The state of the earth: estimating physical parameters from noisy and incomplete earth observation data

Arp, L.R.

2

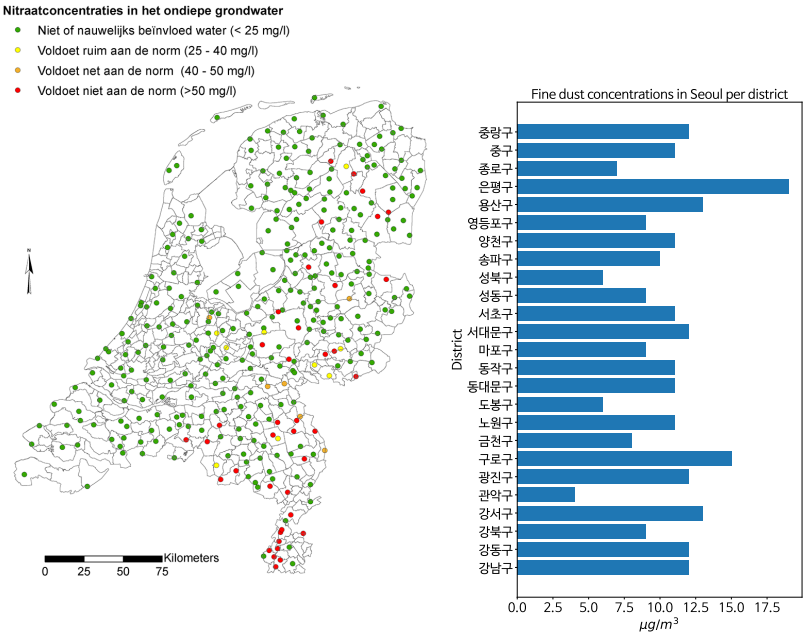
BACKGROUND

The work contained in this dissertation is highly inter-disciplinary, combining concepts from remote sensing, Earth science, physics, environmental biology, and artificial intelligence (AI). The intersection of these fields entails specific challenges that may be unfamiliar to an audience of experts specialised in one of these fields. The information contained in this chapter is intended to improve readability for experts from these different fields, enabling them to more easily follow the content from other fields in the main chapters of the dissertation.

First, Section 2.1 contains information about Earth observation data, which is the main type of input data for parameter estimation tasks. Second, Section 2.2 elaborates on radiative transfer models (RTMs) and the motivation for using them for parameter estimation. Finally, Section 2.3 covers the basics of the main computational methods that can be applied to parameter estimation from Earth observation data.

2.1. EARTH OBSERVATION DATA

In this section, we will cover the background of Earth Observation (EO) data. Section 2.1.1 will cover the main types of EO data available and motivate our goal of leveraging satellite data for physical parameter estimation. Next, we provide some general information on optical EO data, which we will be focusing on throughout this dissertation, in Section 2.1.2, and highlight some of the challenges arising in this context.



(a) Netherlands ground water

(b) Seoul fine dust

Figure 2.1: Examples of sensor network-based in-situ datasets. Figure 2.1a shows the ground water quality in the Netherlands as assessed by the Dutch National Institute for Public Health and the Environment [15] (image credits RIVM). Figure 2.1b shows the concentrations of fine dust (PM10 particles) in the atmosphere at the measuring stations of various districts in Seoul, South Korea, measured by AIRKOREA [12].

2.1.1. TYPES OF EO DATA

Earth observation data is a collective term for data that has been obtained from sensors measuring some properties of the Earth. For example, the Dutch RIVM, a government organisation responsible for national health and environment, operates a sensor network of 350 measuring stations, spread across the country, to monitor the quality of ground water in the Netherlands [11]. Other examples of this type of EO data includes the atmospheric fine particulate matter (fine dust) measuring stations in South Korea [12], the seismic activity and earthquake monitoring network in Japan [13] and various ecological measurements undertaken by NEON in the United States [14].

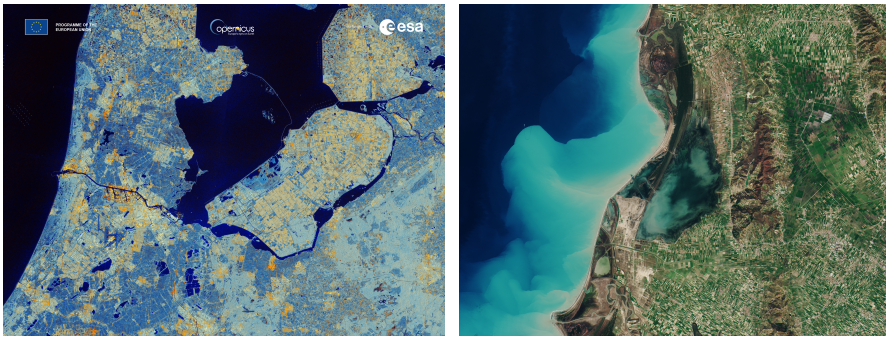
These sensor networks are examples of *in-situ* EO data: they enable direct measurements of the quantities we are interested in. We show two examples of this

type of data, namely the aforementioned ground water and fine dust examples, in Figure 2.1. The data from these sensor networks is typically represented as spatio-temporal point data, where every point is a station where sensors directly measure some quantity. In the ground water example, every measuring station in the sensor network measures water quality at a specific time and location. As a result, this type of data generally needs to be interpolated to enable users to derive estimations of a variable (such as ground water quality) at an arbitrary location; our work in Chapter 3 considers this type of data in more detail and will present a method to effectively perform this interpolation task.

In contrast, the largest part of EO data comes in the form of remote sensing data. In remote sensing, sensors are deployed remotely, which then observe the object of study from a distance. The main sources of remotely sensed EO data are aerial imagery, obtained by airborne sensors mounted on aircraft, and satellite imagery, obtained by spaceborne sensors mounted on satellites orbiting the Earth. The advantage of remotely sensed EO data over in-situ data is that it can be collected at a larger scale and obtain measurements for the full spatial area it is observing (a line determined by an orbit or flight path, with a width that is referred to as its swath). The resulting geo-referenced image can be directly used without requiring interpolation, and remotely sensed data is usually less expensive to set up and maintain at scale than in-situ data.

On the other hand, this type of data (especially aerial data) can still be expensive, is dependent on a sufficient spatial- and temporal resolution to produce meaningful results, and can typically only measure proxy variables such as the intensity of reflected light at different wavelengths (see Section 2.1.2) instead of the actual quantities of interest. This necessitates further processing to be converted into usable *data products*, thereby potentially introducing additional inaccuracies into the process. Most remotely sensed EO data is partially processed prior to data distribution, converting the raw data streams obtained by the sensors into a geo-referenced image with a predetermined map projection (usually *WGS84*), informative metadata and quality flags per pixel. Such images usually contain multiple data bands, where every band is an additional dimension containing an image for a different variable. For example, in optical imagery, the data often contains multiple bands measuring the intensity of reflected light at different spectral wavelengths, and synthetic aperture radar data contains multiple bands for different scene polarisations.

Once in orbit, spaceborne EO sensors can obtain continuous observations without major further interventions from the ground level, providing a large volume of raw data to its operators. For example, the Sentinel satellite platforms alone, operated by the European Space Agency (ESA) for the European Commission's Copernicus programme, transmitted 45 pebibytes (about $5.1 \cdot 10^{16}$ bytes) worth of EO data



(a) Sentinel-1 (b) Sentinel-2

Figure 2.2: Examples of satellite imagery from the Sentinel EO satellites by the European Space Agency (ESA). Figure 2.2a shows an example of Sentinel-1 synthetic aperture radar (SAR) data over the Netherlands (image credits ESA). Figure 2.2b shows an example of Sentinel-2 optical data over the Karavasta Lagoon in Albania (image credits ESA).

over the year 2023, with over 40 000 data products published every day [16]. This large scale makes satellite data an attractive type of EO data to build applications on: if reliable algorithms are created to estimate physical parameters from satellite data, this will provide us with a constant stream of information on the state of the Earth at any given time. Our focus will, accordingly, be on the development of algorithms applicable to spaceborne EO data.

On the other hand, satellite data also comes with significant drawbacks. Compared to aerial data, the increased distance between the sensor and the observation target results in lower spatial resolutions and an increased susceptibility to interference from, e.g., atmospheric conditions. Moreover, a substantial part of the collected data is unusable due to cloud cover (see Section 2.1.3, as well as Chapter 4 for our proposed method to remove clouds from satellite imagery). The validation of satellite-derived estimations can also be challenging, due to their large scale and low resolution, which usually cannot be directly compared to in-situ data. Instead, satellite-derived data products are often validated using derivations from airborne data, which are, in turn, validated using either in-situ data or further intermediate levels of abstraction, such as ground-based spectrometer data.

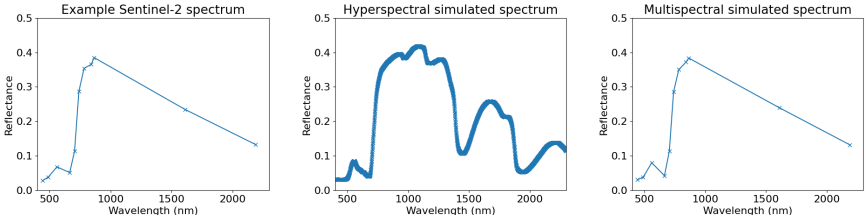
Although this hierarchical validation approach has resulted in the best evaluation of data products currently possible, it entails that the inaccuracies and error rates of lower-level EO data (such as in-situ data) inevitably trickle up to higher-level data (such as satellite data), with an additional information loss at every conversion between two levels of EO data. This can pose challenges for the application

of conventional machine learning and deep learning methods to EO data, because some of the central concepts, such as ground-truth data and reliable metrics to optimise for, are not necessarily available for all types of applications. On the other hand, when building applications that are not reliant on large amounts of accurate ground truth data, certain deep learning techniques can be highly effective, such as transfer learning [17], representation learning [18, 19], self-supervised and semi-supervised methods [20, 21] and foundation models [22, 23, 24]. Furthermore, the models can be trained and evaluated more reliably in problem settings where labels can be reliably generated by human annotators (e.g., land cover classification [25, 26] and segmentation tasks [27, 28, 29]).

2.1.2. OPTICAL EO DATA

The large majority of remotely sensed EO data comes in the form of optical data (shown in Figure 2.2b); a discussion of other forms of remote sensing data, such as radar, lidar and synthetic aperture radar (SAR) (shown in Figure 2.2a), is beyond the scope of this work. The sensors used for optical data are referred to as spectrometers, which measure the intensity of light at certain wavelengths of the electromagnetic spectrum. In addition to the red ($\sim 665nm$), green ($\sim 560nm$) and blue ($\sim 490nm$) wavelengths that are visible to the human eye, spectrometers can measure light intensity for ultraviolet ($\leq 400nm$) and (near-)infrared ($\geq 780nm$) light. This light energy in invisible wavelengths can be highly informative to various applications; for example, red-edge and near-infrared (relatively low wavelength infrared) light is known to be heavily affected by vegetation and photosynthesis [30, 31], while ultraviolet light has applications in, e.g., aerosol detection [32]. A typical light spectrum would not contain the same intensity at every part of the spectrum; for example, light intensity at visible wavelengths tends to be much lower than near-infrared wavelengths; this can be observed in all the spectra visualised in Figure 2.3.

Spaceborne optical sensors measure sunlight reflected by the Earth at a set of pre-determined wavelengths, called spectral bands. Most spaceborne spectrometers produce multispectral data, which contains multiple spectral bands spread around the spectrum. Although there are no strict rules, optical data containing tens of spectral bands is generally referred to as multispectral data. Historically, the NASA Landsat and MODIS satellites have provided multispectral data; over the last decade or so, the ESA Sentinel-2 satellites have gained much traction as a source of optical data. Most of the work in this dissertation will focus on applications based on Sentinel-2 data, since this is the most popular satellite data at the time of writing. An example spectrum measured by Sentinel-2 for a field to the south of Leiden, the Netherlands, on a spring day, can be found in Figure 2.3a.



(a) Sentinel-2 (b) PROSAIL hyperpsectral (c) PROSAIL multispectral

Figure 2.3: Visualisation of example spectra: a) a real-world spectrum measured by the Sentinel-2 multispectral satellite at a field to the south of Leiden, the Netherlands; b) a simulated hyperspectral spectrum simulated by the PROSAIL RTM (see Section 2.2), using the best fitting parameters to the Sentinel-2 observations of Figure 2.3a; c) the same simulation using PROSAIL as in Figure 2.3b, but after converting the hyperspectral output to a multispectral format matching the spectral bands of Sentinel-2 observations.

In addition to multispectral data, there are also sensors producing hyperspectral data. The best-fitting hyperspectral simulation for the Sentinel-2 observations in Figure 2.3a, determined using RTM inversion as described in Section 2.2, can be found in Figure 2.3b, and its corresponding multispectral version can be found in Figure 2.3c. Hyperspectral data is similar to multispectral data, but contains many spectral bands (high tens, hundreds or thousands). The increased spectral resolution may improve parameter estimation approaches by increasing the information content of the data, as illustrated in Figure 2.3 by the extra details in the hyperspectral image in Figure 2.3b compared to the multispectral images in Figures 2.3a and 2.3c. This could potentially reduce ill-posedness in parameter estimation settings; however, the improvements to spectral resolution may come at an expense of lower sensor accuracy, or sacrificing other types of resolution (spatial, temporal). The increased multi-collinearity of the band values may also require additional processing.

Since there are few hyperspectral satellite platforms currently in operation at the time of writing, and the hyperspectral satellite data that is available is often commercial (e.g., the Tanager satellites by Planet Labs or the GHOst satellites by Orbital Sidekick) or otherwise not publicly available for all locations and times (e.g., the PRISMA satellite by the Italian Space Agency), we will focus our methods on multispectral data. Throughout this work, we will specifically focus on Level-2A Sentinel-2 data products, which are optical images that have been atmospherically corrected.

2.1.3. EO DATA CHALLENGES: QUALITY, QUANTITY, AND DIVERSITY

Given the large volumes of EO data available, combined with the scale of EO imagery and the spatial and temporal relationships that are often contained in the data, large deep learning models form an appealing option for many EO tasks, including parameter estimation. However, deep learning models usually require large amounts of labelled data to train. While the EO feature data $\mathbf{x}$ may be plentiful, collecting ground truth data for the physical parameters $\boldsymbol{\theta}$ can be expensive, challenging, or even impossible for parameter estimation in particular, resulting in small dataset sizes for this type of problem. For example, when creating a ground truth dataset for leaf area index (LAI), data collection missions involve sending a team of researchers to an area of interest, who then systematically take measurements at a regular spatial grid that may cover thousands of square metres. Even if such a full grid is sampled, the study area is only a fraction of the total area covered by a single Sentinel-2 data product. Additionally, some parameters may involve extensive lab work (e.g., to test for certain chemicals) which can be destructive to the environment, or may not be directly measurable (for example, even at ground level, LAI is often measured using spectrometers).

The geographical and seasonal diversity of the available data can also be a key factor reducing the effectiveness of parameter estimation models. A low data diversity would not be representative of all possible inputs $\mathbf{x} \in \mathcal{X}$, thereby forcing models to extrapolate beyond training data (a task known to be challenging for deep learning models in particular [33]). For example, in the context of global vegetation monitoring, given the great diversity of species, ecosystems, climates, land cover types, lighting conditions and more, creating a representative dataset of all these conditions would be highly challenging. Additionally, the feasibility and cost of collecting ground truth data may only be acceptable to certain parts of the world (e.g., wealthy countries). This can bias models to only perform well on applications matching the training data, thereby mainly benefiting countries that are already wealthy. The use of physics-aware approaches, such as physics-informed neural networks [33, 34] and simulation-based hybrid models [35, 36] may improve the generalisability of prediction models.

Even if there is a large, representative ground truth dataset available, the reliability of this data must also be considered. Unlike typical machine learning problems with well-defined benchmark datasets containing reliable, human-defined ground truth labels (e.g., MNIST, CIFAR10 [37]), the ground truth data in parameter estimation concerns physical quantities that must be measured, such as the temperature, humidity, LAI and $PM10$ concentrations in the example of Chapter 1. For many parameters, the accuracy of these measurements is limited, and they frequently involve a tradeoff between accuracy and scale (see Section 2.1.1). This type of data is typically not directly derived from true measurements, but rather based

on estimations derived from a measured light spectrum. The measurements (that may be at a microscopic scale) must then be aggregated over a $10 \times 10 m^2$ grid cell (assuming a reasonably high spatial resolution), whose conditions can greatly vary within the spatial coverage of the cell.

Finally, the EO data $\mathbf{x}$ passed to the model will always be noisy, due to limitations of the sensors themselves, atmospheric interference, and spatial aggregation artefacts (e.g., spectral mixing [38]).

CLOUD COVER AND OPTICAL DATA

One of the most important drawbacks of optical data compared to, e.g., synthetic aperture radar (SAR) data, is that the signal measured by the spectrometer can be blocked by cloud cover. When the spectrometer encounters clouds, it becomes impossible to measure light at the ground level, while observing ground-level processes is generally the objective of optical data. At any time, 55% to 72% of the Earth is covered by clouds on average [39], with oceans accounting for the higher end of this range. Moreover, this cloud cover can be affected by spatial- and temporal autocorrelation, exacerbating the issue. For example, tropical regions will experience large amounts of cloud cover at any time of year, while temperate regions may experience long stretches of constant cloudy conditions during winter, followed by long stretches of clear conditions in summer. As a result, cloud cover can be a large obstacle to the use of satellite-based optical data for parameter estimation. In some cases, it can take months before a cloud-free observation can be made, resulting in large temporal gaps in the data. Removing clouds to fill these data gaps is an active area of research (see, e.g., [40, 41, 42]), and existing cloud removal methods are usually limited by a combination of poor scalability, limited reliability of ground truth data and poor transferability between different types of EO data.

Due to these limitations, cloud cover is one of the key challenges for parameter estimation using EO data. Therefore, we propose a novel cloud removal method in Chapter 4, through which we aim to increase the amount of usable optical data, thereby improving parameter estimation. Our approach aims to overcome some of the limitations of existing methods through computational efficiency, and by requiring no model training, thereby avoiding ground truth quality and transferability limitations.

2.2. RADIATIVE TRANSFER MODELS

Given the challenges and limitations of optical EO data described in Section 2.1.3, a conventional application of machine learning models is not always feasible when estimating parameters using Earth observation data: ground truth data may not

be available, may not be accurate enough to train high-quality estimators on, or may not be representative of all conditions in which the model would be deployed. However, the Earth system is studied by numerous scientific disciplines, many of which contain a wealth of scientific domain knowledge on the physical processes affecting this system. This domain knowledge can be used to alleviate some of the drawbacks of purely data-driven approaches, by simulating synthetic data. Since we are interested in parameter estimation from Earth observation data, we would need to use a domain knowledge-based (physical) model to simulate spectral data, corresponding to light spectra as measured by optical Earth observation satellites, for a specified parameter configuration. This type of simulation model is known as a radiative transfer model:

Definition 2.1 (radiative transfer model)**.** A radiative transfer model (RTM) is a simulation model $M : D_P \rightarrow \mathcal{X}$ whose parameterisation $\boldsymbol{\theta} \in D_P$ represents the state of physical parameters on Earth. It simulates a light spectrum $\boldsymbol{x} \in \mathcal{X}$ (where $\mathcal{X}$ is usually a space in $\mathbb{R}^d$ containing d spectral bands) that could be produced by the specified conditions.

RTMs are based on well-studied physical laws and domain knowledge. They model how a beam of light is affected as it is absorbed or reflected by the media it encounters, such as particles and gas concentrations in the atmosphere [1, 2], ocean water and microorganism production [3], or vegetation canopies and leaves [4, 43, 44]. An RTM takes a physical parameter configuration $\boldsymbol{\theta}$, such as the example configuration described in Section 1, as input, and uses this to simulate what a hypothetical beam of light would have looked like under these conditions. These domain knowledge-driven models play a pivotal role in parameter estimation applications where purely data-driven approaches cannot be easily applied.

The RTM we will focus on in Chapters 5 and 6 is PROSAIL, which combines the PROSPECT leaf model [43] and the 4SAIL canopy model [44], and is widely used in state-of-the-art vegetation parameter estimation methods [35, 45, 46]. However, many of our findings, especially those in Chapter 6, are likely to generalise to other RTMs.

2.2.1. RTM INVERSION

An RTM simulates a light spectrum $\mathbf{x} = M(\boldsymbol{\theta})$ based on the input parameters $\boldsymbol{\theta}$ it received. However, the data observed by optical satellites (see Section 2.1.2) already contains the light spectrum $\mathbf{x}$. Instead, the unknown physical parameters are the targets needing estimation, while these form the input parameters $\boldsymbol{\theta}$ of the RTM. Therefore, RTMs must be *inverted* (model inversion) in order to use them for parameter estimation through EO data.

Definition 2.2 (RTM inversion)**.** Given an RTM $M : D_P \rightarrow \mathcal{X}$, where $\mathbf{x} = M(\boldsymbol{\theta})$, RTM inversion refers to computing the inverse function $M^{-1} : \mathcal{X} \rightarrow D_P$ of M such that $\boldsymbol{\theta} = M^{-1}(\mathbf{x})$.

When inverting RTMs for parameter estimation, an analytical inversion is generally not possible to directly formulate M^{-1}. This is because the complex internal structure of the RTM M, often containing highly non-linear relationships and partial- or ordinary differential equations, is ill-suited to the derivation of an inverse function M^{-1} of M. Instead, the RTM inversion problem can be interpreted as a black-box numerical optimisation problem (for details, see Section 2.3.1). Here the task is to find a parameter configuration $\hat{\boldsymbol{\theta}}$ for which, when comparing the associated simulated spectrum $M(\hat{\boldsymbol{\theta}})$ and the observed spectrum $\mathbf{x}$, the difference between these spectra should be minimal:

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathcal{D}_P}{\operatorname{argmin}} \mathcal{L}\big(M(\boldsymbol{\theta}), \mathbf{x}\big) \tag{2.1}$$

Here $\mathcal{L}$ is a loss function measuring the goodness-of-fit between the observed spectrum $\mathbf{x}$ and the simulated spectrum $M(\boldsymbol{\theta})$ for a given parameterisation $\boldsymbol{\theta}$. This traditional approach to RTM inversion has been applied successfully to parameter estimation using PROSAIL and EO data [47, 48], although some authors note that its primary purpose is to validate the RTMs themselves [49]. Any of the black-box optimisation methods described later, in Section 2.3.1, could be used to find $\hat{\boldsymbol{\theta}}$ by optimising the loss function $\mathcal{L}$. In Figure 2.3, the hyperspectral (Figure 2.3b) and multispectral (Figure 2.3c) spectra both contain a PROSAIL simulation using the configuration $\hat{\boldsymbol{\theta}}$ as identified through Equation 2.1.

In recent years, much of the research focus for RTM inversion has been on so-called *hybrid models* [50, 51, 52, 45, 53, 54, 55, 46]. Hybrid models, sometimes referred to as *inverted simulators* or *inverted emulators*, combine knowledge-driven RTM simulations with data-driven machine learning models performing the inversion [35]. First, the RTM is used to generate a look-up table (LUT) of input parameters $\boldsymbol{\theta}$ and the simulated light spectra $\mathbf{x}$; this LUT can optionally be combined with a tabular dataset containing real-world data [56]. A machine learning model can then be trained on this dataset, taking the spectral data $\mathbf{x}$ as input features to predict the parameters $\boldsymbol{\theta}$ used by the RTM to generate that spectrum.

This hybrid modeling approach amortises the main computational cost of RTM inversion to a machine learning model training procedure. Therefore, parameter estimations can be performed much more efficiently in this manner compared to an approach requiring a new, high-dimensional optimisation procedure for every new problem instance, which may be highly relevant for, e.g., global mapping applications. Recent work on hybrid models often focuses on different sampling

strategies and heuristics through active learning [51, 45], enabling efficient and effective training in parts of the space where, for example, uncertainty is the highest. Although hybrid models can be an effective tool for efficiently performing the inversion of RTMs, they, like other parameter estimation methods, are limited by the ill-posedness of the problem. We explore this further in Chapters 5 and 6.

2.2.2. ILL-POSEDNESS

The inversion of RTMs is generally considered an ill-posed problem [57, 58, 59]. Ill-posed problems are problems that do not meet the requirements of well-posedness; the following must hold for an arbitrary problem to be considered well-posed [60]:

1. **The problem has a valid solution.** In the context of parameter estimation, this means that there exists a configuration of target parameters $\boldsymbol{\theta}$ that explains the observed light spectrum $\mathbf{x}$: $\exists \boldsymbol{\theta} : M(\boldsymbol{\theta}) \approx \mathbf{x}$.

2. **The solution to the problem is unique.** There should be only one configuration $\boldsymbol{\theta}$ that explains the observed light spectrum $\mathbf{x}$: $|\operatorname{argmin}_{\boldsymbol{\theta} \in D_P} \mathcal{L}\big(M(\boldsymbol{\theta}), \mathbf{x}\big)| = 1$.

3. **The solution moves continuously with regard to the inputs.** In parameter estimation, when visualising a point moving through the space $\mathcal{X}$ of possible input spectra $\mathbf{x}$, every movement in this space should correspond to a smooth movement of the solution $\hat{\boldsymbol{\theta}}$ in the parameter space D_P, with no sudden jumps to other parts of the space or other discontinuities. If f is the function mapping $\mathbf{x}$ to $\hat{\boldsymbol{\theta}}$, this function should be continuous: $\forall \mathbf{x}' \in \mathcal{X} : f(\mathbf{x}') \in D_P \wedge \lim_{\mathbf{x}' \to \mathbf{x}} f(\mathbf{x}') \in D_P \wedge \lim_{\mathbf{x}' \to \mathbf{x}} f(\mathbf{x}') = f(\mathbf{x})$.

Although RTM inversion is generally considered an ill-posed problem, due to a violation of requirement 2 (unique solution), this ill-posedness is not yet well understood, and, to our knowledge, there had been no structured, formal analysis of the phenomenon. Therefore, we aimed to fill this knowledge gap in Chapter 5 by systematically evaluating the ill-posedness of PROSAIL inversion, i.e., the inversion of an RTM that is widely used for vegetation parameter estimation applications.

2.3. ESTIMATION METHODOLOGIES

Within the scope of this work, two main strategies are considered for performing parameter estimation: black-box optimisation and machine learning. As a result, many of the chapters in this thesis assume some knowledge of these techniques, which form the backbone of our methodological contributions, but do not provide much background on these methods for readers who may not be familiar with

them. In this section, we will explain the relevant methods in more detail, such that the rest of the chapters can be more easily understood.

2.3.1. BLACK-BOX OPTIMISATION

Optimisation problems are pervasive in many different fields, including logistics [61], operations research [62] and industrial design [63]. Mathematically, suppose we are interested in a set of variables $P_1, P_2, ..., P_d$, whose domain is denoted as D_P (in continuous settings, this would be the d-dimensional space of real numbers $\mathbb{R}^d$). When assigned a specific value, these variables form a *configuration* $\boldsymbol{\theta}$:

Definition 2.3 (configuration)**.** A d-dimensional vector $\boldsymbol{\theta} \in D_P$ containing concrete value assignments $\theta_1, \theta_2, ..., \theta_d$ for the variables $P_1, P_2, ..., P_d$, representing the current state of a physical system in parameter estimation.

We also have access to an *objective function* $g(\boldsymbol{\theta})$:

Definition 2.4 (objective function)**.** A function $g : D_P \to \mathbb{R}$ mapping an input configuration $\boldsymbol{\theta}$ to a scalar $g(\boldsymbol{\theta})$, indicating the quality of the configuration (typically indicated by the distance between $M(\boldsymbol{\theta})$ and $\mathbf{x}$ in parameter estimation).

A practitioner may be interested in finding the *optimum* for g; that is, a configuration $\boldsymbol{\theta}^*$ for which the value of g is either maximised (for example, the best environmental conditions to make crops grow as fast as possible) or minimised (for example, the best water management approaches to ensure the risk of forest fires is as low as possible). In the context of physical parameter estimation, this optimum $\boldsymbol{\theta}^*$ would become the prediction $\hat{\boldsymbol{\theta}}$. Assuming the objective function should be minimised, the goal of optimisation is to find:

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta} \in D_P}{\operatorname{argmin}}\, g(\boldsymbol{\theta}) \tag{2.2}$$

In ideal cases, the optimum $\boldsymbol{\theta}^*$ can be computed analytically; for example, by solving for $\boldsymbol{\theta}$ after setting the derivative $g'(\boldsymbol{\theta}) = 0$. Unfortunately, in most practical optimisation problems, it is not possible to analytically compute $\boldsymbol{\theta}^*$, because i) the objective function g may be unknown entirely (black-box optimisation), ii) the objective function may be known, but not differentiable (for example, if g involves complex simulations), or iii) there may be an unknown number of *local* optima $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_n^*$ (as opposed to a single *global* optimum $\boldsymbol{\theta}^*$) where $g'(\boldsymbol{\theta}_1^*) = g'(\boldsymbol{\theta}_2^*) = ... = g'(\boldsymbol{\theta}_n^*) = 0$.

Black-box optimisation refers to general-purpose optimisation methods where the objective function g is unknown. In this case, the optimisation task requires a *search* over the *search space* D_P, where only the output of the objective function

g, but not the function itself, can be used to guide the search. Examples of such methods include stochastic local search methods [64], evolutionary algorithms [65], metaheuristic algorithms such as particle swarm optimisation [66] and ant colony optimisation [67], as well as surrogate model-based Bayesian optimisation [68]. Black-box optimisation often involves a tradeoff between *exploration* (covering as much of the search space as possible) and *exploitation* (quickly reaching a local optimum for promising regions in the search space).

A typical black-box optimisation setting is limited by its reliance on the objective function g. Although there are methods that are more robust to noisy objective function evaluations [69], such methods cannot make a difference in cases where the signal from the objective function itself is unreliable. For example, the choice of objective function could be inappropriate for the problem instance, or may be a loosely correlated proxy function to an unknown true objective function, which may converge to a solution that is incorrect for the problem. Therefore, even if the global optimum $\boldsymbol{\theta}^*$ can be found reliably, this may not always mean that the identified optimum is also the true solution. If the objective function g is not fully reliable, there may be other points in the search space that consistently evaluate to a worse objective function value, but are actually the true solution. We explore and address this problem in detail in Chapter 5, where black-box optimisation is used extensively in the experiments to characterise the loss landscape of RTM inversion, and Chapter 6, where black-box optimisation is an important component in our proposed method to for approximating the set of potential solutions to inference problems (including RTM inversion).

In addition, Chapters 3, 4 and 6 all rely on black-box optimisation techniques for *automated algorithm configuration*; in this problem setting, the hyperparameters of the methods used (such as machine learning models, whose automated configuration is also known as AutoML) are automatically tuned using optimisation approaches based on their performance on validation data.

2.3.2. SUPERVISED MACHINE LEARNING

Much of machine learning (ML) consists of supervised ML, which refers to a wide range of predictive models whose model parameters can be tuned (trained) via data. Popular examples of traditional supervised machine learning models include linear regression, support vector machines (SVM) and Gaussian processes [70]. The training of supervised machine learning models is an optimisation problem, like those described in Section 2.3.1, where the objective function g to be minimised consists of a *loss function* $\mathcal{L}$, such as the mean squared error (MSE), mean absolute error (MAE), or accuracy. The loss function measures the predictive performance of the machine learning model by comparing the predictions made by

the model (under some parameterisation) to the ground truth values they should approximate (thus 'supervising' the model). Unlike black-box optimisation, the optimisation procedure for training a machine learning model can usually be performed efficiently via gradient-based methods.

ML models can be used to estimate parameters from the EO data $\mathbf{x}$, referred to as *features*, to predict $\boldsymbol{\theta}$. At the same time, the training of ML models is itself an inference problem, where the parameterisation $\boldsymbol{\theta}$ of a machine learning model must be inferred from the observed training data.

In the context of EO, deep learning models (a type of ML model using large neural networks containing many artificial neurons) are often preferred, due to their ability to take advantage of the large volumes of (largely unlabelled) training data, as well as the scale of their predictions (e.g., full images). Given their suitability for image data and modelling local spatio-temporal patterns, convolutional neural networks (CNNs) are particularly popular in EO settings, and have been applied successfully to various problems, such as land cover classification [71, 72], crop classification [73], semantic segmentation [74] and super-resolution [75]. More recently, many proposed models incorporate some form of attention mechanism [76, 77, 78], and transformer architectures have seen a rise in popularity, particularly in zero-shot settings [79, 80], where a model, trained on a problem with a certain set of classes, needs to make accurate predictions for a problem with a different set of classes, without any additional training for this new task. Deep neural networks are a natural choice for *data fusion* approaches, where information from multiple sensors (e.g., optical data from Sentinel-2 and SAR data from Sentinel-1) can be combined automatically in a latent representation [40, 41, 81].

Unlike the examples above, parameter estimation is a more difficult inference problem to perform using conventional machine learning and deep learning approaches (see Section 2.1.3 for details). Although there are deep learning-based data products [82], and prediction models [83, 84] available to perform the estimation for, e.g., LAI, it is difficult to train and evaluate such models for global applications when the ground truth data used for this may be insufficient in terms of quantity and/or quality (see Section 2.1.3). Because the training and validation data would have similar drawbacks, inaccuracies and biases in the model would be difficult to diagnose with the available data. This may, in part, explain why RTM inversion-based hybrid approaches (see Section 2.2.1) remain popular, often using traditional machine learning methods such as Gaussian processes.

In summary, deep learning-based approaches can be highly effective at a large number of typical Earth observation problems, can exploit the large volumes of input data available, and, once trained, scale well to rapidly process large image datasets and produce predictions for, e.g., all the pixels in such images. On the other hand, these approaches run into similar problems as traditional methods,

such as ill-posedness (see Section 2.2.2), while diagnosing those problems may be more difficult for deep learning models compared to traditional approaches. Our work in Chapter 4 may help improve the input data consistency for datasets to be used by deep learning methods, while our work in Chapter 6 can shed light on the nature of the parameter estimation problem, to help diagnose ill-posedness that deep learning models would also be affected by.