# Data Retrieval in Earth Observation

Course material prepared by

**Wolfgang Wagner**
**Alexander Gruber**
**Sebastian Hahn**
**Mariette Vreugdenhil**
**Thomas Melzer**
**Raphael Quast**

May 2019

# Preface

This script serves as supportive teaching material for the graduate course "Parameter Retrieval from Earth Observation" held at the Technische Universität Wien (TU Wien).

# Contents

# 1 Introduction

The aim of Earth observation (EO) is to obtain information about processes and interactions of the Earth system (e.g., atmosphere, hydrosphere, biosphere, cryosphere, etc.) on a global scale using remote sensing technology supplemented by surveying and in situ techniques. Such processes and interactions can be characterized by various **geophysical data** representing different properties of the land (e.g., surface temperature, biomass, leaf area coverage), water (e.g., ocean color, salinity, suspended matter), and atmosphere (e.g., temperature and moisture profiles at different altitudes). In this lecture you will learn how these geophysical data can be retrieved from calibrated sensor measurements, which is the key to correctly interpret and use EO data.

## 1.1 Notation

This script covers topics from different fields like physics, statistics, numerical mathematics and earth observation, each employing its own terminology and notation. We have endeavoured to maintain a consistent notation throughout this text, though in some cases this was not possible because it would have interfered with established notation or simply would have become too cumbersome. We will use

| Example(s) | Meaning | Font type /Symbol Description |
|---|---|---|
| $x$, $x_i$ | scalar quantity, vector element | small roman |
| $\mathbf{x}$ | vector | small bold |
| $\mathbf{A}$ | matrix | capital bold |
| $X$ | random variable | capital roman |
| $\vec{X}$ | random vector | capital roman + vector arrow |
| $\hat{x}$,$\hat{\mathbf{x}}$ | estimated or observed quantity | symbol + circumflex |
| $\hat{X}$ | estimator (random variable) | capital roman + circumflex |

Table 1.1: Symbols and notation used frequently in this text.

the above conventions whenever possible, in particular when discussing things at an abstract level, e.g., the general forward and backward model in Sec. 1.4. However, when discussing specific models or implementations, we will sometimes switch to a different notation.

## 1.2 Data Collection

The primary data source in EO is **remote sensing**, i.e. measurements of electromagnetic waves mainly from satellites or air planes. Objects of the Earth system can be sensed either through the radiation

emitted by an object itself or radiation of external sources like the sun that is reflected by the object (passive case), or through radiation emitted by an artificial source (e.g., lidar, radar) and reflected by the object (active case). Such measurements can be taken in the visible domain as well as in other regions of the electromagnetic spectrum.

Most physical processes that affect the emission or reflection of electromagnetic waves are frequency dependent. This means that they affect the radiation itself or the interaction between radiation and matter in the various regions of the electromagnetic spectrum in different ways. This **frequency- or wavelength dependency** enables us to derive various geophysical parameters from observations acquired at different wavelengths. For example, Planck's radiation law teaches us that every object radiates waves over the entire electromagnetic spectrum with a total energy proportional to the object's temperature. Wien's displacement law furthermore relates the object's temperature to a radiation intensity peak at a certain wavelength. If we would now observe the radiated energy of an object over a range of different wavelengths and search for the occurring radiation maximum, we could derive the object's temperature. In other words we can retrieve the state of a geophysical variable (in this case the physical temperature) by converting information that is contained in the observed radiation emitted by the object.

Of course, there are many geophysical processes and variables that cannot be observed with remote sensing and/or ground observation techniques due to a lack of tangible measurement capabilities. However, thanks to advances in sensing technologies and our increasing capabilities to combine data from different sensors, platforms and models, more and more EO data products are become available. For instance, multispectral sensors such as Sentinel-2 are sensitive to the reflectance of objects in the visible and the infrared domain. Optical information can be of great use to classify land cover types (e.g., the CORINE land cover map) and the additional information in the infrared domain even allows us to retrieve some information about the healthiness of vegetation (e.g., leaf area index). Other applications of multi- or hyperspectral sensors are, for example, the observation of ocean color, land-, sea-, and cloud temperature measurements, and atmospheric water vapour and aerosol concentration measurements. However, optical waves penetrate most natural objects (but not, e.g., water) only to a depth of the size or smaller than their wavelength, which is in the optical and infrared domain in the order of $nm$ to $\mu m$. Hence, they have the big disadvantages that they can only be used for sensing the very surface of the Earth and objects above it, and that they cannot look through clouds. To overcome this limitation one can use microwave sensors. The most common bands for which microwave sensors are available in space are Ku-, X-, C-, and L-band, whose wavelengths are somewhere in the ranges of 1-2 cm, 2-3 cm, 5-7 cm, and 15-30 cm respectively. Sensors in these frequency domain enable us to observe the Earth independent of cloud coverage and external illuminating sources such as the sun, and also to penetrate the Earth's surface and objects such as vegetation to a certain degree. Microwave sensors are e.g. used for soil moisture and vegetation biomass mapping, the measurement of sea wind directions and speeds, the measurement of snow water equivalent, and sea ice mapping.

## 1.3 Data Pyramid

At the very basic level, EO sensors measure and record electric quantities such as voltage, current or power. In order to become useful these sensor raw data need to be converted in a series of processing steps into geophysical quantities that can be physically interpreted and ingested into models. As illustrated in Figure 1.1 the sensor raw data, usually referred to as Level 0 data, are stepwise converted into geometrically and radiometrically correct measurements in sensor units (Level 1), geophysical variables at the same resolution and location (Level 2), and finally value-added data products mapped on uniform space–time grid scales (Level 3 and Level 4). In this process, the size of the output data in general decreases in each processing step. When novel EO satellite missions are considered, the volume of Level 1 data easily makes up tens of Petabyte, the derived Level 2 to Level 4 data sets are in the more practical range from Giga to Terabyte.

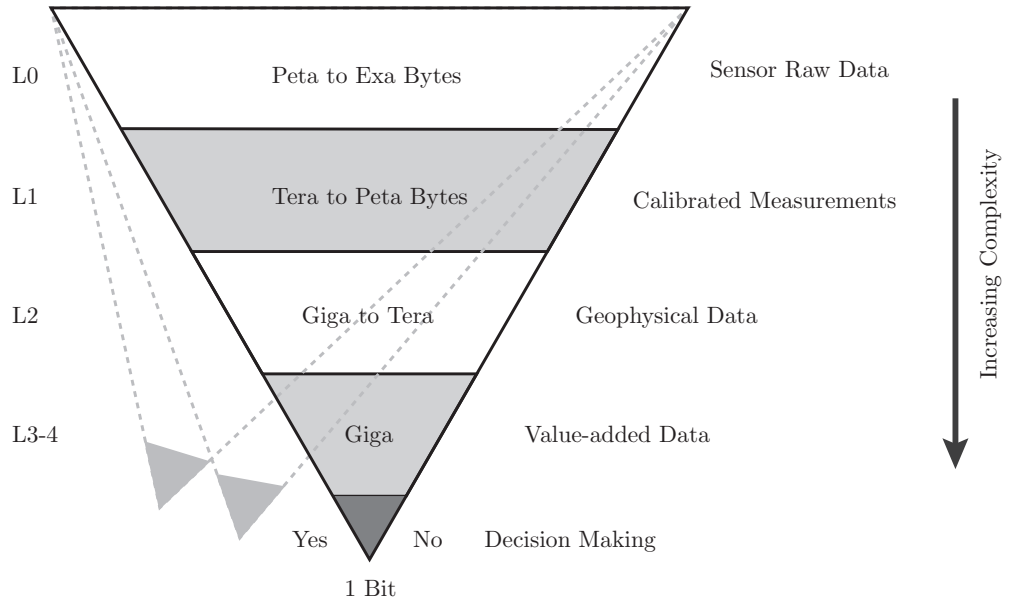



Figure 1.1: Earth observation data pyramid. Note that the same base data may build up diverse pyramids.

The process of transforming the Level 0 data into Level 3 or 4 data is usually very complex, and peculiarities in each processing step may have significant impact on all the higher level data products. This is in fact why experts are needed who are capable of understanding the complete processing chain, starting from the technical details of how the sensors work to the way of how the data products are being used in applications. The high complexity of most EO data processing chains is one reason why one should be careful of thinking of EO data as 'direct measurements'. In fact, for some of EO 'data products' there is so much modelling involved that the usual distinction between data versus models may not be meaningful any longer (e.g. evaporation, gross primary productivity, etc.).

In this course, we **focus on the conversion of Level 1 into Level 2 data**. Examples for Level 1 data are top-of-the atmosphere radiance $\left[W \cdot m^{-2} \cdot sr^{-1}\right]$ in the case of optical multispectral imagers, backscatter coefficient $\left[m^2 \cdot m^{-2}\right]$ in the case of radar or lidar, and brigthness temperature $[K]$ in the

case of microwave radiometers. Examples for Level 2 data are the physical temperature of an object $[K]$, the volumetric soil moisture content $\left[m^3 \cdot m^{-3}\right]$, and the stem volume of a forest stand $\left[m^3 \cdot ha^{-1}\right]$. While Level 1 data are usually well defined and quite accurate, Level 2 data are more uncertain. This is due to the fact that the **derivation of the Level 1 data represents an engineering problem**, whereas one only needs to have accurate information about the sensor and measurement geometry. On the other hand, when deriving Level 2 data one needs to make simplifying assumption about the shape and properties of natural objects (e.g. composition of the atmosphere, geometric structure of the vegetation, etc.).

## 1.4 Data Modelling

The conversion of Level 1 sensor measurements into Level 2 geophysical data requires two complementary modelling steps, namely the construction of a **forward model** and its **inversion** (see Figure 1.2). Particularly the inversion is of central interest in applications (even though in data assimilation the forward model – in this case also called "forward operator" – may be used to simulate the Level 1 data and compare them to the observations).

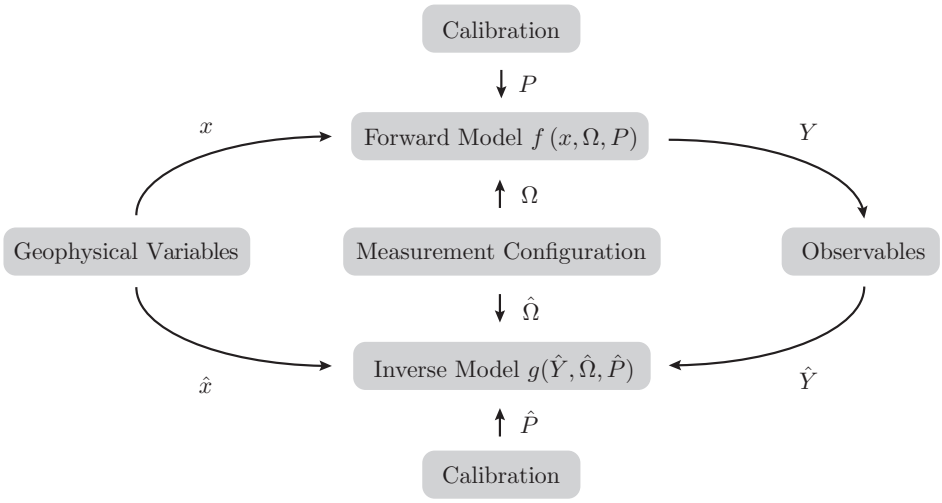



Figure 1.2: Forward and inverse modelling for parameter retrieval.

The forward modelling step summarizes all processes that are involved in the propagation of the radiation from the source to the receiver, including interactions of the waves with illuminated matter such as soil, vegetation canopy or atmospheric particles. Mathematically, a **forward model** can be expressed as

$$\mathbf{y} = f(\mathbf{x}, \mathbf{\Omega}, \mathbf{p}) \tag{1.1}$$

where $\mathbf{y}$ is the state vector of the observable quantities (i.e., Level 1 data), $\mathbf{x}$ is the state vector of the geophysical variables in which we are interested (i.e., Level 2 data), $f(\cdot)$ is a function that relates $\mathbf{x}$ to $\mathbf{y}$ (i.e., the formulation of the wave propagation and interaction processes), $\mathbf{\Omega}$ is a set of controllable

measurement conditions (e.g., the wavelengths, viewing direction, time, sun position, polarization, ...), and $\mathbf{p}$ is a set of model parameters characterizing the forward model[1].

In the second step, the forward model is inverted in order to retrieve the parameters of interest from the observables. This step is called **model inversion** and can be written as

$$\hat{\mathbf{x}} = g(\hat{\mathbf{y}}, \hat{\mathbf{\Omega}}, \hat{\mathbf{p}}) \tag{1.2}$$

$\hat{\mathbf{x}}$ are estimates of the geophysical parameters obtained by a function $g(\cdot)$ that uses the actual measurements $\hat{\mathbf{y}}$ and approximations $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{x}}$ of the measurement conditions and model parameters as input. Note that the function $g(\cdot)$ may be the inverse of the function $f(\cdot)$, i.e. $f^{-1}(\cdot)$, but does not necessarily need to be so. In the latter case, the set of parameters $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{p}}$ used to describe the measurement conditions and model status may be quite different from the ones used in (1.1).

## 1.5 Retrieval Errors

Considering that the interactions of the electromagnetic waves with the atmosphere and Earth's surface are quite complex, the mathematical formulation of a forward model and its inversion is often quite challenging. From adjustment theory we know that the inversion of a system that is driven by $n$ independent parameters $\mathbf{x}$ requires at least $n$ independent measurements $\mathbf{y}$ and a known or approximated relationship between the measurements and the driving variables. In most cases, EO data are very limited in their spatial, temporal, radiometric, and spectral resolution due to limited power supply, available orbit geometries, and various constraints in the sensor design (e.g. the choice of the antenna size). In addition, the atmosphere allows the penetration of electromagnetic waves only in certain spectral regions called atmospheric windows. Hence, the information content of satellite observations and thus the available number of independent measurements is very limited. Furthermore, the huge distance of the satellites and their sometimes very large footprint size involves a large number of geophysical variables and interactions between them in the wave propagation process. This can be, for example, the propagation of waves through a humid or polluted atmosphere, and the mixing of different land cover types (vegetation, soil, urban areas, water surfaces, etc.) in coarse-scale data. In other words, the forward model may get very complex, including a lot of geophysical variables whereas only very few independent measurements are available. Therefore the inversion of the forward model is often an **ill-posed problem** or may even be impossible. In practice, one always has to use auxiliary data or assume empirically observed relationships or conditions, in order to approximate the data formation processes and reduce the number of unknowns.

It is thus obvious that the estimates $\hat{\mathbf{x}}$ as given by (1.2) will never capture the true state of the geophysical variable but will be related to it through some **systematic and random error** components:

$$\hat{\mathbf{x}} = \mathbf{x} + b(\mathbf{x}) + \mathbf{n} \tag{1.3}$$

where $b(\cdot)$ is a systematic bias that can be also of higher order, and $\mathbf{n}$ is a random noise component

[1] Please note that the distinction between 'variables' and 'parameters' is not always straight forward. As a guideline, a variable is an entity that changes with respect to another entity in a given system, and a parameter is any entity that helps characterizing the system.

(whose distribution is normally assumed to be independent of $\mathbf{x}$ and to have zero mean). These retrieval errors partly originate from errors in the measurement process, i.e., deviations of the measurements of the observables $\hat{\mathbf{y}}$ from their true states $\mathbf{y}$, which have the same shape as (1.3). However, a much bigger source for errors are **model imperfections**. To minimize these uncertainties, both forward models $f(\cdot)$ and their inversions $g(\cdot)$ must be calibrated.

## 1.6 Model Calibration

The goal of model calibration is finding a set of model parameters that minimize the difference between the model predictions and the observed values of interest. The model calibration process is typically an iterative procedure of parameter evaluation and refinement, but due to model imperfections and unpreventable errors during the measurement process a perfect equality between the model output and the available data is virtually impossible. Therefore, a pragmatic approach is to pre-define acceptable ranges of deviations, which still ensure an adequate reproduction of the observed process behaviour.

How well the model output fits to the observed data is usually analyzed using qualitative and quantitative evaluation methods. In that respect, goodness-of-fit criteria are part of quantitative evaluation methods, whereas qualitative assessment methods are typically based on visual inspection using various kinds of illustrations (e.g., scatterplots, histograms, cumulative distribution functions,...). This evaluation process is usually referred to as **model validation**, which represents an important step in the course of finding the optimal set of model parameters. Ideally, the validity of model predictions are tested against an independent data set and therefore, usually a subdivision into a calibration and validation (Cal/Val) set is made. The calibration set is used for estimating the model parameters and the validation set is used for accuracy evaluation. The latter will be topic of Chapter 6.

Model calibration needs to be considered whenever the predictive validity of the model is threatened by spatial and/or temporal variations and heterogeneities, regardless of whether it's a numerical or an analytical model. In this way the model predictions will be valid under a wide range of circumstances. However, this does not mean that non-calibrated models are useless. They still can be suitable for an initial assessment and to obtain first results.

### 1.6.1 Model Parameters

Model parameters are quantitative values that characterize the model. They can represent **physical** (e.g., freezing point, relaxation frequency, soil porosity at plot scale), **"effective"** (e.g., soil porosity at 50 km scale) or even **fictitious** quantities acting on different temporal and/or spatial scales. Of course, the ideal scenario is when the model parameters are physical quantities that can also be measured in the field or laboratory. Yet, in most cases model parameters are effective or fictitious quantities that cannot be assessed by independent means. In these two cases they can be thought of as adjustment buttons or tuning knobs of the model. The decision and definition of these parameters is important for a complete and relevant specification of the model. Model parameters are sometimes the result of replacing processes that are, for example, too small or complex to be physically represented. The overall model complexity is usually linked to the number of model parameters (more parameters =

higher complexity). And when a model is excessively complex **overfitting** may occur, which can lead to a poor predictive model performance.

Model parameters can be determined in various ways: (i) experience, educated guess, look-up tables or literature review; (ii) field and laboratory experiments; (iii) by fitting a model to observed data (e.g., linear regression); (iv) measured or estimated from another data set; (v) function of other parameters. In most of these cases prior information or knowledge on the model parameters (e.g., range, constraints) is already given, but not all properties are known from the beginning and may depend on the model design or quality of the calibration data set (e.g., uncertainty, sensitivity). It is important to note that model parameters do not have universal validity and may easily change if (i) data come at a different spatio-temporal resolution, (ii) the model changes; (iii) the calibration data set is different; (iv) the model validation approach changes.

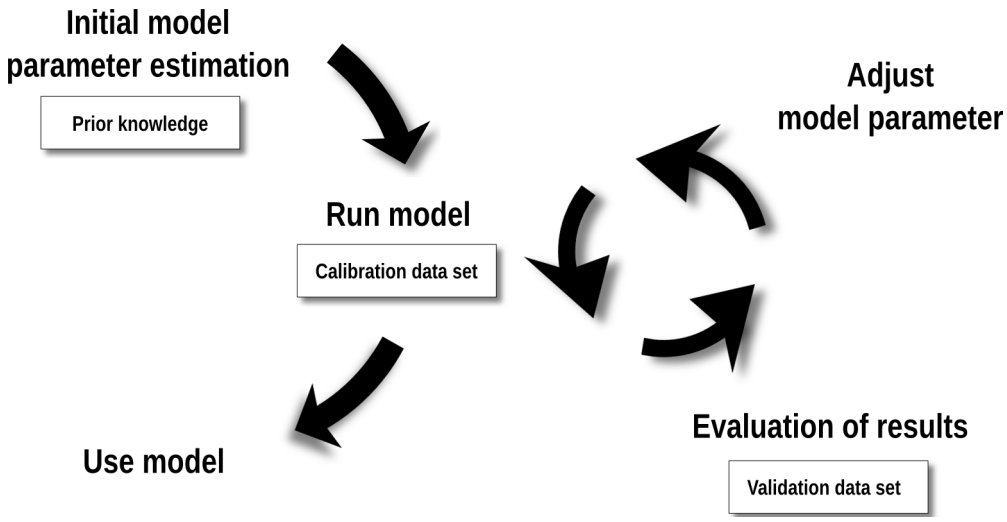



Figure 1.3: Diagram of the model calibration process.

## 1.6.2 Model Calibration Process

The basic steps of model calibration are shown in Figure 1.3. Prior knowledge on the model quantities (e.g., ranges, constraints) should be explicitly incorporated in the model calibration process, which typically allows an initial estimation of the parameters. A sufficiently large and representative calibration data set is important for meaningful model simulations. In any case it is essential to detect and remove any discrepancies (e.g., outliers and/or trends) and noise in the calibration and validation data set before starting the calibration procedure. The model performance will be evaluated after each model run either qualitatively by visually inspecting the agreement between the model predictions and the validation data set or quantitatively in terms of goodness-of-fit measures. The deviation $C$ between a set of model predictions $P^k = \{P_1^k, \ldots, P_N^k\}$ and the corresponding validation data set $O^k = \{O_1^k, \ldots, O_N^k\}$ for the $k$th model run can be expressed as:

$$C^k = \delta\left(P^k, O^k\right), \tag{1.4}$$

whereby $\delta$ denotes a goodness-of-fit criterion. Typically, we have $P_i^k = \hat{x}_i^k$ with

$$\hat{\mathbf{x}}^k = g(\hat{\mathbf{y}}^k, \hat{\mathbf{\Omega}}^k, \hat{\mathbf{p}}^k). \tag{1.5}$$

After each model run various **goodness-of-fit criteria** can be calculated and evaluated. Table 1.2 gives an overview of common performance metrics, each of them highlighting only a specific aspect. Consequently, the choice of the goodness-of-fit criterion should reflect the intended use of the model and at the same time consider the characteristics of the data.

| **Criterion** | **Symbol** | **Formulation** |
|---|---|---|
| Average error (or bias) | AE | $\frac{1}{n}\sum_{i=1}^{N} P_i - O_i = \bar{P} - \bar{O}$ |
| Normalised average error | NAE | $\left(\bar{P} - \bar{O}\right)/\bar{O}$ |
| Fractional mean bias | FB | $\left(\bar{P} - \bar{O}\right) / \left(0.5\left(\bar{P} + \bar{O}\right)\right)$ |
| Variance ratio | VR | $s_P^2/s_O^2$ |
| Fractional variance ratio | VR | $\left(s_P^2 - s_O^2\right) / \left(0.5\left(s_P^2 + s_O^2\right)\right)$ |
| Root mean square error | RMSE | $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - O_i\right)^2}$ |
| Normalised RMSE | NRMSE | $RMSE/\bar{O}$ |
| Mean absolute error | MAE | $\frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$ |
| Normalised mean absolute error | NMAE | $MAE/\bar{O}$ |
| Median absolute error | MedAE | $\text{median}\left(\lvert P_i - O_i\rvert\right)$ |
| Upper quartile absolute error | UppAE | $75^{\text{th}}\text{percentile}\left(\lvert P_i - O_i\rvert\right)$ |
| Maximal absolute error | MaxAE | $\max\left(\lvert P_i - O_i\rvert\right)$ |
| Ratio of scatter | RS | $\sum_{i=1}^{N}\left(O_i - \bar{O}\right)^2 / \sum_{i=1}^{N}\left(P_i - \bar{O}\right)^2$ |

Table 1.2: An overview of goodness-of-fit criteria comparing model prediction and observations [10]. $P_i$ and $O_i$ represent the predicted and observed value and $\bar{P}$, $\bar{O}$ and $s_P^2$, $s_O^2$ are their means and variances.

In practice, a frequently used goodness-of-fit criterion is the Root Mean Square Error (RMSE), as well as the Average Error (AE). However, due to their lack of robustness (i.e., the ability to tolerate outliers) quantile-based criteria (like MedAE, UppAE) have an advantage. Thus, it is recommended to compute robust goodness-of-fit criteria as well, particularly when the distribution and characteristics of the data are not well known.

How the parameters are tuned after each model run may be deduced from the experience gained in the previous model simulations or in a trial-and-error procedure. However, also (semi-) automatic approaches, like random searches or directed-search algorithms, can be used. It is obvious that during the entire process insight, intuition and sound judgement are important aspects. The overall success of calibrating the model generally depends on the amount and quality of the available data in relation to the complexity of the model. But also the limits which can be reached can vary from application to application [10].

At the end, if an acceptable model has been obtained it can be used for the intended purposes, e.g.,

uncertainty analysis, decision making, scenario analysis, prediction. And finally, all information gained during the calibration of the model (e.g., parameter uncertainty, model strengths and deficiencies) should be considered for these applications to infer reliable conclusions.

## 1.7 Data Descriptions

When all input data are defined and the inversion model is known and calibrated, then the targeted geophysical variable can be retrieved from the Level 1 data. However, the derived Level 2 data set shall not only contain the geophysical variable itself, but must inform the users about uncertainties and caveats of the data product as good as possible. Therefore, a Level 2 data product shall in the optimum scenario contain the following data groups (each containing one or more data fields):

- **Acquisition Time**: Day and time given e.g. in Coordinated Universal Time
- **Location**: latitude/longitude or $x/y$ in the chosen cartographic projection and grid
- **Geophysical variable**: Retrieved variable in its physical units
- **Uncertainty estimate**: Estimates of the uncertainty using e.g. error propagation or other error models
- **Mask**: Indicates that the retrieval is not possible due to lack of sensitivity of the measured signal to the variable of interest
- **Quality Flags**: Flags indicating environmental conditions under which the retrieval is likely to fail
- **Auxiliary Data**: Selected data fields describing the satellite and sensor, and the retrieval algorithm and its input data

The retrieval process and output data groups/fields are illustrated in 1.4. Note that model calibration is done off-line, when setting up the processing system. For the calibration one typically needs historic data, which is why the use of **data cubes** is advantageous.

## 1.8 Aims of this Course

This course shall teach you of how to correctly use higher level data products derived from Earth observation data. For this it is necessary that you understand the entire data retrieval process, from the definition of a forward model, the selection to an inversion technique, the modelling of retrieval uncertainties, to the validation of the derived geophysical variables.

Rather than providing a comprehensive description of all available sensing techniques, modelling approaches and retrievable variables, this course aims to illustrate the principles of the problem by going into depth for one distinct problem: The retrieval of surface soil moisture data from the Advanced Scatterometer (ASCAT), which is an active microwave sensor operating in C-band (5.3 GHz). But of

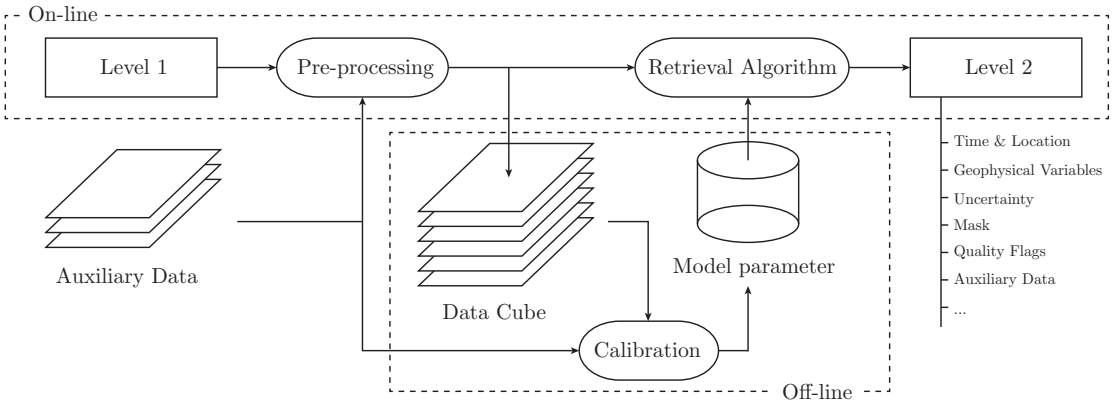



Figure 1.4: Level 2 data retrieval process.

course, the basic considerations hold for any other geophysical parameter.

This script is structured as follows:

- Chapter 2 discusses how science approaches the problem of constructing, selecting and validating models
- Chapter 3 provides some background on the ASCAT instrument and basic physical concepts needed in the following chapters
- Chapter 4 illustrates different approaches to the forward modelling problem by describing theoretical, empirical and hybrid approaches to model backscatter from vegetated soil.
- Chapter 5 discusses a direct inversion scheme and non-linear optimisation techniques
- Chapter 6 gives an overview of strategies to estimate the impact of the imperfectness and simplifications of modelling and retrieval techniques.

# 2 Philosophy of Science

## 2.1 Why it Matters

The day-to-day work of an earth observation expert engaged in the retrieval of geophysical parameters is shaped by the scientific and technical challenges that need to be solved (algorithm development, software implementation, data issues, etc.). This work is seemingly free of any philosophical considerations. However, this is not true since many of the basic decisions that need to be made when developing forward models and retrieval services are strongly influenced by the philosophical pre-dispositions of the involved scientist and engineers. How strongly different philosophical world views shape the seemingly objective and quite technical nature of earth observation becomes often only clear when representatives of different scientific schools start engaging in a scientific controversy about whose method is the "best". Such controversies may be latent for many years, but at latest they erupt openly during validation activities that aim to inter-compare and quantity the errors of competing data products.

While, without question, some retrieval techniques may work better than others, it is nonetheless often the case that quite different models and algorithms give comparable results. This is known as the **equifinality** principle which states that, given only a limited number of measurements available for model validation, there may well be different model structures and parameter data sets that explain the remote sensing observations equally well [26].

When two or more methods give quasi indistinguishable results then the question is whether there are any additional criteria that allow selecting one method over the others? From a practical point of view the answer is obvious: there are good reasons to select the simplest algorithm because simplicity translates directly into ease and lower costs of implementation. Even though it is not an irrefutable principle of logic, one may also refer to a principle of parsimony known as **Occam's razor**. Occam's razor states that among competing hypotheses, the hypothesis with the fewest assumptions should be selected, i.e. one should stick to simpler theories until simplicity can be traded for greater explanatory power.

Some knowledge about important discourses in the **philosophy of science** is thus helpful for scientists and even engineers working in Earth observation. This is particularly the case during controversies that question on a fundamental level the validity of the selected retrieval methods. Students attending this course are encouraged to make themselves familiar with the topic beyond that what is discussed in the rest of this chapter. Recommended readings are the following three highly influential books:

- *The Logic of Scientific Discovery* by Karl Popper[1] ("Falsifiability")

[1]The book *Logik der Forschung* was first published in 1935 by the Julius Springer Verlag, Wien. The first English version was published in 1959 by Hutchinson & Co.

- *The Structure of Scientific Revolutions* by Thomas S. Kuhn[2] ("Scientific Paradigms")
- *Against Method* by Paul Feyerabend[3] ("Anything goes")

## 2.2 Images of Science

As discussed by Barker and Kitcher, philosophers have tried to elaborate general conceptions of science – "images of science" as they call it [2] : »Many of those images are popular among scientists (although scientists tend to be unworried about the details of their favourite image), and they are prominent in newspaper discussions of science, as well as in books written for the general public. You are probably familiar with most of them.

The images are often taken to describe how science actually is. When someone (historian, sociologist, or journalist) discovers that a piece of work fails to fit the preferred image, though, there is often a significant shift in perspective. The image is no longer seen as descriptive, but as normative: This is how science should be. Despite this shift, a connection with description usually remains. The problematic work is a deviation from the proper course of scientific activity, a course taken to be exemplified in the overwhelming majority of scientific investigations.

**Image 1:** One of the most enduring images, stemming from Bacon and the charter of the early Royal Society, and persisting into the present, views science as a reliable means of accumulating useful knowledge. The *scientific method starts with observations*, prudently generalizing from them to yield more general conclusions. The patterns that emerge from this activity can be applied to predict and control the course of nature in ways that improve the human lot. Over the years, decades, and centuries, the community of scientists builds a vast edifice of useful knowledge.

**Image 2:** A different image, made popular by the influence of the twentieth century philosopher Karl Popper, opposes the thought of beginning from observation. Scientists inevitably need some guiding idea – they cannot simply obey the command "Observe!" If you were told to observe, you would have to ask just what you were supposed to be observing! They *begin from conjectures* (the bolder, the better), embarking on a daring adventure – an exploration of territory beyond the frontiers of knowledge. Yet their voyages are tempered by thorough self-criticism. No hypothesis, however plausible or apparently well supported by the available evidence, is beyond criticism. With luck, the adventurers may arrive at useful truths, valuable because they contribute to human goals or because they are worth apprehending for their own sake, yet *acceptance of these must always be provisional.* Science is to be celebrated both for its delivery of provisional truths and because it is an expression of human striving and human freedom.

**Image 3:** In a third image, science is the *epitome of rationality*, the best way of using our minds to make sense of the world. Although our scientific conclusions are always revisable, there are objective logical relations between evidence (on the one hand) and hypotheses and theories (on the other). So it is legitimate to speak of some of the great scientific achievements – the fundamental principles of quantum mechanics, our knowledge of laws of chemical combination, the molecular basis of inheritance –

[2]First published in 1962 by The University of Chicaco Press.
[3]First published in 1975 by New Left Books.

as objectively well established: not so certain that they are immune to revision, but as close to certainty as we can ever come. Science aims at a systematically unified and complete account of nature, and the history of the past four centuries can be understood as making significant progress toward reaching that goal. ...

**Image 4:** According to that image, *science is a thoroughly human activity*, one of many in which people engage. As in other domains of human affairs, the practice of science is shaped by social interaction. Scientific choices are affected by scientists' social context, personal interest, and ambition, and by their broader beliefs about everything, including religion and politics. On some versions of this image science is neither cumulative nor progressive. It is better viewed as a series of exercises in articulating dogma, punctuated by changes in doctrinal fashion. ...«

## 2.3 Some Important Thinkers

A historical treatment of the main ideas in the philosophy of science would have to start from the Greek philosophers Plato (428/7–348/7 BC) and Aristotle (384–322 BC), include scholars from the Medieval period such as Roger Bacon (c.1214–92) and William of Ockham (c.1280–1349), and continue with some of the scientist and philosophers who triggered the scientific revolution such as Nicolaus Copernicus (1473–1543), Johannes Kepler (1571–1630), Galileo Galilei (1564–1642), René Descartes (1596–1650), and Isaac Newton (1642–1727). Such a comprehensive discussion is clearly out of scope for this course. But in order to make the students familiar with at least some important philosophical arguments that continue to be of relevance in today's scientific discourses, several (subjectively selected) important thinkers are shortly introduced in the following. All text is taken from the excellent book *A Historical Introduction to the Philosophy of Science* by John Losee [11], which is highly recommended for further reading. Most images are taken from *Wikimedia Commons*.

#### Gottfried Wilhelm Leibniz (1646-1716)

*"Gottfried Wilhelm Leibniz was the son of the Professor of Moral Philosophy at the University of Leipzig. An omnivorous reader, Leibniz studied philosophy at his father's university, and jurisprudence at Jena. Leibniz spent much of his adult life at court, first at Mainz and later at Hannover. During this service he was entrusted with diplomatic missions which enabled him to establish contacts with numerous political and intellectual leaders. Leibniz worked tirelessly for legal reform, for Protestant religious unification, and for the advancement of science and technology. He maintained extensive correspondences with the leading thinkers of his day and actively promoted scientific co-operation by means of his membership in the Royal Society, the French Academy, and the Prussian Academy. It is ironic that his later years were marked by bitter polemics with the followers of Newton over priorities in the invention of the calculus."* [p. 86]

"Leibniz was a practising scientist who made important contributions to mathematics and physics. And he confidently extrapolated from his scientific findings to metaphysical assertions. Indeed, Leibniz set up a two-way commerce between scientific theories and **metaphysical principles**. Not only did

he support his metaphysical principles by analogical arguments based on scientific theories, he also employed metaphysical principles to direct the search for scientific laws." [p. 89]

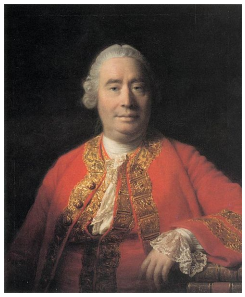
"Leibniz sought to interpret the universe in such a way that the mechanistic world-view, which focuses on material and efficient causation, is supported by teleological considerations. **Extremum principles**, **conservation principles**, and the principle of continuity were well suited to effect the desired integration of the mechanistic and teleological standpoints. In the case of extremum principles, for example, the teleological connotation is that natural processes occur in certain ways in order that certain quantities achieve a minimum (or maximum) value. It is a short step, and one that Leibniz was anxious to take, to the position that a Perfect Being created the universe in such a way that natural processes satisfy these principles." [p. 90]

"For instance, he argued that because nature always selects the easiest, or most direct, course of action from among a set of alternatives, the passage of a light ray from one medium into another obeys Snell's Law. Leibniz derived Snell's Law by applying the differential calculus which he had developed to the condition that the "path difficulty" of the ray (the path length times the resistance of the medium) is a minimum. And he took his success in this enterprise as support for the metaphysical principle that God governs the universe in such a way that a **maximum of "simplicity"** and **"perfection"** be realized." [p. 89]

"Locke had bemoaned the fact that we cannot advance from a knowledge of the association of qualities to a knowledge of the internal consitutions or "real essences" of things. Leibniz took quite a different attitude towards this epistemological gap. He conceded that, at the level of phenomena, **scientists can reach only probability**, or "moral certainty". But he was convinced that the general metaphysical principles he had formulated were necessary truths." [p. 90]

### David Hume (1711-76)

*"David Hume enrolled to study law at the University of Edinburgh, but left without receiving a degree. He neglected his legal studies for the pursuit of philosophy. Hume spent several years at Rheims and La Flèche, where he completed work on the Treatise of Human Nature (1739–40). Hume was greatly disappointed with the reception accorded this book which "fell deadborn from the press". Undaunted, he revised and popularized the Treatise in An Enquiry Concerning Human Understanding (1748). Hume also published an Enquiry Concerning the Principles of Morals (1751), and a lengthy History of England (1754–62). Hume was rebuffed in his attempts to secure positions at the Universities of Edinburgh and Glasgow. His opponents alleged heresy and even atheism. In 1763 Hume was appointed secretary to the British ambassador to France, and subsequently was lionized by Parisian society." [pp. 87]*



"Hume consistently denied that a knowledge of atomic configurations and interactions—even if it could be achieved—would constitute a necessary knowledge of nature. According to Hume, even if our faculties were "fitted to penetrate into the internal fabric" of bodies, we could gain no knowledge of a necessary connectedness among phenomena. The most we could hope to learn is that certain configurations and motions of atoms have been constantly conjoined with certain macroscopic effects. But knowing that a constant conjunction has been observed is not the same thing as knowing that a

particular motion must produce a particular effect." [pp. 91-92]

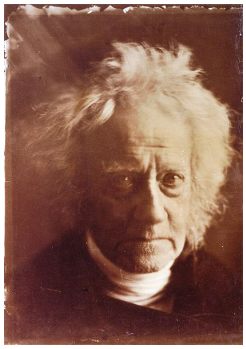
"Hume maintained that Descartes was wrong to hold that we possess innate ideas of mind, God, body, and world. According to Hume **sense impressions** are the **sole source of knowledge** of matters of fact. He thus echoed Aristotle's dictum that there is nothing in the intellect which was not first in the senses. Hume's version was that "all our ideas are nothing but copies of our impressions, or, in other words, that it is impossible for us to think of anything, which we have not antecedently felt, either by our external or internal senses."" [p. 93]

"Hume undertook to examine our idea of a **"causal relation"**. He noted that if we mean by a 'causal relation' both 'constant conjunction' and 'necessary connection', then we can achieve no causal knowledge at all. This is because we have no impression of any force or power by means of which an A is constrained to produce a B. The most that we can establish is that events of one type invariably have been followed by events of a second type. Hume concluded that the only "causal" knowledge that we can hope to achieve is a knowledge of the de facto association of two classes of events." [p. 94].

"Hume conceded that we do feel that there is something necessary about many sequences. According to Hume, this feeling is an impression of the "internal sense", an impression derived from custom. He declared that "after a repetition of similar instances, the **mind is carried by habit**, upon the appearance of one event, to expect its usual attendant, and to believe that it will exist." Of course, the fact that the mind comes to anticipate a B upon the appearance of an A is no proof that there is a necessary connection between A and B." [p.94]

### John Herschel (1792-1871)

*"John Herschel was the son of the great astronomer William Herschel. The achievements of the elder Herschel included the discovery of Uranus and the compilation of valuable data on double stars and nebulae. John Herschel studied at Cambridge, and thereafter devoted his life to the pursuit of science. His scientific achievements included studies of double refraction in crystals, experiments in photography and photo-chemistry, a method of computing binary-star orbits, and numerous astronomical observations. Herschel spent the period 1834–8 at the Cape of Good Hope, where he successfully extended his father's survey of double stars and nebulae to the Southern skies." [pp. 103]*



"One of Herschel's important contributions to the philosophy of science was a clear distinction between the **"context of discovery"** and the **"context of justification"**. He insisted that the procedure used to formulate a theory is strictly irrelevant to the question of its acceptability. A meticulous inductive ascent and a wild guess are on the same footing if their deductive consequences are confirmed by observation." [p. 104]

"Despite the fact that too much significance has been attributed to certain experiments in the evaluation of competing theories, the general attitude which promotes a search for falsifying instances has been most important in the history of science. Herschel encouraged this attitude. He demanded that the

scientist assume the role of **antagonist against his own theories**, and seek both direct refutations and exceptions which limit the range of application of these theories. Herschel believed that the worth of a theory is proved only by its ability to withstand such attacks." [p. 108]

### William Whewell (1794-1866)

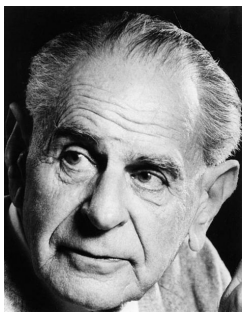
*"William Whewell graduated from Trinity College, Cambridge, where he was appointed Professor of Minerology (1828), Professor of Moral Philosophy (1838), and Vice-Chancellor (1842). He was instrumental in introducing into England the Continental version of the calculus, and was largely responsible for broadening the course of study at Cambridge. Whewell performed extensive researches on the tides, and was recognized— by Lyell and Faraday, among others—as an authority on scientific nomenclature. He completed his extensive History of the Inductive Sciences in 1837, and based his Philosophy of the Inductive Sciences (1840) on the results of this historical analysis." [pp. 103]*

"William Whewell, a contemporary of Herschel, sought to base his philosophy of science on a comprehensive survey of the history of science. Whewell proposed to examine the actual process of discovery in the various sciences in order to see if any patterns are displayed therein. Whewell claimed originality for his approach, pointing out that previous writers on the philosophy of science had regarded the history of science as a mere storehouse of examples which may be cited to illustrate particular points about scientific method. Whewell proposed to invert this relationship which had made the history of science dependent on the philosophy of science." [p. 108]

"The pattern of scientific discovery which Whewell claimed to see in the history of the sciences was a three-beat progression comprising a prelude, an inductive epoch, and a sequel. The prelude consists of a collection and decomposition of facts, and a clarification of concepts. An inductive epoch arises when a particular conceptual pattern is superinduced on the facts. And its sequel is the consolidation and extension of the integration thus achieved." [p. 109]

### Karl Popper (1902-94)

*"Karl Popper was Professor of Logic and Scientific Method at the University of London. In the influential Logic of Scientific Discovery (German 1934, English 1959), Popper criticized the Vienna Circle's search for a criterion of empirically meaningful statements, and suggested instead that empirical science be demarcated from pseudoscience with respect to methodology practised. He has reaffirmed and augmented this position in Conjectures and Refutations (1963). During World War II, Popper published The Open Society and its Enemies, an attack on Plato, Hegel, Marx, and all thinkers who would impose inexorable laws on history." [p. 144]*



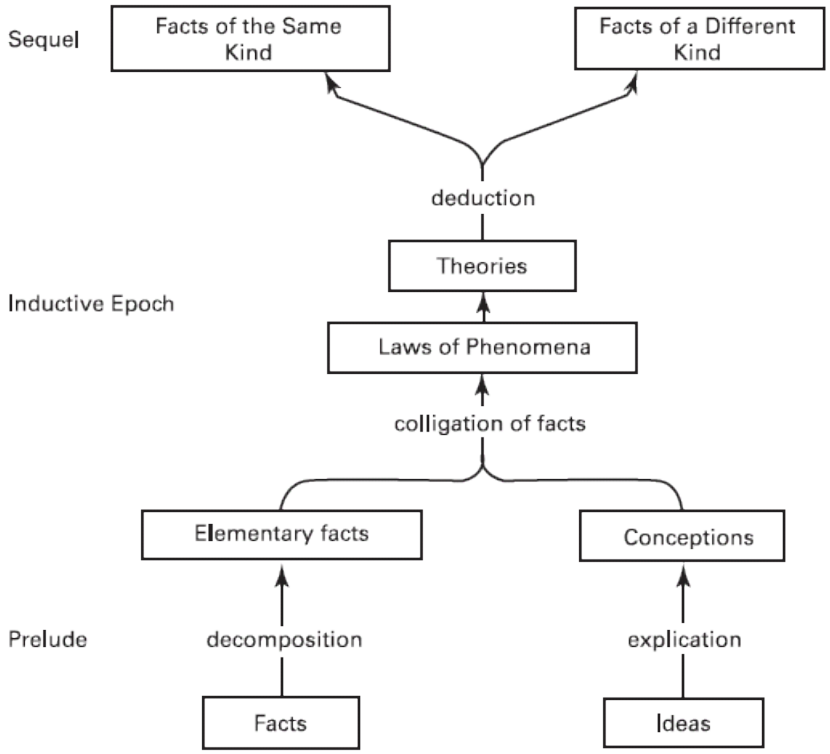



Figure 2.1: Whewell's Pattern of Discovery. [p. 110]

"Karl Popper ... noted that it always is possible to achieve agreement between a theory and observational evidence. If certain evidence is inconsistent with consequences of the theory, a number of strategies may be pursued to **"save" the theory**. The evidence may be rejected outright, or it may be accounted for either by adding auxiliary hypotheses or by modifying the rules of correspondence. These strategies may introduce a staggering degree of complexity into a theoretical system. Nevertheless, evasion of falsifying evidence in these ways always is possible." [p. 153]

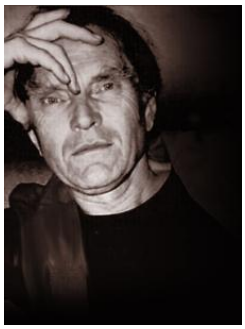
"According to Popper, proper empirical method is continually to expose a theory to the possibility of being falsified. He concluded that the way to combat conventionalism is to make a decision not to employ its methods. Consistent with this conclusion, he proposed a set of methodological rules for the empirical sciences. The supreme rule is a criterion of adequacy for all other rules, much as Kant's categorical imperative is a criterion of adequacy for moral norms. This supreme rule states that all rules of empirical method 'must be designed in such a way that they do not protect any statement in science against falsification.'" [p. 153]

"A hypothesis that is exposed to the possibility of **falsification** satisfies Popper's demarcation criterion. It has qualified to be included in the realm of permissible scientific discourse. To be acceptable, a hypothesis must satisfy a further requirement. It must withstand tests designed to refute it." [p. 154]

"Popper viewed the history of science as a sequence of conjectures, refutations, revised conjectures, and additional refutations. Proper scientific procedure is to expose conjectures to the most severe tests that can be devised. If a conjecture passes a test, then it has received **"corroboration"**. Popper insisted that corroboration is a "backward-looking" appraisal. The achievement of corroboration does

not justify a belief that a hypothesis is true, or approximately true. Popper consistently has opposed the appeal to inductive arguments to justify hypotheses. On his view, it is incorrect to argue that because hypothesis H passed tests $t_1 ... t_n$, it is probable that H will pass test $t_{n+1}$." [p. 155]

### Paul Feyerabend (1924-98)



*"Paul Feyerabend received a Ph.D. from the University of Vienna and taught at the University of California. He was a self-professed "anarchist" who opposed the search for rules of theory-replacement and "rational reconstructions" of scientific progress. Feyerabend's position was that "anything goes" and that the mark of creativity in science is a proliferation of theories. Consistent with this orientation, his major work is titled Against Method (1975)." [p. 177]*

"Basic to the Logical Reconstructionist philosophy of science is a claim about the theory-independence of observation reports. Orthodox theorists assumed that the truth or falsity of observation reports can be decided directly without appeal to sentences of the theoretical level. It was the orthodox position that theory-independent sentences of the observational level provide bona fide tests of theories. It also was the orthodox position that the sentences of the theoretical level acquire empirical meaning from the sentences of the observational level. Thus the theoretical level is parasitic upon the observational level. Paul Feyerabend suggested that the dependence had been misconstrued. It is **observation** reports that are parasitic **[dependent] on theories.**" [p. 178]

"Observation reports have no status apart from the theoretical context in which they occur." [p. 180]

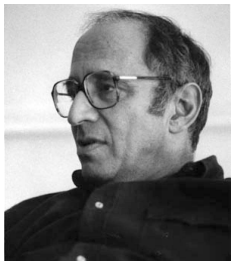
"Feyerabend maintained that there is no reason for a practising scientist to consult the philosophy of science. There is nothing in the philosophy of science which can help him solve his problems. In particular, theories of confirmation do not help the scientist to decide which theories to accept. This is because **theories of confirmation are based on two false assumptions**. The first false assumption is that there is a theory-independent observation language with respect to which theories may be evaluated. The second false assumption is that it is possible for a theory to agree with all the known facts in its domain. But in practice there always is some evidence that counts against a theory." [p. 189]

### Thomas Kuhn (1922-96)

"The numerous criticisms of orthodoxy had a cumulative effect. Many philosophers of science came to believe that something vital is lost when science is reconstructed in the categories of formal logic. It seemed to them that the proposed orthodox analyses of 'theory', 'confirmation', and 'reduction' bear little resemblance to actual scientific practice." [p. 197]

"Thomas Kuhn's The *Structure of Scientific Revolutions* was a widely discussed alternative to the

*"Thomas Kuhn received a Ph.D. in physics from Harvard, taught for many years at Princeton, and then at MIT. He contributed important historical studies of the Copernican Revolution and the origins of Quantum Mechanics. His widely influential book The Structure of Scientific Revolutions directed attention to the role of paradigms in the historical development of science." [p. 197]*



orthodox account of science. Kuhn formulated a "rational reconstruction" of scientific progress, a reconstruction based on his own interpretation of developments in the history of science. But Kuhn's reconstruction is not simply another history of science. Rather, it includes a second-order commentary—a philosophy of science—in which he presents normative conclusions about scientific method." [p. 198]

"Kuhn developed this emphasis into a model of scientific progress in which periods of **"normal science"** alternate with periods of **"revolutionary science"**." [p. 198]

"Normal science is a conservative enterprise. Kuhn characterized it as "puzzle-solving activity". The pursuit of normal science proceeds undisturbed so long as application of the paradigm satisfactorily explains the phenomena to which it is applied. But certain data may prove refractory. If scientists believe that the paradigm should fit the data in question, then confidence in the programme of normal science has been shaken. The type of phenomena described by the data is then regarded as an anomaly." [p. 198]

"The presence of an anomaly or two is not sufficient to cause abandonment of a **paradigm**. Kuhn maintained that a logic of falsification is not applicable to the case of paradigm rejection. A paradigm is not rejected on the basis of a comparison of its consequences and empirical evidence. Rather paradigm rejection is a three-term relation which involves an established paradigm, a rival paradigm, and the observational evidence." [p. 199]

"Thus paradigm replacement resembles a ***gestalt*-shift**. Competing paradigms are not wholly commensurable. Given a particular problem, two paradigms may differ with respect to the types of answer deemed permissible." [p. 199]

### Imre Lakatos (1922-1974)

"Lakatos agreed with Kuhn that refutation neither is nor should be followed invariably by rejection. Theories should be allowed to flourish even within an "ocean of anomalies". But after awarding Kuhn high marks for his emphasis on continuity, Lakatos criticized him for treating revolutionary episodes as instances of "mystical conversion"." [pp. 202-3]

"Lakatos maintained that unless a rational reconstruction of theory replacement can be given, the interpretation of scientific change must be left to historians and psychologists. Popper had produced a rational reconstruction, according to which scientific progress is a sequence of conjectures and attempted

*"Imre Lakatos, a native of Hungary, was a victim of Nazi persecution who subsequently spent three years in jail during the era of Stalinist repression. In 1956 he left Hungary for England where he pursued investigations in philosophy of mathematics and philosophy of science at Cambridge and the London School of Economics." [p. 197]*

refutations. Lakatos sought to improve upon this reconstruction. In particular, he urged that the basic unit for appraisal should be **"research programmes"** rather than individual theories. According to Lakatos, a research programme consists of methodological rules: some tell us what paths of research to avoid (negative heuristic) and others what paths to pursue (positive heuristic)." [p. 203]

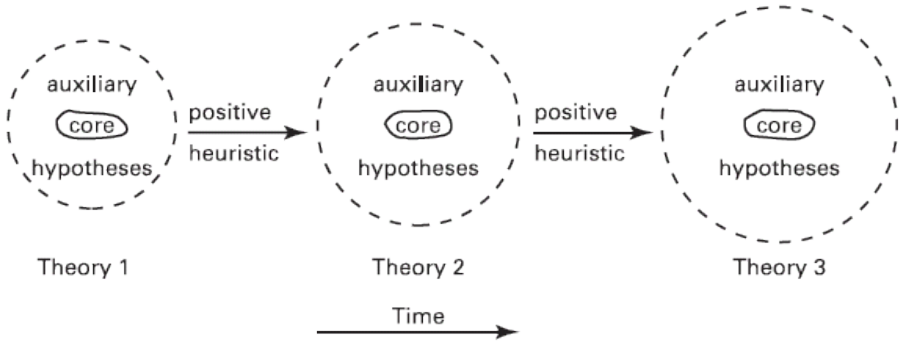



Figure 2.2: Lakatos's Scientific Research Programme. [p. 203]

# 3 Some Background

## 3.1 ASCAT Soil Moisture

Throughout the course we will use the example of the ASCAT surface soil moisture data product to illustrate the complexity involved in the retrieval of Level 2 data. Therefore, in the following subsections, we will firstly describe the ASCAT instrument (Section 3.1.1), the ASCAT soil moisture data service as provided in near real-time by EUMETSAT's Satellite Application Facility in Support to Operational Hydrology and Water Management (H SAF) (Section 3.1.2), and the Level 2 data product itself (Section 3.1.3)

### 3.1.1 Advanced Scatterometer

The Advanced Scatterometer (ASCAT) is a real aperture radar system operating in C-band (VV polarization) and provides day- and night-time measurement capability almost unaffected by cloud cover. The instrument measures the Normalized Radar Cross Section (NRCS), or so-called **backscatter coefficient** $\sigma^\circ$ expressed in $m^2m^{-2}$ or dB, of the Earth's surface. The design and performance specification is based on the experience of the scatterometer flown on the ERS-1 (1991-2000) and ERS-2 (1995-2011) satellites. ASCAT can work in two different modes: Measurement or Calibration. The Measurement Mode is the only mode that generates science data for users. The instrument uses a ranging technique called "chirp", at which long pulses with linear frequency modulation are generated at the carrier frequency of 5.255 GHz. After receiving and de-chirping of the ground echoes, a Fourier transform is applied in order to relate different frequencies in the signal to slant range distances. A pre-processing of the noise and echo measurements is done already on board to reduce the data rate to the ground stations, e.g. various averaging takes place reducing the raw data by a factor of approximately 25. Within each pulse repetition interval, after all pulse echoes have been decayed, the contribution of the thermal noise is monitored in order to perform a measurement noise subtraction during the ground processing. A side effect of on board processing is a degree of spatial correlation between different and within received echoes, but this is taken into account later on by the Level 1b processing. The radar pulse repetition frequency (PRF) is approximately 28.26 Hz, which yields to 4.71 Hz for the beam pulse repetition frequency in the sequence fore-, mid- and aft beam [7].

Three **Metop** (Meteorological Operational) satellites, launched in 2006 (Metop-A), 2012 (Metop-B) and 2018 (Metop-C), are flying in a sun-synchronous orbit (29 day repeat cycle) with an ascending/descending node at 9:30 p.m./a.m. and a minimum orbit height of 822 km. Each satellite makes a little more than 14 orbits per day with this orbit configuration. The advanced measurement geometry of ASCAT allows twice the coverage, compared to its predecessors flying on ERS-1 and ERS-2. Thus,

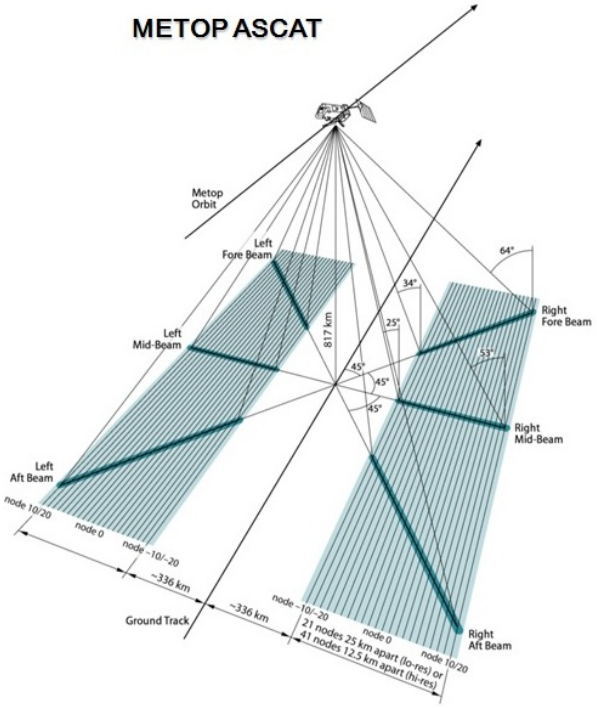



Figure 3.1: Antenna configuration of Metop ASCAT.

the **daily global coverage** increased from 41% to 82%. The **spatial resolution** was also improved from 50 km to 25-34 km, whilst maintaining the same radiometric accuracy compared to the ERS-1/2 scatterometer. The advanced global coverage capability is accomplished by two sets of three fan-beam antennas, compared to only one set at the ERS satellites. The fan-beams are arranged broadside and $\pm 45°$ of broadside, thus, allowing to observe three azimuthal directions in each of its two 550 km swaths (see Figure 3.1). The swaths are separated from the satellite ground track by about 360 km for the minimum orbit height. Each point on the Earth's surface that falls within one of the two swaths are seen by all three antennas, and a so-called $\sigma°$ **triplet** can be observed. The **incidence angles** for those two antennas which are perpendicular to the flight direction intersect the surface between 25° and 53.3°, whereas the other four antennas have an incidence angle range from 33.7° to 64.5° [6]. The multi-angle viewing capability of ASCAT is needed for an effective and precise geophysical retrieval of surface variables (e.g. wind speed and direction over ocean, vegetation conditions over land). Otherwise remaining ambiguities hamper or render impossible a successful retrieval of such variables.

### 3.1.2 Soil Moisture Data Service

The first global satellite-based soil moisture data set freely shared with the user community was derived the scatterometer on-board ERS-1 and ERS-2 and released in 2002 [17]. Since then, progress in algorithmic research and improvements in sensor technologies has lead to an increasing number of satellite products with operational potential. Soil moisture derived from Metop ASCAT became the first operational soil moisture product in 2008. The ASCAT soil moisture data service is implemented and operated by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT).

EUMETSAT offers a suite of soil moisture products that all share the same physical basis, but come in different flavours (spatio-temporal sampling, data latency, data format, consistency). For example, the near real-time products (available within 1-6 hours after sensing) serve first and foremost the needs of the Numerical Weather Prediction (NWP) user community, but typically lack of consistency due to calibration or software updates. On the other hand, long-term data records are processed irregularly (2-5 years) but consistency is among the most important properties.

Challenges faced by the ASCAT soil moisture data service include instrument calibration, product validation, algorithmic improvements, quality flagging and application development. In an **operational environment** stability of the data quality needs to be sustained in time. A continuous monitoring is in place detecting and correcting possible instrument drifts or other product degeneration effects. Over time, several algorithmic updates have become available. These updates need careful checking and testing in off-line and parallel processing environments before transferring them to the operational processing framework. A controlled migration is time consuming and users needs to be informed and prepared for upcoming modifications in the product.

### 3.1.3 Soil Moisture Data Product

The ASCAT soil moisture product represents relative **surface soil moisture** of the topmost soil layer ($< 5$ cm). Soil moisture is expressed in degree of saturation ranging from 0% (completely dry) to 100% (fully saturated). The accuracy and reliability of the soil moisture product can be different depending on the prevailing surface conditions, i.e. density of vegetation, presence of open water and urban areas, as well as, the occurrence of snow and frozen soil conditions. Therefore, additional information is needed. The spatial resolution of the product is 25-34 km and defined by the spatial resolution of the Level 1b backscatter product. A trade-off between spatial resolution and radiometric accuracy was necessary in the processing of the Level 1b backscatter product by applying a variable spatial filter width. As a result, the spatial resolution of the Level 1b backscatter product ranges from 25 km in the near swath to 34 km in the far swath. The TU Wien soil moisture retrieval algorithm is implemented in a Python software package called soil WAter Retrieval Package (WARP) and used to derive surface soil moisture information. In practice, the latest Metop ASCAT Level 1b Fundamental Climate Data Record (FCDR) and the latest operational Level 1b data are manually combined to a common data set ensuring a consistent Level 1b calibration which is used as input. The ASCAT soil moisture data product contains quite long list of data fields informing the user not just about the soil moisture value per se but about many other aspects as well, such as information about the satellite, sensor, geo-location, retrieval uncertainty, quality flags, and many more (see Table 3.1 and the following sub-sections). Users should use this information to pre-select data which are suited for their applications. Depending on the user requirements and environmental conditions the spatial extent of usable soil moisture retrievals may be quite limited (Figure 3.2).

#### 3.1.3.1 Satellite parameters

The satellite parameters consist of location, speed and health information of the spacecraft. Orbit maneuvers are quite common to avoid obstacles in the pathway of the satellite. Monitoring power con-

Table 3.1: Selection of data fields from the METOP ASCAT surface soil moisture data product. For detailed information see the ASCAT Data Product Guide [5].

| Name | Scaling factor | Units | Type | Byte size |
|---|---|---|---|---|
| product_name | n/a | n/a | str | 67 |
| product_parent_name_1 | n/a | n/a | str | 67 |
| instrument_id | n/a | n/a | enum | 4 |
| instrument_model | n/a | n/a | enum | 3 |
| product_type | n/a | n/a | enum | 3 |
| processing_level | n/a | n/a | enum | 2 |
| spacecraft_id | n/a | n/a | enum | 3 |
| sensing_start | n/a | n/a | time | 15 |
| sensing_end | n/a | n/a | time | 15 |
| processing_centre | n/a | n/a | enum | 4 |
| processor_major_version | n/a | n/a | uint | 5 |
| processor_minor_version | n/a | n/a | uint | 5 |
| format_major_version | n/a | n/a | uint | 5 |
| format_minor_version | n/a | n/a | uint | 5 |
| orbit_start | n/a | n/a | uint | 5 |
| orbit_end | n/a | n/a | uint | 5 |
| x_position | $10^3$ | n/a | int | 11 |
| y_position | $10^3$ | n/a | int | 11 |
| z_position | $10^3$ | n/a | int | 11 |
| subsat_latitude_start | $10^3$ | deg | int | 11 |
| subsat_longitude_start | $10^3$ | deg | int | 11 |
| subsat_latitude_end | $10^3$ | deg | int | 11 |
| subsat_longitude_end | $10^3$ | deg | int | 11 |
| latitude | $10^6$ | deg | int32 | 4 |
| longitude | $10^6$ | deg | int32 | 4 |
| utc_line_nodes | n/a | UTC | time | 6 |
| swath_indicator | n/a | n/a | int8 | 1 |
| sigma0_trip | $10^6$ | dB | uint32 | 4 |
| f_land | $10^3$ | n/a | uint16 | 2 |
| f_usable | n/a | n/a | int8 | 1 |
| inc_angle_trip | $10^2$ | deg | uint16 | 2 |
| azi_angle_trip | $10^2$ | deg | uint16 | 2 |
| warp_nrt_version | n/a | n/a | uint16 | 2 |
| param_db_version | n/a | n/a | uint16 | 2 |
| soil_moisture | $10^2$ | % | uint16 | 2 |
| soil_moisture_error | $10^2$ | % | uint16 | 2 |
| mean_soil_moisture | $10^2$ | % | uint16 | 2 |
| soil_moisture_sensitivity | $10^6$ | dB | int32 | 4 |
| dry_backscatter | $10^6$ | dB | int32 | 4 |
| wet_backscatter | $10^6$ | dB | int32 | 4 |
| sigma40 | $10^6$ | dB | int32 | 4 |
| sigma40_error | $10^6$ | dB | int32 | 4 |
| slope40 | $10^6$ | dB/deg | int32 | 4 |
| slope40_error | $10^6$ | dB/deg | int32 | 4 |
| correction_flags | n/a | n/a | uint8 | 1 |
| processing_flags | n/a | n/a | uint8 | 1 |
| snow_cover_probability | n/a | n/a | uint8 | 1 |
| frozen_soil_probability | n/a | n/a | uint8 | 1 |
| inundation_or_wetland | n/a | n/a | uint8 | 1 |
| topographical_complexity | n/a | n/a | uint8 | 1 |

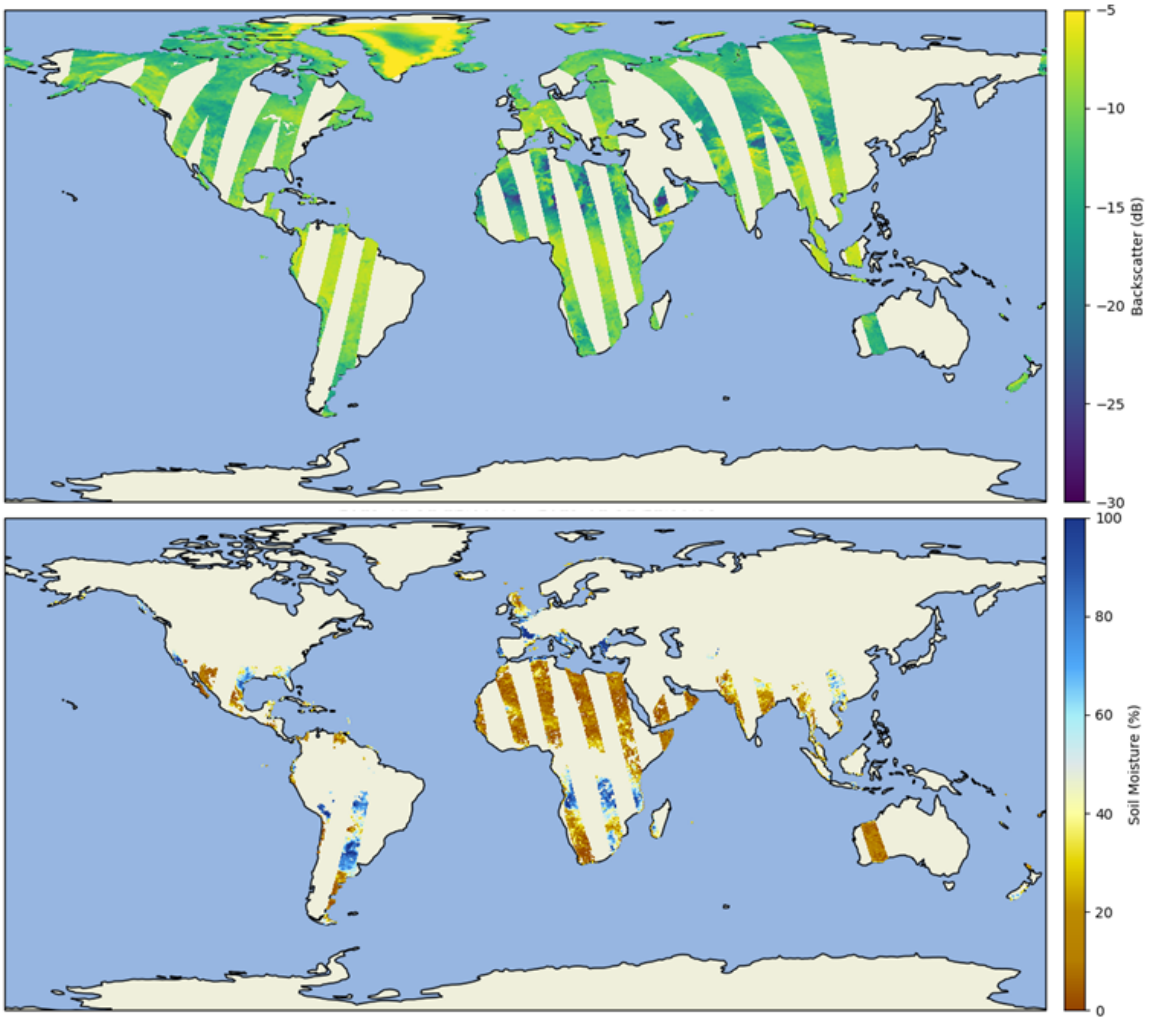



Figure 3.2: ASCAT backscattering measurements normalized to a reference incidence angle of to 40° (top) and surface soil moisture retrievals after masking for frozen ground, snow and other factors that disturb the measurements. The ASCAT data were collected by the instrument on board of METOP-C on 31 Jaunary 2019 between 12:00 and 22:00 UTC.

sumption and other vital parameters are important to keep the satellite mission safe and operational.

#### 3.1.3.2 Geo-location parameters

Latitude and longitude coordinates of the measurements are computed based on the position of the satellite and information of the measurement itself. Exact knowledge about the time is important as well, which is typically given in Coordinates Universal Time (UTC).

#### 3.1.3.3 Geophysical variables and uncertainty

The retrieved variable in physical units (e.g. backscatter, soil moisture, wind speed) and information about the uncertainty of the observation are part of the data product provided to the user. Additional processing or corrections flags indicate that e.g. the retrieval has not been possible due to lack of sensitivity of the measured signal to the variable of interest or because of any other circumstance making it impossible retrieving a valid observation.

#### 3.1.3.4 Quality and auxiliary flags

The quality and auxiliary flags provide an initial assistance on the usability of the geophysical product. Under certain environmental conditions the retrieval is likely to fail, which cannot be captured in every case by the error model. For example, probability flags about snow cover or frozen soil condition shall guide the user to mask invalid observations. However, external data sets with a better quality are more preferable, but usually not available in near real-time.

## 3.2 Radiative Transfer Theory

In the case of ASCAT, which is an active microwave instrument, we are dealing with coherent waves. In principle, one would have to use **electromagnetic theory** to describe the measurement and scattering processes. Yet, constructing forward models that describe the interaction of microwaves with complex objects starting from Maxwell's equation is either very hard or practically impossible. Therefore, one usually applies radiative transfer theory which describes the **propagation of radiation through a medium**. The theory itself only deals with incoherent radiation, where the most important quantity is the intensity of the wave. Yet, coherent phenomena may partly be incorporated in radiative transfer models in an indirect fashion by choosing appropriate "parameterizations".

In radiative transfer theory one can imagine that the radiation is made up of innumerable photons, whereas the energy of one photon is given by

$$q = h \cdot \nu \tag{3.1}$$

where $q$ is the photon energy, $h = 6.626 \cdot 10^{-34}\ m^2 \cdot kg \cdot s^{-1}$ is the Planck constant, and $\nu$ is the frequency of the electromagnetic wave. This equation links the **wave nature** of radiation with its

**particle nature**. The photons flow through a medium where they may be absorbed and scattered. The medium itself may emit photons. The processes of **absorption**, **scattering**, and **emission** are described on a macroscopic level through the use of phenomenological parameters.

### 3.2.1 Radiometric Quantities

A number of quantities are commonly used to characterize electromagnetic radiation and its interaction with matter. Following [4] these are (see also Table 3.2 and Figure 3.3):

- **Radiant energy**: The energy carried by an electromagnetic wave. It is a measure of the capacity of the wave to do work by moving an object by force, heating it, or changing its state. The amount of energy per unit volume is called *radiant energy density*.
- **Radiant flux**: The time rate at which radiant energy passes a certain location. It is closely related to the wave *power*, which refers to the time rate of doing work. The term *flux* is also used to describe the time rate of flow of photons.
- **Radiant flux density**: Corresponds to the radiant flux intercepted by unit area of a plane surface. The density for flux incident upon a surface is called *irradiance*. The density for flux leaving the surface is called *exitance* or *emittance*.
- **Solid angle**: The solid angle $\Omega$ subtended by area $A$ on a spherical surface is equal to the area $A$ divided by the square of the radius of the sphere.
- **Radiant intensity**: The radiant intensity of a point source in a given direction is the radiant flux per unit solid angle leaving the surface in that direction.
- **Radiance**: The radiant flux per unit solid angle leaving an extended source in a given direction per unit projected area in that direction (see Figure 3.3).

The radiance $L$ is probably the most important quantity in radiative transfer theory because it is constant when the radiation travels through vacuum (while the other quantities change due to the diffusive nature of radiation).

Table 3.2: Radiation Quantities [4].

| Quantity | Usual symbol | Defining equation | Units |
|---|---|---|---|
| Radiant energy | $Q$ | | Joule |
| Radiant energy density | $W$ | $W = \frac{dQ}{dV}$ | Joule/m$^3$ |
| Radiant flux | $\Phi$ | $\Phi = \frac{dQ}{dt}$ | Watt |
| Radiant flux density | $E$ (irradiance)<br>$M$ (emittance) | $E, M = \frac{d\Phi}{dA}$ | Watt/m$^2$ |
| Radiance intensity | $I$ | $I = \frac{d\Phi}{d\Omega}$ | Watt/steradian |
| Radiance | $L$ | $L = \frac{dI}{dA\cos\theta}$ | Watt/steradian m$^2$ |

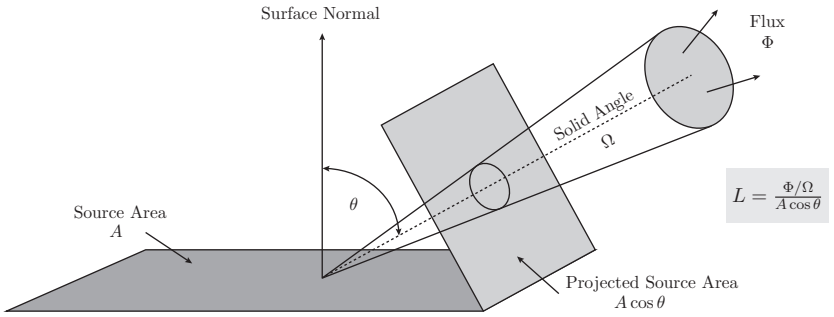



Figure 3.3: Concept of radiance [4].

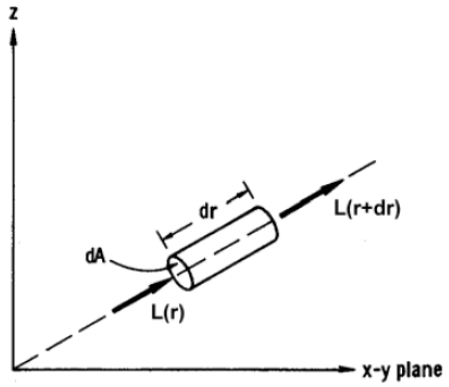



Figure 3.4: Radiation transfer across an infinitesimal cylinder [19].

### 3.2.2 Equation of Radiative Transfer

To understand how radiative transfer theory describes the flow of radiation through a medium let us consider a small cylindrical volume of cross-section $dA$ and thickness $dr$ (see Figure 3.4). When travelling through the medium, a part of the radiation may be absorbed, causing an extinction of the radiation on its way through the cylinder:

$$dL(absorption) = \kappa_a L dr \tag{3.2}$$

where $\kappa_a$ is the absorption coefficient. Similarly, the radiation may increase on its way through the cylinder when the medium emits radiation at the given frequency. The gain in energy is given by

$$dL(emission) = \kappa_a J_a dr \tag{3.3}$$

where $J_a$ is a source function that account for thermal emission into the forward direction. The thermal emission, which is caused by the object's temperature according to Planck's law, is expressed in terms of the absorption coefficient and -source function ($\kappa_a J_a$) because we may assume thermodynamic equilibrium, at which absorption and emission are equal. However, this term is only important for passive microwave remote sensing and can be dropped in active microwave remote sensing.

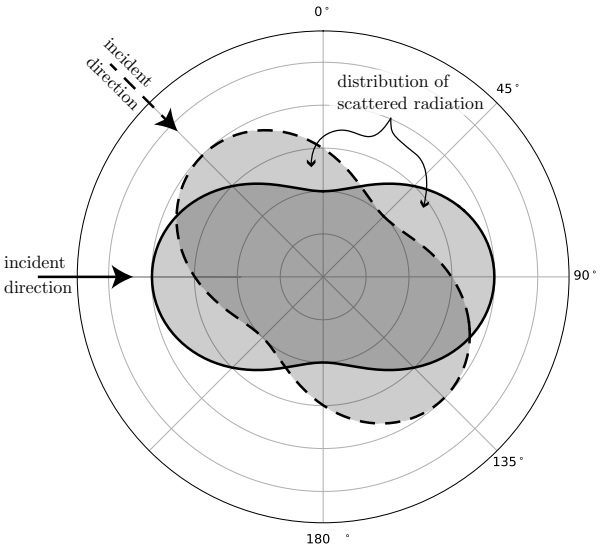



Figure 3.5: Rayleigh phase function

Finally, a part of the radiation may be scattered, causing either a loss of radiance (if the radiation is scattered away from the forward-direction), or a gain of radiance (if incoming radiation from other directions is scattered into the forward-direction).

To describe the angular scattering behaviour of the medium, one can use a **phase function** $P(\theta, \phi, \theta', \phi')$ where $\theta$ and $\theta'$ are the polar angles, and $\phi$ and $\phi'$ the azimuth angles in the incident and scattered directions, respectively. The phase function accounts for the effect that radiation is usually not scattered isotropically in all directions. Integration of the phase function over all space angles must lead to identity due to energy conservation:

$$\frac{1}{4\pi} \iint_{4\pi} P(\theta', \phi', \theta, \phi) d\Omega' = \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} P(\theta', \phi', \theta, \phi) \sin(\theta') d\theta' d\phi' = 1 \tag{3.4}$$

In the most simple example of isotropic scatterers, $P(\theta, \phi, \theta', \phi')$ is equal to 1. Another very common phase function is the Rayleigh phase function (Figure 3.5):

$$P_{Rayleigh}(\Theta) = \frac{3}{4} \left(1 + \cos^2(\Theta)\right) \qquad \cos(\Theta) = \cos(\theta)\cos(\theta') + \sin(\theta)\sin(\theta')\cos(\phi - \phi') \tag{3.5}$$

where $\Theta$ is the angle between the directions $(\theta, \phi)$ and $(\theta', \phi')$

The fraction of an incoming radiance $dL(\Omega)$ within a differential solid-angle $d\Omega$ that is scattered into a differential solid-angle $d\Omega'$ is given by:

$$dL^{d\Omega \to d\Omega'} = \frac{\kappa_s}{4\pi} L(\theta, \phi) P(\theta, \phi, \theta', \phi') d\Omega d\Omega' dr \tag{3.6}$$

where $\kappa_s$ is the scattering coefficient that accounts for the amount of radiation that undergoes a scattering process.

The amount of radiance incoming from all directions other than the original path $(\theta, \phi)$ that is scattered

into the original path (i.e. the scattering-gain) can now be written as:

$$dL^{4\pi\setminus(\theta,\phi)\to(\theta,\phi)} = \frac{\kappa_s}{4\pi} \iint\limits_{4\pi\setminus(\theta,\phi)} L(\theta',\phi')P(\theta',\phi',\theta,\phi)d\Omega' dr \tag{3.7}$$

Similarly, the amount of radiance incoming from the original path that is scattered into directions other than the original path (i.e. the scattering-loss) can be written as:

$$dL^{\Omega\to 4\pi\setminus(\theta,\phi)} = \frac{\kappa_s}{4\pi} L(\theta,\phi) \iint\limits_{4\pi\setminus(\theta,\phi)} P(\theta,\phi,\theta',\phi')d\Omega' dr \tag{3.8}$$

In order to be able to combine the formulas for the scattering-gain and the scattering-loss, we split the integral over the phase-function into two parts and use (3.4) to evaluate the first:

$$dL^{(\theta,\phi)\to 4\pi\setminus(\theta,\phi)} = \kappa_s L(\theta,\phi)dr \underbrace{\frac{1}{4\pi}\iint\limits_{4\pi} P(\theta,\phi,\theta',\phi')d\Omega'}_{=1} - \frac{\kappa_s}{4\pi}L(\theta,\phi) \iint\limits_{(\theta,\phi)} P(\theta,\phi,\theta',\phi')d\Omega' dr \tag{3.9}$$

$$= \kappa_s \left[ L(\theta,\phi) - \frac{1}{4\pi} \iint\limits_{(\theta,\phi)} L(\theta,\phi)P(\theta,\phi,\theta',\phi')d\Omega' \right] dr \tag{3.10}$$

The total change of radiance due to scattering is now given by:

$$\begin{aligned} dL(scattering) &= (\text{scattering-gain}) - (\text{scattering-loss}) = dL^{4\pi\setminus(\theta,\phi)\to(\theta,\phi)} - dL^{(\theta,\phi)\to 4\pi\setminus(\theta,\phi)} \\ &= \frac{\kappa_s}{4\pi} \iint\limits_{4\pi\setminus(\theta,\phi)} L(\theta',\phi')P(\theta',\phi',\theta,\phi)d\Omega dr - \kappa_s \left[ L(\theta,\phi) - \frac{1}{4\pi} \iint\limits_{(\theta,\phi)} L(\theta,\phi)P(\theta,\phi,\theta',\phi')d\Omega' \right] dr \\ &= -\kappa_s L(\theta,\phi) + \frac{\kappa_s}{4\pi} \iint\limits_{4\pi} L(\theta',\phi')P(\theta',\phi',\theta,\phi)d\Omega' dr \end{aligned} \tag{3.11}$$

Considering both gains and losses, the total change in radiance for the **active case** is now given by:

$$dL = dL(emission) - dL(absorption) + dL(scattering) \tag{3.12}$$

Inserting the above representations and summarizing the scattering and the absorption coefficient to the volume **extinction coefficient,** i.e., $\kappa_e = \kappa_s + \kappa_a$, the radiative transfer equation for active microwave remote sensing problems takes on the following form:

$$\frac{dL(\theta,\phi)}{dr} = -\kappa_e L(\theta,\phi) + \frac{\kappa_s}{4\pi} \int_0^{2\pi}\int_0^{\pi} P(\theta',\phi',\theta,\phi)L(\theta',\phi')\sin\theta' d\theta' d\phi' \tag{3.13}$$

Based on this equation, we will introduce vegetation scattering models for the special case of backscattering, i.e., where the scattered direction is the negative incident direction, i.e.:
$\theta_s = \pi + \theta_i$ and $\phi_s = \pi + \phi_i$.

### 3.2.3 Bidirectional Reflectance Function

To describe scattering by an area (Figure 3.6) one can use several different reflectance quantities. The most basic one is the **bidirectional reflectance distribution function (BRDF)**, in $sr^{-1}$, defined by

$$BRDF(\theta_i, \phi_i, \theta_s, \phi_s) = \frac{L_s(\theta_s, \phi_s)}{E_i(\theta_i, \phi_i)} \tag{3.14}$$

where $L_s$ is the radiance scattered into the direction $(\theta_s, \phi_s)$ and $E_i$ is the irradiance from direction $(\theta_i, \phi_i)$. It should be noted that monochromatic, uniform, and isotropic illumination is assumed. Measurements of the BRDF always involve an average over finite intervals, e.g. over the solid angle $\Omega_s$.

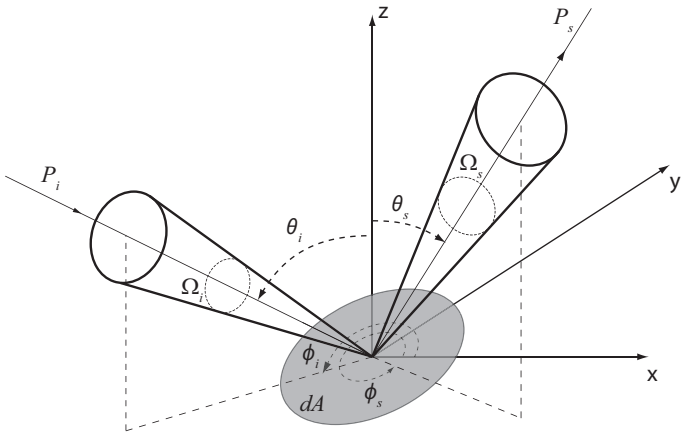



Figure 3.6: Scattering by an areal element. From [24].

## 3.3 Backscattering Coefficient

The backscattering coefficient $\sigma^\circ$ as measured by ASCAT is a measure of the electromagnetic energy intercepted and reradiated at the same wavelength by the Earth's surface. It is defined by electromagnetic theory (Section 3.3.1), can be estimated from ASCAT measurements via the use of the radar equation (Section 3.3.2), and modelled with radiative transfer theory (Section 3.3.3).

### 3.3.1 Definition

The fundamental quantity to describe scattering by an object (Figure 3.7) is the so-called **radar cross section** which has the dimensions of an area ($m^2$). The radar cross section is a complex combination of multiple factors: size, shape, material, edges, wavelength, and polarization. In theory the radar cross section of an object of any shape can be determined by solving Maxwell's equations. Given an incident and reflected electric field, the **bi-static** cross section is defined by

$$\sigma(\theta_i, \phi_i, \theta_s, \phi_s) = \lim_{R \to \infty} 4\pi R^2 \frac{|\mathbf{E}_s(\theta_i, \phi_i, \theta_s, \phi_s)|^2}{|\mathbf{E}_i(\theta_i, \phi_i)|^2} \tag{3.15}$$

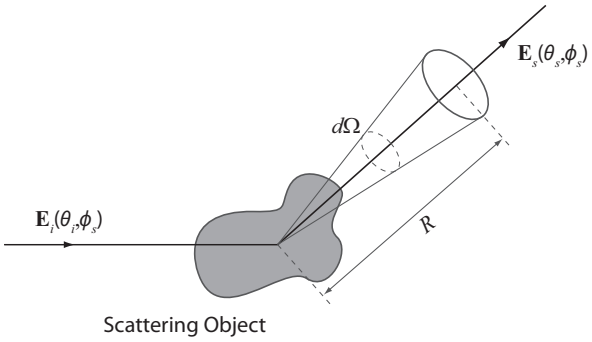



Figure 3.7: A plane electromagnetic wave with an electric field vector $\mathbf{E}_i$ is incident upon a scatterer, resulting in a scattered field $\mathbf{E}_s$. From [24].

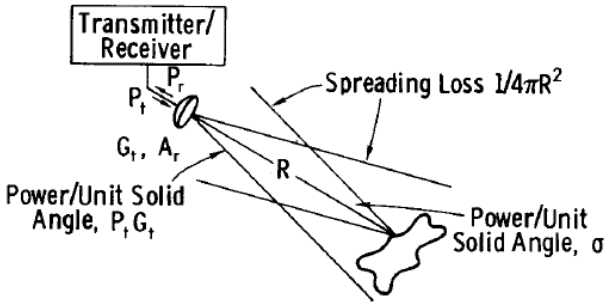



Figure 3.8: Geometry and quantities involved in the monostatic radar equation. From [20].

In the **mono-static** case one speaks of backscatter, and the cross section becomes

$$\sigma = \sigma(\theta_i, \phi_i, \theta_i, \phi_i + \pi) \tag{3.16}$$

To quantify backscatter as recorded by a radar by a variable that is independent of the areal extent of the target we can define the **backscatter coefficient**, which is the cross section normalized with an area $A$:

$$\sigma^\circ = \frac{\sigma}{A} \tag{3.17}$$

Its unit is thus $m^2 \cdot m^{-2}$, but in most cases it is expressed logarithmically:

$$\sigma^\circ[dB] = 10 \log \sigma^\circ \left[m^2 \cdot m^{-2}\right] \tag{3.18}$$

Note that this differential determination of the backscatter coefficient is based on the assumption that the areal distribution of single point scattering targets within the illuminated area is fairly random, i.e. no single scatterer dominates.

### 3.3.2 Measurement

To derive the radar cross section – and consequently the backscattering coefficient – from active microwave measurements one must describe the flow of energy from pulse generation to echo reception

Figure 3.8. This is done via the radar equation which yields the received power $P_r$ of a backscattered signal as as:

$$P_r = \frac{P_t G_t A_r}{(4\pi)^2 R^4} \sigma \tag{3.19}$$

$P_t$ is the total power of the transmitted antenna signal, $G_t$ is the antenna gain that describes how much of the emitted signal propagates in the direction of the target, $A_r$ is the effective antenna area, and $R$ is the range between the sensor and the target. These quantities describe the measurement configuration and are usually well known respectively can be well estimated from calibration measurements. Hence, the cross section can be derived from:

$$\sigma = \frac{(4\pi)^2 R^4 P_r}{G_t A_r P_t} \tag{3.20}$$

Knowing the illuminated area, the backscattering coefficient $\sigma^\circ$ is straight forward to compute.

### 3.3.3 Modelling

Recognizing the similarity of how scattering processes are described by electromagnetic- and radiative transfer theory respectively, it becomes clear that these concepts are related (compare e.g. Figures 3.6 and 3.7). For the bi-static case, the backscattering coefficient $\sigma^\circ$ is linked to the $BDRF$ via:

$$\sigma^\circ(\theta_i, \phi_i, \theta_s, \phi_s) = 4\pi BRDF(\theta_i, \phi_i, \theta_s, \phi_s) cos(\theta_i) cos(\theta_s) \tag{3.21}$$

More generally, the bi-static backscattering coefficient is linked to the radiances of the incident and scattered waves by

$$\sigma^\circ(\theta_i, \phi_i, \theta_s, \phi_s) = 4\pi \frac{L_s(\theta_s, \phi_s)}{L_i(\pi - \theta_i, \phi_i)} cos(\theta_s) \tag{3.22}$$

For the mono-static case the backscatter coefficient can be calculated from

$$\sigma^\circ(\theta, \phi) = 4\pi \frac{L_s(\theta, \phi + \pi)}{L_i(\pi - \theta, \phi)} \cos(\theta) \tag{3.23}$$

These relationships show that the backscattering coefficient $\sigma^\circ$ as measured by ASCAT and other active remote sensing instruments can be estimated from radiative transfer theory when having found a solution for the scattered radiance $L_s$ for a given target and incoming radiance $L_i$.

# 4 Forward Modelling

To illustrate different approaches for constructing forward models we discuss alternative models describing the interaction of microwave pulses with vegetated land surfaces. In general, one distinguishes **theoretical**, **empirical**, and **semi-empirical models** which are, respectively, based upon physical laws, statistical relationships, or a combination of those. Such forward models are the basis for inverting backscatter measurements to retrieve soil moisture and vegetation.

In the forward model for ASCAT land surface backscatter, we have to describe the propagation of the microwave pulses from the antenna, where they are transmitted, to the Earth's surface, where it is reflected, and back again to the antenna, where the remaining fraction of the emitted radiance is measured. Thanks to the fact that the atmosphere is to a large degree transparent in the low-frequency microwave range (1-10 GHz), we can neglect atmospheric effects.

## 4.1 Backscatter from Vegetation

For modelling backscatter from vegetation one usually uses radiative transfer theory. Therefore, the phase information is dropped and only the net energy change of the radiation when passing a certain volume, such as vegetation, is considered. Strictly speaking, this approach is wrong because the detailed scattering behaviour of the various vegetation elements (stems, branches, leaves) can only be understood by considering both the amplitude and phase of the electromagnetic waves. Yet, solving Maxwell's equations for such complex scattering targets is almost impossible. Hence, one reverts to radiative transfer theory and treats its model parameters as effective quantities, i.e. its values describe interference effects indirectly.

As illustrated by 4.1 the three main scattering components are:

- **Volume scattering**: A part of the microwave pulse is scattered directly backward by the vegetation layer. For dense vegetation this may be the only scattering component as seen be the radar.
- **Surface scattering**: A part of the impinging microwave pulse passes through the vegetation layer down to the soil surface, where its backscattered, and further damped by the vegetation on the way back.
- **Interaction terms**: Higher-order scattering terms come from interaction effects, i.e. the scattering of the microwave pulse by both vegetation and soil. Usually only the first-order scattering term (vegetation-soil, soil-vegetation) contributes in important ways to total backscatter.

The total backscatter from the vegetated soil thus becomes

$$\sigma^0 = \sigma_v^0 + \sigma_s^0 + \sigma_{int}^0 = 4\pi \frac{L_v + L_s + L_{int}}{L_i} cos(\theta) \tag{4.1}$$

where the subscripts $v$, $s$, and $int$ indicate the volume, surface, and interaction terms respectively. $L_i$ is the radiance of the radar pulse impinging on the vegetation, and $\theta$ is the incidence angle. Before discussing a first-order solution of radiative transfer theory that treats the vegetation as a continuous and homogeneous layer, we introduce one example of a relatively complex forest backscatter model that aims to quantify the scattering contributions from different vegetation elements (leaves, branches, trunks, etc.) separately.

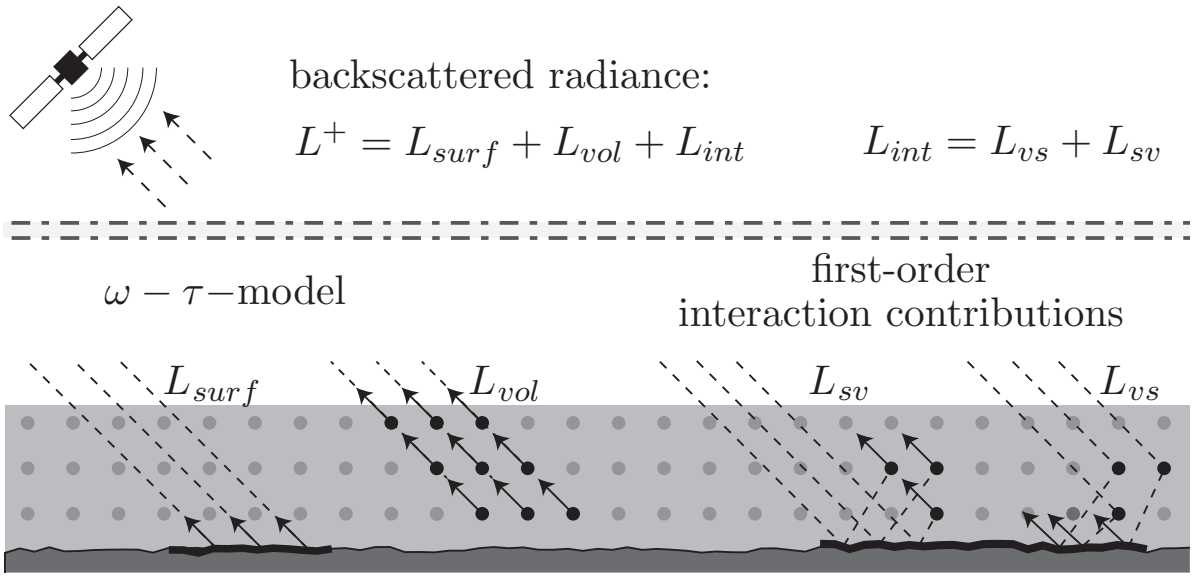



Figure 4.1: Schematic illustration of the main scattering components from a vegetated soil surface. Modified from [16].

### 4.1.1 Vegetation Modelled by its Structural Elements

As an example of a detailed vegetation scattering model we shortly introduce the Michigan Microwave Canopy Scattering Model (MIMICS) model that was developed by Ulaby et al. in 1990 [22] for modelling backscatter of forest in the 1-10 GHz frequency range. MIMICS is a first-order solution of the radiative transfer equation for a tree canopy that contains different structure elements, illustrated in Figure 4.2. Its major attributes are as follows:

1. The canopy is divided into three regions: (i) the crown region, (ii) the trunk region, and (iii) the underlying ground region.

2. The crown regions of individual trees are described by ellipsoids with height $d$ and diameter $D_c$.

3. The crown region may contain branches and needles, modelled as dielectric cylinders and charac-

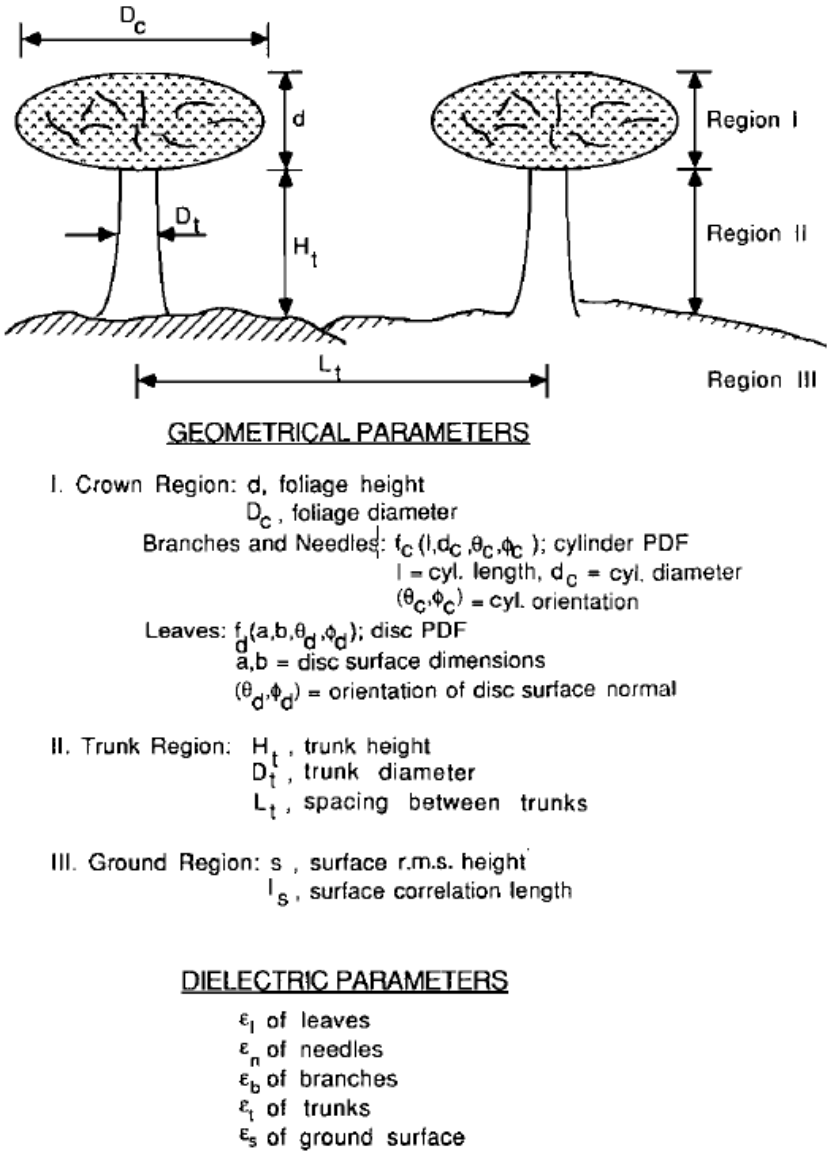



Figure 4.2: Relevant canopy characteristics. [22]

terized by a joint probability density function (PDF) $f_c(l, d_c, \theta_c, \phi_c)$. $l$ and $d_c$ define the cylinder length and -diameter, and the angles $\theta_c$ and $\phi_c$ their orientation.

4. The crown region may also contain leaves, modelled as flat rectangular discs and characterized by a joint PDF $f_d(a, b, \theta_d, \phi_d)$. $a$ and $b$ are the dimensions of the leaves and $\theta_d$ and $\phi_d$ their orientation.

5. The trunk region is characterized by the joint PDF $f_t(H_t, D_t, \theta_t, \phi_t)$. $H_t$ and $D_t$, are the average height and diameter, and $\theta_t$ and $\phi_t$ the orientation of the trunks.

6. The number of trees per unit area is defined by the number density $N$. The trees are assumed to be randomly distributed within the modelled region. A continuous canopy is assumed.

7. The roughness of the underlying ground surface is characterized by a roughness correlation function. It may be soil or water (as in a swamp).

8. The dielectric constants of the branches ($\varepsilon_b$), needles ($\varepsilon_n$), leaves ($\varepsilon_l$), trunks ($\varepsilon_t$), and the soil surface ($\varepsilon_s$) are specified by a set of dielectric models in terms of their water content, microwave frequency, and physical temperature

9. As a first-order model, it includes only scattering processes that involve single scattering by each region and double scattering by pairs of regions.

Obviously, it is not easy to obtain reliable information about all the above described vegetation characteristics. Therefore, the parameterization of MIMICS and similar models usually lacks reference data and over larger scales it is often pragmatic to use a simplified description of the vegetation layer. Such an approach will be described in the following section.

## 4.1.2 Vegetation Modelled as a Homogeneous Medium

### 4.1.2.1 Model Formulation

Let us assume that vegetation can be treated as a homogeneous medium (air) embedded randomly with particles (vegetation elements). To quantity scattering and absorption by the vegetation and the underlying soil we use the following model parameters: the absorption and scattering coefficients $\kappa_a$ and $\kappa_s$, which describe the strength of absorption and scattering of the particles within the volume, the phase function $\mathbf{P}$, which describes the directional scattering behaviour of the vegetation particles, and the ground scattering phase $\mathbf{G}$, which describes the directional scattering behaviour and absorption properties of the soil layer. The geometry of the problem is illustrated in Figure 4.3.

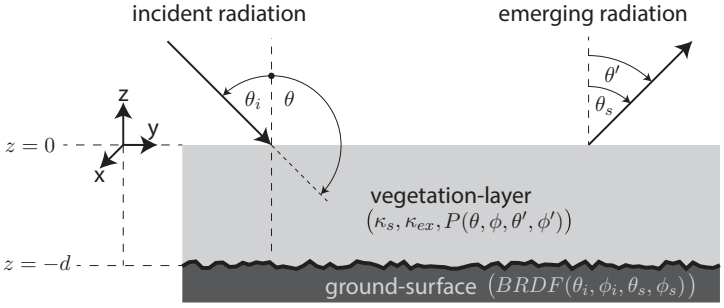



Figure 4.3: Vegetation scattering problem with a homogeneous vegetation layer above a rough soil surface.

In order to find a description for the up-welling radiation emerging from the top of the vegetation-layer, we will now split the radiative-transfer equation (3.13) into two parts: one for up-welling radiation and one for down-welling radiation. This will help us to specify the associated boundary-conditions of the above problem-geometry in a clear way. To perform the desired sparation of equation (3.13), we introduce an explicit notation for up-welling and down-welling radiation: (For the sake of compactness, the azimuth-angle ($\phi$) and the distance $r$ (or later $z$) will be omitted as functional arguments if they

are clearly deducible within the given context.)

$$L(\theta) = L^+(\theta) + L^-(\pi - \theta) \quad \text{with}$$

$$L^+(\theta) = \begin{cases} L(\theta) & \theta \in \left[0, \frac{\pi}{2}\right] \\ 0 & \text{else} \end{cases} \qquad L^-(\theta) = \begin{cases} L(\pi - \theta) & \theta \in \left[\frac{\pi}{2}, \pi\right] \\ 0 & \text{else} \end{cases}$$

The reason why $L^-(\theta)$ is introduced in that particular way is to ensure that the arguments of both $L^-(\theta)$ and $L^+(\theta)$ are defined as zenith-angles.

$$\begin{aligned} \frac{\partial L^+(\theta)}{\partial r} + \kappa_e L^+(\theta) + \frac{\partial L^-(\pi-\theta)}{\partial r} + \kappa_e L^-(\pi-\theta) =& \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_0^{\pi/2} L^+(\theta') P(\theta', \theta) d\Omega' \\ &+ \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_{\pi/2}^{\pi} L^-(\pi - \theta') P(\theta', \theta) d\Omega' \end{aligned} \tag{4.2}$$

We can now re-combine the separated integrals, by substituting $\tilde{\theta} = \pi - \theta'$ in the second integral:

$$\begin{aligned} \frac{\partial L^+(\theta)}{\partial r} + \kappa_e L^+(\theta) + \frac{\partial L^-(\pi-\theta)}{\partial r} + \kappa_e L^-(\pi-\theta) =& \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_0^{\pi/2} L^+(\theta') P(\theta', \theta) d\Omega' \\ &- \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_{\pi/2}^{0} L^-(\tilde{\theta}) P(\pi - \tilde{\theta}, \theta) d\tilde{\Omega} \end{aligned} \tag{4.3}$$

Using the minus in front of the second integral to switch the boundaries, and re-naming the dummy-variable $\tilde{\theta}$, we thus find:

$$\frac{\partial L^+(\theta)}{\partial r} + \kappa_e L^+(\theta) + \frac{\partial L^-(\pi-\theta)}{\partial r} + \kappa_e L^-(\pi-\theta) = \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_0^{\pi/2} L^+(\theta') P(\theta', \theta) + L^-(\theta') P(\pi - \theta', \theta) d\Omega'$$

Now we can replace the path-length $r$ with the distance within the canopy $z = r\cos(\theta)$ and write down two coupled differential-equations, one for up-welling radiation and one for down-welling radiation:

**Up-welling radiation:** $\theta \in [0, \frac{\pi}{2}]$:

$$\cos(\theta)\frac{\partial L^+(\theta)}{\partial z} + \kappa_e L^+(\theta) + = \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_0^{\pi/2} L^+(\theta') P(\theta', \theta) + L^-(\theta') P(\pi - \theta', \theta) sin(\theta') d\theta' d\phi'$$

Introducing a description with respect to the cosines of the polar-angle i.e.
$\mu = \cos(\theta) \quad \Rightarrow \quad d\Omega = \sin(\theta) d\theta d\phi = -d\mu d\phi$ we finally arrive at:

$$\mu\frac{\partial L^+(\mu)}{\partial z} + \kappa_e L^+(\mu) = \frac{\kappa_s}{4\pi} \int_0^{2\pi} \int_0^{1} L^+(\mu')(\mu, \mu') + L^-(\mu') P(-\mu', \mu) d\mu' d\phi'$$

**Down-welling radiation:** $\theta \in [\frac{\pi}{2}, \pi]$

$$\cos(\theta)\frac{\partial L^-(\pi-\theta)}{\partial z} + \kappa_e L^-(\pi-\theta) = \frac{\kappa_s}{4\pi}\int_0^{2\pi}\int_0^{\pi/2} L^+(\theta')P(\theta',\theta) + L^-(\theta')P(\pi-\theta',\theta)sin(\theta')d\theta' d\phi'$$

Substituting $\pi - \theta \rightarrow \theta$ and similarly introducing the description with respect to the cosines we find:

$$-\mu\frac{\partial L^-(\mu)}{\partial z} + \kappa_e L^-(\mu) = \frac{\kappa_s}{4\pi}\int_0^{2\pi}\int_0^{1} L^+(\mu')P(\mu',-\mu) + L^-(\mu')P(-\mu',-\mu)d\mu' d\phi'$$

Finally, the separation of the radiative transfer equation can now be written in a very compact form as (where $\mu$ represents the cosine of the zenith-angle, i.e. $\mu = \cos(\theta)$):

$$\pm\mu\frac{\partial L^\pm(\mu)}{\partial z} + \kappa_e L^\pm(\mu) = F^\pm(\mu) \qquad F^\pm(\mu) = \frac{\kappa_s}{4\pi}\int_0^{2\pi}\int_0^{1} L^+(\mu')P(\mu',\pm\mu) + L^-(\mu')P(-\mu',\pm\mu)d\mu' d\phi' \tag{4.4}$$

For a complete description of our scattering problem in Figure 4.3 two more definitions remain, namely the boundary conditions at $z = 0$ and $z = -d$, i.e., at the top of the vegetation layer and the vegetation-soil boundary. At the canopy top ($z = 0$), no scattering or absorption process has been taken place yet. Hence, the downward radiation must be equal to the total incident radiation. Since we consider incident-radiation from a single direction, the top-layer boundary condition reads:

$$L^-(z=0,\mu,\phi) = L_0\ \delta(\mu-\mu_i)\delta(\phi-\phi_i) \tag{4.5}$$

where $\delta$ denotes the Dirac-delta function and $(\mu_i, \phi_i)$ is the zenith- and azimuth angle of the incoming radiation.

At the soil boundary ($z = -d$), the upward and downward radiances are related through the ground scattering properties, which are expressed through the **bidirectional reflectance distribution function** $BRDF$ (whose arguments are defined as zenith-angles):

$$L(-d,\theta,\phi) = \int_0^{2\pi}\int_{\frac{\pi}{2}}^{\pi} L(-d,\theta',\phi')\cos(\pi-\theta')BRDF(\pi-\theta',\phi',\theta,\phi)\sin(\theta')d\theta' d\phi' \tag{4.6}$$

Expressed in terms of the up- and down-welling radiances (defined with respect to the cosines of the zenith-angles) this condition reads:

$$L^+(-d,\mu,\phi) = \iint_{0\,0}^{2\pi\,1} \mu' BRDF(\mu',\phi',\mu,\phi)L^-(-d,\mu')d\mu' d\phi' \tag{4.7}$$

The $BRDF$ can be seen as the bare soil scattering model for the soil layer underneath the vegetation, similar to the scattering models that were discussed in Section 4.2, but now built by means of radiative

transfer theory. Similar to the phase function $P$, the $BRDF$ determines the fraction of radiation incoming from a certain direction $(\theta', \phi')$, which is scattered into the propagation direction $(\theta, \phi)$.

If we consider an incident brightness $L_0$ incoming from a single direction (i.e. we set $L^-$ in (4.7) to $L^-(\mu', \phi') = L_0\delta(\mu' - \mu_i)\delta(\phi' - \phi_i)$, we find:

$$L^+(\mu, \phi) = \mu_i \, BRDF(\mu_i, \phi_i, \mu, \phi)L_0 \tag{4.8}$$

We can see that, unlike for $P$, the integral of the $BRDF$ over all space angles is not equal to one, but equal to the "directional hemispherical reflectance" of the surface. That is, it depends on the fraction of incident radiation that is transmitted and absorbed by the soil medium. It can be written as:

$$r = \frac{1}{L_0} \int_0^{2\pi} \int_0^1 L^+(\mu', \phi')d\mu' d\phi' = \int_0^{2\pi} \int_0^1 \mu_i \, BRDF(\mu_i, \phi_i, \mu, \phi)d\mu \, d\phi \tag{4.9}$$

Finally it shall be noted that the $BRDF$ is expected to obey (Helmholtz)-reciprocity, i.e. it is invariant when interchanging the roles of the incoming- and outgoing direction.

#### 4.1.2.2 First-Order Solution

Let us now seek a first-order solution for the volume scattering coefficient $\kappa_s$, i.e., terms $\kappa_s^2$ and higher order will be neglected. For this, we have to convert the differential radiative transfer equations into integral equations, integrate the boundary conditions and finally seek for an iterative solution. To achieve this, we have to consider the unknown integral terms in (4.4) as **source terms $\mathbf{F}^{\pm}$**, since these are responsible for an energy gain due to scattering from upward and downward directions into propagation direction and solve the differential equations as though they are known functions.

A set of formal solutions to the equations 4.4 is given by:

$$L^+(z, \mu) = e^{-\frac{\kappa_e}{\mu}(z+d)} \; L^+(-d, \mu) + \frac{1}{\mu} \int_{-d}^{z} e^{-\frac{\kappa_e}{\mu}(z-z')} F^+(z', \mu)dz' \tag{4.10}$$

$$L^-(z, \mu) = e^{\frac{\kappa_e}{\mu}z} \; L^-(0, \mu) + \frac{1}{\mu} \int_{z}^{0} e^{\frac{\kappa_e}{\mu}(z-z')} F^-(z', \mu)dz' \tag{4.11}$$

where $L^+(-d, \mu)$ and $L^-(0)$ are given by the boundary-conditions (4.5) and (4.7). Since the source terms $F^{\pm}$ are again functions of the upward and downward radiance $L^{\pm}$, the above equations are not real solutions but can be used as starting point for deriving an approximate solution.

The principal way to go for deriving such an approximate solution is to iteratively insert the above representations for $L^{\pm}$ into the $F^{\pm}$ integrals. Since the $F^{\pm}$ integrals are itself proportional to the scattering-coefficient $\kappa_s$, the process will result in a series whose terms are successively multiplied by $\kappa_s^n$, i.e. in a so-called "successive orders of scattering series".

Under the assumption that the scattering-coefficient is small (i.e. $\kappa_s << 1$), only the first terms of this series will contribute a significant contribution to the derived representation, and all other terms can

be neglected. If we omit all terms with orders of $\kappa_s$ higher than 1, only the following five terms will remain in the final representation for the up-welling radiance at the top of the canopy $L^+(z=0,\mu)$:

$$L^+ = L^+_{vol} + L^+_{surf} + L^+_{sv} + L^+_{vs} + L^+_{svs} + \mathcal{O}(\kappa_s^2) \tag{4.12}$$

The **volume-contribution** $L_v$, is the contribution to total radiance which is due to direct scattering of the incoming wave by the vegetation elements.

$$L^+_{vol}(z=0,\mu) = \frac{1}{4\pi}\int_{-d}^{0}\left[L_0\frac{\kappa_s}{\mu}e^{\frac{-\kappa_e}{\mu}z'}e^{-\frac{\kappa_e}{\mu}(z-z')}P(-\mu_i,\phi_i,\mu,\phi)\right]dz' \tag{4.13}$$

under the assumption that $\kappa_s$, $\kappa_e$ and $P$ are constant throughout the canopy, the z-integration can directly be evaluated and we find:

$$L^+_{vol}(z=0,\mu) = L_0\,\frac{\omega}{4\pi}\frac{\mu_i}{\mu_i+\mu}\left(1-e^{-\frac{\tau}{\mu_i}}e^{-\frac{\tau}{\mu}}\right)P(-\mu_i,\phi_i,\mu,\phi) \tag{4.14}$$

where the **single-scattering albedo** $\omega$ and the **optical depth** $\tau$ have been introduced via:

$$\omega = \frac{\kappa_s}{\kappa_e} \qquad\qquad \tau = \kappa_e d \tag{4.15}$$

The **surface-contribution** $L_s$, represents scattering from the soil surface, attenuated by the overlying vegetation.

$$L^+_{surf}(z=0,\mu) = L_0\,e^{-\frac{\tau}{\mu_i}}e^{-\frac{\tau}{\mu}}\mu_i BRDF(\mu_i,\phi_i,\mu,\phi) \tag{4.16}$$

The **volume-surface-term** $L^+_{vs}$, the **surface-volume-term** $L^+_{sv}$ and the **surface-volume-surface-term** $L^+_{svs}$ are contributions to the total radiance that originate from multiple-scattering events.

$L^+_{vs}$ represents incoming radiation that is scattered once by the vegetation-elements towards the surface, and then reflected back at the surface into the direction of the detector. Likewise, $L^+_{sv}$ represents the attenuated incoming radiation that is scattered first by the surface and then scattered again by the vegetation elements into the direction of the detector. Finally, $L^+_{svs}$ represents the attenuated incoming radiation that is scattered first by the surface, then scattered again by the vegetation towards the surface and then scattered another time by the surface towards the direction of the detector. Even though $L_{svs}$ is proportional to $\kappa_s$, it will in the following be neglected since it involves two scattering-events by the soil-surface.

The surface-volume and volume-surface term are given by:

$$L^+_{vs} = L_0\,\frac{\omega}{4\pi}\,e^{-\frac{\tau}{\mu}}F_{vs}(\mu,\phi) \qquad\qquad L^+_{sv} = L_0\,\frac{\omega}{4\pi}\,e^{-\frac{\tau}{\mu}}F_{sv}(\mu,\phi) \tag{4.17}$$

with

$$F_{vs}(\mu,\phi) = \mu_i \int_0^{2\pi}\int_0^1 \frac{\mu'}{\mu_i - \mu'}\left(e^{-\frac{\tau}{\mu_i}} - e^{-\frac{\tau}{\mu'}}\right) BRDF(\mu',\phi',\mu,\phi)P(-\mu_i,\phi_i,-\mu',\phi')d\mu' d\phi'$$

$$F_{sv}(\mu,\phi) = \mu_i \int_0^{2\pi}\int_0^1 \frac{\mu'}{\mu_i - \mu'}\left(e^{-\frac{\tau}{\mu}} - e^{-\frac{\tau}{\mu'}}\right) BRDF(\mu_i,\phi_i,\mu',\phi')P(\mu',\phi',\mu,\phi)d\mu' d\phi'$$

Together, they represent the (first-order) surface-volume interaction-term denoted as $L^+_{int}$.

$$L^+_{int}(\mu,\phi) = L^+_{vs}(\mu,\phi) + L^+_{sv}(\mu,\phi) = L_0 \frac{\omega}{4\pi} e^{-\frac{\tau}{\mu}}\left[F_{vs}(\mu,\phi) + F_{sv}(\mu,\phi)\right] \tag{4.18}$$

#### 4.1.2.3 $\omega$-$\tau$ Model

In practice, we are usually dealing with measurements of monostatic ($\theta_s = \theta_i, \phi_s = \phi_i + \pi$) backscattering-coefficients ($\sigma^0(\theta,\phi)$). Using equation (3.21) and (3.23) we can express the backscattering coefficient as follows:

$$\sigma^0 = \sigma^0_{vol} + \sigma^0_{surf} + \sigma^0_{int} \tag{4.19}$$

$$\sigma^0_{vol} = \mu_i \frac{\omega}{2}\left(1 - e^{-\frac{2\tau}{\mu_i}}\right) P(-\mu_i,\phi_i,\mu_i,\phi_i+\pi) \tag{4.20}$$

$$\sigma^0_{surf} = \underbrace{4\pi\mu_i BRDF(\mu_i,\phi_i,\mu_i,\phi_i)\mu_i}_{\sigma^0_{soil}} e^{-\frac{2\tau}{\mu_i}} \tag{4.21}$$

$$\sigma^0_{int} = \mu_i\, \omega\, e^{-\frac{\tau}{\mu}}\left(F_{vs}(\mu_i,\phi_i) + F_{sv}(\mu_i,\phi_i)\right) \tag{4.22}$$

If we now choose $P$ to be represented by a Rayleigh phase function as defined in Equation (3.5), neglect all multiple-scattering contributions and introduce the **two-way transmissivity** $\gamma^2$ as:

$$\gamma^2(\theta) = e^{-\frac{2\tau}{\cos\theta}} \tag{4.23}$$

we arrive at a very compact representation of the monostatic backscattering coefficient:

$$\sigma^0(\theta) = \cos(\theta)\frac{3\omega}{4}\left(1 - \gamma(\theta)^2\right) + \sigma^0_{soil}(\theta)\gamma^2(\theta) \tag{4.24}$$

One can see that $\gamma^2$ controls the relative contributions of the vegetation layer and the soil surface, i.e. when $\gamma^2$ is large then backscatter is dominated by the soil rather by the vegetation, and vice versa.

The formulation (4.24) represents the zero-order solution of the radiative transfer equation, and is in essence identical to the **Water Cloud Model** which will be discussed in Section 4.3.1. It is a **parsimonious** model that uses only two parameters ($\omega$ and $\tau$ respectively $\gamma^2$) to describe the complex interactions of microwaves with vegetation. It is no surprise that this model is not able to deal with

all types of scattering phenomena in vegetation (even after calibrating the model parameters). Yet, a multitude of studies have demonstrated that this simple model (in one of its many variants) captures the key effects of vegetation on backscatter quite well. Therefore, it is at present the most widely used model to describe backscatter of vegetated soil.

## 4.2 Backscatter from Bare Soil

Let us now turn to the question of how to model backscatter from bare soils, i.e. $\sigma^0_{soil}(\theta)$, using theoretical (Section 4.2.1) and empirical (Section 4.2.2) approaches. Practically all widely used bare soil backscatter models assume that scattering takes predominately place at the air-soil interface. Yet, when the soil is dry the microwave radiation may penetrate several centimeters to decimeters into the soil, potentially leading to strong subsurface signals from rocks and stones (Section 4.2.3). Under such conditions the standard models fail to correctly describe backscatter.

### 4.2.1 Theoretical Models

When an electromagnetic wave impinges from above upon the boundary between two semi-infinite media, a portion of the incident wave is scattered backwards and the rest is transmitted forward into the lower medium. When the radiation penetrates only little into the sub-surface medium or when the lower medium is homogeneous then scattering takes only place at the surface boundary. In such a situation one speaks of **surface-scattering** and assumes a sharp border between the two media. Scattering is then dependent only on the geometric (roughness) and dielectric (soil moisture, soil composition) properties of the soil surface.

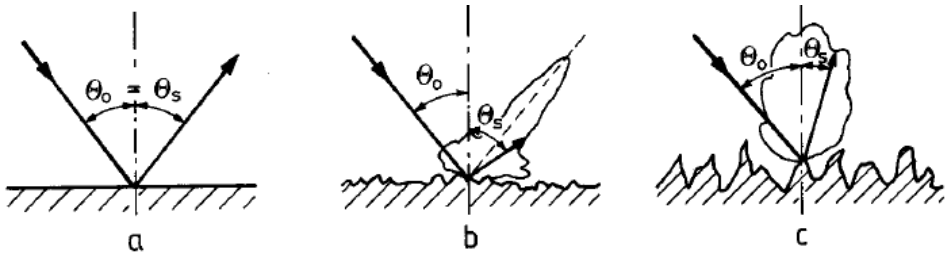



Figure 4.4: Specular and diffuse components for a a) perfect plane, b) slightly rough, and c) very rough surface.

Depending on the roughness of the soil surface (as compared with the radar wavelength), the incoming radar pulse is scattered quite uniformly into all directions (Figure 4.4c) or predominately in the forward direction (Figure 4.4b). Only when the soil surface is completely flat, specular reflection occurs (Figure 4.4a). This latter case probably never occurs in the real world, but it is still worth considering it because the reflection and transmission coefficients of a perfect plane can be precisely derived from electromagnetic theory. According to **Fresnel**, the reflection coefficients $\Gamma$ for the vertically ($\perp$) and

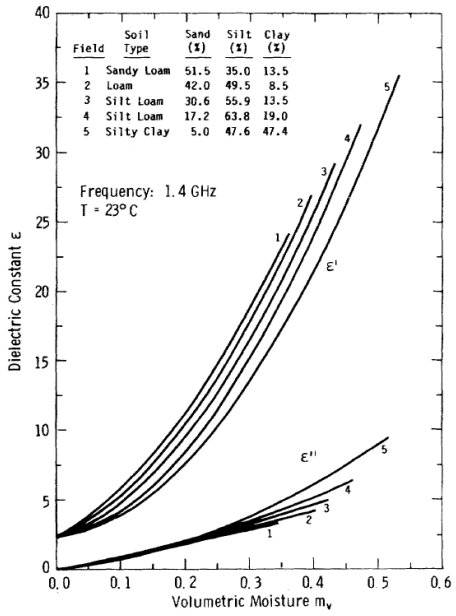



Figure 4.5: Dielectric constant for different soil types in dependency on the water content [9].

horizontally ($\|$) polarized electric field are

$$\begin{aligned} \Gamma_{\|} &= \frac{E^s_{0,\|}}{E^i_{0,\|}} = \frac{\cos\theta - \sqrt{\epsilon - \sin^2\theta}}{\cos\theta + \sqrt{\epsilon - \sin^2\theta}} \\ \Gamma_{\perp} &= \frac{E^s_{0,\perp}}{E^i_{0,\perp}} = -\frac{\epsilon\cos\theta - \sqrt{\epsilon - \sin^2\theta}}{\epsilon\cos\theta + \sqrt{\epsilon - \sin^2\theta}} \end{aligned} \tag{4.25}$$

where $\theta$ is the incidence angle, and $\epsilon = \epsilon' + i\epsilon''$ is the complex soil dielectric constant. Both the real part $\epsilon'$ (a measure of the polarizability of the medium) and imaginary part $\epsilon''$ (a measure of the absorption losses of the medium) part of $\epsilon$ strongly increase with increasing soil moisture (Figure 4.5). Therefore, also the Fresnel reflectivities strongly increase with soil moisture (Figure 4.6). In other words, the wetter the soil surface becomes the more energy is reflected upward rather than transmitted into the lower medium (i.e. soil). This is not only true for a perfectly flat soil surface, but is generally true for any rough soil.

To a first approximation, the effects of soil moisture and surface roughness on the scattered wave can be treated independently from each other. While soil moisture mostly affects the overall energy of the scattered wave, surface roughness affects its distribution. As the distribution shows pronounced patterns over all scattering directions $(\theta_s, \phi_s)$, surface roughness becomes the dominating factor when considering the mono-static case (i.e. backscatter).

Given the geometry of a soil surface is considerably easier to describe than that of vegetation, efforts to model scattering by a rough surface starting from Maxwell's equations have been more successful than for vegetation. Theoretical models such as the Geometric Optics Model, the Small Perturbations Model, or the **Integral Equation Method** (IEM) are widely used. Yet, even for rough surfaces important approximations must be made that are hardly fulfilled in nature. Therefore, also theoretical

bare soil backscatter models must be used with care, and it should not be expected that their model parameters can be matched with field measurements.

To quantify the roughness of a soil surface, theoretical models usually assume that the surface height profile $z(x)$ (Figure 4.7) can be represented by a Gaussian random process with zero mean. Its surface height probability distribution is thus given by

$$p(z) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{z^2}{2s^2}} \tag{4.26}$$

where is $s$ the standard deviation. It is called the **root mean square height** and is the most commonly used parameter for characterizing surface roughness. A second important roughness parameter is the **correlation length** which is a measure for the spatial correlation between surface points. It can be estimated from the normalized surface autocorrelation function which is defined through

$$\rho(x') = \frac{\int z(x)z(x+x')dx}{\int z^2(x)dx} \tag{4.27}$$

Note that the surface correlation function is characterized in a distinct direction but in most cases it is assumed to be isotropic. The surface correlation length is now defined as the distance between two distinct points, at which the normalized surface correlation function has dropped to $\frac{1}{e}$ (see Figure 4.7).

Note that the described parameterization of surface roughness represents a purely geometric concept. It does not account for the fact that for real soil surfaces the microwave pulses always penetrate to some extent into the soil, interacting with soil particles and water inclusions on a microscopic level. Therefore, it would be more correct to think of roughness as a kind of "dielectric roughness" rather than the roughness of the soil surface as perceived by the eye and optical and lidar measurement techniques.

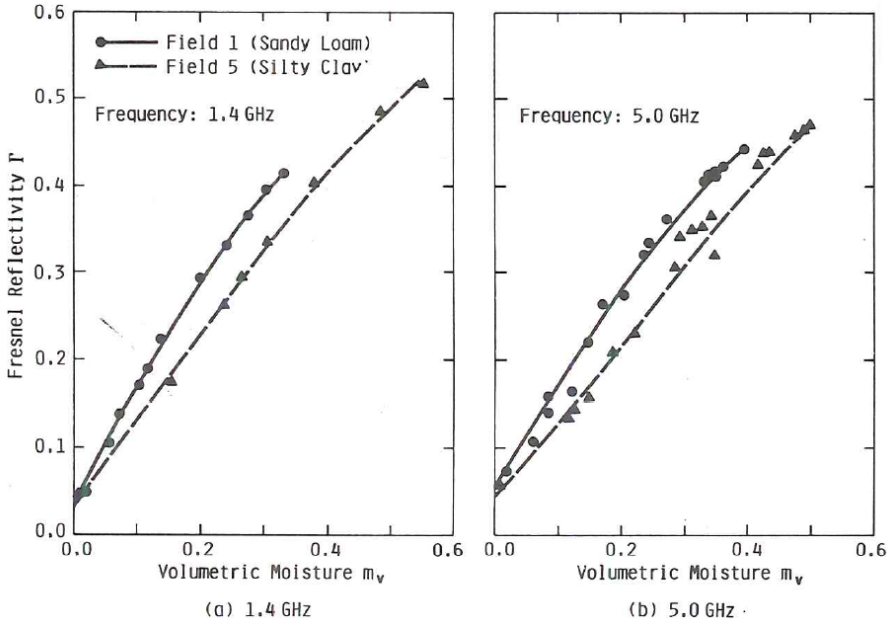



Figure 4.6: Fresnel reflectivity at normal incidence as a function of soil moisture. From [21] p. 1817.

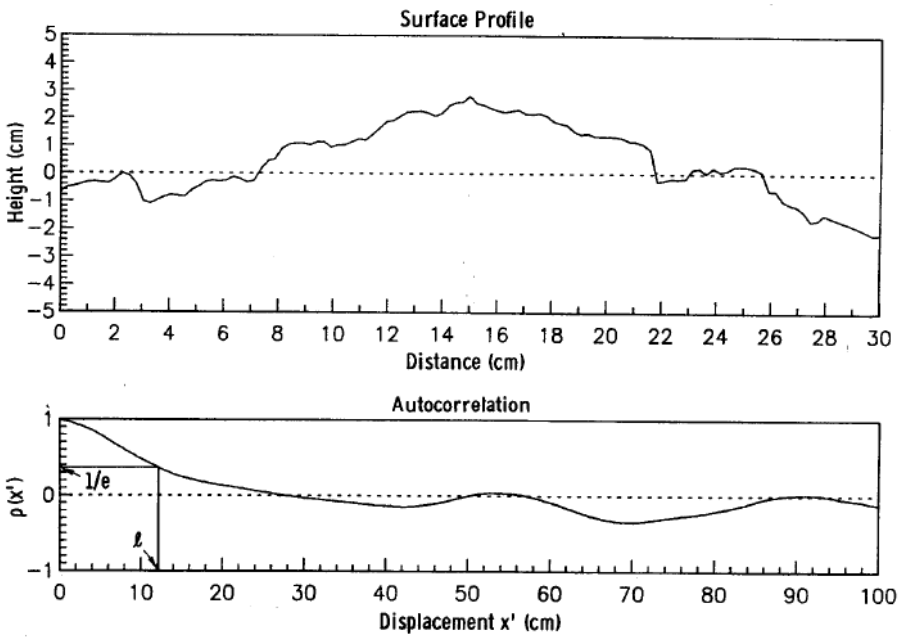



Figure 4.7: Example of a surface height profile (top) and the corresponding autocorrelation function (bottom) with correlation length $l$ [20].

Last but least, real soils may exhibit multi-scale roughness patterns, which is why multi-scale or even fractal representations would be needed to correctly parameterize the roughness of a soil surface.

### 4.2.2 Empirical models

Empirical models can be constructed from experimental data. The experiments must be informed by theory and collect sufficient and comprehensive reference data for calibrating the models. For constructing bare soil backscatter models one must collect data describing the soil surface properties (soil moisture, surface roughness, etc.) and the measurement configuration (frequency, polarization, incidence angle, etc.). As there are limits to how many data can be collected during one experiment, empirical models have by their very nature a limited validity range. Yet, when discussing the validity range of empirical models, one should distinguish the model structure and the concrete values of the model parameters. While the model parameters may indeed be valid only for the experimental data from which they were derived, the model structure may have a much wider validity range.

#### 4.2.2.1 Linear Model

The most basic model for describing backscatter from bare soils is a **linear model**. As numerous field experiments with ground-based radar devices have demonstrated, the backscattering coefficient $\sigma^0_{soil}$ exhibits a linear relationship with soil moisture when expressed in dB (Figure 4.8). Therefore, one can write

$$\sigma^0_{soil}[dB] = A + Bm_v \tag{4.28}$$

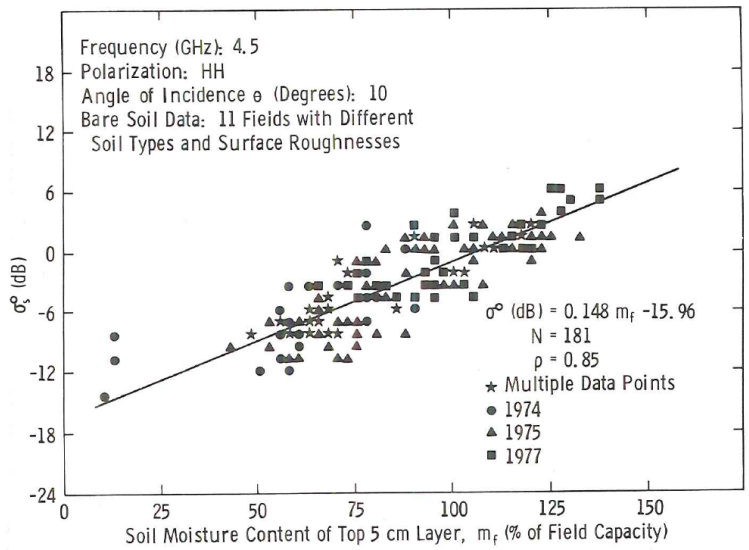



Figure 4.8: Measured backscattering coefficient of bare soil as a function of soil moisture for a variety of roughness conditions and soil textural compositions. From [20] p. 861.

where $m_v$ is the **volumetric soil moisture content** in $m^3m^{-3}$, and $A$ and $B$ are two empirical parameters that depend on the sensor (frequency, polarization) and soil surface (roughness, soil texture). The parameter $A$ represents the backscattering coefficient for a dry soil ($m_v = 0m^3m^{-3}$) and typically takes on values in the range from -20 to -5 dB. The parameter $B$ represents the sensitivity of the observable (i.e. $\sigma^0_{soil}$) to the target variable (i.e. $m_v$). As Figure 4.8 shows, $\sigma^0_{soil}$ changes by more than 15 dB when going from dry to wet conditions. The linear model turns into an exponential relationship when converting the backscattering coefficient to linear units $[m^2m^{-2}]$:

$$\sigma^0_{soil}[m^2m^{-2}] = 10^{\frac{A+Bm_v}{10}} = e^{\frac{ln10(A+Bm_v)}{10}} = ae^{bm_v} \tag{4.29}$$

Here we have made use of the formula $10^x = e^{ln10x}$.

Note that this simple linear model has been rejected by some scientists because theoretical bare soil backscatter models as discussed in Section 4.2.1 also exhibit non-linear behaviour. For example, IEM predicts that the $\sigma^0$ looses its sensitivity to soil moisture when the soil is wet. This conflict is essentially unresolved, and in practice it is up to the model developer to decide which approach to take.

#### 4.2.2.2 Oh-Model

Another widely used empirical model is the one developed by Oh et al. [14]. The **Oh-model** was developed for predicting the root mean square height and soil moisture from multi-polarized (VV, HH, HV) multi-incidence (10-70°) radar observations. Figure 4.9 shows the angular dependency of the VV and HV responses on soil moisture as well as the ratio between the HH and VV response. The co-polarized measurements $\sigma^0_{vv}$ have a significantly higher magnitude than the cross-polarized measurements $\sigma^0_{hv}$. The reason for this is that 90 degree polarization changes can only occur for very distinct scattering geometries that cause multiple reflections from oblique surface points. Backscatter

increases with increasing soil moisture for all polarization modes. The difference between VV and HV measurements decreases with increasing incidence angle, whereas it increases between the VV and HH response. For VV and HH polarization, the backscatter response to soil moisture is larger for large incidence angles, whereas for HV polarization it is larger for small incidence angles.

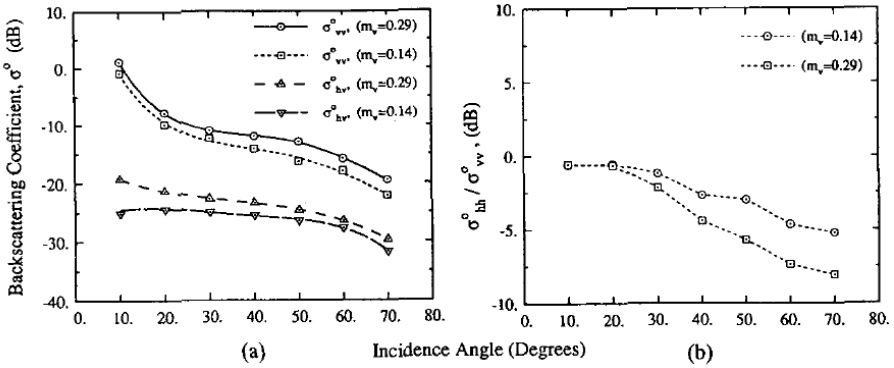



Figure 4.9: Angular responses of (a) $\sigma^0_{vv}$ and $\sigma^0_{hv}$, and (b) the ratio $\sigma^0_{hh}/\ \sigma^0_{vv}$ for different soil moisture conditions (modified from [14]).

Oh et al. modelled the different backscatter polarization responses indirectly using polarization ratios. Hence, in a first step the cross- and co-polarization rations were modelled separately and were then fitted to estimate the backscatter responses themselves. The polarization ratios are given as:

$$\begin{aligned} q &= \frac{\sigma^0_{hv}}{\sigma^0_{vv}} = 0.23\sqrt{\Gamma_0}(1 - e^{-ks}) \\ \sqrt{p} &= \sqrt{\frac{\sigma^0_{hh}}{\sigma^0_{vv}}} = 1 - \left(\frac{2\theta}{\pi}\right)^{\frac{1}{3\Gamma 0}} e^{-ks} \end{aligned} \tag{4.30}$$

$\Gamma_0$ is the Fresnel reflection coefficient in nadir direction, $ks$ is the root mean square height multiplied with the wave number, and $\theta$ is the incidence angle. The models are illustrated in Figure 4.10. The curves were empirically determined based on the observations and the shapes are defined as exponential functions, given through $f = \alpha\,(1 - \beta e^x)$. The main exponential shape of the function is given through the roughness $ks$. The dependency on soil moisture is indirectly given through the dielectric constant, which influences the Fresnel reflection coefficient. For the cross-polarization ratio $q$, soil moisture acts as a constant factor for the entire roughness range with an additional shape factor of 0.23, (i.e., $\alpha = 0.23\sqrt{\Gamma_0}$). That is, the higher the soil moisture content, the lower is the difference between $\sigma^0_{hv}$ and $\sigma^0_{vv}$. There is no incidence angle dependency on $q$ and no roughness dependent shape factor ($\beta = 1$). The shape of the cross-polarization ratio is slightly different. Both, the incidence angle and soil moisture define the shape of $q$ as a function of the roughness through $\beta = \left(\frac{2\theta}{\pi}\right)^{\frac{1}{3\Gamma 0}}$. In the possible range of the incidence angle between 0 and 90 degrees ($\theta \in \left[0, \frac{\pi}{2}\right]$), the factor $\left(\frac{2\theta}{\pi}\right)$ is below one. That is, a higher incidence angle means a stronger increase in the polarization difference with increasing soil moisture. This incidence angle and soil moisture dependency decreases with increasing surface roughness. In other words, increasing soil moisture content leads to stronger roughness-dependent scattering for VV

polarization at low incidence angles than for the HH polarization. There is no additional roughness independent shape factor ($\alpha = 1$), but note that the definition $f = \alpha\,(1 - \beta e^{x})$ in the co-polarized case refers to the square root the polarization ratio, i.e., $f = \sqrt{p}$.

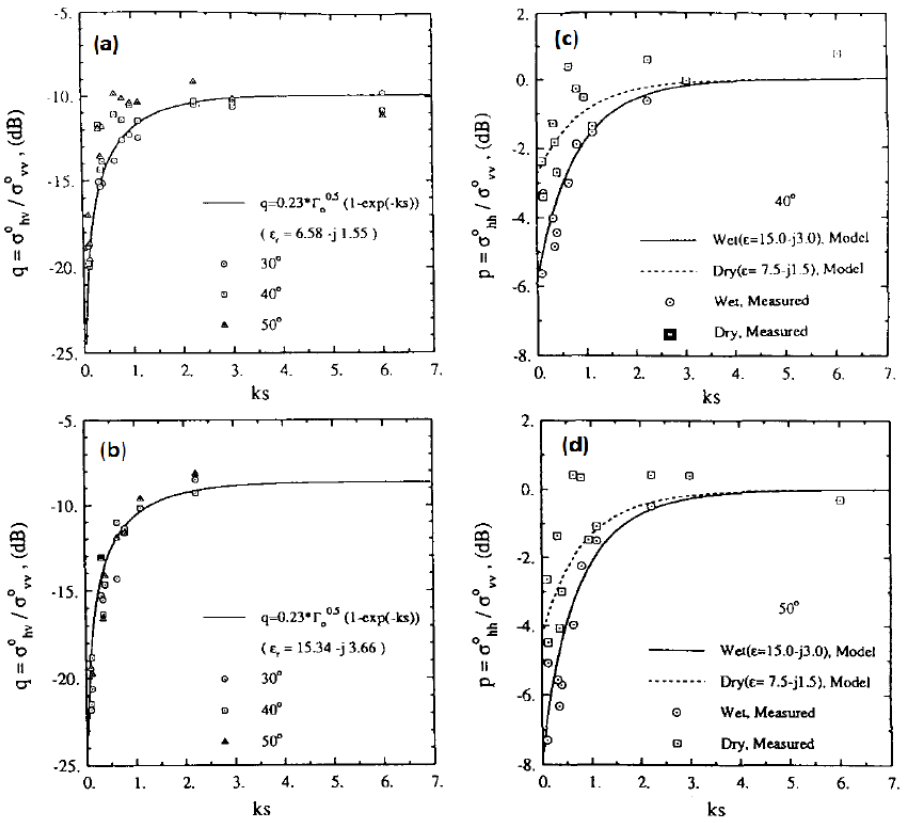



Figure 4.10: Roughness dependency of the cross-polarized ratio for (a) dry and (b) wet soils, and the co-polarized ratio at 40° (c) and 50° (d) (modified from [14]).

By defining a polarization ratio dependent functional form of $\sigma^0_{vv}$, and by using the ratios $p$ and $q$, the respective polarization responses are modelled through:

$$\begin{aligned}
\sigma^0_{vv}(\theta, \varepsilon, ks) &= \frac{g\cos^3\theta}{\sqrt{p}}[\Gamma_v(\theta) + \Gamma_h(\theta)] \\
\sigma^0_{hh}(\theta, \varepsilon, ks) &= p\,\sigma^0_{vv}(\theta, \varepsilon, ks) \\
\sigma^0_{hv}(\theta, \varepsilon, ks) &= q\,\sigma^0_{vv}(\theta, \varepsilon, ks)
\end{aligned} \tag{4.31}$$

with $g = 0.7(1 - e^{-0.65(ks)^{1.8}})$. $\sigma^0_{vv}$ and $\sigma^0_{hh}$ are both proportional to the average of the horizontally and vertically polarized Fresnel reflection coefficients $\Gamma_h$ and $\Gamma_v$. The HH component is smaller than the VV component by a multiplying factor $p$, which accounts for the difference in level between $\sigma^0_{vv}$ and $\sigma^0_{hh}$ for smaller values of $ks$, and introduces the relative soil moisture dependency.

Figure 4.11 shows the modelled backscatter values compared with measurements of three different surfaces with different roughness and moisture conditions. One can see a very good agreement between the model and the observations, except for some outliers in surface (a) at 10 degrees incidence angle. These are caused by a very strong coherent scattering component at near-nadir for very smooth surfaces,

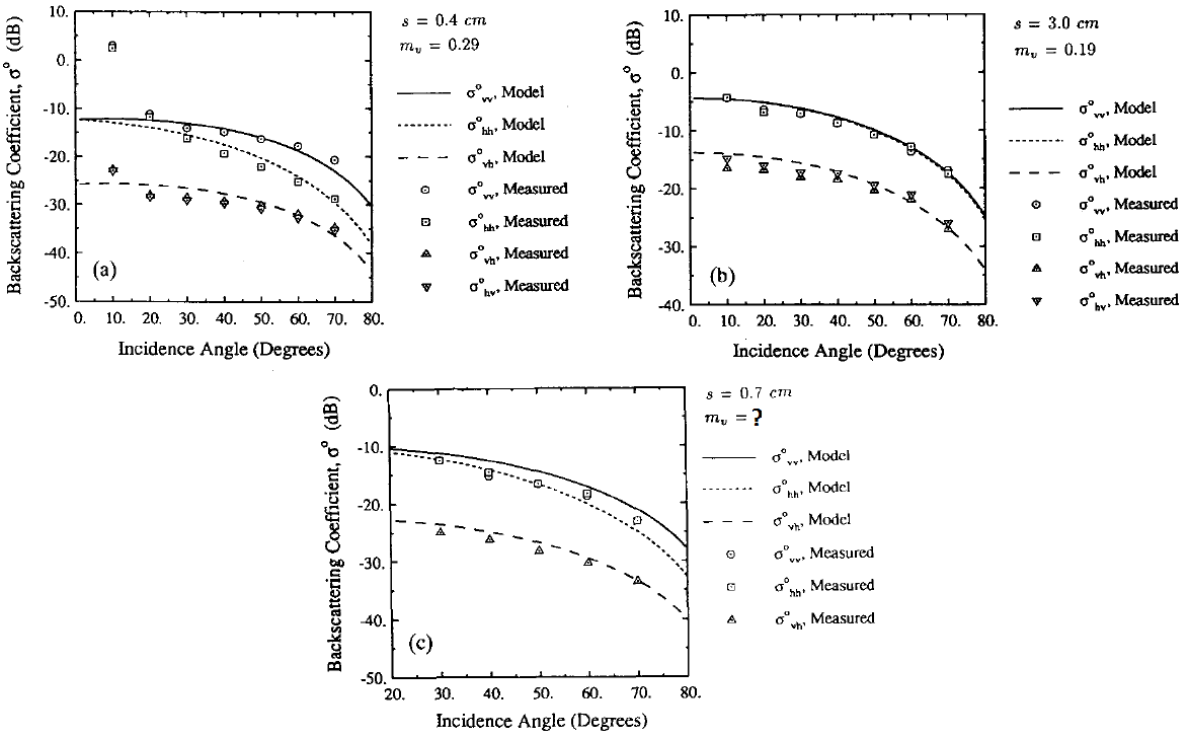



Figure 4.11: Comparison between modelled backscatter values and measurements for the different polarization modes at three surfaces with different roughness and moisture conditions. (modified from [14]).

which are not accounted for in the model.

### 4.2.3 Subsurface Scattering

Backscatter models may fail for many reasons. And there are situations where all models – independent of their nature – fail in the same way. To illustrate this point let us discuss one physical phenomenon which causes major problems in ASCAT soil moisture retrievals. Various validation studies with ASCAT have found that over desert regions and other arid environments, the C-Band backscatter data may exhibit an anomalous behaviour where backscatter increases with decreasing soil moisture ([27]). None of the bare soil backscatter models discussed in the previous two section can explain this. One hypothesis for explaining this phenomenon is that this may be related to the presence of brightly reflecting sub-surface scatterers that become visible when the soil dries. This hypothesis was recently confirmed by laboratory experiments conducted by Morrison [12]. In these experiments, one soil consisted of a gravel layer overlain by a soil with a smooth surface, whilst in a second realization the gravel was randomly mixed through the soil volume, and which presented a rough surface (Figure Figure 4.12). Rainfall was simulated by the addition of several millimetres of water. High-resolution, C-band **tomographic profiling** captured the backscattering drying curves, and their behaviours with incidence angles, polarization, and soil structure. The anomalous behaviour was observed in both co- and cross-polarized data, with the addition of water often producing immediate and strong diminution of backscatter. Backscattering profiles through the soil captured by the tomographic profiling imagery

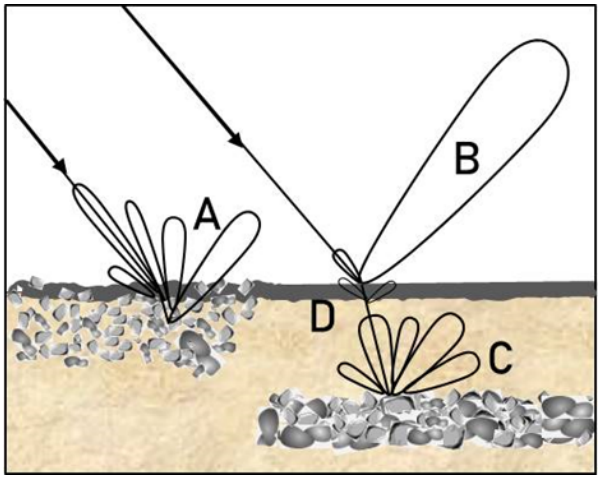



Figure 4.12: Summary of the scattering from the various soil types in a cross-sectional view through a soil. Left shows that the rough-surfaced mixed soil provides backscattering from both the surface and volume (A). Right shows that the smooth-surfaced layered soil forward scatters strongly (B), whereas the rough-surfaced buried layer provides backscatter (C). The figure also shows that wetting of the surface can cause the appearance of a 'rough surface interface' at the wetting front which can backscatter (D). From [12].

showed that the onset of anomalous behaviour was always associated with a shift of the dominant backscattering from the surface to the sub-surface.

## 4.3 Parsimonious Land Surface Backscatter Models

By combining the different vegetation- and bare soil backscatter models as discussed in the previous two sections one can create a multitude of models describing the interaction of C-band microwave pulses with the land surface. Here, we discuss only two parsimonious models, namely the Water Cloud Model (Section 4.3.1) and the TU Wien backscatter model (Section 4.3.2).

### 4.3.1 Water Cloud Model

The Water Cloud Model, developed by Attema and Ulaby in 1978 [1], is based on the fact that a vegetation canopy is usually composed of more than 99% air and that the dielectric constant of the vegetation is dominated by the dielectric constant of the water which is held in place by the organic vegetation matter. The assumptionis that a vegetation layer can be regarded as a cloud of randomly distributed water droplets which are held in place by the vegetation. The scattering geometry of such a problem is illustrated in Figure 4.13. The relevant parameters are the incidence angle $\theta$, the effective illuminated area $A_{ill}$, the vegetation layer height $h$, the volume of the illuminated vegetation $V$, and the water content per unit volume $W$. The original formulation of the Water Cloud Model is

$$\sigma^0 = C\cos\theta\left(1 - e^{-\frac{DWh}{\cos\theta}}\right) + A\cos\theta e^{Bm_v - \frac{DWh}{\cos\theta}} \tag{4.32}$$

where $A$, $B$, $C$, and $D$ are unknown model parameters, and $m_v$ is the volumetric soil moisture content. Note that the product $Wh$ represents the total water content in the vegetation canopy per surface area.

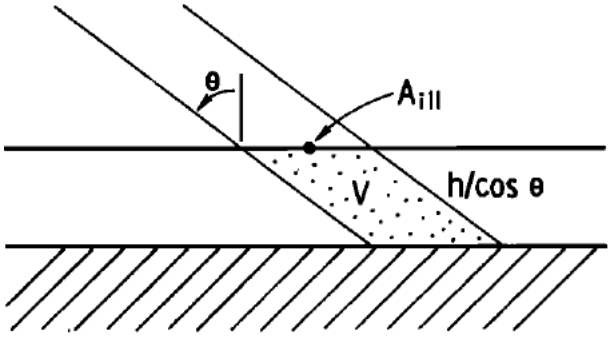



Figure 4.13: Assumed scattering geometry for the water cloud model [1].

Essentially, the Water Cloud Model is a zero-order radiative transfer solution for vegetation coupled with the linear bare soil backscatter model. This can be seen by substituting (4.29) into (4.24) and comparing the resulting terms to (4.32). Since the original publication [1] a multitude of modifications of the Water Cloud Model have been proposed. For example, many authors proposed to replace the term $Wh$ by more readily observable vegetation variables such as the Normalized Difference Vegetation Index (NDVI) or the Leaf Area Index (LAI).

### 4.3.2 TU Wien Backscatter Model

The TU Wien backscatter model aims to describe backscatter over vegetated surfaces in an easier way than radiative transfer models. A number of assumptions are made in order to develop a mathematically parsimonious model. One basic assumption of the model is that backscatter, when expressed in decibels, is linearly related to surface soil moisture content. We are already familiar with this approach for bare soil surfaces (cf Equation (4.28)). The TU Wien model generalizes this assumption to the case of a vegetated soil surface.

The TU Wien backscatter model takes into account that backscatter is strongly dependent on incidence angle and assumes that the slope and curvature of the relationship between backscatter and incidence angle are affected only by vegetation and surface roughness, but not by changes in soil moisture. This is illustrated by Figure 4.14 where a change in soil moisture from dry to wet soils shifts backscatter equally over all incidence angles. However, an increase in vegetation changes the slope of the backscatter and incidence angle relationship. As we have seen in previous sections, the dominant mechanisms contributing to the backscattering coefficient of vegetation are volume scattering in the vegetation canopy and surface scattering from the underlying soil surface. The incidence angle behavior of the volume and the surface scattering terms are distinctly different. With the exception of very rough surfaces, backscatter

from soil covered with dormant vegetation decreases rapidly with the incidence angle. In contrast, the contribution from the vegetation canopy is fairly uniform over a large range of incidence angles. Hence, over a fully grown vegetation the slope of the backscatter incidence angle relationship becomes less steep. Consequently, temporal changes in vegetation state can be parameterized using the slope and curvature of the backscatter incidence angle relationship.

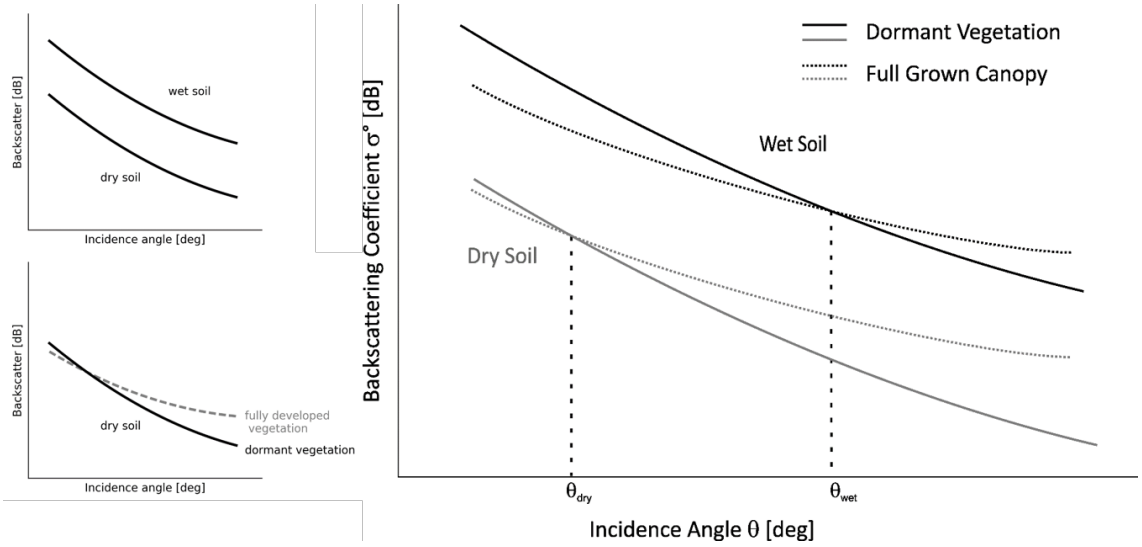



Figure 4.14: Relation between backscatter and incidence angle with relation to soil moisture and vegetation.

Another important assumption of the TU Wien backscatter model is the existence of so called dry- and wet **cross-over angles** ($\theta_{dry}$ and $\theta_{wet}$), at which changes in vegetation do not affect backscatter (Figure 4.14). The physical basis for this assumption is that vegetation tends to attenuate the signal from the soil surface at low incidence angles, resulting in lower total backscatter than for bare soils. At higher incidence angles the backscatter is higher compared to bare soil conditions, due to volume scattering from the vegetation. Therefore, there should be incidence angles where the attenuation of the observed bare soil backscatter and the contribution from the vegetation are in equilibrium, the so-called cross-over angles. This behaviour is also depicted by the $\omega$-$\tau$-model when assuming – as is often done – that vegetation growth/decay only changes the vegetation optical depth ($\tau$), but not the single-scattering albedo ($\omega$). Differentiating Equation (4.24) with respect to $\gamma^2$ shows that the cross-over angle can be found where

$$\frac{3\omega cos(\theta)}{4} = \sigma^0_{soil}(\theta) \tag{4.33}$$

Let us now come to the mathematical formulation of the TU Wien backscatter model which, from its conception, is a **change detection algorithm**. Changes in backscatter are modeled by defining a so-called dry reference, the lowest observed backscatter which depends on incidence angle, surface roughness and vegetation state, and a shift in backscatter which is mainly driven by changes in soil moisture and vegetation state. Consequently, backscatter is assumed to be dependent on the incidence

angle ($\theta$), soil moisture ($m_s$), surface roughness ($s$), and vegetation ($V$):

$$\sigma^0(\theta, m_s, s, V) = \sigma^0_{dry}(\theta, s, V) + \Delta\sigma^0(m_s, s, V) \tag{4.34}$$

The dry reference $\sigma^0_{dry}$ represents the backscatter that one would obtain under completely dry conditions. $\Delta\sigma^0$ is the shift in backscatter that we would get with increasing soil moisture and/or changing vegetation. Naturally, there is an upper boundary to the shift in backscatter, as soil moisture saturates. This upper boundary is the so called wet reference $\sigma^0_{wet}$, which is the equivalent to the dry reference, but for saturated soil conditions. Hence, we can express this backscatter shift $\Delta\sigma^0$ for a specific incidence angle as the difference between the dry and wet reference at this angle, scaled with the degree of saturation ($m_s$). The degree of saturation ranges between 0 and 1 and is just another unit to represent soil moisture:

$$\Delta\sigma^0(\theta, m_s, s, V) = m_s[\sigma^0_{wet}(\theta, s, V) - \sigma^0_{dry}(\theta, s, V)] \tag{4.35}$$

By substituting (4.35) into (4.34) we obtain:

$$\sigma^0(\theta, m_s, s, V) = \sigma^0_{dry}(\theta, s, V) + m_s[\sigma^0_{wet}(\theta, s, V) - \sigma^0_{dry}(\theta, s, V)] \tag{4.36}$$

The incidence angle behaviour of $\sigma^0$ is usually modelled using a second-order Taylor series expansion (i.e. a second-order polynomial function):

$$\sigma^0(\theta, s, V) = \sigma^0(\theta_{ref}, s, V) + \sigma'(\theta_{ref}, s, V)(\theta - \theta_{ref}) + \frac{1}{2}\sigma''(\theta_{ref}, s, V)(\theta - \theta_{ref})^2 \tag{4.37}$$

where $\theta_{ref}$ is the reference incidence angle, and $\sigma'$ and $\sigma''$ are the slope and curvature at the reference angle. The reference angle is normally chosen such as to lie in the middle of the measurement swath (e.g. 40° for ASCAT). Equation (4.36) in combination with Equation (4.37) constitutes the TU Wien backscatter model. This model has been applied to predict backscatter as measured by scatterometers (ASCAT, etc.) and Synthetic Aperture Radars (ENVISAT, Sentinel-1, etc.) (whereby for the latter one normally uses a linear incidence angle relationship instead of the second order polynomial as given by Equation (4.37)).

As one can see from Equations (4.36) and (4.37), the model has in total four model parameters, namely $\sigma^0_{dry}$, $\sigma^0_{wet}$, $\sigma'$, and $\sigma''$. All four model parameters depend on soil surface roughness ($s$) and vegetation status ($V$), and hence vary in both space and time. The standard route for estimating them is by fitting the model to actual observations. Hence, they become effective parameters which describe the backscattering behaviour of vegetation and soil in an implicit, surrogate manner. Note that for ASCAT the slope and curvature can be derived from the multi-angular measurements in a rather direct manner (Section 5.2.2). Hence, in case of ASCAT $\sigma'$ and $\sigma''$ can be regraded as observables rather than model parameters obtained via calibration.

# 5 Model Inversion

## 5.1 Approaches to the Inversion Problem

In the previous chapters we have learned that the forward models $f(\cdot)$ aim to predict the observable quantities $\mathbf{Y}$ based on a set of geophysical variables $\mathbf{x}$, controllable measurement conditions $\mathbf{\Omega}$, and model parameters $\mathbf{P}$ according to equation (1.1) (repeated here for ease of reading)

$$\mathbf{y} = f(\mathbf{x}, \mathbf{\Omega}, \mathbf{p}) \tag{5.1}$$

Depending on their origins and scope, forward models may take on vastly different forms. This has important implications when trying to invert them in order to estimate $\mathbf{x}$ from actual measurements $\mathbf{y}$.

In advanced **data assimilation systems** the mathematical complexity of the forward models $f(\cdot)$ is usually not such a big issue as the forward models (in this context often called forward operator) are coupled directly to the application models (e.g. numerical weather prediction, runoff forecasting) which are usually process oriented and hence simulating numerous variables that can used as input into the forward operator. Such an approach provides estimates $\hat{\mathbf{x}}$ that are consistent within the chosen application model. For the targeted application this may of course be highly beneficial. However, for other applications that rest on different model formulations, this may be problematic as assimilation based estimates $\hat{\mathbf{x}}$ may be more strongly driven by the chosen application model rather than by the measurements.

To obtain estimates of $\mathbf{x}$ that are agnostic of specific applications and as close as possible to the actual measurements $\hat{\mathbf{y}}$, forward models $f(\cdot)$ have to be inverted according to (1.2) (again repeated here for ease of reading)

$$\hat{\mathbf{x}} = g(\hat{\mathbf{y}}, \hat{\mathbf{\Omega}}, \hat{\mathbf{p}}) \tag{5.2}$$

using an inversion model $g(\cdot)$ that uses the actual measurements $\hat{\mathbf{y}}$ and approximations $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{p}}$ of the measurement conditions and model parameters as input. Depending on the mathematical complexity of $f(\cdot)$, we can distinguish three approaches to model inversion:

- **Direct inversion**: When $f(\cdot)$ has a mathematical simple form, with unambiguous relationships between the observables $\mathbf{y}$ and the geophysical variables $\mathbf{x}$, then it is possible to analytically invert $f(\cdot)$, i.e. $g(\cdot) = f^{-1}(\cdot)$.
- **Iterative nonlinear optimization**: When $f(\cdot)$ becomes more complex, which usually encom-

passes that ambiguities between $\mathbf{y}$ and $\mathbf{x}$ arise, a direct inversion is not possible any longer. In this case one may use iterative nonlinear optimization techniques to optimally fit the predictions of the forward model $f(\cdot)$ to the observed values $\hat{\mathbf{y}}$. In this case, as one works only with $f(\cdot)$, $g(\cdot)$ is not determined in practice. Nonetheless, conceptually, $g(\cdot) = f^{-1}(\cdot)$ also holds in this case.

- **Approximation**: When $f(\cdot)$ becomes so complex that also iterative optimization techniques fail then the only route left is to introduce additional assumptions to simplify the problem (e.g. by restricting the validity range of the model) and to find approximate functions of $f(\cdot)$. A popular approach is to train a neural network using $f(\cdot)$ and use the trained neural network for the retrieval. In this case $g(\cdot)$ represents a completely new mathematical model that resembles the hypothetic function $f^{-1}(\cdot)$ only in terms of its functional behaviour.

Section 5.2 illustrates the direct inversion approach based on the example of retrieving soil moisture from ASCAT measurements using the TU Wien backscatter model. Section 5.3 discusses various iterative nonlinear optimization techniques, and Section 5.4 provides a brief discussion of approximation techniques, focusing on neural networks.

## 5.2 Direct Inversion

A mathematically simple model that can be inverted directly is the linear bare soil backscatter model given by Equation (4.28). When its parameters $A$ and $B$ are known, it is straight forward to estimate the surface soil moisture content $m_v$ over bare soils from the backscatter measurements (expressed in dB):

$$m_v = \frac{\sigma^0_{soil} - A}{B} \tag{5.3}$$

In practice, the challenge is to find good estimates of the parameters $A$ and $B$, which must vary in space and time in order to correctly depict the strong influence of surface roughness on the backscatter measurements.

Similarly, the TU Wien backscatter model (4.36) can be inverted to estimate soil moisture over vegetation from backscatter measurements [25]:

$$m_s = \frac{\sigma^0(\theta, m_s, s, V) - \sigma^0_{dry}(\theta, s, V)}{\sigma^0_{wet}(\theta, s, V) - \sigma^0_{dry}(\theta, s, V)} \tag{5.4}$$

Again, the challenge is to find spatially and temporally varying estimates of the two model parameters $\sigma^0_{dry}$ and $\sigma^0_{wet}$, reflecting spatio-temporal patterns in surface roughness, land cover and vegetation status. For ASCAT, this is done in a stepwise manner. In the following, the most important steps of this procedure are described.

### 5.2.1 Slope and Curvature

Thanks to the fact that ASCAT measures the backscattering coefficient from different viewing directions, it becomes possible to estimate the slope and curvature as given by the second order polynomial (4.37) in a rather direct manner. As illustrated by Figure 3.1, the ASCAT instrument consists of three antennas on each side, called fore-, mid-, and aft-beam. At each overpass of a particular point, the sensor collects three backscatter measurements taken at different azimuth and incidence angles. Each of these **measurement triplets** can be used to obtain two slope estimates:

$$\begin{aligned} \sigma'\left(\frac{\theta_{mid}-\theta_{fore}}{2}\right) &= \frac{\sigma^0_{mid}(\theta_{mid})-\sigma^0_{fore}(\theta_{fore})}{\theta_{mid}-\theta_{fore}} \\ \sigma'\left(\frac{\theta_{mid}-\theta_{aft}}{2}\right) &= \frac{\sigma^0_{mid}(\theta_{mid})-\sigma^0_{aft}(\theta_{aft})}{\theta_{mid}-\theta_{aft}} \end{aligned} \tag{5.5}$$

$\sigma^0_{fore}$, $\sigma^0_{mid}$, and $\sigma^0_{aft}$ are the backscatter measurements from the fore-, mid-, and aft-beam, and $\theta_{fore}$, $\theta_{mid}$, and $\theta_{aft}$ the corresponding incidence angles, respectively. When having a sufficiently large number of measurements evenly distributed over the entire incidence angle range, the slope and curvature parameters at the chosen reference angle, i.e., $\sigma'(\theta_{ref})$ and $\sigma''(\theta_{ref})$, can be obtained by empirically fitting a linear function to the measurements:

$$\sigma'(\theta) = \sigma'(\theta_{ref}) + \sigma''(\theta_{ref})(\theta-\theta_{ref}) \tag{5.6}$$

Recall that the slope and curvature parameters are dependent on land cover and vegetation state. Therefore, these parameters are estimated for each pixel and each day of the year separately. That is, the slope and curvature parameters represent the incidence angle dependency of backscatter on the average vegetation conditions of the respective day of the year.

### 5.2.2 Incidence Angle Normalization

To directly compare ASCAT backscatter measurements across time and space it is necessary to normalize them to a reference angle. As ASCAT's incidence angle range is from about 25° to 64.5°, a reference angle of 40° was chosen to minimize interpolation errors. As $\sigma'(\theta_{ref})$ and $\sigma''(\theta_{ref})$ are already known, Equation (4.37) can be inverted to estimate the backscattering coefficient at the reference angle:

$$\sigma^0(\theta_{ref}) = \sigma^0(\theta) - \sigma'(\theta_{ref})(\theta-\theta_{ref}) - \frac{1}{2}\sigma''(\theta_{ref})(\theta-\theta_{ref})^2 \tag{5.7}$$

### 5.2.3 Cross-Over Angles

The only two parameters of the TU Wien model that are impossible to estimate from ASCAT measurements in a rather direct manner – and hence require careful calibration – are the cross-over angles $\theta_{dry}$ and $\theta_{wet}$. As discussed in Section 4.3.2, backscatter at these incidence angles is assumed to be insensitive to (temporal) changes in vegetation cover. Once these parameters are known, the effect of vegetation on backscatter at any other incidence angle can be estimated from the knowledge of the slope and curvature.

Based on an analysis of global long-term backscatter data of ASCAT and its predecessor ERS SCAT, the cross-over angles were fortunately found to remain stable over time, and appeared to be also independent of the geographic location. Their empirically derived values are $\theta_{dry}$=25° and $\theta_{wet}$=40°, respectively (see Figure 5.1).

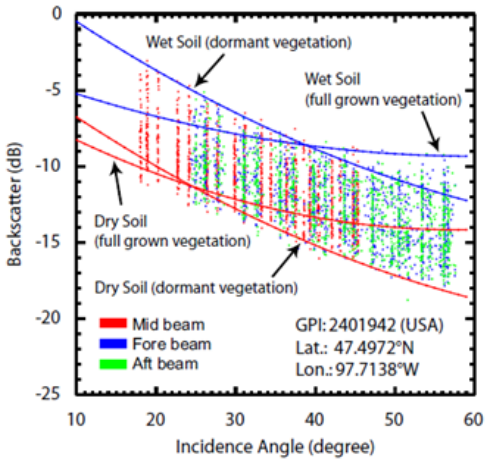



Figure 5.1: Empirical derivation of the dry and wet crossover angle.

### 5.2.4 Dry- and Wet Reference

Our initial problem was the estimation of the dry and wet reference parameters $\sigma^0_{dry}(\theta, s, V)$ and $\sigma^0_{wet}(\theta, s, V)$, in order to apply the inverse model (5.4). This can be done at the dry- and wet cross-over angles, because backscatter is at these angles is not sensitive to vegetation phenology. That is, backscatter changes are only caused by soil moisture changes. Therefore, to derive the dry reference, we can use the normalization function (5.7) to normalize historic ASCAT measurements to $\theta_{dry}$, which was, as mentioned above, empirically found to be 25°:

$$\theta_{ref} = \theta_{dry} = 25° \tag{5.8}$$

If we assume that the soil was completely dry at least once during the entire calibration period, then the lowest observed backscatter value normalized to the dry cross-over angle, must represent the dry reference. In order to account for possible outliers, not the lowest backscatter, but a certain percentile of the data (e.g. 5th percentile) or the median of the smallest measurements (e.g. of the 5 percent lowest values) is to be used.

The wet reference can be obtained in a completely analogue fashion. First, all measurements are normalized to $\theta_{wet}$, and then the wet reference is estimated assuming that the soil was at least a few times completely saturated. Note that the dry- and wet reference parameters are again calculated for each day of the year.

### 5.2.5 Soil Moisture Retrieval

After we have now calibrated the model, i.e., estimated all model parameters, we can finally apply the inverse model (5.4) in order to obtain the soil moisture retrievals. The retrieval is carried out at the chosen reference angle, which by pure chance is equal to the wet cross-over angle, i.e. $\theta_{ref}$=$\theta_{wet}$=40°. For this reason, only the $\sigma^0_{dry}(\theta, s, V)$ changes over time, whereas $\sigma^0_{wet}(\theta, s, V)$ is stable in time (Figure 5.2).

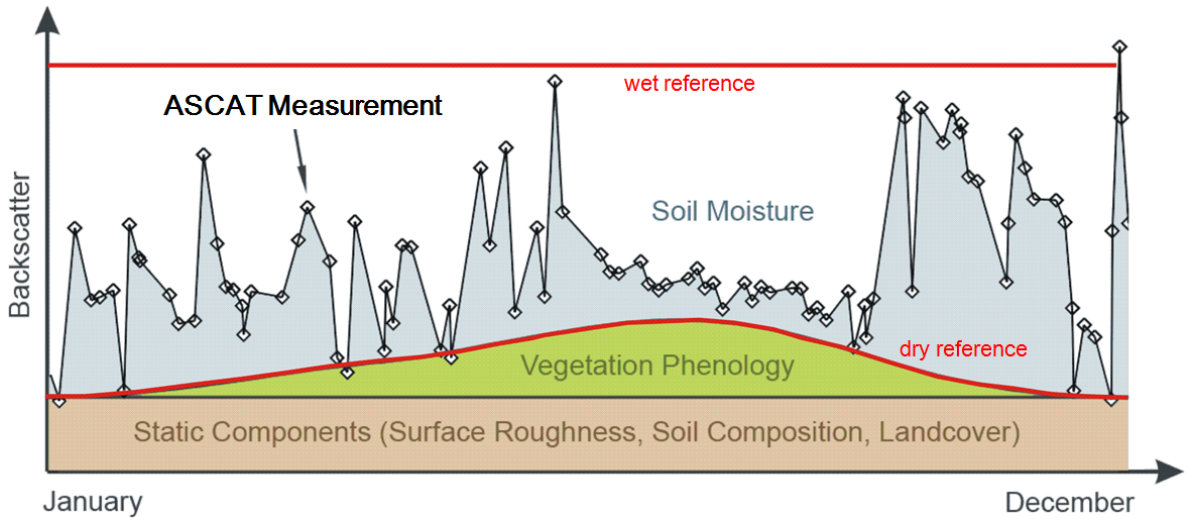



Figure 5.2: Illustration of the behaviour of the dry and wet reference.

In summary, in this section we have used the TU Wien backscatter model to demonstrate how a parsimonious model can be directly inverted through a series of sequential processing steps. The great advantages of such an approach are (i) that a direct analytic inversion is possible, and (ii) that the model parameters can be estimated from the historical backscatter observations themselves. The following sections will introduce more complex inversion techniques for forward models that can't be inverted analytically, as most of the models that were shown in Chapter 4.

## 5.3 Iterative Nonlinear Optimization

In this section, we discuss how nonlinear function minimization techniques can be used to compute model parameters (calibration), and/or predict the values of the geophysical variable(s) of interest (model inversion) from observations. The typical – and most important – scenario is the **minimization of a sum of squared differences**

$$h(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}, \mathbf{\Omega}_i, \mathbf{p}) - y_i)^2, \mathbf{x} \in \mathbb{R}^p \tag{5.9}$$

between observations $\mathbf{y} = (y_1, \ldots, y_N)$ and predictions based on the corresponding unknown realization of the variable, $\mathbf{x}$, and parameters $\mathbf{p}$, according to our forward model $f(\cdot)$. $N$ is the number of observations, and $i$ is the observation index and is applied as subscript, e.g. $\mathbf{\Omega}_i$ (the part of the measurement conditions that applies to the $i$-th observation). In the rest of this section, we will drop, for the sake of clarity of presentation, the distinction between parameters and variables, since it is of no relevance to the technicalities of non-linear optimization, and will only obfuscate the discussion by cluttering up the notation. For the same reason, we drop the use of the circumflex for observed or estimated quantities.[1]

In the context of function minimization, $h(\cdot)$ is often referred to as **cost** or **objective function**. Now, given a scalar function $h(\mathbf{x}), \mathbf{x} \in \mathbb{R}^p$, we want to compute $\mathbf{x} = \arg\min_{\mathbf{x}} h(\mathbf{x})$. For certain problem classes (e.g., linear least squares, see below), there exists a closed-form (i.e., one step) solution. If a closed-form solution does not exist or is not known, one can try to determine the minimum iteratively, by starting from an initial guess $\mathbf{x}^0$ and computing a series of increments $\Delta\mathbf{x}^k$ such that

$$h(\mathbf{x}^{k+1}) = h(\mathbf{x}^k + \Delta\mathbf{x}^k) < h(\mathbf{x}^k). \tag{5.10}$$

$k$ is the time or step index and is applied as superscript, e.g. $\mathbf{x}^k$.
$\mathbf{x} = (x_1, .., x_p)^T \in \mathbb{R}^p$ is the parameter vector, with $p$ being the number of variables/parameters and $\mathbf{x}^0$ the initial parameter vector estimate.
$\Delta\mathbf{x}^k = \mathbf{x}^{k+1} - \mathbf{x}^k$ is the $kth$ change in the parameter vector. It is proportional to the $kth$ descent direction.

By minimizing Equation (5.9), we hope to recover the the values of the variables and parameters which are, in the least squares sense, most compatible with the observed values. The choice of **least squares as goodness-of-fit criterion** has both historical and mathematical reasons; other criteria are possible, but least square techniques are among the best understood and most widely used.

Below we only discuss analytic methods that minimize a cost function, such as Gradient descent and Levenberg-Marquard. For difficult cost functions with many local minima and / or high curvature, **stochastic approaches** like *Simulated Annealing* are sometimes used. But note that there are other methods, such as the *Simplex method*, that use completely different approaches (i.e. that neither

[1] Note that all quantities involved in the actual optimization process are either observed or estimated. Moreover, in the context of linear and non-linear optimization, $\mathbf{y}$ usually refers to the observed values, and the "hat" quantity $\hat{\mathbf{y}}$ to the prediction of the observables obtained by applying the forward model $f(.)$ to the current estimate of the parameters/varaiables, $\mathbf{x}$. So the distinction we established in the introductory chapter (and which made good sense there) can not be maintained in this section.

compute nor use derivatives of the cost function). A good book on nonlinear optimization is [3]. The 'bible' of numerical programming [15] also contains some useful material, in particular with respect to the implementation of standard optimization algorithms.

### 5.3.1 Gradient Descent

It can be show that as long as the gradient does not vanish[2], i.e. $\nabla h(\mathbf{x}^k) \neq \mathbf{0}$, there exists $\alpha^k \in \mathbb{R}^+$ such that Eq. 5.10 holds with

$$\Delta\mathbf{x}^k = -\alpha^k \nabla h(\mathbf{x}^k). \tag{5.11}$$

The negative of the gradient $-\nabla h(\mathbf{x}^k)$ is known as **direction of steepest descent**: this is the direction along which $h(\cdot)$ decreases fastest if it were linearly continued from $\mathbf{x}^k$, i.e., the direction with the smallest directional derivative. Accordingly, optimization algorithms based on Eq. 5.11 are known as **steepest descent** methods.

A decrease in the cost function can be obtained not only for the *direction of steepest descent*, but for any direction $\boldsymbol{d}^k$ that fulfills the descent condition: $\nabla h(\mathbf{x}^k)^T \boldsymbol{d}^k < 0$. Such directions that 'point away' from the gradient are called *descent directions*. In many methods, the descent direction is obtained by pre-multiplying the steepest descent vector by a positive definite matrix $\boldsymbol{D}^k$, i.e.,

$$\boldsymbol{d}^k = -\boldsymbol{D}^k \nabla h(\mathbf{x}^k). \tag{5.12}$$

Thus, the descent condition becomes

$$\nabla h(\mathbf{x}^k)^T \boldsymbol{d}^k = -\nabla h(\mathbf{x}^k)^T \boldsymbol{D}^k \nabla h(\mathbf{x}^k) < 0.$$

Note that for positive definite $\boldsymbol{D}^k$, this condition always holds.

Steepest descent according to Equation (5.11) is a reliable standard method for minimization, as long as a suitable step-size is chosen. However, convergence tends to be slow compared to 2nd-order methods like Newton's method.

### 5.3.2 Newton's Method

By providing the 2nd derivatives of the cost function, a faster rate of convergence can be attained.

Given the 2nd order Taylor approximation of $h(\mathbf{x})$ about point $\mathbf{x}^k$:

$$h(\mathbf{x}) \approx h(\mathbf{x}^k) + \nabla h(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^k)^T \nabla^2 h(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) \tag{5.13}$$

[2] Remember that the gradient must become zero at the position of a minimum.

with Hessian matrix

$$\nabla^2 h(\mathbf{x}) = \left( \frac{\partial^2 h(\mathbf{x})}{\partial x_i \partial x_j} \right)_{1<=i,j<=p} \tag{5.14}$$

an approximation of its gradient is obtained by differentiating w.r.t. $\mathbf{x}$

$$\nabla h(\mathbf{x}) \approx \nabla h(\mathbf{x}^k) + \nabla^2 h(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) \tag{5.15}$$

A necessary condition for a minimum is that the gradient vanishes, i.e., $\nabla h(\mathbf{x}) = \nabla h(\mathbf{x}^{k+1}) = 0$. Thus, we arrive at the equation

$$-\nabla h(\mathbf{x}^k) = \nabla^2 h(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k), \tag{5.16}$$

which, provided the Hessian is invertible, can be solved for $\mathbf{x}^{k+1}$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\nabla^2 h(\mathbf{x}^k))^{-1} \nabla h(\mathbf{x}^k). \tag{5.17}$$

Equation (5.17) constitutes a step of the *pure Newton* method. Note that this is a special case of Equation (5.12). Hence, provided the Hessian is positive definite, the increment according to Equation (5.17) is indeed a descent direction.

Newton's method will lead to super-linear convergence when the approximation Equation (5.13) holds with positive definite Hessian. In the case of a quadratic function with constant positive definite Hessian, Newton's method will even jump to the optimal solution in a single step. However, convergence can be slow far away from the minimum; also, the method may fail to make any progress at all if the Hessian is not positive definite or even break down completely if the Hessian becomes singular. A practical issue is that the computation (and inversion) of the Hessian tends to be costly computationally, especially when the number of parameters is large. There exist several approaches that try to build the Hessian or the inverse Hessian incrementally from gradient estimates, for example the so called *conjugate gradient* methods. Another method, that addresses the issues mentioned above for the special case of least squares problems, is the **Levenberg-Marquardt** algorithm (see below).

### 5.3.3 Least Squares and the Gauss-Newton Method

If $h$ is of the form given in Equation (5.9):

$$h(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{N} (f(\mathbf{\Omega}_i, \mathbf{x}) - y_i)^2 = \frac{1}{2} \sum_{i=1}^{N} (f_i - y_i)^2 = \frac{1}{2} \|\mathbf{f} - \mathbf{y}\|^2, \mathbf{x} \in \mathbb{R}^p$$

we have a linear least squares problem. Its gradient and and Hessian are given by:

$$\nabla \mathbf{f}(\mathbf{x}^k) = \left( \frac{\partial \mathbf{f}}{\partial \mathbf{x}} | \mathbf{x} = \mathbf{x}^k \right)^T \in \mathbb{R}^{p \times N} \tag{5.18}$$

$$\begin{aligned} \nabla h(\mathbf{x}^k) &= \left( \left( \frac{\partial h}{\partial \mathbf{f}} | \mathbf{f} = \mathbf{f}(\mathbf{x^k}) \right) \left( \frac{\partial \mathbf{f}}{\partial \mathbf{x}} | \mathbf{x} = \mathbf{x}^k \right) \right)^T \\ &= \nabla \mathbf{f}(\mathbf{x}^k)(\mathbf{f}(\mathbf{x^k}) - \mathbf{y}) \in \mathbb{R}^p \end{aligned} \tag{5.19}$$

$$\nabla^2 h(\mathbf{x^k}) = \nabla \mathbf{f}(\mathbf{x}^k) \nabla \mathbf{f}(\mathbf{x}^k)^T + \sum_{i=1}^{N} (f_i(\mathbf{x^k}) - y_i) \nabla^2 f_i(\mathbf{x^k}). \tag{5.20}$$

By substituting these values into Equation (5.17), we could compute a step of the Newton method for the least squares problem Equation (5.9). The *Gauss-Newton method* is essentially such a Newton step, but in addition it assumes that the second term in Equation (5.20) - which contains the second order derivates of $f(\cdot)$ w.r.t. the parameters $\frac{\partial f}{\partial x_i \partial x_j}$ - is zero. In other words, the Gauss-Newton method is the Newton method applied to a least square problem, which replaces the true Hessian Equation (5.20) with

$$\nabla^2 h(\mathbf{x^k}) := \nabla \mathbf{f}(\mathbf{x}^k) \nabla \mathbf{f}(\mathbf{x}^k)^T. \tag{5.21}$$

Although the omission of the 2nd derivatives of $f(\cdot)$ from the Hessian can also be justified from a numerical perspective, its main advantage is computational speed: the Hessian according to Equation (5.21) is simply obtained as outer product of the gradient of $f()\dot{}$ with itself, i.e., all that is required to compute the Hessian once the gradient is known is one additional matrix multiplication.

If $f(\cdot)$ is linear - i.e., $f(\mathbf{\Omega}_i, \mathbf{x}) = \mathbf{\Omega}_i{}^T \mathbf{x}$ - we have a linear least squares problem with $\nabla \mathbf{f} = (\mathbf{\Omega}_1, .., \mathbf{\Omega}_N)$. In this case $h(\cdot)$ is a quadratic cost function, the aprroximation Equation (5.21) becomes the true Hessian, and Equation (5.17) becomes the *pseudo inverse* solution to linear least squares. Provided the matrix $(\mathbf{\Omega}_1, .., \mathbf{\Omega}_N)$ has full row rank, this will move us to the minimum in a single step.

### 5.3.4 The Levenberg-Marquardt Method

The Gauss-Newton method usually converges very fast when near a non-singular minimum. However, when the objective function $h(\cdot)$ is locally not well approximated by a quadratic function with positive definite Hessian, convergence may become slow or the method may even optimization down. In such cases, gradient descent - although slower - is the preferred method, which should be used until the parameter vector $\mathbf{x}^k$ has moved into a more 'benign' region. The **Levenberg-Marquardt** method - or LM for short - tries to move in an **adaptive fashion between** the extremes of the **Gauss-Newton** method on the one **and** the **gradient descent** approach on the other hand: when good progress is being made (as indicated by a substantial reduction in the cost function), LM moves toward the Gauss-Newton approach. However, when the cost function decreases too slowly or even increases, LM will behave more like gradient descent. This movement between the two methods is mediated by a

regularization parameter $\lambda \geq 0 \in \mathbb{R}$, in a modified version of Equation (5.21)

$$\nabla^2 h(\mathbf{x^k}) := \nabla \mathbf{f}(\mathbf{x}^k) \nabla \mathbf{f}(\mathbf{x}^k)^T + \lambda \mathbf{I}. \tag{5.22}$$

For small values for $\lambda$ move the approach into the Gauss-Newton direction. Note that if the Hessian Equation (5.21) becomes singular or negative definite, it can be made positive definite by using a sufficiently large value for $\lambda$ in Equation (5.22). In other words: if good progress is being made, $\lambda$ should be decreased, otherwise, it should be increased. The choice of the thresholds and increments involved in the implementation of the LM algorithm are problem dependent; some pointers are given in [15], chapter 15.5.

## 5.4 Approximations

Thus far, we have assumed that the forward model $y = f(\mathbf{\Omega}, \mathbf{x})$ is known and can be evaluated for any given input pair $\mathbf{\Omega}, \mathbf{x}$. The nonlinear optimization methods used to invert the model - as discussed in the previous section - in addition require that the first or even second order derivatives of $f$ can be determined. As you have seen, forward models can become very complex, and although a given geophysical model may serve as a basis for scientific discussion and contribute much to our understanding of the phenomenon under investigation, it must not necessarily also be well suited for actual computation, be it due to performance considerations or numerical problems. An alternate approach is to use a very general class of functions $\hat{f}_{MLP}$ that can approximate any sufficiently smooth target function. One such function class is the *Multi Layer Perceptron (MLP)*, a popular example of a so-called **Artificial Neural Network (ANN)**. It consists of a hierarchy of several layers of simple processing units or 'neurons', as illustrated in Figure 5.3. Such a neuron simply computes the inner product between its inputs and its associated *weight vector* $\mathbf{w}_i$, and then computes its output by applying a non-linear function to the inner product. The outputs of each layer serve as input to the succeeding layer. The actual mapping computed by such a MLP is determined by its weight vectors (or weights). The weights $\mathbf{w}$ are computed by minimizing an expression akin to Eq. 5.9

$$h(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (\hat{f}_{MLP}(\mathbf{\Omega}_i, \mathbf{x}, \mathbf{w}) - y_i)^2, \mathbf{w} \in \mathbb{R}^m \tag{5.23}$$

Note that here, $h$ is a function of $\mathbf{w}$, while the parameters $\mathbf{x}$ and inputs $\mathbf{\Omega}$ are assumed to be fixed. The network is then 'trained' so that the above least squares expression becomes minimal. *Training* of the network is in practice tantamount to minimizing the above expression w.r.t $\mathbf{w}$ using *back-propagation*, a special kind of gradient descent algorithm, that exploits the structure of the MLP. By exchanging the role of $y_i$ and $\mathbf{x}_i$, we can use a MLP also to approximate the inverse of the forward model.

Still another approach to inverting the forward model are so called **lookup tables**. These simply tabulate the output $Y_i$ of the forward model for a discrete subset of parameters $\mathbf{x}_i$ and features $\mathbf{\Omega}_i$. For a given observation $y$, the nearest tabulated output $j = \arg\min |y_i - y|$ is found, and its associated parameter vector $\mathbf{x}_i$ is returned as solution. An obvious advantage of the lookup-table approach is operational speed, since once the lookup table has been computed, only a simple linear nearest neighbour search has to be performed. However, depending on the number of parameters and the required accuracy

of the result, the storage required to hold the discrete samples can become very large, even prohibitively so.

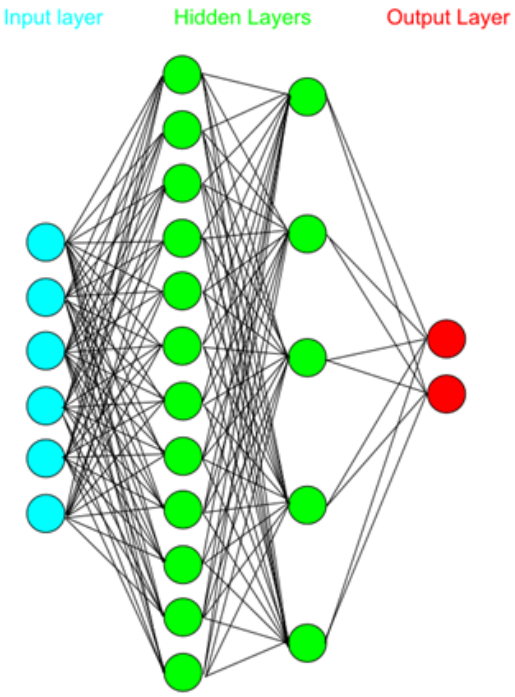



Figure 5.3: Illustration of the structure of a Multi Layer Perceptron (MLP). It consists of 2 interface layers – the input and output layer – and one or several hidden layers. The input layer will hold any measurement or available auxiliary parameters that contribute to the process being modelled, (e.g., the vegetation state, soil texture, or the incidence angle); it is used to feed observed data into the MLP. The hidden layers consist of internal (hence, hidden) processing units. Finally, the output layer computes and holds the estimated function value $f(\mathbf{x})$, based on the activations of the units in its immediately preceeding hidden layer. The weights connecting the layers are determined during the training phase.

# 6 Error Modelling

After estimating the state of a geophysical parameter the most important question is how accurate the estimate is. When characterizing the accuracy of geophysical parameters they are often considered as stationary random variables. The accuracy of the retrieval can then be characterized by: 1) a systematic (predictable) error, often referred to as "bias", which is the deviation of the expected value from the "truth", and 2) a random (non-predictable) error, often referred to as "variance", which is the spread of the random variable around its expected value. (The inverse of bias and variance are also known as "trueness" and "precision", respectively.) This characterization can be done *a-priori*, *a-posteriori*, or simultaneously with the retrieval process.

In a-priori error characterization, assumptions or estimates of the errors of the actual antenna measurement are made *before* applying the retrieval model (e.g., through the measurement of reference targets, or the estimation of thermal sensor noise). To estimate the accuracy of the final retrieved parameter, the impact of the retrieval algorithm on the measurement errors is modelled by propagating the error through the algorithm. Note that this only applies to random errors since a-priori known systematic errors are usually corrected before applying the retrieval algorithm.

In the a-posteriori case, reference data is used to evaluate the accuracy of the retrieved parameter estimates. Often it is assumed that the reference data approximate the "true" state of the parameter with a much higher accuracy than our measurements. Hence, the deviations between those two can be seen as errors of our parameter estimates.

From adjustment theory we know that the accuracy of parameters can be estimated simultaneously with the parameter when having an over-determined measurement system with a certain set of independent measurements. This is the case for so-called "data assimilation" frameworks. However, they are only mentioned for completeness, their detailed discussion is beyond the scope of this lecture.

## 6.1 Statistical parameters

As previously mentioned, in error characterization the geophysical parameters under validation are usually treated as continuous random variables. Hence, we focus on their distribution as given by their *probability density function* (pdf) $p_X(x)$. Before treating different error metrics, the following section briefly summarizes the most important statistical parameters that can be used to characterize a distribution of continuous random variables. In the following, we denote random variables by capital latin letters, e.g. $X$, except for estimators of distribution parameters with established symbols, e.g, $\hat{\mu}$ for the estimate for the mean.

### Mean

The mean $\mu_X$ – also referred to as expectation or expected value of $X$– is the first moment of the distribution and is the integral over all possible values of $X$ weighted by their probabilities, and is given by :

$$\mu_X = E[(X)] = \int_{-\infty}^{\infty} x p_X(x) dx. \tag{6.1}$$

If we add $N$ independent observations of $x_1, .., x_N$ that all come from the same population which are independent and identically distributed (i.i.d.), according to the law of large numbers we can estimate the central tendency as the arithmetic mean, the so called *sample mean*:

$$\hat{\mu}_X = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{6.2}$$

Note that the *estimator* $\hat{\mu}_X$ as defined above is again a random variable. It can be shown that as $N \to \infty$, $\hat{\mu}_X$ gets arbitrarily close to the true parameter value $\mu_X$; such estimators are called *asymptotically consistent*. If we replace the random variables $X_i$ in Equation (6.2) with actual observations $x_i$

$$\hat{m}_X = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{6.3}$$

we obtain a realisation of $\hat{\mu}_X$, which is called an *estimate*. Estimates of the moments are typically used instead of the unknown moments in order to characterise a distribution.

### Variance and Standard Deviation

The variance is the central second moment of the distribution and represents the spread around its mean:

$$\sigma_X^2 = Var[X] = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - \mu_x)^2 p_X(x) dx. \tag{6.4}$$

An asymptotically consistent estimator of the variance is given by:

$$\hat{\sigma}_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu}_X)^2. \tag{6.5}$$

An inconvenience of the variance is its unit, which is the square of the data unit. Hence, it is common to use the square root of the variance, called the standard deviation:

$$\sigma_X = \sqrt{\sigma_X^2} \tag{6.6}$$

The use of $\sigma$ to denote standard deviation is well established in statistics; unfortunately, the $\sigma$ is also used for backscatter-related quantities in physics , e.g., $\sigma_0$. Normally, the difference should be clear

from the context.

#### Covariance

The covariance is a function of two random variables characterized by their joint pdf $p_{XY}(x, y)$:

$$\sigma_{XY} = E[(X - E[X])(Y - E[Y])] = \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_X) p_{XY}(x, y) dxdy. \tag{6.7}$$

The covariance expresses how well one variable can be predicted as a linear function of the other (the higher the covariance, the better linear predictability). For Gaussian variables (but not in general!), a covariance of zero implies that they are independent. An asymptotically consistent estimator of the covariance from $N$ pairs of measurements is given by:

$$\hat{\sigma}_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y). \tag{6.8}$$

Variances and covariances of two or more variables are often arranged in a so called *covariance matrix*. For our 2 variable example, the *covariance matrix* $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}. \tag{6.9}$$

Note that the diagonal contains the variances, while the off-diagonal contains the covariances of the variables. Also, since, $\sigma_{XY} = \sigma_{YX}$, covariance matrices are always symmetric. They are also positive semi-definite.

## 6.2 A-priori error characterisation and error propagation

In error propagation, the uncertainties (or errors) associated with the parameters are represented by their variances. The covariance matrix of the parameters is estimated by propagating along the processing chain the known covariance matrix of the input variables. Note that error propagation deals solely with the variance (i.e., precision), but not the bias of the estimated parameters.

### 6.2.1 Theory

Let $\mathbf{x} = (x_1, ..., x_p)$ be a p-dimensional vector. $\mathbf{x}$ is assumed to be an instance of a p-dimensional random vector $\vec{X} = (X_1, \ldots, X_p)$, with covariance matrix $\mathbf{\Sigma}_{\vec{X}}$. We are interested in how the covariance transforms under a mapping $\mathbb{R}^p \to \mathbb{R}^q$, $\vec{Y} = f(\vec{X})$, i.e., given $\mathbf{\Sigma}_{\vec{X}}$ and $f(.)$, we would like to know the covariance of $\vec{Y}$, $\mathbf{\Sigma}_{\vec{Y}}$.

#### 6.2.1.1 Affine Case

If $f$ is an affine mapping of the form:

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{A} \in \mathbb{R}^{q \times p}, \mathbf{b} \in \mathbb{R}^{q}, \tag{6.10}$$

then the covariance matrix transforms like:

$$\mathbf{\Sigma}_{\vec{Y}} = \mathbf{A}\mathbf{\Sigma}_{\vec{X}}\mathbf{A}^{T}. \tag{6.11}$$

#### 6.2.1.2 General (Non-Linear) Case

If $f$ is a non-linear mapping, we first linearise it by its first order Taylor approximation about the operation point $\mathbf{x}_0$:

$$\mathbf{y} = f(\mathbf{x}) = f(\mathbf{x}_0) + \left(\frac{\partial f}{\partial \mathbf{x}}\right)(\mathbf{x} - \mathbf{x}_0), \tag{6.12}$$

whereby $\left(\frac{\partial f}{\partial \mathbf{x}}\right)$ is the Jacobian of $f$ at $\mathbf{x}_0$. Putting everything together, we finally obtain:

$$\mathbf{\Sigma}_{\vec{Y}} = \left(\frac{\partial f}{\partial \mathbf{x}}\right)\mathbf{\Sigma}_{\vec{X}}\left(\frac{\partial f}{\partial \mathbf{x}}\right)^{T}. \tag{6.13}$$

### 6.2.2 Example

We shall illustrate error propagation using the incidence angle normalization step in the TU Wien method, which was given by Equation (5.7):

$$f(\mathbf{x}) = \sigma^0(\theta_{ref}) = \sigma^0(\theta) - \sigma'(\theta_{ref})(\theta - \theta_{ref}) - \frac{\sigma''}{2}(\theta_{ref})(\theta - \theta_{ref})^2. \tag{6.14}$$

This step computes from a backscatter measurement $\sigma^0(\theta)$ taken at an arbitrary incidence angle $\theta$ the corresponding backscatter at $\sigma^0(\theta_{ref})$ at the reference angle. In our case, the inputs are the actual measurement $\sigma^0(\theta)$ and the *slope* and *curvature* parameters $\sigma'(\theta_{ref})$ and $\sigma''(\theta_{ref})$ computed earlier in the processing chain:

$$\mathbf{x} = [\sigma^0(\theta), \sigma'(\theta_{ref}), \sigma''(\theta_{ref})] \tag{6.15}$$

If we assume that the backscatter, slope, and curvature – i.e., the elements of the $\mathbf{x}$ – are mutually uncorrelated, the covariance matrix of $\mathbf{x}$ is simply:

$$\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{I}_{3\times 3}\left[\mathrm{Var}[\sigma^0(\theta)], \mathrm{Var}[\sigma'(\theta_{ref})], \mathrm{Var}[\sigma''(\theta_{ref})]\right]^{T}, \tag{6.16}$$

whereby $\mathbf{I}_{3\times 3}$ is the $3 \times 3$ unity matrix and the variances of the elements have been determined during previous processing steps. The Jacobian is obtained by computing the derivatives of Equation (6.14)

with respect to $\mathbf{x}$:

$$\left(\frac{\partial f}{\partial \mathbf{x}}\right) = [1, (\theta - \theta_{ref}), -0.5(\theta - \theta_{ref})^2]. \tag{6.17}$$

Hence, according to Equation (6.13), the noise variance of the normalised backscatter is given by

$$\mathrm{Var}[\sigma^0(\theta_{ref})] = \mathrm{Var}[\sigma^0(\theta)] + (\theta - \theta_{ref})\mathrm{Var}[\sigma'(\theta_{ref})^2] + 0.25(\theta - \theta_{ref})^4\mathrm{Var}[\sigma''(\theta_{ref})]. \tag{6.18}$$

The complete algorithm that applies the TU Wien method (named WAter Retrieval Package - WARP - Version 5.5), including the error propagation, is discussed in [8].

## 6.3 A-posteriori error characterization and validation

In the a-posteriori case we aim to characterize the accuracy of the parameter estimates by comparing them with reference measurements. Conventional metrics used for this purpose are difference metrics which characterize the agreement or disagreement between two data sets $X$ and $Y$. Since no reference data set will ever perfectly measure the true state of any geophysical parameter, these so-called relative error metrics cannot tell anything about absolute error levels with respect to the "truth" and must be interpreted as relative errors of some data set $X$ with respect to a reference data set $Y$. There are, however, metrics that aim to overcome the dependency of error metrics on the availability of high-accuracy reference data. The development and improvement of such metrics is an ongoing research topic.

In addition, large differences can exist between the spatio-temporal scale of Earth observation data and the available reference data. This can introduce scaling errors, which can significantly influence the actual measurement error estimates. Therefore, before error estimation methods will be introduced, the following section will treat scale related issues and provide tools on how they can be resolved.

### 6.3.1 Scaling considerations

When speaking of scales, we mean a *characteristic length* or a *characteristic time*, referring either to the geophysical process under consideration itself (**process scale**), to the measurement process (**measurement scale**), or to the output of a retrieval model (**model scale**). In order to retrieve a geophysical parameter through modelling of observations, the process scale, model scale, and measurement scales should be similar.

Geophysical parameters are usually not driven by a single process, but by various interlinked underlying driving processes. For soil moisture for instance, the most obvious driving process is meteorological forcing, i.e., precipitation and evaporation. However, also surface properties can lead to differences in soil moisture. For example, land cover type lead to differences in water uptake, topographic features such as hill slopes cause different vertical moisture redistribution, and soil texture is responsible for differences in infiltration rates. Subsequently, soil moisture processes take place at many different

temporal and spatial scales. While rainfall can homogeneously occur over large areas, soil texture changes can be very local. Rainfall induced wetting can happen very rapidly whereas solar radiation driven evaporation can be very slow.

Hence, one has to carefully choose a proper measurement scale in order to resolve the required soil moisture features at the respective process scales. The measurement scale is in the temporal domain determined through the measurement frequency or, in case of satellites, the overpass time. In the spatial domain the measurement scale can be determined through the **support**, **spacing**, and **extent**. The support refers to the area over which the measurement integrates, i.e., the spatial resolution, the spacing refers to the distance between single measurements, and the extent refers to the overall sampled area (see Figure 6.1). ASCAT for example, has a 550 km wide swath (extent) that provides measurements observed at 25 km resolution, resampled using a 36 km hamming window (support) to a 12.5 km grid (spacing).

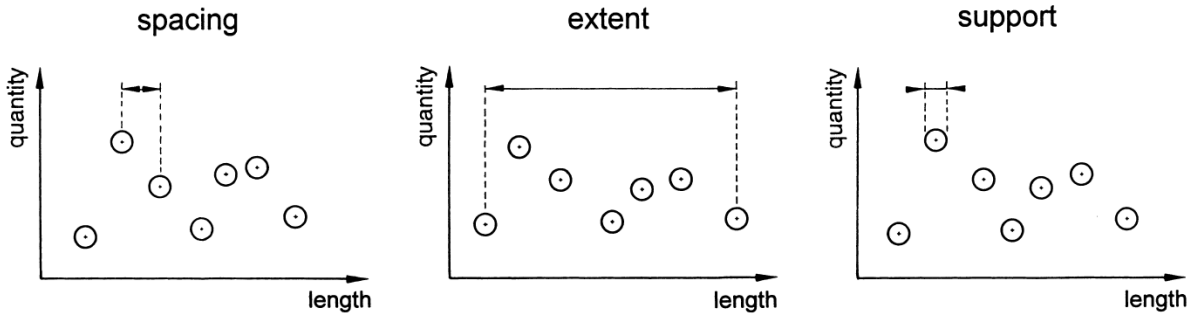



Figure 6.1: Illustration of extent, support, and spacing of a measurement process (Adapted from [28]).

Crucial to understand is that measurements from different instruments taken at different measurement scales will also resolve different process scales. For example, an in situ monitoring station for soil moisture might be able to capture local soil texture or topography induced soil moisture features within a certain area, which are averaged out in a 25 km wide satellite footprint. Therefore, there will be always discrepancies when comparing measurements taken at different spatio-temporal scales, referred to as **scaling errors**. Such scaling errors can be of a systematic nature (i.e., predictable) and of a random nature (i.e., non-predictable with a zero-mean).

Systematic scaling differences originate from stable surface properties that are responsible for local soil moisture variations, i.e., topography, soil texture and land cover. Even though absolute soil moisture levels vary within a certain area, there might be spatial soil moisture patterns which persist in time. Point measurements taken at such temporally stable locations can thus be representative for the large-scale soil moisture conditions, even though biased in a systematic way. This **temporal stability concept** was first introduced by Vachaud in 1985 [23].

Random scaling errors, on the other hand, are induced by soil moisture variations that affect the measurement scales of different instruments in different and unpredictable ways. This can be for example a precipitation event, if the rain cloud covers only some parts of a satellite footprint, maybe even without covering a therein placed in situ sensor, but also if the rain is not falling with an homogeneously intensity within the footprint. Such errors might also origin from variations in solar radiation driven evaporation rates due to cloud coverage. Random scaling errors are also referred to as **representa-**

**tiveness errors**, because they limit small-scale measurements in being representative for coarse-scale processes.

Since some parts of the scaling errors are systematic, they can be corrected for. Therefore, we make the assumption of stationarity of the measured quantity (i.e., temporal stability of the probability density function), and require a sufficient number of measurements taken at the same location at different points in time. A large variety of methods for removing systematic differences between the data sets exist, referred to as **rescaling methods**. Only the most common ones shall be introduced in the following.

Let us therefore first assume a linear relationship between two data sets $X$ and $Y$, i.e., only linear systematic differences:

$$Y = a + bX \tag{6.19}$$

#### Linear regression

In the linear regression, the additive and multiplicative biases $a$ and $b$ are determined in a least-squares fashion, i.e., by minimizing the squared differences between the measurements. Data set $Y$ rescaled against data set $X$ can then be calculated through:

$$Y^X = \frac{Y - a}{b} \tag{6.20}$$

The superscript denotes the scaling reference.

#### Minimum - maximum fit

Another possible condition for the determination of the scaling coefficients is that the absolute minimum and maximum values should match:

$$Y^X = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \cdot [\max(X) - \min(X)] + \min(X) \tag{6.21}$$

#### Mean - standard deviation fit

Slightly more robust against outliers and small non-linearities than the minimum-maximum fit is the fitting of the mean and the standard deviation:

$$Y^X = \frac{(Y - \overline{Y})}{\sigma_Y} \cdot \sigma_X + \overline{X} \tag{6.22}$$

#### CDF-matching

In some cases, e.g., in complex terrain, it might be appropriate to assume a non-linear relationship across scales:

$$Y = a + bX + cX^2 + dX^3 + ... \tag{6.23}$$

The most common way to account for such non-linearities between data sets is the matching of their cumulative distribution function (cdf). In most cases this is done either by fitting a higher-order polynomial (e.g., of fifth order) or by sorting the data after ascending values, splitting it into percentiles, and then rescaling the percentiles through a linear regression. This concept is illustrated in Figure 6.2.

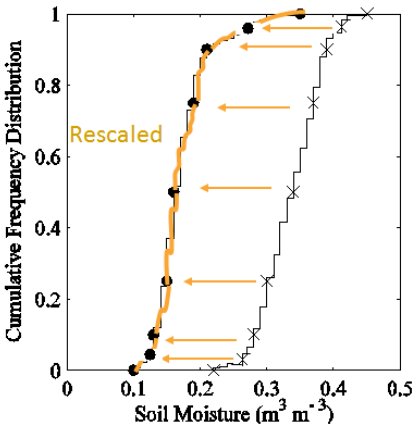



Figure 6.2: Illustration of the CDF matching technique.

#### Summary

Measurements from different sensors taken over coinciding locations but at different measurement scales will inherently contain discrepancies, referred to as scaling errors, independent of actual measurement errors. These origin from the fact that physically not the same soil- or therein contained water volume is sampled. However, driving soil moisture processes show a large degree of temporal stability and spatial autocorrelation, hence usually the highest fraction of scaling errors is of a systematic nature and can be corrected for using appropriate rescaling techniques, provided that data is available over a sufficiently large time span. The following sections will introduce the most common error characterization methods and discuss also how they are influenced by systematic and random scaling errors.

### 6.3.2 Characterization of statistical dependency

Several metrics can be used to characterize a statistical relationship between two data sets; the correlation between data sets. Positive correlation means that if the values of one data set increase (or decrease), the values of the other data set increase (or decrease). Negative correlation means that if

the values of one data set increase, the values of the other data set decrease and vice versa. As only the behaviour of change in the measurements is compared, the correlation is to a large degree independent of systematic biases,e.g. measurement- or scaling biases. However, both measurement noise and representativeness errors (random scaling errors), will lead to a slight reduction in the correlation, depending on the magnitude of total signal variability (i.e., the signal-to-noise ratio). The correlation is commonly used to characterize the *temporal agreement* of parameter estimates under validation with a reference data set.

#### 6.3.2.1 Linear correlation

The most common way to characterise statistical dependency between two data sets is the linear Pearson correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{6.24}$$

It represents the covariance between two data sets normalized with their respective standard deviation, and thus assumes values between -1 and 1. 0 means no statistical dependency, 1 means the joint distribution lies on a line with positive slope, -1 means the joint distribution lies on a line with negative slope.

For the population correlation coefficient it is common to use the symbol $\rho$. Correlation coefficients estimated from a sample are normally indicated with $R$ or $r$ (without hat):

$$R = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{x})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{6.25}$$

Sometimes also the squared value of Person's R is used, referred to as Coefficient of Determination. Although its values are always closer to zero, their meaning is identical. The $R^2$ tell us the fraction of the variance of $Y$ that is explained by $X$:

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \tag{6.26}$$

#### 6.3.2.2 Rank correlation

In some cases data sets might show a non-linear statistical dependency (e.g., through an increasing bias between two data sets in the summer months due to a higher vegetation influence on one of them). Such non-linear correlation is better characterized through Spearman's rank correlation coefficient, which is calculated similar to Pearson's R, but using the rank of the values (i.e., the sort order index when sorting the data after increasing values) instead of the values themselves.

$$\rho = \frac{\sigma_{r_X r_Y}}{\sigma_{r_X} \sigma_{r_Y}} \tag{6.27}$$

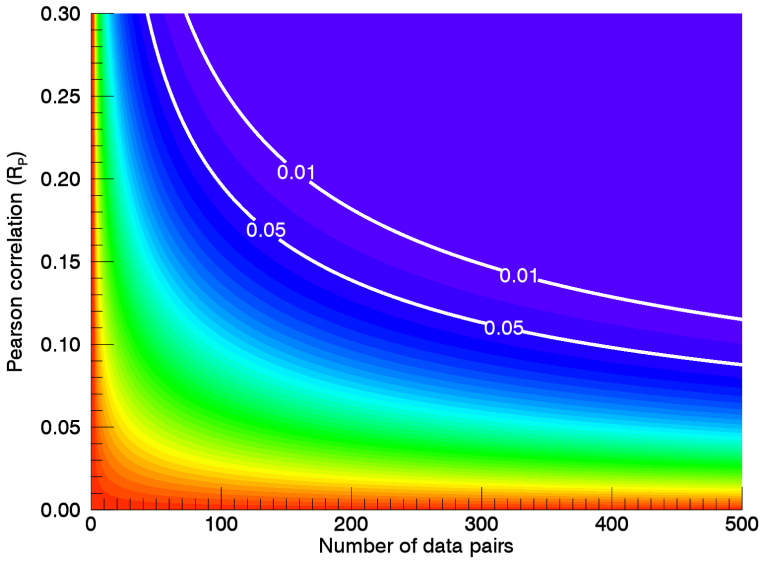



Figure 6.3: Relationship between the Pearson correlation, the number of data pairs and the correlation significance. White lines are the most common probability thresholds (0.05 and 0.01).

#### 6.3.2.3 Correlation significance

A certain probability exists that the calculated correlation level was achieved by a coinciding pattern in the distribution of the two data sets which is not caused by a real relationship between them. Such correlation is not significant. The significance level of a correlation can be estimated using statistical hypothesis tests. The common test for correlation significance is the Student's t-test in which the test statistic follows a Student's t distribution. In this test a probability level is calculated that links the observed correlation level to the number of available data pairs. That is, if a correlation level is achieved using a certain number of data pairs it is much more likely to be significant than the same correlation level achieved using much less data pairs. This behaviour is illustrated in Figure 6.3. Common thresholds for considering a correlation level to be significant are 0.05 or 0.01 what means that there is a 5% or 1% probability that the correlation is a coincidence. The probability is calculated for a test value $t$ in the Student's t distribution.

$$t = R \cdot \sqrt{\frac{N-2}{1-R^2}} \tag{6.28}$$

where $R$ is the correlation level (this can be both the Spearman or the Pearson correlation coefficient) and $N$ is the number of data pairs.

### 6.3.3 Characterization of absolute deviations

While the correlation characterizes how similar two data sets behave over time, one might also want to estimate how well the actual measurement values agree. Conventional metrics used for this purpose are **difference metrics**, that is, they characterize the agreement of one data set with a chosen reference data set, assuming that this reference is a good approximation for the "truth".

Even tough very accurate reference data sets might exist, one has to take care of scaling errors, since they directly contribute to difference metrics and might be significantly larger than actual measurement errors ofthe data set under validation. Therefore, when comparing data sets with different spatial measurement scales, one should always apply a rescaling prior to calculating the error metrics, in order to correct for such scaling errors. Random errors, on the other hand, will always persist, even though their absolute magnitude is also influenced by multiplicative coefficients in the rescaling.

Unfortunately, a-priori rescaling always corrects for differences between the data sets which originate from systematic measurement errors in either of the data sets. Hence, if one wants to characterize such errors, reference data at the same spatial scale as the data set under validation are required, which in Earth observation is usually not available.

The most common difference metrics will be introduced in the following.

#### 6.3.3.1 Bias

The bias is the difference between the mean values of two data sets and is used to characterize a systematic over- or underestimation of the estimated parameter states with respect to the reference data set.

$$bias = \overline{X} - \overline{Y} \tag{6.29}$$

As mentioned before, the calculation of the bias makes only sense if the spatial scale of the data sets matches and no rescaling has been applied, or, if one wants to particularly estimate systematic scaling errors using two data sets which are assumed to have negligible systematic measurement errors themselves.

#### 6.3.3.2 Standard deviation ratio

The standard deviation ratio is, like the bias, a measure for systematic differences between two data sets, but this time of second-order. That is, it compares the variability of the data sets instead of their mean values.

$$SDR = \frac{\sigma_X}{\sigma_Y} \tag{6.30}$$

The same considerations as for the bias apply.

#### 6.3.3.3 Root Mean Squared Difference (RMSD)

The RMSD is the square root of the averaged squared differences between two data sets. In other words the RMSD represents the average absolute deviation between single measurements. Sometimes also the Mean Squared Difference (MSD) is used, calculated without taking the square root. In this

case, the unit is again the square of the data unit.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2} \tag{6.31}$$

By mathematically decomposing the RMSD it can be shown that deviations between the data sets can origin from three sources, namely from a decorrelation, from a bias, and from a difference in the variance [13].

$$RMSD = \sqrt{MSD_{corr} + MSD_{bias} + MSD_{var}} \tag{6.32}$$

where

$$\begin{aligned} MSD_{corr} &= 2\sigma_X\sigma_Y(1 - R_P) \\ MSD_{bias} &= (\overline{X} - \overline{Y})^2 \\ MSD_{var} &= (\sigma_X - \sigma_Y)^2 \end{aligned} \tag{6.33}$$

We can see that the individual terms can get negligibly small while the other terms might still significantly contribute to the total RMSD. That is, the RMSD is a composition of different error sources. The decorrelation term is mainly driven by random fluctuations of the data sets, containing also representativeness errors, the bias term originates from additive and multiplicative systematic differences and scaling errors, and the variance term is caused by multiplicative systematic differences and scaling errors between the data sets. Higher order systematic differences influence all three terms.

Usually we want to characterize random and systematic errors of a data set individually, or completely correct for the latter when expecting large scaling errors. Hence, the bias term is often subtracted from the RMSD. The resulting metric is called **centered RMSD** (cRMSD).

$$cRMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(X_i - \overline{X}) - (Y_i - \overline{Y})^2} \tag{6.34}$$

The cRMSD only accounts for additive biases between the data sets. If one expects also systematic differences of higher order, rescaling techniques as those discussed in Section 6.3.1 can be applied. The resulting metric is called **unbiased RMSD** (ubRMSD).

$$ubRMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i^X)^2} \tag{6.35}$$

the superscript indicates that $Y$ was rescaled against $X$. Both the cRMSD and the ubRMSD are usually considered to represent random errors of the data set under evaluation with respect to the reference data set, plus additional representativeness errors. However, even though bias corrected, they still show a high dependency on the variability of the reference data set, as it can be seen from (6.33). This might lead to difficulties in the interpretation when comparing the RMSD estimates for geophysical parameters in different geographic regions because their variability might be significantly different (e.g., for soil moisture variations in desert areas and in rain forest).

### 6.3.4 The triple collocation method

The RMSD was introduced in the previous section as averaged squared differences from a reference data set. Obviously, if considered as an absolute error metric, the reliability of the RMSD strongly depends on the quality of the reference data set. [18] proposed a method to overcome this problem. Instead of squaring the differences between two products, like it was done in (6.31), we can cross multiply the differences between three collocated data sets of the same geophysical parameters:

$$\begin{aligned} RMSE_X &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(X_i - Y_i)(X_i - Z_i)]} = \sqrt{\langle (X_i - Y_i)(X_i - Z_i)\rangle} \\ RMSE_Y &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(Y_i - X_i)(Y_i - Z_i)]} = \sqrt{\langle (Y_i - X_i)(Y_i - Z_i)\rangle} \\ RMSE_Z &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(Z_i - X_i)(Z_i - Y_i)]} = \sqrt{\langle (Z_i - X_i)(Z_i - Y_i)\rangle} \end{aligned} \tag{6.36}$$

We then get individual error estimates for all three data sets ($X$, $Y$, and $Z$) which are independent of each other and do not rely on an approximately error-free reference data set. Hence, we now refer to them as Root Mean Square Error (RMSE) since they are no longer difference metrics. The Gauss brackets $\langle ... \rangle$ denote the averaging step. The success of this method requires several assumptions to be fulfilled, which will be explained in the following derivation of the error model.

First, we assume that all three data sets are related to the same geophysical parameter and contain both a systematic and a random error term:

$$\begin{aligned} X &= \alpha_X + \beta_X \Theta + \varepsilon_X \\ Y &= \alpha_Y + \beta_Y \Theta + \varepsilon_Y \\ Z &= \alpha_Z + \beta_Z \Theta + \varepsilon_Z \end{aligned} \tag{6.37}$$

$\alpha_i$ and $\beta_i$ are the additive and a multiplicative systematic error terms of the respective data sets. Note that there might be also higher-order terms or the apparent terms might be negligibly small. $\varepsilon_i$ are the random error terms of the data sets which are assumed to be Gaussian (i.e., normal distributed with an expected value of zero). $\Theta$ is the true state of the parameter that is estimated by the data sets. Since all three data sets will probably have a different systematic error with respect to the unknown truth, we randomly choose one data set (in the following this will be data set $X$, i.e., $\alpha_i \equiv 0$ and $\beta_i \equiv 1$) as a reference to rescale the other two against, in order to get rid of relative systematic differences between the data sets:

$$\begin{aligned} X^X &= \hat{\Theta}^X + \varepsilon_X^X \\ Y^X &= \hat{\Theta}^X + \varepsilon_Y^X \\ Z^X &= \hat{\Theta}^X + \varepsilon_Z^X \end{aligned} \tag{6.38}$$

$\hat{\Theta}$ is the common observed parameter state of all three data sets that still contains the systematic error of the reference with respect to the unknown truth. The superscript $X$ denotes the chosen scaling

reference and should emphasize that both the rescaled measurements and the remaining random error terms are still biased with the systematic errors of the reference data set. Only relative systematic differences have been corrected for. If we now cross-multiply differences of (6.38), we obtain

$$\begin{aligned}(X^X - Y^X)(X^X - Z^X) &= {\varepsilon_X^X}^2 - \varepsilon_X^X\varepsilon_Y^X - \varepsilon_X^X\varepsilon_Z^X + \varepsilon_Y^X\varepsilon_Z^X \\ (Y^X - X^X)(Y^X - Z^X) &= {\varepsilon_Y^X}^2 - \varepsilon_Y^X\varepsilon_X^X - \varepsilon_Y^X\varepsilon_Z^X + \varepsilon_X^X\varepsilon_Z^X \\ (Z^X - X^X)(Z^X - Y^X) &= {\varepsilon_Z^X}^2 - \varepsilon_Z^X\varepsilon_X^X - \varepsilon_Z^X\varepsilon_Y^X + \varepsilon_X^X\varepsilon_Y^X\end{aligned} \tag{6.39}$$

If we further average over a sufficient number of measurement triplets, then the assumptions of zero-mean Gaussian distributed and mutually independent random errors lead to

$$\begin{aligned}\langle {\varepsilon_i^X}^2 \rangle &\rightarrow Variance(\varepsilon_i^X) = MSE_i^X \\ \langle \varepsilon_i^X \varepsilon_j^X \rangle &\rightarrow Covariance(\varepsilon_i^X \varepsilon_j^X) = 0 \\ i, j &\in [X, Y, Z] \quad , \quad i \neq j\end{aligned} \tag{6.40}$$

and further to the final error estimates

$$\begin{aligned}RMSE_X^X &= \sqrt{\langle [(X_i^X - Y_i^X)(X_i^X - Z_i^X)] \rangle} = StdDev(\varepsilon_X^X) \\ RMSE_Y^X &= \sqrt{\langle [(Y_i^X - X_i^X)(Y_i^X - Z_i^X)] \rangle} = StdDev(\varepsilon_Y^X) \\ RMSE_Z^X &= \sqrt{\langle [(Z_i^X - X_i^X)(Z_i^X - Y_i^X)] \rangle} = StdDev(\varepsilon_Z^X)\end{aligned} \tag{6.41}$$

$RMSE_i$ are thus estimates of the standard deviations $StdDev()$ of the random errors $\varepsilon_i$ of the individual data sets. The superscript $X$ indicates that those error estimates still contain the systematic error of the chosen reference $X$ with respect to the unknown truth. The rescaling applied to obtain (6.38) only removes relative systematic differences between the data sets. The therefore computed scaling parameters are known and could be used to transform the obtained error estimates of (6.41) back into their own data space, what means that the error estimates are robust against errors in the chosen reference. However, the back-transformed error estimates would still contain the systematic error of the individual data sets with respect to their common "truth" and would further be biased against each other with their relative systematic difference. Hence, for a meaningful intercomparison of the random error constituents, the errors are usually kept in the data space of one chosen reference.

In the triple collocation several assumptions are made and the interpretation of the error estimates must be made very carefully. The most important considerations are summarized in the following.

- The triple collocation requires three data sets that are collocated in space and time. In Earth observation one will never have three data sources with the exact same spatial and temporal resolution and sampling points. As discussed in Section 6.3.1, deviations from a jointly observed soil volume will introduce systematic and random scaling errors, of which the systematic ones are corrected for in the a-priori applied rescaling. The jointly observed signal will be interpreted by the triple collocation method as a "common truth", and every deviation from this signal as an error with respect to this "truth". That is, the triple collocation error estimates of the data sets will contain their individual measurement noise plus the representativeness error of the respective data set with respect to the joint signal.

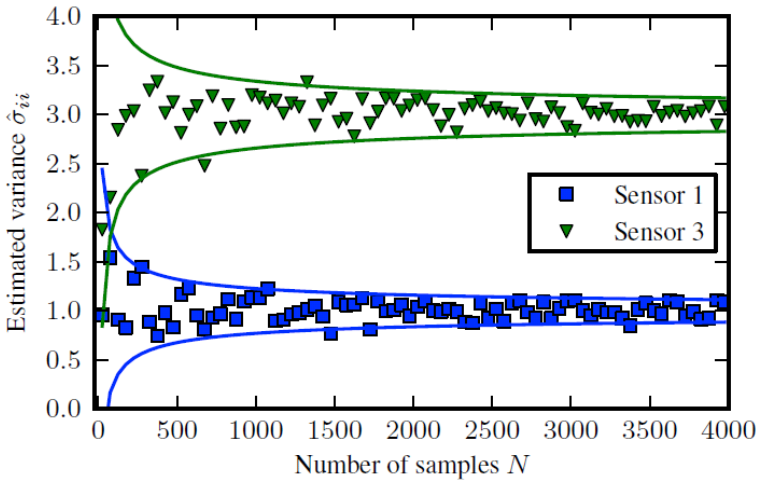



Figure 6.4: Numerical simulation of the impact of the finite number of samples on the triple collocation error estimates [29].

- Systematic differences between the data sets are removed in (6.38). Depending on the choice of the rescaling function, the systematic relationship between the data sets might not be captured with a sufficient accuracy. Remaining systematic differences might then propagate through the error model and can cause significant biases in the estimates of the random errors. Care must therefore be taken in selecting an appropriate rescaling method.

- One very crucial assumption is made on the errors of the individual data sets to be mutually independent. This causes the covariance terms in (6.41) to vanish. Due to their different signs, apparent error correlations can cause both an inflation or a deflation of the total error estimates. The validity of this assumption highly depends on the choice of the data set combination. If, for instance, two passive microwave sensors observing in the same frequency band are used, radio frequency interference (RFI; that is the pollution of the Earth surface's microwave signal with man-made microwave signaEquationin the same frequency such as those from wireless networks) will introduce correlated errors in both sensors.

- According to (6.40), the triple collocation requires theoretically an infinite number of triplets for the covariances to converge to zero and for the total result to converge to the error variance. In reality the finite number of available observation triplets will always introduce noise in the error estimates. An estimate of the magnitude of this noise is shown in Figure 6.4.

- Even though the relative systematic differences between the data sets are corrected, one should keep in mind that the error estimates will still contain the systematic error of the chosen reference with respect to the "common truth", which will always remain unknown. This systematic error can significantly change the variability of the data sets, depending on the choice of the reference. Since the triple collocation error estimates show a dependency on the data set variability similar to those of the RMSD (see (6.33)), the interpretation of the results can significantly change when using a different data set as the reference, even though the error estimates themselves are independent on the error level of the chosen reference.

Due to the robustness against errors in a reference data set the triple collocation has evolved as one of

the most important error modelling tools for estimates of geophysical parameters in Earth observation. However, its configuration and interpretation requires a high level of expertise and is still an open research topic.

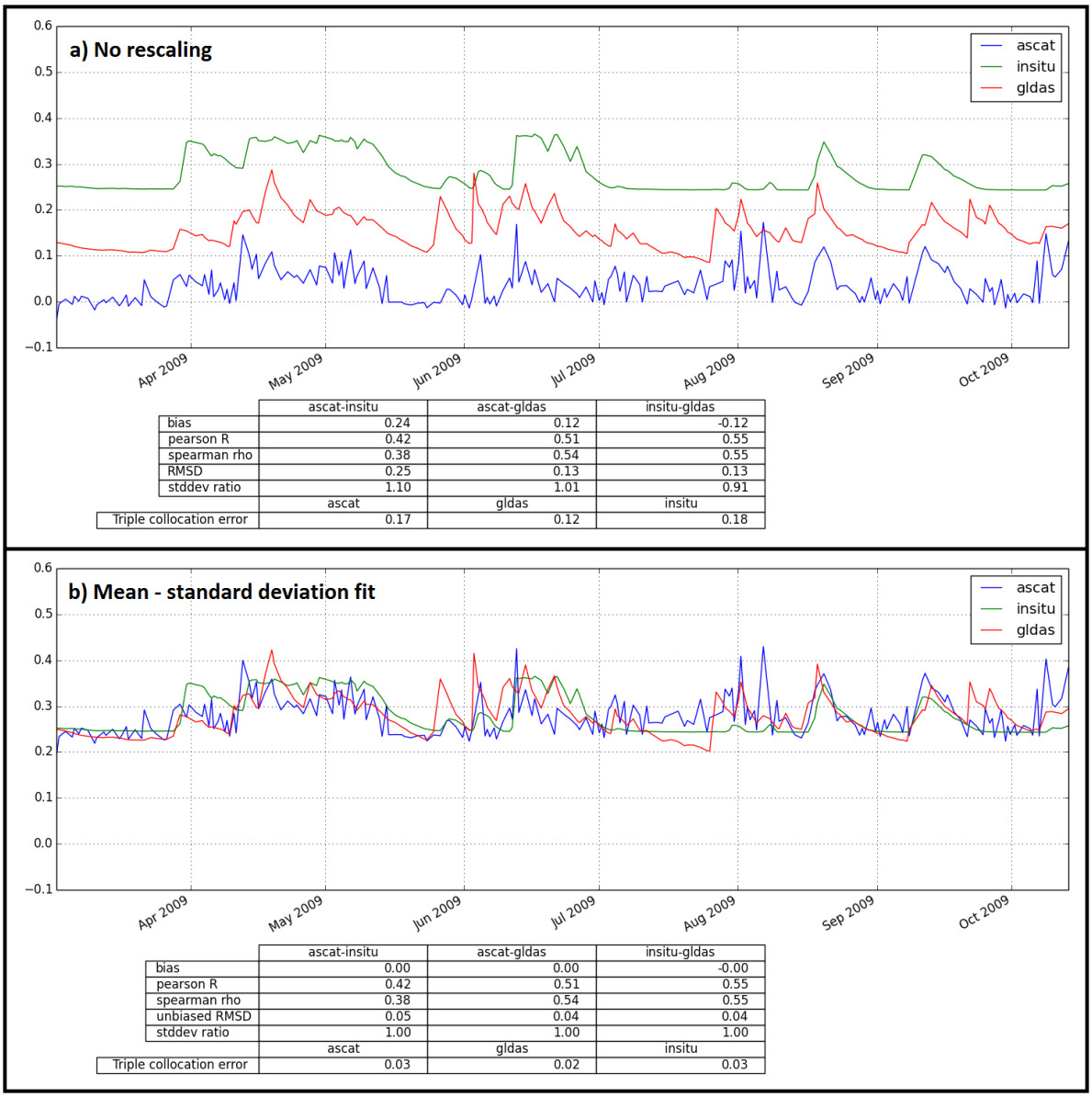
### 6.3.5 Example

Figure 6.5 shows two examples for a comparison between MetOp ASCAT soil moisture retrievals, in situ soil moisture measurements, and soil moisture estimates from a land surface model (GLDAS-Noah), one time without an a-priori rescaling, one time after applying a mean-standard deviation fit.

We can see the clear difference in the mean and the standard deviation between the data sets, reflected in biases above zero and standard deviation ratios (SDR) away from one. Note that the bias between ASCAT and the land surface model (GLDAS) is lower than between ASCAT and the in situ measurements. Also the SDR between ASCAT and GLDAS is close to one. This is because their spatial resolution is almost the same, even though the pixeEquationare not entirely collocated.

We also see that the RMSD before applying the rescaling is significantly larger than after the rescaling, and also higher than the bias. This is again because it inherently contains both systematic and random deviations. The correlation coefficients, on the other hand, remain unchanged since they are insensitive to systematic differences.

The triple collocation error estimates before the rescaling are meaningless, since the rescaling is mandatory for the underlying error model (see Equation (6.38)). After rescaling, they are always lower than the unbiased RMSD estimates, since they are estimates for random errors of the individual data sets themselves, whereas the unbiased RMSD contains the random errors of both data sets for which the difference is calculated.




| | ascat-insitu | ascat-gldas | insitu-gldas |
|---|---|---|---|
| bias | 0.24 | 0.12 | -0.12 |
| pearson R | 0.42 | 0.51 | 0.55 |
| spearman rho | 0.38 | 0.54 | 0.55 |
| RMSD | 0.25 | 0.13 | 0.13 |
| stddev ratio | 1.10 | 1.01 | 0.91 |
| | ascat | gldas | insitu |
| Triple collocation error | 0.17 | 0.12 | 0.18 |

| | ascat-insitu | ascat-gldas | insitu-gldas |
|---|---|---|---|
| bias | 0.00 | 0.00 | -0.00 |
| pearson R | 0.42 | 0.51 | 0.55 |
| spearman rho | 0.38 | 0.54 | 0.55 |
| unbiased RMSD | 0.05 | 0.04 | 0.04 |
| stddev ratio | 1.00 | 1.00 | 1.00 |
| | ascat | gldas | insitu |
| Triple collocation error | 0.03 | 0.02 | 0.03 |

Figure 6.5: Example of a comparison between MetOp ASCAT soil moisture retrievals, in situ soil moisture measurements, and soil moisture estimates from a land surface model (GLDAS-Noah), one time without an a-priori rescaling (a), one time after applying a mean-standard deviation fit (b).

# Bibliography


[1] Attema, E., and F. T. Ulaby (1978), Vegetation modeled as a water cloud, *Radio Science*, **13**(2), p. 357–364.

[2] Barker, G., and P. Kitcher (2013), *Philosophy of Science: A New Introduction*, New York Oxford: Oxford University Press.

[3] Bertsekas, D. (1999), *Nonlinear Programming*, 2nd ed., Athena Scientific.

[4] Elachi, C., and J. van Zyl (2006), *Introduction To The Physics and Techniques of Remote Sensing*, Wiley Series in Remote Sensing and Image Processing, Wiley.

[5] EUMETSAT (2015), ASCAT Product Guide Version 5, *Tech. rep.*

[6] Figa-Saldana, J., J. J. W. Wilson, E. Attema, R. Gelsthorpe, M. R. Drinkwater, and A. Stoffelen (2002), The advanced scatterometer (ASCAT) on the meteorological operational (MetOp) platform: A follow on for European wind scatterometers, *Canadian Journal of Remote Sensing*, **28**(3), p. 404–412, doi:10.5589/m02-035.

[7] Gelsthorpe, R. V., E. Schied, and J. J. W. Wilson (2000), ASCAT – Metop's advanced scatterometer, *ESA Bulletin (ISSN 0376-4265)*, **102**, p. 19–27.

[8] Hahn, S., T. Melzer, C. Paulik, C. Reimer, S. Hasenauer, and C. Steiner (2013), Algorithm Theoretical Baseline Document (ATBD) for HSAF product H25/SM-OBS-4.

[9] Hallikainen, M. T., F. T. Ulaby, M. C. Dobson, M. A. El-Rayes, and L.-K. Wu (1985), Microwave dielectric behavior of wet soil - part 1: Empirical models and experimental observations, *Geoscience and Remote Sensing, IEEE Transactions on*, p. 25–34.

[10] Janssen, P., and P. Heuberger (1995), Calibration of process-oriented models, *Ecological Modelling*, **83**(1), p. 55–66.

[11] Losee, J. (2001), *A Historical Introduction to the Philosophy of Science*, Oxford: Oxford University Press.

[12] Morrison, K., and W. Wagner (2019), Explaining anomalies in sar and scatterometer soil moisture retrievals with sub-surface scattering in dry soils, *IEEE Transactions on Geoscience and Remote Sensing*, (submitted).

[13] Murphy, A. H. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, *Monthly weather review*, **116**(12), p. 2417–2424.

[14] Oh, Y., K. Sarabandi, and F. T. Ulaby (1992), An empirical model and an inversion technique for radar scattering from bare soil surfaces, *Geoscience and Remote Sensing, IEEE Transactions on*, **30**(2), p. 370–381.

[15] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in C*, Cambridge University Press.

[16] Quast, R., and W. Wagner (2016), Analytical solution for first-order scattering in bistatic radiative transfer interaction problems of layered media, *Applied Optics*, **55**(20), p. 5379–5386.

[17] Scipal, K., W. Wagner, M. Trommler, and K. Naumann (2002), The global soil moisture archive 1992-2000 from ERS scatterometer data: First results, p. 1399–1401, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Toronto, Canada, doi:10.1109/IGARSS.2002.1026129.

[18] Stoffelen, A. (1998), Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, *J. Geophys. Res.*, **103**(C4), p. 7755–7766.

[19] Ulaby, F., R. Moore, and A. Fung (1981), *Microwave Remote Sensing: Active and Passive, Vol. II - Microwave Remote Sensing Fundamentals and Radiometry*, Addison-Wesley, Advanced Book Program, Reading, Massachusetts.

[20] Ulaby, F., R. Moore, and A. Fung (1982), *Microwave Remote Sensing: Active and Passive, Vol. II - Radar Remote Sensing and Surface Scattering and Emission Theory*, Addison-Wesley, Advanced Book Program, Reading, Massachusetts.

[21] Ulaby, F., R. Moore, and A. Fung (1986), *Microwave Remote Sensing, Active and Passive, Volume III: From Theory to Applications*, Artech House, Incorporated.

[22] Ulaby, F. T., K. Sarabandi, K. McDonald, M. Whitt, and M. C. Dobson (1990), Michigan microwave canopy scattering model, *International Journal of Remote Sensing*, **11**(7), p. 1223–1253.

[23] Vachaud, G., A. Passerat De Silans, P. Balabanis, and M. Vauclin (1985), Temporal stability of spatially measured soil water probability density function, *Soil Sci. Soc. Am. J.*, **49**(4), p. 822–828, doi:10.2136/sssaj1985.03615995004900040006x.

[24] Wagner, W. (2010), Radiometric calibration of small-footprint full-waveform airborne laser scanner measurements: Basic physical concepts, *ISPRS Journal of Photogrammetry and Remote Sensing*, **65**, p. 505–513.

[25] Wagner, W., G. Lemoine, and H. Rott (1999), A method for estimating soil moisture from ers scatterometer and soil data, *Remote Sensing of Environment*, **70**(2), p. 191–207, doi:10.1016/S0034-4257(99)00036-X.

[26] Wagner, W., N. Verhoest, R. Ludwig, and M. Tedesco (2009), Editorial'remote sensing in hydrological sciences', *Hydrology and earth system sciences*, **13**(6), p. 813–817.

[27] Wagner, W., S. Hahn, R. Kidd, T. Melzer, Z. Bartalis, S. Hasenauer, J. Figa-Saldaña, P. de Rosnay, A. Jann, S. Schneider, J. Komma, G. Kubu, K. Brugger, C. Aubrecht, J. Züger, U. Gangkofner,

S. Kienberger, L. Brocca, Y. Wang, G. Blöschl, J. Eitzinger, and K. Steinnocher (2013), The ascat soil moisture product: A review of its specifications, validation results, and emerging applications, *Meteorologische Zeitschrift*, **22**(1), p. 5–33.

[28] Western, A. W., R. B. Grayson, G. Blöschl, G. R. Willgoose, and T. A. McMahon (1999), Observed spatial organization of soil moisture and its relation to terrain indices, *Water resources research*, **35**(3), p. 797–810.

[29] Zwieback, S., W. Dorigo, and W. Wagner (2013), Estimation of the temporal autocorrelation structure by the collocation technique with an emphasis on soil moisture studies, *Hydrological Sciences Journal*, **58**(8), p. 1729–1747.