overarching trends very consistently.



(a) Timeseries of backscatter-signal



(b) Location of gridpoints



(c) Annual percentage change of backscatter in Segment 3

Figure 5.14: (a) Time series and (b) location of a representative grid point from Segment 3. (c) Annual percentage changes in backscatter for all grid points in Segment3.

5.3 Discussion

At certain grid points or in specific regions, the model does not perform satisfactorily, while at the same time, there are regions where we achieved very good results. Evidently, a crucial feature is missing that would sufficiently explain the variability of the backscatter signal in these cases. While it is not yet clear which specific feature is responsible, it is evident that the currently considered explanatory variables — vegetation and soil moisture + a time component — are not sufficient to fully model the observed backscatter in the whole study area. This issue will be addressed in more detail in the chapter “Future Work”. If a trend in backscatter is indeed present at these locations, this suggests that the missing factor is dynamic over time. It is certainly something that is not captured by the current soil moisture and vegetation datasets. Although it could still be related to soil moisture or vegetation in principle, this relationship is not adequately represented in the available data. This issue will also be discussed in greater detail in the chapter on limitations.

Model errors Notably high model errors (rRMSE, rMAE) were observed at grid points located in urban regions such as Sydney and Melbourne. This is likely due to the complex surface characteristics in cities, where built structures cause irregular radar backscatter that cannot be captured by vegetation and soil moisture inputs alone.

Correlations The observed spatial differences in correlation (Figure 5.1) can be explained by the nature of land surface conditions and their interaction with radar signals. In areas dominated by sparse vegetation, cropland, shrubland, or grassland, the moderate to good correlation between LAI and backscatter likely results from seasonal vegetation changes that affect canopy structure and surface roughness — factors to which radar is sensitive. In contrast, weak or even negative correlations in the arid region between Lake Torrens and Lake Frome may be caused by dry, rocky soils and subsurface scattering effects that distort the relationship between vegetation and radar response.

Evergreen forest areas show consistently low or negative correlation between LAI and backscatter, which can be attributed to signal saturation in dense canopies. In such conditions, increases in leaf area no longer lead to measurable changes in backscatter. Moreover, volume scattering and complex structural interactions dominate the radar signal, making LAI a poor predictor.

The correlation between backscatter and soil moisture is strongest in agricultural areas, where well-defined moisture cycles and relatively sparse vegetation allow the radar to respond clearly to surface wetness. In contrast, the same arid region between the salt lakes again shows weak or negative correlation values, highlighting the difficulty of capturing meaningful soil moisture dynamics under such conditions.

5.3.1 Environmental Characteristics of Key Locations

In the following, three selected locations are examined in more detail to highlight their specific characteristics and how these may influence the interpretation of input variables and backscatter observations.

In and around Salt Lakes In the region between Lake Torrens and Lake Frome (this area is shown in Figure 5.7), the vegetation cover is extremely sparse and consists primarily of salt-tolerant species such as samphire. Due to their small or absent leaf surfaces, these plants contribute only minimally to the Leaf Area Index (LAI). Combined with the extreme soil conditions — characterized by a thin salt crust overlying soft, water-impermeable clay — this results in overall very low and weakly dynamic LAI time series, which are of limited value for modelling or interpreting soil moisture dynamics.

Similarly, in salt lakes such as Lake Torrens and Lake Frome, the C-band backscatter signal recorded by the ASCAT sensor is almost entirely reflected at the surface, especially when dry salt crusts are present. As a result, the signal provides little to no information about subsurface moisture or hydrological changes, which severely limits its usefulness for soil moisture estimation in these environments.

As outlined in the background chapter on remote sensing, specularly reflecting surfaces such as dry salt crusts or calm open water bodies tend to deflect most of the microwave energy away from the sensor. This leads to very low backscatter values (σ^0) that do not correlate with either soil moisture or vegetation conditions.

Taken together, neither the input variables — soil moisture and leaf area index, nor the target variable backscatter provide valid or interpretable measurements in this region.

Urban regions In the areas shown in Figure 5.12 and Figure 5.8, a clear temporal trend in the backscatter signal can be observed: it increases continuously and, in some cases, quite markedly over the study period. The model predictions are only able to partially reproduce this trend — while there is a slight increase in some locations, in others the predictions remain stable over time. This suggests that the input data, particularly soil moisture and Leaf Area Index, reflect the actual surface changes, but only to a limited extent (or in some cases hardly at all).

During the study period, Melbourne experienced substantial urban growth. The metropolitan population increased from approximately 3.6 million in 2007 to over 4.9 million by 2020, with reaching around 5.3 million by 2024 [Aus21] [Mac24]. In response to this demographic development, the Victorian state government implemented significant changes to the Urban Growth Boundary (UGB) in 2010 and again in 2012, adding approximately 6,000 hectares of land to accommodate residential and commercial expansion [Vic17].

To investigate this situation, optical satellite imagery from 2007 and 2020 (Figure 5.15) was examined for the Melbourne region. The comparison reveals urban expansion, including new residential and industrial areas, a denser road network, and increased surface sealing. These developments result in greater surface roughness and the emergence of radar-relevant structures (e.g., building corners) that enhance C-band radar backscatter. However, these structural changes are not explicitly captured by standard input variables such as soil moisture or LAI, which explains why the model predictions only partially reflect the observed backscatter trends.

The hypothesis, therefore, is that the observed increase in backscatter in that case is primarily driven by ongoing urbanization and the sealing of surfaces in and around Melbourne.

Responsive open landscape The investigated grid point at 32.37 ° S / 144.22 ° E (Figure 5.10) is located in western New South Wales (Australia), between the towns of Ivanhoe and Wilcannia. It comprises a mixture of four land cover types: mosaic of trees and shrubs with herbaceous vegetation, grassland, sparse vegetation, and pure herbaceous cover. These types share the characteristic of open to semi-open vegetation structures with a high proportion of herbaceous plants. As a result, they are particularly sensitive to climatic variability such as rainfall events or drought periods.

The low to moderate vegetation height allows for direct observation of the soil surface by radar, facilitating a strong coupling between soil moisture and backscatter. At the

(a) Melbourne 31.12.2007 (b) Melbourne 31.12.2020

Figure 5.15: Optical imagery of urban Expansion in and around Melbourne 2007 vs 2020 (Source: Google Earth Pro [Goo24])

same time, the open structure supports pronounced seasonal dynamics in vegetation development, as reflected in the LAI. This combination of factors makes the site particularly suitable for analyzing the relationships between soil moisture, vegetation, and radar backscatter. In the following, the time series shown in Figure 5.10 is analyzed more closely.

The years 2010–2012 and 2016 show a consistent pattern in the interaction between soil moisture, vegetation, and radar backscatter: In both periods, above-average rainfall led to a marked increase in soil moisture, accompanied by vegetation growth (as reflected in the LAI) and a corresponding rise in the backscatter signal. While soil moisture and LAI declined relatively quickly after the end of the wet phases, the backscatter signal decreased much more slowly. This delayed response can likely be explained by the radar's sensitivity to near-surface moisture residues, increased surface roughness due to vegetation remains, or structural changes in the soil surface. Such effects can sustain elevated backscatter levels even after soil moisture and LAI have already returned to lower values.

In contrast, the period from 2020 to 2023 exhibits a different pattern: although soil moisture increased in the form of numerous short-term peaks without reaching a sustained high level, the backscatter signal shows a continuous upward trend throughout this period. At the same time, LAI follows a clear seasonal cycle but reaches increasingly higher peak values over the years, particularly in 2023. A possible explanation for this behavior lies in the cumulative impact of repeated rainfall events during consecutive La Niña years (a climatic phase characterized by above-average rainfall in parts of Australia due to changes in Pacific Ocean and atmospheric circulation). Even if the soil periodically dries out, near-surface moisture residues, an increasing cover of organic material, or structural changes to the soil surface — such as erosion, surface crusting, or increased roughness — may contribute to a steadily rising backscatter signal. The combination of these factors could explain the observed continuous increase in backscatter, even though a similar

trend is not visible in the soil moisture data.

Overall, this development highlights the complex interactions between rainfall patterns, vegetation dynamics, soil surface structure, and radar backscatter over longer time scales.

5.4 Limitations and Critical Reflection

Despite the careful design and implementation of this study, certain limitations should be acknowledged when interpreting the results.

First, all input data used, such as ASCAT backscatter and ERA5-Land soil moisture, are based on remote sensing systems or model simulations. These data sets are highly valuable but subject to measurement errors, modeling assumptions, and system-specific constraints. For example, the soil moisture product of ERA5-Land is bounded by a minimum threshold, which can result in under-representation of extreme dry conditions in arid environments.

From a methodological perspective, the selection of models was intentionally limited to three established regression approaches, and hyperparameter values were fixed globally. More complex or locally adapted models could not be tested due to computational and time constraints. Similarly, only one third of the available grid points within the defined spatial domain were included in the analysis to maintain feasible runtime. However, the selected sample was spatially balanced and representative of the overall region.

The interdisciplinary nature of this research also presents challenges. As a data scientist, I do not have in-depth expertise in geoscience or remote sensing. However, a solid understanding of the physical processes involved — as well as of satellite sensor characteristics — is essential for correctly interpreting and modeling environmental variables. Ongoing collaboration and exchange between disciplines is therefore crucial to ensure the validity and robustness of scientific conclusions.

Lastly, caution is needed when interpreting local results. Individual grid points may exhibit striking or unusual patterns, but these cannot be readily generalized to other areas. The high spatial heterogeneity of land surface characteristics — such as land cover, climate, and land use — makes direct extrapolation difficult. It is therefore essential to keep the generalizability of interpretations in mind when analyzing localized phenomena and to always consider them within a broader spatial context.

The results presented in this study apply exclusively to the selected study area and observation period. Generalization to other regions or time frames should be approached with caution.

CHAPTER 6

Conclusion

The aim of this study was to model the ASCAT backscatter signal using a machine learning approach and to investigate how trends in observations, model predictions, and residuals can reveal spatial and temporal drift patterns. The two research questions guided the analytical framework and are answered below.

Research Question 1: To what extent can the ASCAT backscatter signal be estimated with a machine learning model, evaluated based on rRMSE, rMAE, Pearson correlation coefficient, and runtime?

The results show that the selected machine learning model is capable of approximating the backscatter signal with satisfactory accuracy. The average error metrics (rRMSE and rMAE) fall within an acceptable range, and there is a clear correlation between predicted and observed values. Moreover, the computational performance remains practical, even for larger datasets, supporting the scalability of the approach to broader regional analyses.

Research Question 2: To what extent can Mann-Kendall test results for ASCAT backscatter, model predictions, and residuals be used to identify spatial and temporal drift patterns that may reflect long-term land cover change?

The findings demonstrate that combining Mann-Kendall trend tests across observations, predictions, and residuals is an effective method for distinguishing different types of drift. While trends in the backscatter signal alone often reflect environmental change, only the inclusion of residual trends allows for an assessment of model stability over time. The resulting trend categories and segmentation enable a spatially explicit identification of areas potentially affected by long-term shifts in vegetation structure, land cover, or soil conditions. Analyses of selected example time series confirm that the method can capture both gradual changes and abrupt events and help in interpreting their potential causes.

In conclusion, this study shows that the integration of machine learning and trend analysis provides a robust framework for both modeling radar backscatter and detecting

longer-term environmental dynamics. The proposed approach was successfully applied across diverse landscapes ranging from open grasslands and shrublands to forested areas, agricultural land, and urban regions, and offers valuable potential for systematic monitoring of environmental change using remote sensing time series.

6.1 Future Work

While this study provides valuable insights into long-term trends in the backscatter signal and its relationship with soil moisture and LAI, several avenues for future research remain.

One important aspect concerns the seasonal dynamics, which were clearly present in the data but not the primary focus of this work. Future studies could apply models such as SARIMAX or other approaches capable of explicitly capturing seasonal components to better reflect seasonal variability. Due to computational constraints and limited time, such models were not explored in this study. Moreover, alternative resampling or aggregation strategies for input data or residuals could help to reveal additional seasonal structures and assess their impact on trend detection.

Another promising direction involves the analysis of outliers. Outliers are common in environmental datasets but were not systematically addressed here. A dedicated examination of outliers – or their exclusion in certain contexts – could improve model accuracy and reduce potential distortions in trend estimations.

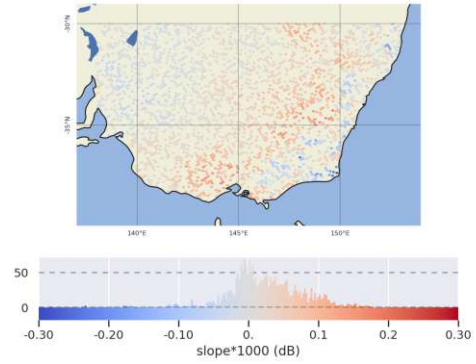
Additionally, the thresholds used for trend classification in this study were manually selected. Although sensitivity analyses suggest that moderate changes in thresholds have only minor effects on overall results, automated or dynamic thresholding techniques could enhance the robustness and reproducibility of future classifications.

This analysis was conducted on a selected study area in Australia, chosen primarily for computational feasibility. A global-scale analysis would be a highly interesting extension. Preliminary explorations prior to defining the study area indicated that some regions – such as parts of France – show almost no long-term backscatter trends, while other regions may exhibit even stronger changes than those observed in the selected area.

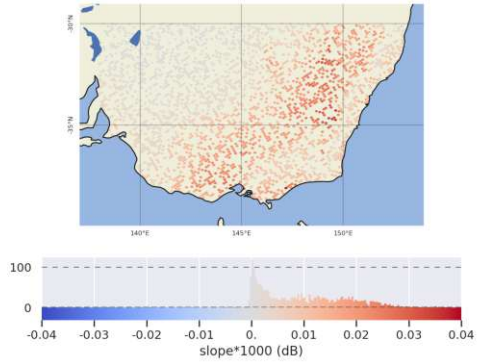
Finally, deeper feature selection and expansion could further improve the prediction of backscatter. Incorporating additional explanatory variables may lead to better model performance, particularly at grid points where current predictions are weak or where strong trends in the observations are not adequately captured by the model. This could support a more precise understanding of the underlying drivers and improve model generalizability.

CHAPTER 7

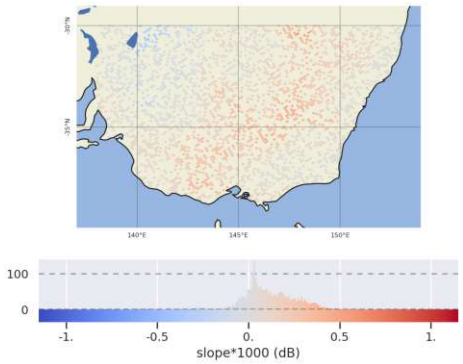
Appendix



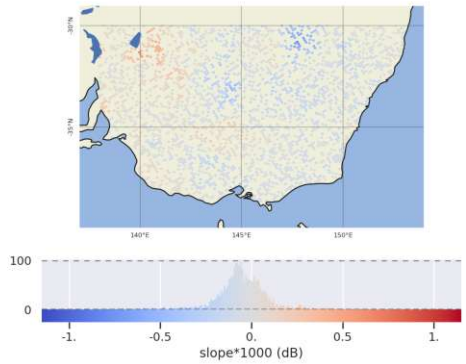
(a) Slope – LAI



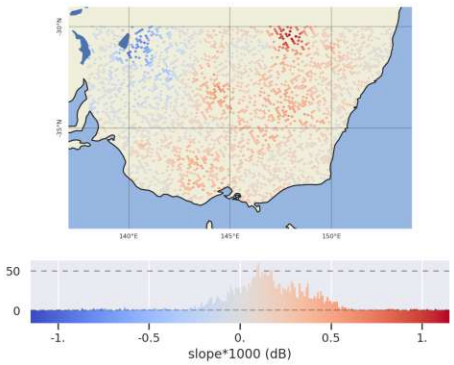
(b) Slope – SM



(c) Slope – Prediction



(d) Slope – Residuals



(e) Slope – backscatter

Figure 7.1: Sen's slope for each variable across the study area.

Overview of Generative AI Tools Used

In the course of this work, AI-based tools were used to support the writing and editing process. For translation tasks, the web-based tool *DeepL Translator* (accessed July 2025) was employed due to its high accuracy and fluency [Dee25]. Furthermore, *ChatGPT-4o* by OpenAI (May 2024 release) was used via the platform `chat.openai.com` to assist with rephrasing statements, ensuring grammatical correctness, and eliminating spelling mistakes [Ope24]. These tools served as support and did not replace critical human judgment.

List of Figures

List of Tables

Bibliography

[Aus21] Australian Bureau of Statistics. Regional population – victoria. https://www.abs.gov.au/statistics/people/population/regional-population, 2021. Accessed: July 2025.

[BDPCV+13] Jose M. Bioucas-Dias, Antonio Plaza, Gustavo Camps-Valls, Paul Scheunders, Nasser Nasrabadi, and Jocelyn Chanussot. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, June 2013.

[BFLS98] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E. Schapire. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5), October 1998.

[BG09] Feras Batah and Sharad Gore. Ridge regression estimator: combining unbiased and ordinary ridge regression methods of estimation. *Surveys in Mathematics and its Applications*, 10 2009.

[Bre96] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Bre01] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[BSJ+19] Ebrahim Babaeian, Morteza Sadeghi, Scott B. Jones, Carsten Montzka, Harry Vereecken, and Markus Tuller. Ground, Proximal, and Satellite Remote Sensing of Soil Moisture. *Reviews of Geophysics*, 57(2):530–616, June 2019.

[CD14] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, June 2014.

[CEB+07] D. Richard Cutler, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology*, 88(11):2783–2792, November 2007.

[DDJC+12] Wouter Dorigo, Richard De Jeu, Daniel Chung, Robert Parinussa, Yi Liu, Wolfgang Wagner, and Diego Fernández-Prieto. Evaluating global trends (1988–2010) in harmonized multi-satellite surface soil moisture. *Geophysical Research Letters*, 39(18):2012GL052988, September 2012.

[Dee25] DeepL GmbH. DeepL Translator (Web Version). `https://www.deepl.com/translator`, 2025. Accessed: July 15, 2025.

[DGD+15] W.A. Dorigo, A. Gruber, R.A.M. De Jeu, W. Wagner, T. Stacke, A. Loew, C. Albergel, L. Brocca, D. Chung, R.M. Parinussa, and R. Kidd. Evaluation of the esa cci soil moisture product using ground-based observations. *Remote Sensing of Environment*, 162:380–395, 2015.

[DMK+24] Xulong Duan, Ahsen Maqsoom, Umer Khalil, Bilal Aslam, Talal Amjad, Rana Faisal Tufail, Saad S. Alarifi, and Aqil Tariq. Enhancing soil moisture retrieval in semi-arid regions using machine learning algorithms and remote sensing data. *Applied Soil Ecology*, 204:105687, December 2024.

[DUADA06] Ramón Díaz-Uriarte and Sara Alvarez De Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, December 2006.

[ESDS21] NASA Earth Science Data Systems. Data Latency | NASA Earthdata, May 2021. Publisher: Earth Science Data Systems, NASA.

[FSWA+02] J. Figa-Saldaña, J J.W. Wilson, E. Attema, R. Gelsthorpe, M R Drinkwater, and A. Stoffelen. The advanced scatterometer (ASCAT) on the meteorological operational (MetOp) platform: A follow on for European wind scatterometers. *Canadian Journal of Remote Sensing*, 28(3):404–412, January 2002.

[Fu98] Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[GBME08] R. Grimm, T. Behrens, M. Märker, and H. Elsenbeer. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. *Geoderma*, 146(1-2):102–113, July 2008.

[Goo24] Google. Google earth pro. `https://earth.google.com`, 2024. Imagery accessed July 2025.

[GSW00] R V Gelsthorpe, E Schied, and J J W Wilson. ASCAT – Metop's Advanced Scatterometer. 2000.

[GZB+14] João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.

[Gé17] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2017.

[Ha70] Arthur E. Hoerl and Robert W. Kennard and. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[HK70] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HK06] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, October 2006.

[HMS+13] R. Hollmann, C. J. Merchant, R. Saunders, C. Downy, M. Buchwitz, A. Cazenave, E. Chuvieco, P. Defourny, G. De Leeuw, R. Forsberg, T. Holzer-Popp, F. Paul, S. Sandven, S. Sathyendranath, M. Van Roozendael, and W. Wagner. The ESA Climate Change Initiative: Satellite Data Records for Essential Climate Variables. *Bulletin of the American Meteorological Society*, 94(10):1541–1552, October 2013.

[HMvdW+20] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[HTF09a] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[HTF09b] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[Hun07] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[jup] jupyterlab. GitHub - jupyterlab/jupyterlab: JupyterLab computational environment.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013.

[LKC04] Thomas M. Lillesand, Ralph W. Kiefer, and Jonathan Chipman. Remote Sensing and Image Interpretation. *John Wiley and Sons*, 2004.

[LLD+18] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[LW01] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *Forest*, 23, 11 2001.

[M+10] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[Mac24] Macrotrends LLC. Melbourne metro area population 1950–2024. `https://www.macrotrends.net/global-metrics/cities/206168/melbourne/population`, 2024. Accessed: July 2025.

[MKP21] Ranga Myneni, Yuri Knyazikhin, and Taejin Park. MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V061, 2021.

[MSDAP+21] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, Brecht Martens, Diego G. Miralles, María Piles, Nemesio J. Rodríguez-Fernández, Ervin Zsoter, Carlo Buontempo, and Jean-Noël Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, September 2021.

[noa] Home Page - H SAF Official web site.

[noa25] SciTools/cartopy, August 2025. original-date: 2012-08-03T07:43:59Z.

[Ope24] OpenAI. ChatGPT-4o (May 2024 release). `https://chat.openai.com`, 2024. Accessed: July 15, 2025.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[PYR+24] Jiabin Pu, Kai Yan, Samapriya Roy, Zaichun Zhu, Miina Rautiainen, Yuri Knyazikhin, and Ranga B. Myneni. Sensor-independent LAI/FPAR CDR: reconstructing a global sensor-independent climate data record of MODIS and VIIRS LAI/FPAR from 2000 to 2022. *Earth System Science Data*, 16(1):15–34, January 2024. Publisher: Copernicus GmbH.

[pyt] python. cpython/Lib/os.py at 3.13 python/cpython.

[Qua25] Raphael Quast. raphaelquast/EOmaps, July 2025. original-date: 2021-09-27T09:54:17Z.

[Sec10] UNFCCC. Secretariat. Implementation plan for the global observing system for climate in support of the unfccc. Technical report, the Global Climate Observing System secretariat, 2010.

[Sha22] Amin Shaban, editor. *Satellite Monitoring of Water Resources in the Middle East*. Springer Water. Springer International Publishing, Cham, 2022.

[Sho25] Md Manjurul Hussain Shourov. mmhs013/pyMannKendall, August 2025. original-date: 2019-03-08T08:09:22Z.

[SP10] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[SS02] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[SS04] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[SSDH+22] Xu Shan, Susan Steele-Dunne, Manuel Huber, Sebastian Hahn, Wolfgang Wagner, Bertrand Bonan, Clement Albergel, Jean-Christophe Calvet, Ou Ku, and Sonja Georgievska. Towards constraining soil and vegetation dynamics in land surface models: Modeling ASCAT backscatter incidence-angle dependence with a Deep Neural Network. *Remote Sensing of Environment*, 279:113116, September 2022.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[UAY+20] University of Duhok, Fawaz Aswad, Ali Yousif, University of Duhok, Sayran Ibrahim, and University of Duhok. Trend Analysis Using Mann-kendall And Sen's Slope Estimator Test for Annual And Monthly Rainfall for Sinjar District, Iraq. *The Journal of the University of Duhok*, 23(2):501–508, December 2020.

[Vap95] Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren

Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[Vic17] Victorian Planning Authority. Urban growth boundary – key facts. `https://vpa.vic.gov.au/strategy-guidelines/strategy-guidelines-2/more-information/urban-growth-boundary-key-facts`, 2017. Accessed: July 2025.

[VR20] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[VRD09] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[WBO+17] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

[WHK+13] 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.

[WLH+23] Wolfgang Wagner, Roland Lindorfer, Sebastian Hahn, Hyingglok Kim, Mariette Vreugdenhil, Alexander Gruber, Milan Fischer, and Miroslav Trnka. Global Scale Mapping of Subsurface Scattering Signals Impacting ASCAT Soil Moisture Retrievals, August 2023.

[WLM+22] Wolfgang Wagner, Roland Lindorfer, Thomas Melzer, Sebastian Hahn, Bernhard Bauer-Marschallinger, Keith Morrison, Jean-Christophe Calvet,

Stephen Hobbs, Raphael Quast, Isabella Greimeister-Pfeil, and Mariette Vreugdenhil. Widespread occurrence of anomalous C-band backscatter signals in arid environments caused by subsurface scattering. *Remote Sensing of Environment*, 276:113025, July 2022.

[YPC02] Sheng Yue, Paul Pilon, and George Cavadias. Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259(1-4):254–271, March 2002.

[ZH05] Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, April 2005.