POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian Deep Learning for Forest Height Estimation from TanDEM-X Single-Pass InSAR Data

Master's Thesis in Space Engineering

Federico Ghio

Advisor:
Prof. Stefano Tebaldini — Politecnico di Milano

Co-Advisors:
Dr. Daniel Carcereri — German Aerospace Center (DLR)
Dr. Paola Rizzoli — German Aerospace Center (DLR)

Academic Year: 2023-24
Student ID: 220481

Non è mai facile iniziare tutto da capo, in una terra sconosciuta e in una lingua straniera, imparare a gestire la solitudine di alcuni momenti. Non è facile lasciare le certezze del tuo mondo abituale per le incertezze di un mondo nuovo.

Aveva ragione Cesare Pavese quando disse: "Viaggiare è una brutalità. Obbliga ad avere fiducia negli stranieri e a perdere di vista il comfort familiare della casa e degli amici. Ci si sente costantemente fuori equilibrio. Nulla è vostro, tranne le cose essenziali — l'aria, il sonno, i sogni, il mare, il cielo — tutte le cose tendono verso l'eterno o ciò che possiamo immaginare di esso".

Ma è proprio per questo che viaggiare, cambiare ambiente e conoscere altre culture è uno straordinario modo per crescere — e per farlo in fretta. Il contatto con un mondo sconosciuto è qualcosa che ti cambia nel profondo perché ti costringe a contare solo sulle tue forze e a superare i tuoi limiti.

Sergio Marchionne

Abstract

Forests are among the planet's most critical ecosystems, shaping climate patterns and sustaining biodiversity. Increasingly threatened by human activities, forests require conservation strategies, for which large-scale monitoring of biophysical parameters, such as forest height and biomass, provides essential insights. Still, existing approaches often compromise accuracy, resolution, time and spatial coverage.

Deep Learning-based methods applied to Interferometric Synthetic Aperture Radar (InSAR) data have shown promising results, exceeding state-of-the-art performance. However, their applicability to long-term monitoring remains hindered, as uncertainty quantification is often neglected despite being essential for assessing the reliability of the estimates and monitoring their changes over time.

This thesis proposes a novel Bayesian Deep Learning-based framework for the estimation of forest height using TanDEM-X single-pass interferometric data. First, a case study in Norway is introduced, leveraging nationwide Airborne Laser Scanning (ALS) data to analyse the impacts of tree species and input-output temporal misalignments on generalisation performance. Second, it investigates how uncertainty arises within the modelling framework and explicitly incorporates aleatoric uncertainty during the estimation process. Additionally, intra and inter-basin methodologies are compared to integrate the notion of epistemic uncertainty. Finally, the study evaluates the model's reliability under out-of-distribution conditions, which may arise in real-world remote sensing applications. Results demonstrate strong generalisation performance alongside the generation of well-calibrated uncertainty maps, supporting the development of uncertainty-aware products for canopy height monitoring through InSAR observations.

Keywords: Bayesian Deep Learning, Earth Observation, Remote Sensing, Forest Monitoring, Canopy Height, InSAR, LiDAR, TanDEM-X

Contents

Introduction

The Importance of Forest Monitoring

Forests cover approximately 31% of the Earth's land areas, hosting nearly 80% of the world's terrestrial biodiversity and acting as the primary ecological and climatic system regulator [1]. Their contribution to essential ecosystem functions — including carbon sequestration, water regulation, soil preservation and habitat provision for species diversity — is intrinsically tied to their structural complexity, defined as the three-dimensional spatial arrangement of vegetation [2]. Moreover, forests play a pivotal role in climate regulation by annually absorbing approximately one-third of anthropogenic CO_2 emissions, functioning as carbon sinks and mitigating the impacts of climate change [3] [4].

Between 1990 and 2020, approximately 420 million hectares of forests were globally lost [1] — an area larger than the European Union — resulting in a significant carbon storage loss, biodiversity decline and habitat destruction, which threaten countless species and undermine the livelihoods of indigenous communities which rely on forest resources [5]. Deforestation, primarily driven by agricultural expansion, urbanisation and logging activities, contributes to nearly 11% of global carbon emissions, more than the entire transportation sector [6].

To support environmental and conservation strategies at global and European levels, the accurate assessment and continuous monitoring of forest structures worldwide are essential to inform stakeholders in decision-making processes, aiming at climate change mitigation and biodiversity conservation.

Research Gap

Traditional in-situ measurements for assessing forest resources, such as tree height and diameter at breast height, provide precise but irregular data due to labour-intensive requirements and logistical constraints [7] [8]. Terrestrial Laser Scanning offers high-resolution three-dimensional canopy structure measurements, extending the coverage to local areas.

However, it is similarly constrained by its limited spatial coverage and occlusions resulting from dense vegetation [2]. In this context, Airborne Laser Scanning (ALS) expands coverage to regional scales with high vertical accuracy; however, its operational costs and logistical demands limit large-scale monitoring, primarily confining its use to acquire reference data for validating remote sensing products.

Satellite-based remote sensing overcomes these limitations through a combination of wide spatial coverage and periodic revisit capabilities enabled by orbital trajectories. For instance, multispectral sensors capture key spectral features relevant to vegetation analysis but are limited by shallow canopy penetration and weather dependency. Differently, Synthetic Aperture Radar (SAR) systems provide weather-independent and high-resolution imaging [9]. However, extracting accurate forest structure information from SAR data is challenging due to the complex interaction between radar waves and on-ground targets, which is influenced by moisture content, canopy structure, terrain topography and seasonal variations [10].

Data-driven approaches, particularly Convolutional Neural Networks (CNNs), have recently emerged as powerful tools for inferring complex non-linear spatial patterns from satellite data [7]. However, generalisation across diverse ecosystems remains an open challenge due to variations in vegetation, climate and terrain features. Consequently, when transitioning to a new environment, training and validating the model within the new target conditions is essential to ensure reliable performance. Additionally, despite their promising performance, most CNN-based studies found in the literature provide punctual estimates without quantifying prediction uncertainty [11] [12] — arguably leading to misleading conclusions — despite the well-documented tendency of CNNs to generate overconfident predictions [13]. Without a reliability measure, stakeholders lack a trustworthy basis for decision-making based on CNN-derived results, which is particularly critical in applications such as change detection and long-term environmental monitoring, where there is no means to determine whether observed variations reflect genuine changes or merely artefacts of the model uncertainty. This gap has sparked a growing interest in the Bayesian interpretation of deep neural networks, which can provide a mathematical framework for uncertainty quantification in predictions [14] [15] [16].

Thesis Objectives

This thesis builds upon the work published in [17] — which presented a novel approach to mapping canopy height in Gabon's tropical forests using TanDEM-X Interferometric Synthetic Aperture Radar (InSAR) data — by introducing a Bayesian perspective into the

fully-supervised CNN-based framework. The model leverages InSAR-derived features as inputs and ALS-derived reference data to generate pixel-level canopy height maps at a 20-metre ground resolution, together with well-calibrated uncertainty estimates. Specifically, this thesis is built around two main objectives, shaping the overall structure of the work:

Generalisation beyond single-nation studies, by evaluating the adaptability of the reference framework to diverse ecosystems and assessing model performance across Norwegian forests; followed by an examination of the implications of temporally misaligned satellite-reference data, an analysis of temporal dynamics, and an investigation into the influence of tree species composition. The ultimate objective is to develop methodologies for reliable large-scale forest monitoring applications.

Enhancing model interpretability and reliability, by integrating Bayesian Deep Learning for uncertainty estimation and leveraging it in the training process to enable more flexible learning dynamics, aiming to improve model performance through a probabilistic interpretation of the data.

Thesis Structure

The structure of this thesis is organised as follows:

Chapter 1 discusses the working principles of Radar, SAR and InSAR technologies, presenting the key equations and theoretical foundations necessary for the thesis understanding.

Chapter 2 introduces deep learning-based regression principles, beginning with the generic concept of feedforward neural networks before focusing on convolutional neural ones. It also explores the role of uncertainty in deep learning, detailing how it can be leveraged through Bayesian Deep Learning approaches.

Chapter 3 reviews the state of the art for estimating canopy height, from traditional in-situ methodologies to more advanced remote sensing techniques, highlighting their advantages and limitations. It examines both physics-based and data-driven approaches.

Chapter 4 presents the datasets used in this study, with a focus on the TanDEM-X mission. Furthermore, it describes the data preprocessing pipeline, including the existing tools leveraged and the necessary steps to ensure proper formatting for deep learning

applications. The chapter concludes by defining the area of interest for the case study, Norway, and characterising its forest coverage in terms of extent and composition.

Chapter 5 investigates the model's generalisation capabilities in the Norwegian environment and establishes a baseline scenario for comparison. It also examines the impact of input-reference temporal misalignment and the influence of various tree species on prediction performance.

Chapter 6 integrates the Bayesian perspective into the original CNN-based framework, analysing its impact on predictive performance and the trade-offs associated with different uncertainty quantification strategies. The chapter also re-evaluates the effects of temporal misalignment and conducts an out-of-distribution analysis to evaluate uncertainty estimation robustness in unfavourable conditions.

Chapter 7 summarises the key findings of this research. It discusses potential improvements and directions for advancing large-scale forest height estimation from InSAR data using deep learning.

1 | Understanding Synthetic Aperture Radar

This chapter provides an overview of Synthetic Aperture Radar (SAR) and Interferometric SAR (InSAR). These technologies are fundamental to understanding the thesis contribution, as they deliver high-resolution, weather-independent imaging and provide accurate topographic information.

1.1. Principles of Radar

Radar (Radio Detection and Ranging) was initially developed for military purposes and proved to be crucial during the Battle of Britain in 1940 by enabling early detection of enemy aircraft, thus providing a strategic advantage in air defence. Its operational mechanism involves the active transmission of a known electromagnetic pulse, which propagates through a medium, interacts with the target and subsequently reflects towards the receiver. The distance (R) to the target can be determined by measuring the round-trip travel time (Δt) between transmission and reception:

$$\mathrm{R} = \frac{c \cdot \Delta t}{2} \tag{1.1}$$

where c is the speed of light in the propagation medium.

Each radar cycle is defined by the *Pulse Repetition Interval* (PRI), consisting of the duration of the transmitted pulse and a subsequent listening period. The latter constrains the maximum detectable range, as echoes arriving beyond this interval may lead to *range ambiguities*, where reflections from preceding pulses are mistakenly associated with new transmissions. The *Pulse Repetition Frequency* (PRF) is the reciprocal of the PRI and represents the number of cycles per second.

1.1.1. Range Resolution

The ability of a radar system to resolve two closely spaced targets along the range direction (i.e., the direction aligned with the antenna's boresight) is referred to as *range resolution* (δ_{rg}). Two targets are resolvable when the reflected signal from the first target ends just as the return from the second target begins, leading to the following expression:

$$\delta_{\text{rg}} = \frac{c \cdot \tau}{2} \tag{1.2}$$

where τ is the pulse duration. Although reducing the pulse duration improves range resolution, it also decreases the pulse energy, leading to a lower Signal-to-Noise Ratio (SNR) at the receiver, which must remain sufficiently high to ensure reliable echo detection.

Pulse compression techniques address this intrinsic trade-off by introducing a *frequency-modulated* transmitted signal, enabling longer pulse durations to improve the SNR while maintaining a constant range resolution. The resulting signal is expressed as:

$$s(t) = e^{i2\pi\left(f_0 t + \kappa_r t^2\right)} \cdot \Pi\left(\frac{t - \frac{\tau}{2}}{\tau}\right) \tag{1.3}$$

where $\kappa_r = \frac{\Delta f}{\tau}$ is the chirp rate and Δf is the total frequency variation over the pulse duration, which corresponds to the signal bandwidth. The rectangular function $\Pi(\cdot)$ defines the temporal extent of the pulse.

Through pulse compression, the achievable range resolution is given by:

$$\delta_{\text{rg}} = \frac{c}{2B} \tag{1.4}$$

demonstrating that wider signal bandwidths lead to finer range resolutions.

1.1.2. Backscatter

When an electromagnetic wave propagates through a medium and encounters an object, it induces oscillating surface currents whose behaviour depends on the object's material composition, geometry and orientation. These currents generate a secondary electromagnetic field, which, as governed by Maxwell's equations, redistributes the incident energy outward with different amplitude and direction relative to the incident wave. This interaction is known as *scattering*, while the portion of energy redirected towards the radar is referred to as *backscatter*.

Point and Distributed Scatterers

A *point scatterer* represents an idealised target whose physical dimensions are significantly smaller than the radar system's angular resolution. Thus, it is modelled as a single-point source of electromagnetic energy, simplifying the analysis of its scattering behaviour.

The interaction between the transmitted electromagnetic wave and the scatterer can be described using fundamental principles of electromagnetic theory. In a bistatic radar configuration, where the transmitting and receiving antennas are spatially separated, and under the assumption of a non-dispersive medium, the power received at the receiver antenna (P_r) depends on multiple parameters: the transmitted power (P_t), the distances from the scatterer to the transmitting and receiving antennas (R_t and R_r, respectively), the wavelength (λ) of the radar signal and the characteristics of both the antennas and the target.

The antenna gain (G) quantifies the effective redistribution of radiated power in space and is given by:

$$G = \frac{4\pi A_\mathrm{eff}}{\lambda^2} \tag{1.5}$$

where $A_\mathrm{eff} = A \cdot K_\mathrm{a}$ denotes the effective aperture, with A representing the geometric aperture and K_a the antenna efficiency.

The Radar Cross-Section (RCS) characterises the scatterer's ability to reflect the incident energy back toward the radar. It is defined as:

$$\sigma = \lim_{\mathrm{R_r}\to\infty} \left(4\pi {\mathrm{R_r}}^2 \frac{S_\mathrm{scatt}}{S_\mathrm{inc}} \right) \tag{1.6}$$

where S_inc is the power density of the incident wave at the scatterer and S_scatt is the power density of the scattered wave at the receiver.

These parameters can then be combined into the *point target bistatic radar equation*:

$$P_\mathrm{r} = \frac{P_\mathrm{t} G_\mathrm{t} G_\mathrm{r} \lambda^2}{(4\pi)^3 \mathrm{R_t^2 R_r^2}} \sigma \tag{1.7}$$

For monostatic radar setup, where the transmitter and receiver are co-located, the equation simplifies due to $\mathrm{R_t} = \mathrm{R_r} = \mathrm{R}$ and $G_\mathrm{t} = G_\mathrm{r} = G$.

In real scenarios, the illuminated area consists of numerous scattering elements within a single resolution cell, collectively forming what is referred to as a *distributed scatterer*. As a result, the received signal is the coherent sum of echoes from multiple unresolved

point scatterers [18]. Assuming no predominant scattering mechanism and considering a monostatic acquisition setup, Eq. (1.7) can be generalised to describe the distributed scattering case as:

$$P_{\mathrm{r}} = \iint_{A_\sigma} P_{\mathrm{t}} \frac{G^2}{(4\pi)^3 \mathrm{R}^4} \sigma^0 \, dA_\sigma \tag{1.8}$$

where A_σ is the illuminated surface area and σ^0 is the *backscattering coefficient*, defined as:

$$\sigma^0 = \frac{\sigma}{A_\sigma} \tag{1.9}$$

This coefficient provides a normalised measure of the scattering strength that is independent of the illuminated area.

1.2. Principles of SAR and InSAR

A SAR is a *side-looking* radar system widely used in remote sensing, typically mounted on air- or spaceborne platforms. By taking advantage of the relative motion between the platform and the targets, it synthesises a larger effective antenna aperture compared to a *Real Aperture Radar* (RAR), enabling high-resolution imaging.

1.2.1. Acquisition Geometry

The SAR acquisition geometry, illustrated in Figure 1.1, is defined by three primary dimensions: the *along-track* or *azimuth* direction, corresponding to the platform's movement direction; the *across-track* or *slant-range* direction, aligned with the radar's line of sight; and the *ground-range*, which is the projection of the slant-range onto the Earth's surface. The vertical dimension denotes the altitude of the radar platform.

The *side-looking* geometry is essential for resolving *left-right ambiguities* found in nadir-looking systems.

1.2.2. SAR Resolution

Ground-Range Resolution The range resolution defined for a chirp signal in Eq. (1.4), refers to the SAR *slant-range resolution* (δ_{sr}). The *ground-range resolution* (δ_{gr}) can be derived from the trigonometric relationships inherent to the acquisition geometry (see Section 1.2.1) as:

$$\delta_{\mathrm{gr}} = \frac{c}{2B \sin(\theta_{\mathrm{inc}})} \tag{1.10}$$

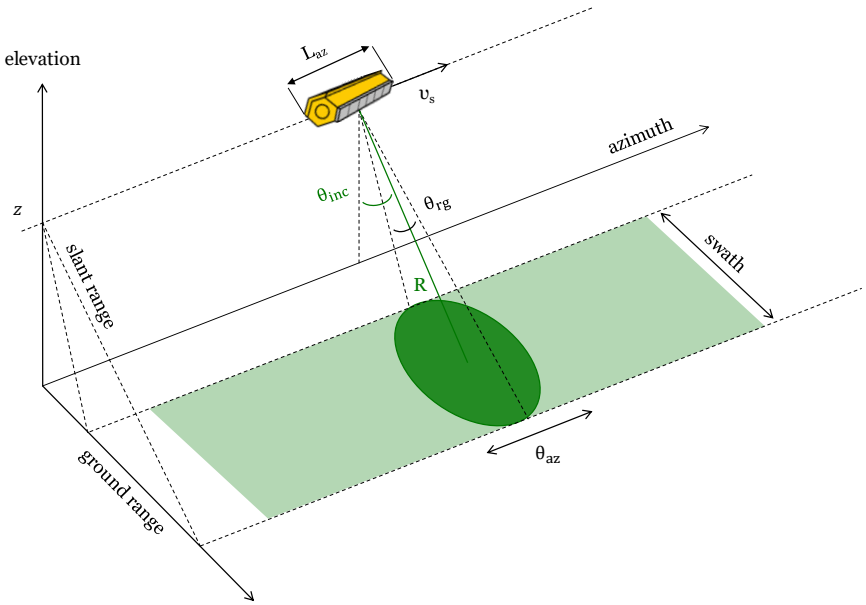


Figure 1.1: Illustration of the SAR acquisition geometry, showcasing a platform positioned at an elevation z, travelling at speed v_s in the azimuthal direction, and observing a scene at a distance R from the platform. The dark green ellipse represents the antenna footprint.

where θ_{inc} is the incidence angle with respect to the ground, which accounts for the looking angle relative to the reference plane (e.g., the ellipsoid) and the local slope of the terrain.

Azimuth Resolution The azimuth resolution (δ_{az}) defines the capability to resolve between two targets close to each other along the azimuth direction. In RAR systems, δ_{az} is constrained by the physical beamwidth (θ_{az}) of the radar antenna, and it is defined as:

$$\delta_{\text{az}} = \text{R}\theta_{\text{az}} = \text{R}\frac{\lambda}{\text{L}_{\text{az}}} \tag{1.11}$$

where λ represents the radar wavelength, L_{az} is the physical length of the antenna and R is the target-antenna distance. Two targets can only be resolved in azimuth if they do not fall simultaneously within the antenna footprint. Consequently, azimuth resolution degrades as the distance from the radar increases, making RARs unsuitable for high-resolution spaceborne imaging.

SAR systems overcome this limitation by exploiting the relative motion between the radar platform and the targets, synthesising a significantly larger effective aperture through the coherent processing of Doppler frequency variations in the received echoes. Targets within the beamwidth exhibit unique *Doppler shifts* depending on their position along

the azimuth direction, enabling SAR to distinguish between closely spaced targets. By integrating these Doppler variations over the synthetic aperture, SAR systems achieves a significantly finer azimuth resolution than RAR systems.

The azimuth resolution of a SAR system is defined as:

$$\delta_{\mathrm{az}} = \frac{\mathrm{L_{az}}}{2} \tag{1.12}$$

with the notable consequence that it no longer depends on the range but only on the antenna's physical dimension.

Image Processing

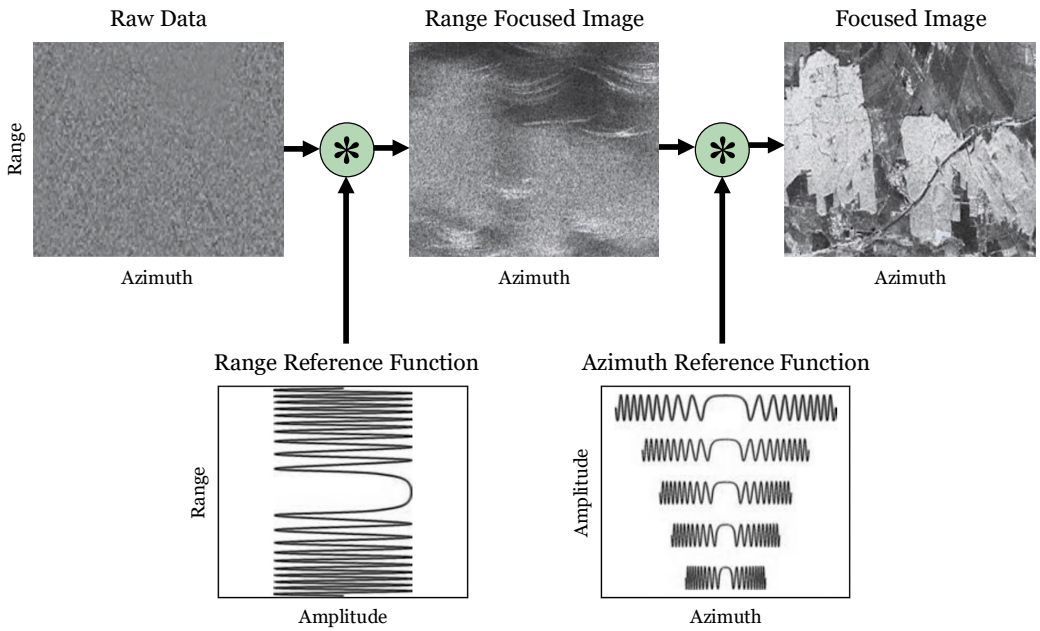


Figure 1.2: Schematic representation of the SAR focusing process [9], where the star symbol denotes the convolution operation.

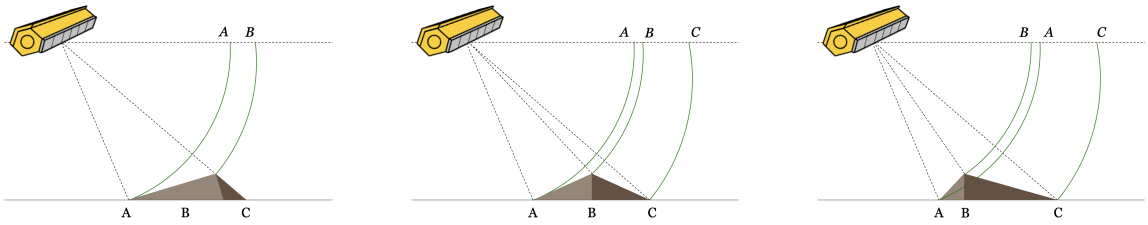
During data acquisition, the radar cyclically transmits and receives pulses, storing the digitised echoes as a two-dimensional data array, typically referred to as *raw data*. The raw data represents the superposition of smeared echoes from individual targets, and additional processing is required to refocus each target's energy into its corresponding resolution cell along both the slant-range and azimuth dimensions. Figure 1.2 schematically illustrates the SAR focusing process.

In the slant-range dimension, focusing is achieved through *range compression* or *focusing*,

which involves applying a matched filter to the received signal. This filtering operation correlates the received echo with a replica of the transmitted chirp, compressing the dispersed signal into a sharp response, as described in [19]. For computational efficiency, range compression is typically performed in the frequency domain and applied independently to each column of the raw data, under the assumption that azimuth *cell migration effects* are negligible [18].

Azimuth focusing, analogous to range focusing, is accomplished by multiplying each azimuth line in the frequency domain by the complex conjugate of the azimuth chirp spectrum. This spectrum, unique to a given slant range, ensures that the target's energy is properly focused at its correct azimuthal position.

Prospective Deformation

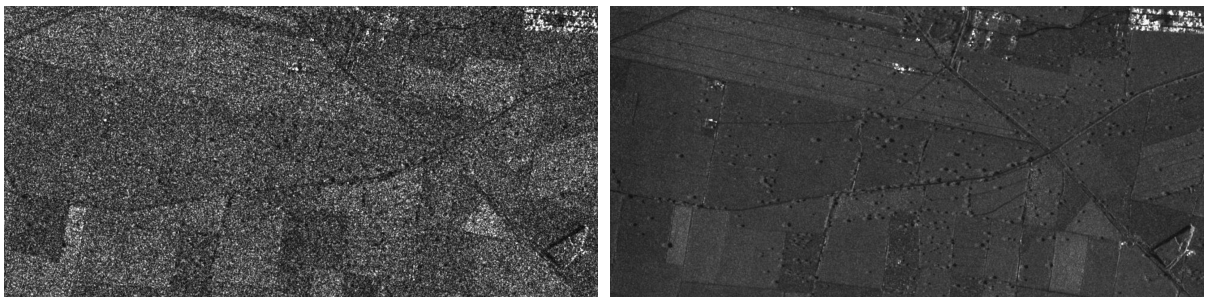


(a) (b) (c)

Figure 1.3: Illustration of SAR images' geometric distortions: (a) shadow, (b) foreshortening, and (c) layover.

The combination of regular distance sampling in the slant range and the side-looking acquisition geometry introduces variability in ground sampling, which follows Eq. (1.4). This variability results in significant geometric distortions, particularly in areas with pronounced topography, manifesting as *shadows*, *foreshortening* and *layover*.

Foreshortening occurs when a slope faces the sensor, compressing multiple targets within a single-resolution cell and resulting in a brighter return. When the slope inclination (α) exceeds the sensor's incident angle (θ_{inc}), a phenomenon called layover occurs. This makes the top of the slope look closer to the sensor than the bottom, inverting the relative spatial order of the scene. Additionally, a slope that faces away from the sensor and has an inclination steeper than θ_{inc} falls into shadow. This area appears as a dark region due to the lack of backscatter.



(a) (b)

Figure 1.4: A SAR image of an agricultural area: (a) illustrates the characteristic speckle associated with distributed targets, while (b) depicts the same area following applying a temporal multi-looking filter to a set of 32 images, effectively reducing speckle noise.

Speckle

The concept of *speckle* refers to the intensity fluctuations observed between adjacent pixels in a SAR image, even though the underlying scene exhibits uniform characteristics. This phenomenon arises from the coherent summation of backscattered waves from multiple unresolved point scatterers within each resolution cell, producing varying interference patterns across the scene. Although deterministic and theoretically repeatable, these patterns are highly sensitive to minor variations in the geometrical arrangement of the scatterers. An example of this effect is illustrated in Figure 1.4.

To mitigate the effects of speckle, *multilooking* techniques are employed, incoherently averaging multiple realisations sampled from the same distribution to reduce the noise variance by a factor of $1/\sqrt{\mathrm{N_L}}$, where $\mathrm{N_L}$ is the *number of looks*. Multilooking can be implemented through spatial averaging of neighbouring pixels within the same image or temporal averaging across repeated acquisitions of the same area, with the latter preserving spatial resolution.

Standard Acquisition Modes

SAR systems employ several acquisition modes to balance resolution and coverage, catering to various imaging requirements. The conventional *StripMap* mode operates by continuously transmitting radar pulses with a fixed antenna orientation, achieving swath widths ranging from 30 to 100 kilometres. While suitable for standard imaging applications, its resolution is limited by the fixed beam geometry.

To enhance resolution, the *Spotlight* mode steers the antenna beam in the azimuth direction, thereby increasing the synthetic aperture and capturing finer details at the cost of

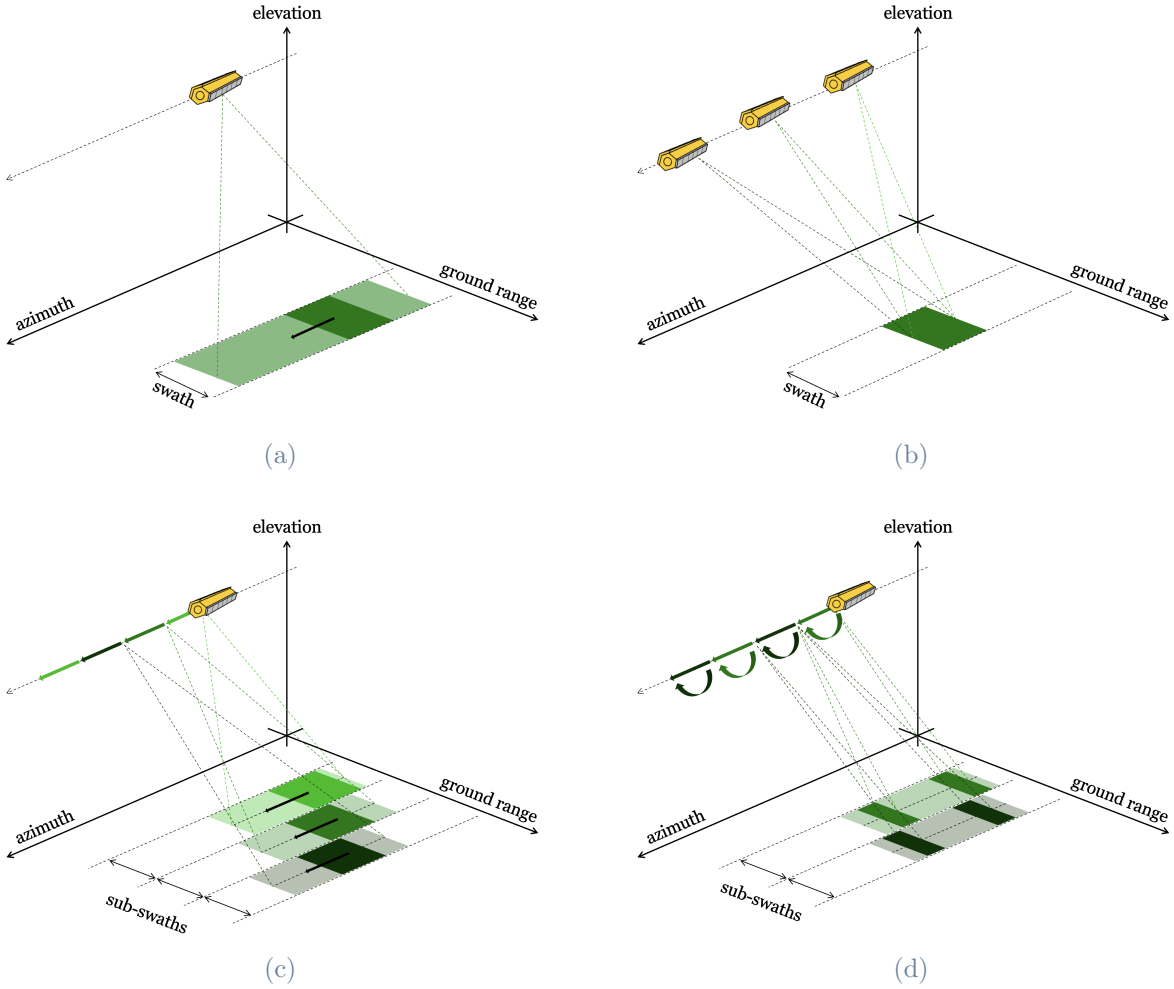


Figure 1.5: Illustration of SAR acquisition modes: (a) Stripmap, (b) Spotlight, (c) ScanSAR, and (d) TopSAR.

reduced coverage, typically limited to approximately 15 kilometres. For broader swath coverage, burst modes such as *ScanSAR* and *TopSAR* are employed. ScanSAR alternates illumination between multiple parallel sub-swaths, enabling swath widths of up to 500 kilometres; however, this comes at the expense of resolution and SNR uniformity. In contrast, TopSAR mitigates these limitations by employing azimuth beam steering, which enhances both resolution and overall image quality. These acquisition modes are illustrated in Figure 1.5.

1.2.3. SAR Interferometry

InSAR is a remote sensing technique designed to extract topographic information by combining radar images of the same area acquired from different sensor positions — spatially,

temporally, or both. In single-pass interferometry, data acquisition co-occurs, whereas in repeat-pass interferometry, images are captured at different times. Simultaneous acquisitions can be classified into: monostatic configurations, where each sensor independently transmits and receives, and bistatic configurations, where one sensor transmits while both act as receivers.

The signals received by two distinct single-channel SAR measurements can be mathematically represented as:

$$s_1 = A_1 e^{-i\frac{4\pi}{\lambda}\mathrm{R}_1+\phi_{\mathrm{obj},1}}, \quad s_2 = A_2 e^{-i\frac{4\pi}{\lambda}\mathrm{R}_2+\phi_{\mathrm{obj},2}} \tag{1.13}$$

where A_1 and A_2 are the received signal amplitudes, $\frac{4\pi}{\lambda}\mathrm{R}_{1,2}$ represents the round-trip propagation phase, and $\phi_{\mathrm{obj},1,2}$ captures the scattering properties of the targets.

Assuming $\phi_{\mathrm{obj},1} = \phi_{\mathrm{obj},2}$ and defining $\Delta\mathrm{R} := |\mathrm{R}_1 - \mathrm{R}_2|$, the scattering phase terms cancel out when s_1 is multiplied by the complex conjugate of s_2:

$$s_1 s_2^* = A e^{-i\frac{2m\pi}{\lambda}\Delta\mathrm{R}} \tag{1.14}$$

where $m = 1$ for a bistatic configuration and $m = 2$ for a monostatic setup. The resulting interferometric phase is defined as:

$$\phi_{\mathrm{if}} = \angle s_1 s_2^* = -\frac{2m\pi}{\lambda}\Delta\mathrm{R} + 2\pi k, \quad k \in \mathbb{Z} \tag{1.15}$$

where the term $2\pi k$ accounts for the phase ambiguity, leading to the *phase unwrapping problem* [20].

InSAR techniques can be broadly classified into three main types: *across-track*, *along-track* and *differential* interferometry [20]. This classification is based on the acquisition geometry and the orientation of the separation between imaging points, referred to as the *interferometric baseline*. This thesis focuses exclusively on across-track interferometry, whose acquisition geometry is illustrated in Figure 1.6, from which the baseline's perpendicular component ($B_\perp$) — defined in the plane normal to the flight direction (i.e., perpendicular to azimuth) — can be derived.

A key parameter for assessing the quality of a SAR interferometer is its vertical resolution, expressed through the height of ambiguity (h_{amb}), which represents the height difference corresponding to a full 2π phase cycle. It is defined as [21]:

$$h_{\mathrm{amb}} = \frac{2\pi}{\kappa_z} \tag{1.16}$$

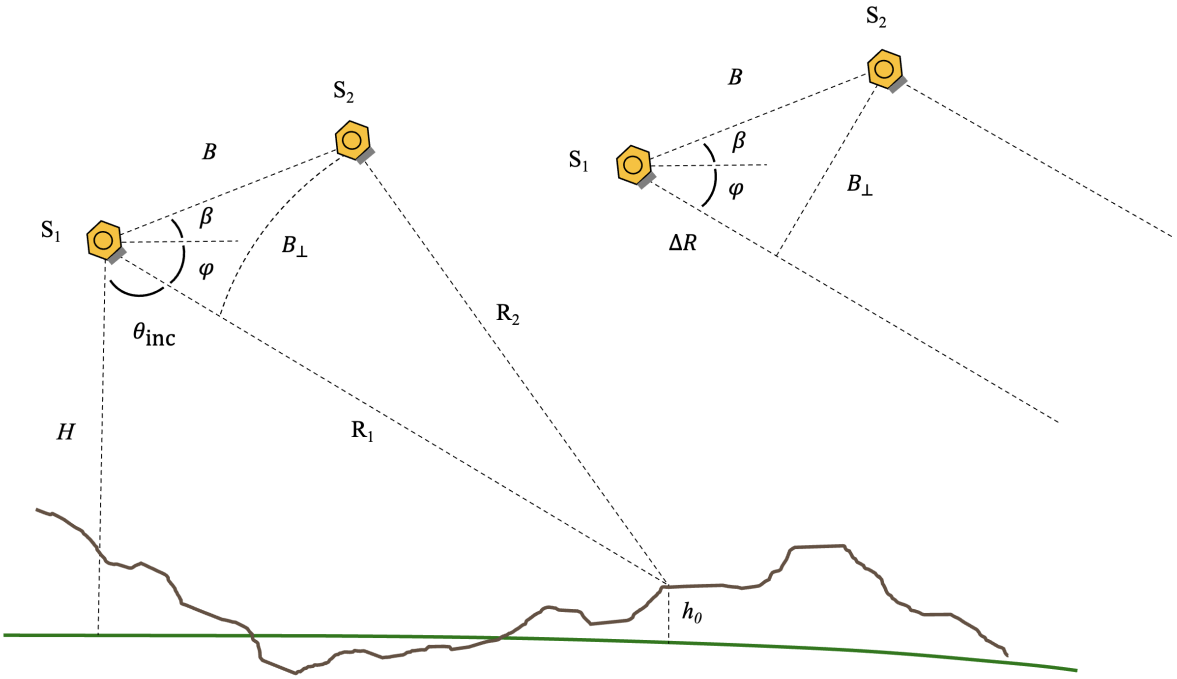


Figure 1.6: Schematic of the across-track SAR interferometry acquisition geometry. The top-right view illustrates the simplified version under the *far-field approximation*, where $R_1 - R_2 \ll R_2$. The green line represents the shape of the ellipsoid, while the brown is the topography.

where κ_z is the vertical wave number. Substituting the expression for κ_z, the height of ambiguity can be rewritten as:

$$h_{\text{amb}} \approx \frac{\lambda R_1 \sin\theta_{\text{inc}}}{mB_\perp} \tag{1.17}$$

where λ is the radar wavelength, R_1 is the slant-range distance, $B_\perp$ is the perpendicular baseline, and m depends on the system configuration.

This formulation highlights that h_{amb} is inversely proportional to $B_\perp$, meaning that larger perpendicular baselines enhance vertical sensitivity and vice versa. However, this improvement is limited by the *critical baseline*, which defines the maximum allowable perpendicular separation between the sensors. Beyond the critical baseline, interferometry becomes infeasible because the spectral overlap of the received signals is lost.

The Interferogram

The *interferogram* is a two-dimensional image depicting the interferometric phase which arises from a pair of SAR images, commonly referred to as *master* and *slave* images.

As shown in [20], the measured interferometric phase consists of three components:

$$\phi_{\text{if}} = \phi_{\text{topo}} + \phi_{\text{flat}} + \phi_{\text{noise}} \tag{1.18}$$

where ϕ_{topo} is the *topographic phase*, encoding the target's elevation information, and ϕ_{flat} is the *flat-Earth phase*, accounting for the interferometric phase increase along the slant-range direction due to the side-looking geometry. The latter can be estimated and compensated for, thereby isolating the topographic information. The term ϕ_{noise} accounts for phase disturbances due to system noise, decorrelation and atmospheric effects.

The subsequent step involves resolving the phase ambiguity, as described in Eq. (1.15). This is achieved through *phase unwrapping algorithms*, which estimate the absolute phase term (ϕ_{abs}) from the wrapped interferometric phase (ϕ_{if}) by adding to each fringe the correct 2π multiple — such a process is called *phase unwrapping* [22].

The unwrapped phase estimate (ϕ_{abs}) can subsequently be employed to generate the final *Digital Elevation Model* (DEM), representing the vertical elevation coordinate for each resolution cell relative to a defined reference height [23].

The Interferometric Coherence

Up to this point, it was assumed that the scattering phase terms $\phi_{\text{obj},1}$ and $\phi_{\text{obj},2}$ in Eq. (1.13) perfectly represent the same process, making the interferometric phase dependent solely on the topography. In reality, this assumption does not necessarily hold, introducing noise into the measured interferogram. Additionally, non-idealities in the acquisition process contribute further to the noise.

The interferometric coherence (γ_{tot}) is the key parameter for measuring the quality of an interferogram [24]. It represents the correlation coefficient between the two complex SAR images s_1 and s_2 and is given by:

$$\gamma_{tot} = \frac{E[s_1 s_2^*]}{\sqrt{E[|s_1|^2]E[|s_2|^2]}} \tag{1.19}$$

where the phase of the complex coherence $\angle\gamma_{\text{tot}} = \phi_{\text{if}}$ is the interferometric phase, while the magnitude $|\gamma_{\text{tot}}|$ is the degree of coherence and it quantifies the level of noise in the

interferogram.

Under the assumption of stationary and ergodic processes, γ_{tot} is typically approximated using the sample coherence $\hat{\gamma}_{\text{tot}}$ [25], which is derived from local statistical properties of the signal:

$$\hat{\gamma}_{\text{tot}}(i,j) = \frac{\sum_W s_1(i,j)s_2^*(i,j)}{\sqrt{\sum_W |s_1(i,j)|^2 \sum_W |s_2(i,j)|^2}} \tag{1.20}$$

As presented in [26], γ_{tot} can be theoretically decomposed into different contributions as:

$$\gamma_{\text{tot}} = \gamma_{\text{SNR}}\,\gamma_{\text{quant}}\,\gamma_{\text{amb}}\,\gamma_{\text{bl}}\,\gamma_{\text{az}}\,\gamma_{\text{vol}}\,\gamma_{\text{temp}} \tag{1.21}$$

where the different terms on the right-hand side of Eq. (1.21) identify the correlation factors due to limited SNR (γ_{SNR}), quantization (γ_{quant}), ambiguities (γ_{amb}), baseline decorrelation (γ_{bl}), relative shift of the Doppler spectra (γ_{az}), volume decorrelation (γ_{vol}) and temporal decorrelation (γ_{temp}).

The individual factors can be estimated, allowing for their compensation in the total coherence [27].

Volume Decorrelation

Of particular interest to this work is the derivation of the *volumetric decorrelation coefficient*, which results from wave penetration into a volumetric target. The orthogonal baseline separation causes variations in the coherent summation of scatterers' contributions, an effect commonly observed in structures such as sand, snow, ice and vegetation.

Assuming that the variation of scatterers' power in the vertical direction (z) can be described by a generic vertical reflectivity function $F(z)$, the received interferometric signal can be expressed as [21]:

$$s_1 s_2^* = \int_{z_0}^{z_0+h_v} F(z)e^{j\kappa_z z}\,dz = e^{j\beta_z z_0}\int_0^{h_v} F(z',\boldsymbol{w})e^{j\kappa_z z'}\,dz' \tag{1.22}$$

where z_0 represents the ground elevation, h_v is the height of the volumetric scattering profile and $z' = z - z_0$.

From the literature [21], the volumetric decorrelation coefficient can be modelled as:

$$\hat{\gamma}_{\text{vol}}(\boldsymbol{w}) = e^{i\kappa_z z_0}\frac{\int_0^{h_v} F(z',\boldsymbol{w})e^{i\kappa_z z'}\,dz'}{\int_0^{h_v} F(z',\boldsymbol{w})\,dz'} \tag{1.23}$$

Here, the volumetric coherence is directly related to the height of the vertical scattering profile. By compensating for other decorrelation factors, as described in Eq. (1.21), the volume height can be estimated by inverting Eq. (1.23).

2 | Deep Learning for Regression Tasks

Regression is a statistical technique that models the relationship between continuous-valued outputs and real-valued input observations. Linear regression techniques assume a linear relationship between input and output data. This assumption does not typically hold in real-case scenarios; therefore, *non-linear* regression methods are employed, leveraging transformation functions to map inputs into alternative spaces where the underlying relationship can be more easily approximated. However, selecting the appropriate transformation is non-trivial and problem-dependent.

Deep learning automates the mapping process — referred to as *feature extraction* — by recursively applying parametrised transformations to progressively capture more complex patterns in the data, leveraging them for the regression task within a unified framework. This recursive formulation lies at the core of *feedforward neural networks*, which act as adaptive feature extractors by composing non-linear mappings to partition the input space and piecewise approximate complex functions.

This chapter presents the principles of *supervised deep learning* for regression tasks, providing the conceptual groundwork for the methodologies introduced in this dissertation. Connections to classical estimation theory are discussed whenever relevant. Further insights into these topics can be found in [28] and [29].

2.1. Feedforward Neural Networks

2.1.1. Inspired by the Brain

Artificial neural networks date back to the 1940s when Warren McCulloch and Walter Pitts introduced a mathematical framework for modelling the brain as a computational system of neural nets [30], which paved the way for the development of artificial neurons, culminating in the Perceptron, which was proposed by Frank Rosenblatt in 1957 as an

early attempt to replicate biological information processing.

Biological neurons integrate signals received through the *dendrites* and generate an output when the cumulative input exceeds a defined threshold, triggering an *action potential*. Drawing inspiration from biology, the Perceptron employed a weighted sum of inputs and a threshold function to produce a binary output. Mathematically, this can be expressed as:

$$a(\boldsymbol{x}_i, \boldsymbol{\theta}) = H(f(\boldsymbol{x}_i, \boldsymbol{\theta})) = H(\boldsymbol{x}_i^T \boldsymbol{\theta} + b) \tag{2.1}$$

with a being the activation and $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ the pre-activation, where $\boldsymbol{x}_i$ represents the input feature vector, $\boldsymbol{\theta}$ the parameter vector, and b an additive bias term. $H(\alpha)$ is the Heaviside step function, defined as:

$$H(\alpha) = \begin{cases} 1, & \text{if } \alpha \geq 0 \\ 0, & \text{elsewhere} \end{cases} \tag{2.2}$$

The introduction of a learning rule allowed the Perceptron to iteratively update its parameters — known as *weights* — based on labelled training data, marking a step towards supervised learning. However, its reliance on linear separability restricts its applicability to problems that can be solved with a linear decision boundary, making it unsuitable for tasks requiring the modelling of non-linearly separable data [31].

The limitations of single-layer Perceptrons were later addressed by Multi-Layer Perceptrons (MLPs), which stacked multiple layers of neurons and employed non-linear activation functions, thereby expanding the representational capacity of artificial networks. These architectures belong to the class of feedforward neural networks, in which data propagates sequentially from input to output layers through intermediate hidden ones. The computation performed by each neuron involves a weighted summation of its inputs, an additive bias term, and the application of a non-linear activation function:

$$a_{p,l,i} \;=\; \Phi\left(\sum_{j=1}^{m_{l-1}} w_{p,l,j}\; \boldsymbol{a}_{j,l-1,i} + \beta_{p,l} \right) \tag{2.3}$$

where (p, l) denotes the p-th neuron in the l-th layer, with $w_{p,l,j}$ and $\beta_{p,l}$ representing its the weights and bias, respectively, and Φ denoting the activation function.

This hierarchical structure enables deep architectures to progressively extract abstract features, where lower-level patterns are combined into higher-order representations, enhancing the expressive power of the networks. The *Universal Approximation Theorem* formally establishes that, given a sufficient number of hidden units and appropriate activation functions, MLPs can approximate any continuous function with arbitrary precision

[32], providing a strong theoretical foundation for developing modern deep learning frameworks.

2.1.2. The Backpropagation Algorithm

Scientific progress is often driven by breakthroughs that bridge theories with real-world applications, and the 2024 Nobel Prize in Physics recognised such a contribution in Artificial Intelligence (AI). A central challenge in developing AI systems is enabling them to learn from data by refining their weights over time, which is achieved using gradient-based techniques that adjust the weights based on a cost function's change with respect to each parameter. Early neural network research, however, was constrained by the lack of an efficient method for computing these gradients in deep architectures. The introduction of the *backpropagation algorithm* in 1986 [33] enabled the systematic computation of gradients by leveraging the network structure, reigniting interest in neural network research.

To ensure its applicability, two conditions must be met. First, the computational operations must be represented as a *Directed Acyclic Graph* (DAG), ensuring that dependencies are non-cyclic and well-defined. Second, each function within the network must be differentiable, allowing for the application of the chain rule of calculus:

$$\frac{d}{dx} f(u(x)) = \frac{d}{dx} \frac{df(u)}{du} \tag{2.4}$$

The algorithm proceeds in two main stages: the *forward pass* evaluates the network's output by sequentially applying transformations to the input data, storing intermediate values needed for later use; the *backward pass* propagates error backwards from the last layer towards the first one, using the stored values to compute gradients efficiently. This mechanism significantly reduces computational redundancy by avoiding repeated evaluations of the same expressions, making training deep neural networks computationally feasible.

2.1.3. Activation Function

As discussed in Section 2.1.1, the activation mechanism shapes the model's representational capacity. Early neural network research historically relied on the Heaviside step function, which enforces a binary threshold-based activation. However, its non-differentiability represents a fundamental limitation for modern neural network training, as it prevents the update rule from leveraging first- or higher-order derivatives.

The introduction of the backpropagation algorithm forced the adoption of differentiable

activation functions, such as the *sigmoid* and *hyperbolic tangent* functions, formally defined as follows:

$$\sigma(\alpha) = \frac{1}{1 + e^{-\alpha}} \tag{2.5}$$

$$\tanh(\alpha) = \frac{e^{2\alpha} - 1}{e^{2\alpha} + 1} \tag{2.6}$$

However, both functions saturate for extremely low or high input values, causing their derivatives to approach zero. This phenomenon, known as the *vanishing gradient problem*, severely restricts weight updates, ultimately slowing down or stopping the training process.

Modern networks have largely transitioned to the *Rectified Linear Unit* (ReLU) [34] [35], defined as:

$$\mathrm{ReLU}(\alpha) \;:=\; \max(0, \alpha) \tag{2.7}$$

which introduces a piecewise linear transformation that retains positive values while mapping negative inputs to zero. This simple yet effective formulation mitigates gradient-related issues. Consequently, ReLU has become the standard activation function for hidden layers in contemporary architectures.

2.1.4. Convolutional Neural Networks

MLPs process inputs independently, lacking built-in awareness of spatial relationships. Their fully connected design imposes fixed input dimensions and rapidly increases parameters as the input size rises. In this work, SAR images spanning extensive spatial areas are employed as input data, where pixels exhibit statistical correlation with their neighbours, making MLPs unsuitable.

Convolutional Neural Networks (CNNs) address these limitations by replacing dense layers with convolutional ones, facilitating the extraction of spatial patterns without constraining the input size.

The Convolutional Layer

A convolutional layer processes the input by sliding a small matrix of learnable parameters — known as a *kernel* — across it, computing a weighted sum over each overlapping region. The kernel's parameters, analogous to those in MLPs, function as the network's weights.

When processing multi-channel data, represented as a 3D tensor — typically consisting of two spatial and one channel dimensions — the resulting transformation at a given spatial

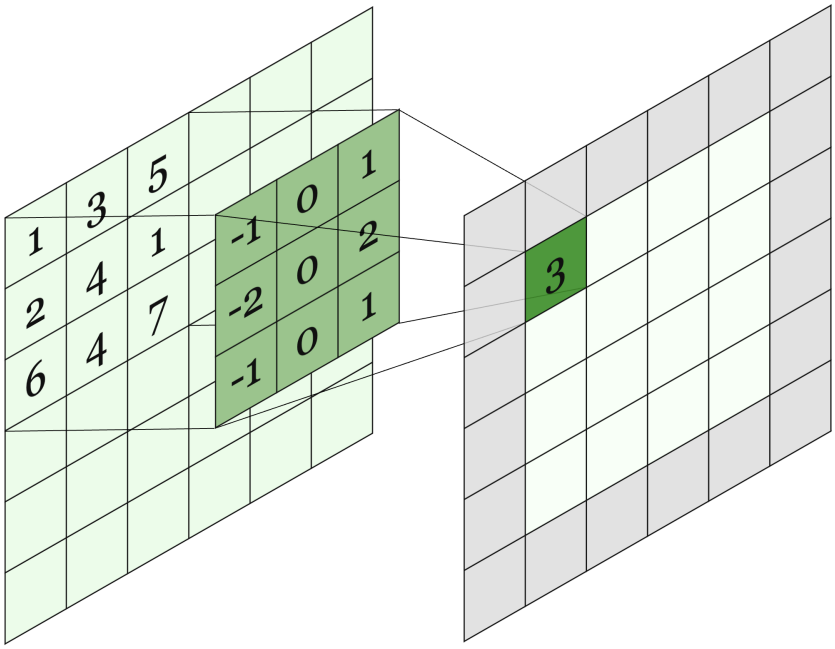


Figure 2.1: The representation depicts the 2D convolution process used in CNNs. The highlighted output value results from applying the high-pass kernel (central matrix) to the input data (left matrix). The greyed-out areas in the output matrix represent pixels where convolution cannot be computed due to a lack of surrounding information.

coordinate (i, j) is expressed as:

$$y(i,j) = \beta + \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \sum_{c=0}^{C-1} w_{(u,v,c)} \cdot x_{(i+u,j+v,c)} \tag{2.8}$$

where H and W define the kernel size, and C corresponds to the number of input channels. Multiple filters are typically applied in parallel, spanning all the input channels — meaning that each filter comprises $H \cdot W \cdot C$ learnable parameters — determining the output feature map's depth. The kernel size defines how many neighbouring pixels influence each output value: larger kernels capture broader contextual information, while smaller kernels focus on finer local details.

A key aspect of convolutional layers is their effect on spatial dimensions. As the kernel operates within a fixed window, it requires full coverage of the input region, inherently reducing the output size. As illustrated in Figure 2.1, a 3×3 kernel decreases the output dimensions by one pixel along each edge, which becomes more pronounced with larger kernel sizes.

To preserve input dimensions, a common approach involves *padding* the input image —

either with constant values (e.g., zeros) or replicating or mirroring the boundary values — before performing the convolution.

Receptive Field

CNNs expand the spatial region influencing each prediction by progressively increasing the input's spatial context, capturing broader spatial dependencies as information flows through successive layers. This cumulative spatial coverage is formally defined as *receptive field*, in analogy to the receptive fields in biological vision systems.

Depending on the application, a broader spatial context is often desirable — making expanding the receptive field an architectural choice — typically achieved by stacking multiple layers with small kernels, providing a more parameter-efficient alternative to using larger kernels.

2.2. Supervised Learning

While the backpropagation algorithm enables the computation of the gradients of the cost function with respect to each parameter, it does not prescribe how these gradients should be used to navigate the loss landscape during learning. A framing of the broader optimisation problem is therefore required.

The following sections formalise the optimisation problem, introduce the *Stochastic Gradient Descent* (SGD) algorithm, and explore their role in the end-to-end learning process.

2.2.1. The Optimization Problem

The core of *machine learning* is determining m weights based on n available observations. From a mathematical perspective, this procedure requires solving a system of equations with m unknowns. In practice, it is often the case that the number of observations exceeds the number of parameters (i.e., $n > m$), resulting in an *overdetermined* system. As a consequence, an exact solution may not exist, necessitating the estimation of the weights ($\hat{\boldsymbol{\theta}}$) that *best fit* the data, i.e., that minimise an *objective function* ($\mathcal{L}$):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \tag{2.9}$$

The objective function to be minimised is typically referred to as *loss function* or *cost function*, as it quantitatively measures the discrepancy between the model's predictions and the expected outcomes, providing a mathematical tool for guiding the parameters

update.

The solution that strictly satisfies Eq. (2.9) corresponds to the *global optimum*, a parameter configuration that yields the most favourable evaluation of the objective function. However, searching for the global optimum is often unfeasible due to the computational challenges of traversing the high-dimensional parameter spaces. Consequently, the optimisation process typically aims to determine a *local optimum*, denoted by $\boldsymbol{\theta}^*$, such that:

$$\forall \boldsymbol{\theta} \in B,\ \exists \delta > 0 \in \mathbb{R}\ :\ |\boldsymbol{\theta} - \boldsymbol{\theta}^*| < \delta,\ \mathcal{L}(\boldsymbol{\theta}^*) < \mathcal{L}(\boldsymbol{\theta}) \quad \text{with} \quad \boldsymbol{\theta}^* \in B \tag{2.10}$$

where $B \subseteq \mathbb{R}^m$ defines the solution space spanned by the m trainable parameters of the model. Eq. (2.10) states that $\boldsymbol{\theta}^*$ minimises the loss within its local neighbourhood, thereby qualifying it as a *locally optimal solution*.

2.2.2. Cost Function

Maximum Likelihood Estimation

In regression tasks, the *Mean Squared Error* (MSE) is a widely employed cost function due to its probabilistic foundation. From a statistical perspective, minimising the MSE follows naturally from *Maximum Likelihood Estimation* (MLE), which emerges as a parameter estimation method in scenarios where uncertainty — inherent from the problem and its modelling — makes achieving perfect prediction unfeasible.

Formally, given a dataset consisting of n observations, denoted as $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, the MLE objective is to estimate the model parameters by maximising the likelihood of the observed data given the model's weights ($\boldsymbol{\theta}$):

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta}) \tag{2.11}$$

Assuming that the observations are *independent and identically distributed* and that the output variable follows a *Gaussian distribution*, the likelihood function factorises as:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{n} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \right\} \quad \text{with} \quad p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \sim \mathcal{N}(y_i \mid f(\boldsymbol{x}_i, \boldsymbol{\theta}), \sigma^2) \tag{2.12}$$

To simplify the mathematical formulation, the problem is often expressed in the log space,

leading to the minimisation of the Negative Log-Likelihood (NLL):

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \left\{ -\sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \right\} \tag{2.13}$$

which under the assumption of *homoscedasticity* (i.e., $\sigma^2 = const.$), simplifies to:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 \right\} \tag{2.14}$$

Since the first term is independent of $\boldsymbol{\theta}$, and the constant multiplicative factor in the second term does not affect the optimisation process, they can be omitted, leading to:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 \right\} \tag{2.15}$$

This result shows that — under the assumption of a homoscedastic Gaussian output distribution — MLE results in the minimisation of the MSE, providing a probabilistic foundation for using MSE-based loss function in regression problems.

Regularisation Techniques

In overdetermined systems, specific parameters' configurations can make the optimisation process unstable — leading to solutions that closely fit the training data — weakening generalisation and ultimately reducing the model's performance on unseen data, a phenomenon known as *overfitting*.

Regularisation techniques reshape the solution landscape, imposing constraints on the parameter space to counteract this issue.

L2-Norm The $\mathcal{L}_2$-norm constrains the magnitude of the model parameters. From a probabilistic perspective, this corresponds to imposing a *zero-mean Gaussian prior* on the weights — $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, \sigma_\theta^2)$ — which naturally leads to a *Maximum a Posteriori* (MAP) estimate of the weights given the observed data:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left\{ \prod_{i=1}^{n} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \right\} \tag{2.16}$$

which on the logarithm-space yields the equivalent minimisation problem:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 + \frac{1}{2\sigma_{\boldsymbol{\theta}}^2} \sum_j \omega_j^2 \right\} \tag{2.17}$$

The MSE-based cost function, when including $\mathcal{L}_2$-regularisation, is reformulated as:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 + \lambda \sum_j \omega_j^2 \right\} \tag{2.18}$$

where ω_j denotes the j-th model parameter, and $\lambda > 0$ controls the regularisation strength, which is inversely related to the variance of the prior on the weights: $\lambda = \frac{1}{2\sigma_{\boldsymbol{\theta}}^2}$. When bias terms are excluded from regularisation, this technique is also referred to as *weight decay*.

Stochastic Gradient Descent

Given a differentiable objective function, *Gradient Descent* (GD), a class of *first-order optimisation methods*, iteratively updates the parameters by following the direction of steepest descent, as determined by the local gradient of the objective function. The process starts from an initial parameter set and, at the k-th iteration, the update rule is given by:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + r_k \boldsymbol{d}_k \tag{2.19}$$

where r_k represents the *learning rate* or *step size*, and $\boldsymbol{d}_k$ denotes the update direction.

Computing the full gradient of the loss function over an entire dataset in high-dimensional settings quickly becomes computationally impractical. Therefore, a widely adopted solution is to approximate the gradient using a randomly sampled subset of the data — referred to as *minibatch*. This subset, denoted as $\mathcal{D}_k \subset \mathcal{D}$, consists of $n_k \ll n$ samples and is updated at each iteration. The empirical mean derivative at the k-th iteration is then approximated as:

$$\nabla \mathcal{L}(\boldsymbol{\theta}_k) \;=\; \frac{1}{n} \sum_{i \in \mathcal{D}} \nabla_{\boldsymbol{\theta}_k} \mathcal{L}_i(\boldsymbol{\theta}_k) \;\approx\; \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} \nabla_{\boldsymbol{\theta}_k} \mathcal{L}_i(\boldsymbol{\theta}_k) \tag{2.20}$$

This method, known as *Stochastic Gradient Descent* (SGD), balances gradient precision and computational efficiency, as the reduced sample size introduces noise into the estimate. While noisier updates require more iterations to converge, the lower computational cost per iteration often leads to a faster overall optimisation process.

Built upon SGD, the *Adaptive Moment Estimation* (ADAM) optimiser enhances the optimisation efficiency by incorporating *adaptive learning rates* and *momentum*. The update rule in ADAM is defined as:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - r_0 \frac{\hat{\boldsymbol{m}}_k}{\sqrt{\hat{\boldsymbol{s}}_k} + \epsilon} \tag{2.21}$$

where the uncorrected momentum terms — representing exponentially weighted averages of past gradient states — are defined as:

$$\boldsymbol{m}_k = \alpha_1 \boldsymbol{m}_{k-1} + (1 - \alpha_1) \nabla \mathcal{L}(\boldsymbol{\theta}_k) \tag{2.22}$$

$$\boldsymbol{s}_k = \alpha_2 \boldsymbol{s}_{k-1} + (1 - \alpha_2) \nabla \mathcal{L}^2(\boldsymbol{\theta}_k) \tag{2.23}$$

While the learning rate r_0 is fixed, ADAM effectively adapts the step size by scaling it inversely to the accumulated squared gradients ($\sqrt{s_k}$). This ensures that parameters with large gradients receive smaller updates, whereas those with smaller gradients are adjusted with larger steps — stabilising parameter updates and accelerating convergence. When $\alpha_{1,2} = 0$, the update simplifies to a basic SGD step, corresponding to a lack of memory.

2.2.3. Batch Normalization

As briefly discussed in Section 2.1.3, a key challenge in training deep neural networks is ensuring stable gradient propagation. While the vanishing gradient problem leads to a progressive reduction in gradient magnitudes, the opposite issue — *gradient explosion* — can also occur. This phenomenon occurs when gradients grow uncontrollably during backpropagation, leading to training instability and, in extreme cases, numerical overflow. Both issues are tied to network depth, as gradients propagate recursively through multiple layers.

A widely adopted strategy to ensure stable gradient propagation while preserving the network's representational capacity is to incorporate *batch normalisation* layers [36]. During training, mini-batches are first standardised to enforce zero mean and unit variance and then rescaled:

$$\tilde{\boldsymbol{x}} = \boldsymbol{\gamma} \odot \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \epsilon}} + \boldsymbol{\theta} \tag{2.24}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the batch-wise mean and variance, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are the trainable scale

and shift parameters, while $\epsilon > 0$ ensures numerical stability.

During inference, batch-wise statistics are replaced by exponentially weighted moving averages of the mean ($\boldsymbol{\mu}_i$) and variance ($\boldsymbol{\sigma}_i^2$), accumulated during training:

$$\boldsymbol{\mu}_i = \alpha \boldsymbol{\mu}_{i-1} + (1-\alpha)\boldsymbol{\mu} \tag{2.25}$$

$$\boldsymbol{\sigma}_i^2 = \alpha \boldsymbol{\sigma}_{i-1}^2 + (1-\alpha)\boldsymbol{\sigma}^2 \tag{2.26}$$

where α — referred to as *momentum* — controls the past states' influence. This ensures a deterministic behaviour when processing unseen data.

Beyond stabilising gradients, batch normalisation has been shown to facilitate the use of higher learning rates, potentially improving convergence efficiency. Despite its widespread adoption, its necessity across all architectures remains an open research question [37].

2.2.4. Training

The complete training procedure can now be outlined. Given a training data-subset comprising n_{train} samples, the minibatch approach — introduced in Section 2.2.2 — is employed to partition the training cycle into

$$n_{\text{steps}} = \left\lfloor \frac{n_{\text{train}}}{BS} \right\rfloor \tag{2.27}$$

steps, where BS denotes the *batch size* (i.e., the number of samples within each minibatch). An *epoch* is then defined as a single complete pass through n_{steps} batches. At the end of each epoch, the training samples are randomly shuffled to produce a new sequence of minibatches for the next epoch.

The training consists of three iterative steps, repeated until a predefined *convergence criterion* is satisfied or a fixed number of epochs is completed. In the *forward pass*, the model processes input data to generate predictions, retaining intermediate computations for subsequent updates as required by the backpropagation algorithm. The outputs are then compared against the target values through the *cost function*. During the *backpropagation* step, the optimisation algorithm employs the stored computations to update the model parameters by evaluating the gradient of the cost function with respect to each weight.

2.2.5. Validation

During the learning phase, the model performance improves up to a point, after which it degrades. This behaviour is assessed by tracking the loss function on both the training set and an independent *validation set* throughout the process. A typical learning pattern involves an initial phase where both losses decrease monotonically, followed by a point where the validation loss reaches a minimum before stabilising or increasing. Extending optimisation beyond this stage often leads to *overfitting*, while prematurely stopping the weights' update may result in *underfitting*, where the model fails to adequately capture the underlying data structure.

Although theoretically the lowest validation loss may occur at any weight update step, pinpointing this exact moment would require validation after every minibatch processing, which is computationally impractical. As fluctuations in the validation loss gradient tend to diminish near the optimal point, evaluation is conventionally performed at the end of each epoch, balancing computational efficiency with reliable performance monitoring.

Early Stopping Rather than relying on a fixed number of training iterations — potentially leading to underfitting or overfitting — the training process can be governed by the *early stopping* strategy. This approach employs a parameter — known as *patience* — which defines the number of consecutive validation loss evaluations that can occur without improvement before stopping the training. The model parameters corresponding to the lowest recorded validation loss are then saved.

2.2.6. Testing

Assessing the model's deployment performance on unseen data is essential; therefore, a separate *test set*, representative of the full range of conditions the model is expected to handle, is reserved to provide unbiased performance estimates.

In the context of this thesis, eight evaluation metrics are employed. Specifically, the Mean Error (ME), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) — along with their normalised version — and the coefficient of determination (R^2), mathe-

matically defined as follows:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \qquad \text{ME\%} = \frac{\text{ME}}{\bar{y}}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad \text{RMSE\%} = \frac{\text{RMSE}}{\bar{y}}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid \hat{y}_i - y_i \mid \qquad \text{MAE\%} = \frac{\text{MAE}}{\bar{y}}$$

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

where $\hat{y}_i$ denotes the predicted value, y_i represents the corresponding reference value and $\bar{y}$ is the mean of the reference observed values.

Each metric offers complementary insights into model strengths and limitations: the ME detects the bias by measuring the average prediction deviations, while the MAE captures the overall accuracy regardless of error direction; the RMSE, being more sensitive to large errors, emphasises the impact of outliers, whereas the R^2 evaluates the goodness-of-fit by quantifying the proportion of variance the model explains.

2.3. Bayesian Neural Networks

In its most general form, the objective of deep learning methods is to predict a target quantity ($\boldsymbol{y}$) from a given input ($\boldsymbol{x}$) using a parametric function (f), such that $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta})$. The parameters ($\boldsymbol{\theta}$) are learned from a labelled training set $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) : n = 1, \ldots, N\}$, where $\boldsymbol{x}_n \in \mathcal{X} \subseteq \mathbb{R}^D$ and $\boldsymbol{y}_n \in \mathcal{Y} \subseteq \mathbb{R}^C$.

Because of the high-dimensional complexity inherent in both the data and the neural networks used to model the underlying phenomena, assessing the uncertainty associated with the model's predictions emerges as a matter of paramount importance — particularly in scenarios where reliability and interpretability are of critical concern.

In this context, the Bayesian framework provides an intuitive approach for modelling and quantifying uncertainty.

2.3.1. Sources of Uncertainty

A thorough understanding of the sources of uncertainty inherent in predictive modelling is essential for assessing model reliability. Two categories of uncertainty are typically distinguished:

Aleatoric Uncertainty arising from inherent randomness or noise in the data generation process;

Epistemic Uncertainty reflecting limitations of the model itself due to insufficient training data or modelling capacity.

In the following, it is explored how these uncertainties emerge mathematically, laying the foundation for advanced uncertainty quantification methods.

Aleatoric Uncertainty

As discussed in Section 2.2.2, a common assumption in regression tasks is that the target variable is influenced by an intrinsic degree of randomness — referred to as *stochastic uncertainty* — which can be modelled as an additive, zero-mean *noise* contribution with covariance matrix $\Sigma_{\boldsymbol{y}}(\boldsymbol{x})$. Formally, the conditional probability distribution of the output is expressed as $p(\boldsymbol{y} \mid \boldsymbol{x})$, with its mean vector and covariance matrix defined as:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{y}}\big[\boldsymbol{y}\big] &&= \int \boldsymbol{y}\ p(\boldsymbol{y} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{y} \\ \Sigma_{\boldsymbol{y}}(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{y}}\Big[\big(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big)\big(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big)^{\top}\Big] &&= \int \big(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big)\big(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big)^{\top} p(\boldsymbol{y} \mid \boldsymbol{x})\, \mathrm{d}\boldsymbol{y} \end{aligned}$$

A common simplification assumes that the noise is homoscedastic, implying that the covariance does not depend on $\boldsymbol{x}$ and is isotropic across output dimensions. This can be expressed as:

$$\Sigma_{\boldsymbol{y}}(\boldsymbol{x}) = \sigma^2 \mathbf{I}_C \tag{2.28}$$

where σ^2 represents a constant noise variance, and $\mathbf{I}_C$ is the $C \times C$ identity matrix.

Consider a linear regression model $f(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathbb{R}^C$, with parameters $(\boldsymbol{\theta})$ estimated via the least squares criteria. The expected loss for a given input can be written as:

$$\mathbb{E}_{\boldsymbol{y}}[\mathcal{L}(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{y}}\Big[\big\|f(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{y}\big\|^2\Big] \tag{2.29}$$

and can be decomposed into the squared deviation from the expected mean and the

inherent noise:

$$\mathbb{E}_{\boldsymbol{y}}\Big[\big\|f(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{y}\big\|^2\Big] \;=\; \big\|f(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big\|^2 \;+\; \mathrm{tr}\big(\Sigma_{\boldsymbol{y}}(\boldsymbol{x})\big) \tag{2.30}$$

The term $\mathrm{tr}\big(\Sigma_{\boldsymbol{y}}(\boldsymbol{x})\big)$ captures the *aleatoric uncertainty*, or *data uncertainty*, which arises from intrinsic noise in the data generation process. It is irreducible and cannot be mitigated by modifying the model or collecting additional data under the same process.

Epistemic Uncertainty

In most practical scenarios, the full data-generating distribution remains unknown. Instead, the model parameters are estimated from a finite training set ($\mathcal{D}$), considered a random sample from the underlying distribution. To capture this dependency, one may define the expected model output as:

$$\boldsymbol{\mu}_f(\boldsymbol{x}) \;=\; \mathbb{E}_{\mathcal{D}}\Big[f\big(\boldsymbol{x},\boldsymbol{\theta}_{|\mathcal{D}}\big)\Big] \tag{2.31}$$

where $\boldsymbol{\mu}_f(\boldsymbol{x})$ represents the mean prediction over all possible realisations of $\mathcal{D}$.

As a consequence, the expected value of the squared model deviation is defined as:

$$\mathbb{E}_{\mathcal{D}}\Big[\big\|f\big(\boldsymbol{x},\boldsymbol{\theta}_{|\mathcal{D}}\big) \;-\; \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big\|^2\Big] \;=$$

$$= \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\big\|f\big(\boldsymbol{x},\boldsymbol{\theta}_{|\mathcal{D}}\big) \;-\; \boldsymbol{\mu}_f(\boldsymbol{x})\big\|^2\Big]}_{\text{Variance}} + \underbrace{\Big[\big\|\boldsymbol{\mu}_f(\boldsymbol{x}) \;-\; \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big\|^2\Big]}_{\text{Bias}^2} \tag{2.32}$$

Both terms — *variance* and *bias*^{2} — capture distinct aspects of *epistemic uncertainty* or *model uncertainty*. The first term (i.e., the variance) quantifies the fluctuation in the model's predictions across different training sets. This reflects uncertainty due to data lack or incomplete sampling of the input space. The second term (i.e., the bias) measures the systematic error introduced by the model's inability to fully capture the underlying data distribution.

A Unified Framework

The overall uncertainty in the predictions can be understood as the combined effect of three distinct components: *variance*, *bias* and *noise power*. Mathematically, this decomposition is expressed as:

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{\boldsymbol{y}}\big[\mathcal{L}(\boldsymbol{x})\big]\Big] \;=$$

$$= \underbrace{\mathbb{E}_{\mathcal{D}}\Big[\big\|f(\boldsymbol{x}, \boldsymbol{\theta}_{|\mathcal{D}}) - \boldsymbol{\mu}_f(\boldsymbol{x})\big\|^2\Big]}_{\text{Variance}} + \underbrace{\Big[\big\|\boldsymbol{\mu}_f(\boldsymbol{x}) - \boldsymbol{\mu}_{\boldsymbol{y}}(\boldsymbol{x})\big\|^2\Big]}_{\text{Bias}^2} + \underbrace{\Big[\mathrm{tr}\big(\Sigma_{\boldsymbol{y}}(\boldsymbol{x})\big)\Big]}_{\text{Noise Power}} \tag{2.33}$$

Together, these components provide a comprehensive framework for analysing prediction errors, enabling a deeper understanding of both reducible and irreducible sources of uncertainty. In subsequent sections, advanced Bayesian approaches are discussed to explicitly model uncertainty and produce well-calibrated predictions.

2.3.2. Bayesian Model Average

A central focus in Bayesian inference is the derivation of the *posterior predictive distribution* — also referred to as the *marginal predictive distribution*. This distribution defines the probability density function of the output ($\boldsymbol{y}$) conditioned on the input ($\boldsymbol{x}$) and the training dataset ($\mathcal{D}$):

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta} \tag{2.34}$$

where $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ denotes the predictive conditional distribution given a set of parameters ($\boldsymbol{\theta}$), and $p(\boldsymbol{\theta} \mid \mathcal{D})$ represents the posterior distribution over the model parameters.

This expression formalises the process of *marginalisation* over the parameter space, effectively computing a weighted average of predictions across an infinite ensemble of models. Such an approach — commonly referred to as *Bayesian Model Averaging* (BMA) — provides a noble framework to capture the *epistemic uncertainty* associated with the suitability of different parameter configurations in representing the data.

By averaging over the posterior distribution, the BMA mitigates the risk of overfitting, as its predictions do not depend on any single, deterministic parameter estimate but rather reflect the variability induced by uncertainty in the parameters themselves.

Epistemic and Aleatoric Components While the BMA primarily targets *epistemic uncertainty* — reflecting a lack of knowledge about the optimal $\boldsymbol{\theta}$ — it also inherently incorporates the effects of *aleatoric uncertainty* through the predictive conditional distribution, which accounts for intrinsic noise in the data generation process. However, it is important to highlight that the BMA does not explicitly model or reduce the aleatoric component; instead, it treats it as fixed and irreducible for each given $\boldsymbol{\theta}$. Thus, the primary strength of BMA lies in addressing epistemic uncertainty by marginalising over likely configurations of $\boldsymbol{\theta}$.

Monte Carlo Approximation

Ultimately, the goal is to compute the BMA, but for Bayesian neural networks, this integral is not analytically tractable. As a result, it is common to approximate the posterior predictive distribution using a *Monte Carlo* approach:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \;\approx\; \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_m) \quad \text{with} \quad \boldsymbol{\theta}_m \sim q(\boldsymbol{\theta} \mid \mathcal{D}) \tag{2.35}$$

where $\boldsymbol{\theta}_m$ are sampled from an approximate posterior distribution $q(\boldsymbol{\theta} \mid \mathcal{D}) \;\approx\; p(\boldsymbol{\theta} \mid \mathcal{D})$.

Despite its theoretical appeal, the Monte Carlo approximation has limitations, particularly in complex, high-dimensional models such as deep neural networks. It is arguable that reliance on the Monte Carlo perspective could be avoided in favour of alternative strategies.

2.3.3. Plug-in Approximation

From the perspective of BMA, *classical* training solutions can be viewed as special-case approximations of the posterior distribution over model parameters. This approach — referred to as *plug-in approximation* — assumes that the posterior distribution is *degenerate*, collapsing to a single point estimate ($\hat{\boldsymbol{\theta}}$), such as the MLE or the MAP solutions:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \;\approx\; \delta\big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\big) \tag{2.36}$$

where $\hat{\boldsymbol{\theta}}$ represents the estimated parameters and $\delta(\cdot)$ denotes the Dirac delta function.

By substituting the approximation into the posterior predictive distribution, the resulting expression is obtained:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \;\approx\; \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, \delta\big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\big)\, \mathrm{d}\boldsymbol{\theta} \;=\; p(\boldsymbol{y} \mid \boldsymbol{x}, \hat{\boldsymbol{\theta}}) \tag{2.37}$$

where the final equality follows from the *sifting property* of the Dirac delta function.

It becomes evident why any non-degenerate approximation of the posterior distribution over model parameters provides a more expressive characterisation than the plug-in approximation. While such approximations may still struggle to fully capture the multimodality and intricate structure of high-dimensional parameter spaces, they generally offer a more flexible and realistic representation than a single Dirac delta mass centred on a point estimate.

BMA and Classical Approaches The predictions obtained via BMA often differ significantly from those produced by classical training methods, particularly in cases where:

- The posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ is spread across multiple modes rather than being sharply concentrated around a single parameter configuration (e.g., when it is multi-modal);
- The predictive conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ exhibits substantial variation across different parameter settings.

Both of these conditions frequently arise in deep neural networks, where the parameter space tends to be highly non-convex and multi-modal. Consequently, Bayesian approaches such as BMA are particularly appealing in such contexts, as they naturally incorporate parameter uncertainty and provide a more calibrated measure of confidence in predictions.

2.3.4. Stochastic Weight Averaging-Gaussian

Deep neural networks often converge to complex regions of the loss landscape, where multiple parameter configurations yield similarly low training loss. Standard point estimates, such as the final iterate of SGD, fail to capture this variability, thereby neglecting uncertainty in parameter values. *Stochastic Weight Averaging-Gaussian* (SWAG) [38] addresses this limitation by fitting a Gaussian approximation around a running average of parameter snapshots, effectively modelling a local region of the posterior distribution. Building on *Stochastic Weight Averaging* (SWA) [39] — which averages parameter vectors upon reaching a predetermined convergence threshold to improve generalisation — SWAG extends this approach by incorporating a covariance structure to characterise *epistemic uncertainty* within the parameter space.

SWA Precursor The SWA algorithm leads to a smoother solution, often lying in a wider basin of the loss surface, periodically updating a running average of parameters:

$$\boldsymbol{\theta}_{t+1}^{\mathrm{SWA}} = \frac{\boldsymbol{\theta}_t^{\mathrm{SWA}} \cdot n + \boldsymbol{\theta}_{t+1}}{n+1} \tag{2.38}$$

where n counts the number of snapshots that have been incorporated.

SWAG Covariance Approximation Beyond averaging parameters, SWAG approximates a local Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}_{\mathrm{SWA}}, \Sigma)$ by maintaining both a running estimate of the first moment and a structured approximation of the parameter covariance. Instead of modelling the full covariance matrix, which is computationally infeasible for high-

dimensional parameter spaces, SWAG employs a *low-rank* plus *diagonal* approximation to capture variability efficiently.

The diagonal component is defined as:

$$\Sigma_{\text{diag}} \;=\; \text{diag}\big(\overline{\boldsymbol{\theta}^2} \;-\; \boldsymbol{\theta}^2_{\text{SWA}}\big), \quad \overline{\boldsymbol{\theta}^2} \;=\; \frac{1}{T}\sum_{t\,=\,1}^{T}\boldsymbol{\theta}_t^2 \tag{2.39}$$

which estimates the variance of individual parameters across the snapshots. To account for correlations between parameters, SWAG also constructs a low-rank component based on deviations from the mean:

$$\Sigma_{\text{low}-\text{rank}} \;=\; \frac{1}{T-1}\sum_{t\,=\,1}^{T}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{SWA}})(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{SWA}})^{\intercal} \tag{2.40}$$

As the value of $\boldsymbol{\theta}_{\text{SWA}}$ is not available during training, this term is further approximated as:

$$\Sigma_{\text{low}-\text{rank}} \;\approx\; \frac{1}{T-1}\boldsymbol{D}\boldsymbol{D}^{\intercal} \tag{2.41}$$

where $\boldsymbol{D}$ is a matrix whose columns contain deviations from the running mean:

$$\boldsymbol{D} \;=\; \big[\ldots,\, (\boldsymbol{\theta}_t - \overline{\boldsymbol{\theta}_t}),\, \ldots\big]$$

To limit the rank of the estimated covariance matrix, K columns are used corresponding to the last K epochs of training.

The final covariance estimate combines the diagonal and low-rank components:

$$\Sigma \;\approx\; \frac{1}{2}\big(\Sigma_{\text{diag}} \;+\; \Sigma_{\text{low}-\text{rank}}\big) \tag{2.42}$$

balancing computational efficiency with sufficient expressiveness to capture local uncertainty. This hybrid representation leverages the diagonal component to model-independent parameter variances and the low-rank term to capture linear correlations, thus approximating the posterior structure without requiring full-rank covariance matrices.

Sampling and Predictive Distribution Once Σ is obtained, SWAG treats the posterior over parameters as:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \;\approx\; \mathcal{N}\Big(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}, \Sigma\Big) \tag{2.43}$$

thus modelling parameter uncertainty with a Gaussian centred at $\boldsymbol{\theta}_{\text{SWA}}$. The predictive distribution for $\boldsymbol{y}$ given $\boldsymbol{x}$ becomes:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \;\approx\; \int p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\big)\, \mathcal{N}\Big(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}, \Sigma\Big) \mathrm{d}\boldsymbol{\theta} \tag{2.44}$$

Since this integral is typically intractable, Monte Carlo integration is used, drawing N parameter samples:

$$\boldsymbol{\theta}_{\text{SWAG}}^{(n)} :=\; \boldsymbol{\theta}_{\text{SWA}} + \Sigma^{1/2}\, \boldsymbol{\epsilon}^{(n)} \quad \text{with} \quad \boldsymbol{\epsilon}^{(n)} \sim \mathcal{N}\big(\mathbf{0}, \mathbf{I}\big) \tag{2.45}$$

and averaging the resulting predictions:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \;\approx\; \frac{1}{N} \sum_{n\,=\,1}^{N} p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\text{SWAG}}^{(n)}\big) \tag{2.46}$$

Each sample ($\boldsymbol{\theta}_{\text{SWAG}}^{(n)}$) represents a plausible set of parameters lying within a local basin of the loss landscape, effectively forming an *implicit ensemble*.

2.3.5. Deep Ensemble

Deep neural networks trained with stochastic optimisers (e.g., SGD) often converge to different local minima, particularly when initialised with distinct random seeds or subjected to varying data orders. In highly non-convex loss landscapes, each training run may settle in a distinct *basin*, yielding different parameter solutions $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$.

Deep Ensembles exploit this phenomenon by treating these solutions as modes of the posterior distribution. By collecting models from separate basins, one effectively samples from multiple modes of the underlying parameter space, thus providing a practical framework for modelling epistemic uncertainty in deep learning.

Posterior Approximation Deep Ensemble extends the *plug-in approximation* (Section 2.3.3) by considering multiple point estimates $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ obtained from independent training runs. Rather than approximating the posterior as a single Dirac delta distribution, it constructs a weighted mixture of Dirac deltas:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \;\approx\; \frac{1}{M} \sum_{m\,=\,1}^{M} \delta\big(\boldsymbol{\theta} - \boldsymbol{\theta}_m\big) \tag{2.47}$$

Intuitively, each local minimum is treated as a distinct mode in the parameter space, providing an empirical approximation of the posterior distribution.

Predictive Distribution Substituting the ensemble-based posterior approximation into the BMA framework yields the following expression for the posterior predictive distribution:

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) &\approx \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \Big(\frac{1}{M} \sum_{m=1}^{M} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_m) \Big) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{M} \sum_{m=1}^{M} \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \, \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_m) \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_m) \end{aligned} \tag{2.48}$$

Hence, it simplifies to the average of the individual model predictions.

More broadly, the ensemble formulation introduces greater flexibility by allowing the assignment of weights (α_m) to each model. When these weights are normalised to sum to one, the resulting combination is referred to as a *mixture of experts*. Conversely, non-normalised weights lead to a formulation, commonly known as *stacking*.

2.3.6. Multi-SWAG

Despite the considerable benefits of SWAG and Deep Ensemble in modelling parameter uncertainty, each approach has inherent limitations. SWAG (Section 2.3.4) focuses on capturing the *local* structure around a single basin in the loss landscape through a Gaussian approximation, thereby neglecting potentially distant modes. Conversely, Deep Ensemble (Section 2.3.5) samples multiple basins by training independently initialised networks yet represents each basin with a single point estimate, thereby omitting the shape of the basin.

Multi-SWAG [40] is designed to reconcile these complementary perspectives by approximating the posterior distribution as an equally weighted *Gaussian mixture model*.

Predictive Distribution Concretely, one first trains multiple SWAG instances, each starting from a different random seed or following a distinct training protocol; this strategy encourages the exploration of separated basins within the loss landscape. Each model then maintains its running mean ($\boldsymbol{\theta}_{\mathrm{SWA}}^{(m)}$) and low-rank plus diagonal covariance ($\Sigma^{(m)}$), thus providing a Gaussian approximation of its local posterior structure. Once these local approximations are obtained, they are combined to form an ensemble of Gaussian posteri-

ors, where each component targets a distinct mode. Through this *ensemble-of-ensembles* viewpoint, Multi-SWAG aims to reconcile the benefits of SWAG — namely, capturing fine-grained parameter uncertainty in a given basin — with the multi-modality exploited by Deep Ensembles. Multi-SWAG jointly captures the *within-basin* (intra-basin) parameter variability of local minima and the *across-basin* (inter-basin) diversity, alleviating the single-basin limitation typically assumed in standard SWAG.

Formally, each SWAG approximation contributes a local Gaussian $\mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}^{(m)}, \Sigma^{(m)})$, leading to a final posterior estimate expressed as an average over M such components:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathcal{N}\Big(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}^{(m)}, \Sigma^{(m)}\Big) \tag{2.49}$$

Substituting it into the BMA framework then yields the desired posterior predictive distribution, expressed by integrating over all the local Gaussians:

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) &\approx \int p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\big) \Big(\frac{1}{M} \sum_{m=1}^{M} \mathcal{N}\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}^{(m)}, \Sigma^{(m)}\big)\Big) \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{M} \sum_{m=1}^{M} \int p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}\big) \, \mathcal{N}\big(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{SWA}}^{(m)}, \Sigma^{(m)}\big) \, \mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{2.50}$$

Since the integral is generally intractable, a common strategy is to draw Monte Carlo samples from each Gaussian component $\mathcal{N}\big(\boldsymbol{\theta}_{\text{SWA}}^{(m)}, \Sigma^{(m)}\big)$. Specifically, one obtains N parameter realisations $\boldsymbol{\theta}_{\text{SWAG}}^{(m,n)}$ computes their individual predictions $p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\text{SWAG}}^{(m,n)}\big)$, and averages the results:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \approx \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} p\big(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\text{SWAG}}^{(m,n)}\big) \tag{2.51}$$

2.3.7. Rethinking Ensembles

In Bayesian inference, the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ encodes the likelihood of each *hypothesis* — a particular set of $\boldsymbol{\theta}$ — given the observed data. Under the *Bayesian Model Averaging* paradigm, the goal is to marginalise over all such hypotheses rather than commit to a single point estimate:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\theta} \tag{2.52}$$

This perspective suggests that, given the finite information available in $\mathcal{D}$, multiple parameter configurations may be consistent with the observed data. Critically, Bayesian Model Averaging assumes the existence of a *single true* — though unknown — hypothesis while recognising that a limited dataset may be insufficient to isolate it with complete confidence. Consequently, the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ serves to re-weight likely parameter configurations based on their alignment with the data. Notably, as the size of the dataset approaches infinity, the hypothesis space is expected to collapse into a degenerate distribution centred around the correct hypothesis. Within this framework, the Monte Carlo integration provides a tractable means of approximating marginalisation by combining the predictive conditional distributions associated with a finite sequence of models sampled from the (potentially approximate) posterior distribution.

By contrast, *model combination* or *ensemble* methods characteristically regard each trained model as an *independent* predictor and combine them by equal or user-selected weights. In other words, the underlying assumption of Deep Ensemble [41] is that the ground truth itself may be better approximated by the *collective* of multiple trained networks — rather than belonging to one distinct parameter setting only. There is no direct attempt to approximate or explore the posterior over parameters; each model reflects a separate training instance, and the final prediction merges them via a simple mixture of outputs.[1]

Although conceptually different from the Bayesian perspective, ensembles can be an effective heuristic to approximate multi-modal posteriors in high-dimensional spaces, as each independently trained model may discover a different local optimum. As shown in [42], the variance across such local solutions can help to explore the hypothesis space better, thereby improving generalisation and calibration relative to a single-point estimate. Nevertheless, ensembles, in their standard form, do not fully correspond to sampling from $p(\boldsymbol{\theta} \mid \mathcal{D})$. Rather, they can be understood as an empirical mixture of point masses centred at each $\boldsymbol{\theta}_m$, whereby each $\boldsymbol{\theta}_m$ is treated as if it were equally plausible a priori. Therefore, although BMA and model combination yield averaged predictions, the *conceptual underpinnings* differ significantly. BMA asserts that, in principle, one correct parameter configuration exists but cannot be identified with certainty given limited data; ensembles posit that the true function can often be better modelled by aggregating a plurality of perspectives.

[1] In practice, heuristics such as different data permutations or slightly varied architectures are often used to encourage model diversity.

3 | Canopy Height Estimation – State of the Art Overview

This chapter details the methodologies available for canopy height mapping, ranging from traditional techniques to advanced Bayesian Deep Learning-based approaches.

The most accurate assessments are obtained through in-situ surveys [43] [44]. However, despite their high precision, field-based measurements are inherently constrained by logistical and economic limitations, rendering them impractical for large-scale applications.

Airborne LiDAR Scanning (ALS) systems offer a scalable alternative for broader forest assessments [45] [46], with centimetre-level vertical accuracy, they enable the direct estimation of canopy height [43]. Over time, ALS data have become a reference source for large-scale forest ecosystem characterisation. Nevertheless, high operational costs limit their feasibility for global-scale applications.

Spaceborne LiDAR systems — including the Global Ecosystem Dynamics Investigation (GEDI) [4] and the Ice, Cloud, and Land Elevation Satellite 2 (ICESat-2) [47] — extend the coverage to a global scale. However, this comes at the expense of sparse observation and limited signal penetration in dense canopies, which may impact data completeness and quality [48].

Satellite-based optical and Synthetic Aperture Radar (SAR) imagery offer a viable alternative to achieve gap-free, large-scale canopy height estimations. The two primary methodological frameworks leveraging these data are *physics-based models*, which rely on electromagnetic interaction principles, and *data-driven approaches*, which infer statistical relationships from large informative datasets.

3.1. Physics-Based Models

SAR data are widely employed in physics-based modelling due to their sensitivity to vegetation structure and dielectric properties. Early models, such as the *Water Cloud*

Model [49], related SAR backscatter to canopy height by representing vegetation as a collection of water droplets. More advanced methods have exploited Polarimetric SAR (Pol-SAR), Interferometric SAR (InSAR), and their combination (Pol-InSAR) [50] with the *Random Volume over Ground* (RVoG) model [51] emerging as a standard approach, leveraging multiple acquisitions or polarimetric data from spaceborne SAR systems. These models benefit from strong physical interpretability but rely on simplifying assumptions about vegetation structure and scattering mechanisms, which may limit their applicability in heterogeneous environments.

Beyond these methods, *Tomographic SAR* (TomoSAR) reconstructs three-dimensional canopy structures by employing multiple interferometric baselines [52]. Applied to L- [53] and P-band [54] airborne SAR data, TomoSAR showed promising results, though operational challenges persist due to high data requirements and decorrelation effects. TanDEM-X bistatic time-series demonstrated the feasibility of spaceborne tomography [55], while the upcoming ESA's BIOMASS mission [56] aims to deliver global-scale tomographic capabilities, albeit at reduced spatial resolution due to the limitations imposed by the P-band wavelength.

3.2. Data-Driven Models

Alongside physics-based approaches, data-driven models provide an alternative for canopy height estimation. Unlike previous methods focused on SAR data, these approaches have also integrated optical images — which had not been considered thus far — given their limited penetration and weaker interaction with vertical canopy structures. Optical data are often combined with SAR data to exploit complementary information through *data fusion*.

The study in [57] evaluated the performance of Sentinel-1 and Sentinel-2, both separately and combined, using *machine learning* algorithms — including ordinary least-squares regression, classification and regression trees, and random forest regression — trained on LiDAR samples from Paraná, Brazil. Sentinel-1 data provided the lowest accuracy, while Sentinel-2 performed better but showed greater temporal variability. The fusion of both datasets yielded the highest performance. Similarly, [58] applied a random forest regression model to estimate forest height across Australia by integrating multi-spectral Landsat data, ALOS/PALSAR L-band backscatter, and ICESat-derived height percentiles. Both studies demonstrated the benefits of fusing conventional SAR with multi-frequency optical products, as the latter achieve high predictive performance but are affected by weather conditions, with performance fluctuating over time due to seasonal and meteorological

factors, ultimately limiting their scalability for continuous monitoring.

Deep learning algorithms have gained attention for their success in computer vision [59][60]. While widely adopted for classification tasks in Earth Observation [61][62], their application to canopy height regression remains relatively under-explored. The study in [7] demonstrated that incorporating spatial context information within a modified "Xception" Convolutional Neural Network (CNN) significantly enhanced forest height estimation from Sentinel-2 data, highlighting the role of spatial patterns and the advantage of CNNs in capturing them. In [63], a complex-valued deep learning model was introduced for forest height estimation using single-baseline L-band Pol-InSAR data from DLR's E-SAR instrument over Traunstein, Germany. The model outperformed the RVoG approach, demonstrating that achieving state-of-the-art performance does not necessarily require very deep architectures, such as those explored in [7]. In [64], a vanilla CNN was employed to estimate forest height from TanDEM-X InSAR data, demonstrating the potential of InSAR measurements for large-scale canopy height mapping. The study focused on generating wall-to-wall canopy height maps across Gabon from single-baseline coherence features. Building upon this work, [65] extended the approach by investigating the temporal evolution of forest height over a 10-year period, demonstrating the applicability of deep learning to monitor long-term canopy dynamics. Additionally, the study highlighted the challenges associated with SAR data acquisition in harsh environments, such as data gaps and terrain-induced distortions. The research is further detailed in [17]. Together, these studies validated the effectiveness of relatively shallow CNN-based architectures for InSAR-derived large-scale canopy height mapping and long-term monitoring. However, models trained on specific environments often lack generalisability across forest structures and climatic conditions. Furthermore, the absence of an associated uncertainty measurement limits the trustworthiness of the framework for downstream applications.

Recent advancements have explored alternative architectures beyond CNNs. In this direction, [66] adopted a Transformer-based model, taking advantage of its ability to capture long-range dependencies and contextual relationships more effectively than convolutional networks, but their application remains computationally demanding and reliant on large-scale training data, posing challenges for generalisation and operational scalability.

3.2.1. Uncertainty-Aware Models

Existing literature rarely provides calibrated uncertainty estimates alongside map products, resulting in model predictions being accepted without a quantified measure of confidence — despite the well-documented overconfidence issue affecting deep learning models

[13].

In this context, [67] leveraged epistemic uncertainty to guide active learning, selecting the most informative samples for labelling by prioritising high-uncertainty and diverse instances. Meanwhile, [13] proposed an ensemble of five independently trained dual-branch CNNs based on the "ResNeXt" architecture, combining Sentinel-1 dual-polarisation SAR data from ascending and descending orbits with twelve Sentinel-2 multi-spectral bands. This approach estimated Normally distributed probability density functions for five forest structure parameters on a national scale in Norway, producing both predictions and well-calibrated uncertainty maps. The findings were further extended in [68], where the CNN ensemble was trained using sparse 25-metre GEDI samples as ground truth to generate a global canopy height map based exclusively on Sentinel-2 data. Validation was conducted with independent LVIS and ALS measurements across North America, Central America, Europe and Africa. Despite the promising performance, the claimed 10-metre resolution remains questionable due to reliance on lower-resolution LVIS data for validation. While uncertainty-aware models enhance reliability and improve learning from noisy datasets, they struggle to generalise beyond the training domain, and current research lacks a systematic assessment of out-of-distribution uncertainty estimation performance.

4 | Datasets, Pre-Processing and Area of Interest

This chapter outlines the datasets, describes the processing pipelines employed, and defines the area of interest analysed in this study.

4.1. TanDEM-X and TerraSAR-X

The TanDEM-X (TDX) and TerraSAR-X (TSX) satellites are German spaceborne Synthetic Aperture Radar (SAR) systems developed through a public-private partnership between the Deutsches Zentrum für Luft- und Raumfahrt (DLR) and Airbus Defence and Space [69] [70]. Launched in mid-2007 (TSX) and mid-2010 (TDX), the satellites operate in a dusk-dawn sun-synchronous orbit at a nominal altitude of 514 kilometres with a repeat cycle of 11 days. Their nearly identical hardware configurations, summarised in Table 4.1, include active phased array antennas and support multiple acquisition modes such as Stripmap, Spotlight, ScanSAR and TopSAR. A dual-receive antenna was integrated to support along-track interferometry and enable the collection of fully polarimetric data (HH, HV, VV, VH).

The TerraSAR-X mission [70] was designed to provide high-quality SAR products for scientific research and commercial applications. To enable synergy with the future TanDEM-X mission, TerraSAR-X was designed with hardware enhancements for precise orbit determination, stable radio frequency signal transmission and inter-satellite synchronisation [70].

The TanDEM-X mission [70] [71], building upon TerraSAR-X [70], was designed to generate a global Digital Elevation Model (DEM) with unprecedented accuracy, achieving a relative height error of 2 metres and a horizontal resolution of 12 metres [72]. Operating in bistatic Stripmap mode, the satellites maintain a "Helix" formation with separations ranging from 100 to 500 metres [71], facilitating single-pass interferometry to minimise temporal decorrelation and atmospheric disturbances. The initial DEM acquisition phase

System Parameters		
Orbit	Inclination	97.4°
	Tube	250 m
Antenna Size	Azimuth	4.8 m
	Elevation	0.7 m
Signal	Center Frequency	9.65 GHz
	Bandwidth	$\leq$ 300 MHz
	PRF	2 kHz – 6.5 kHz
Acquisition	Swath-Width	10 km (Spotlight) – 100 km (ScanSAR)
	Ground Resolution	0.25 m (Spotlight) – 40 m (ScanSAR)
	Look Angle	15° – 60°

Table 4.1: Summary of the main TerraSAR-X and TanDEM-X system parameters.

(2011–2013) produced the first global DEM [73], refined using dual-baseline phase unwrapping techniques and calibrated with ICESat ground control points [74]. Subsequent campaigns (2017–2020) provided the updated Global DEM and the Global DEM Change Map [75].

A key challenge for the mission, relevant to this research, was the impact of volumetric scattering in vegetated areas, where scattering contributions from within the vegetation volume degraded coherence. Short baselines were adopted to minimise these effects in densely vegetated regions.

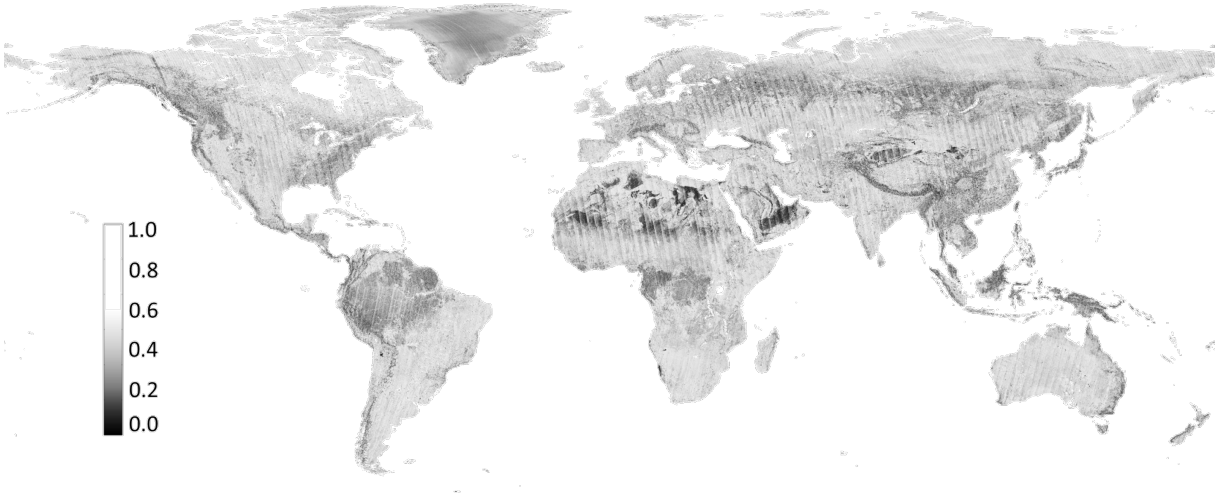


Figure 4.1: Global TanDEM-X coherence mosaic [17].

The global coherence map in Figure 4.1 highlights these effects, showing reduced coherence in regions with dense vegetation, where the signal is primarily influenced by scat-

tering from multiple layers within the canopy. Similar reductions in coherence are also observed over deserts and mountainous areas. The elevation data obtained represent the mean phase centre height of scatterers within the resolution cell [73]. In regions affected by volumetric scattering, penetration of electromagnetic waves into the volume results in systematic elevation offsets [76], which depend on the acquisition geometry and the structure of the scattering volume.

4.1.1. TanDEM-X Bistatic Product Processing Chain

The processing begins with the *Integrated TanDEM-X Processor* (ITP) [77], which divides raw data from each acquisition into scenes of approximately 30 × 50 kilometres, in range and azimuth, respectively. SAR focusing and co-registration of the master and slave image pairs result in the *co-registered single-look complex* (CoSSC) product.

The interferometric processing and raw DEM generation are performed using the *experimental TanDEM-X interferometric processor* (TAXI) [78], developed at the DLR Microwaves and Radar Institute (HR). Coherence estimation employs a conventional multi-looking approach, resulting in a 12-metre resolution, balancing estimation accuracy and spatial detail preservation. The resulting interferogram is subsequently processed to generate the raw acquisition DEM.

Following [79], the volumetric decorrelation coefficient (γ_{vol}) is estimated from the total coherence (γ_{tot}) by inverting Eq. (1.21) and compensating for the remaining decorrelation factors. Additionally, for each scene, a β^0 image is computed using the calibrated amplitude recorded by the transmitting satellite (i.e., the monostatic channel), from which the backscattering coefficient (σ^0) is derived.

Using acquisition metadata and an external reference DEM (e.g., the TanDEM-X Global Edited DEM), various feature maps related to SAR and Interferometric Synthetic Aperture Radar (InSAR) acquisition geometry are generated, including incidence angle maps (θ_{inc}) computed relative to the reference ellipsoid, local incidence angle maps (θ_{LIA}) derived from the topographic slope, geometric distortion classification maps identifying shadow and layover areas [80] [23], and height of ambiguity maps (h_{amb}) computed using satellite positional data, following Eq. (1.17).

4.1.2. TanDEM-X Dataset

The TanDEM-X data used in this investigation consist of 322 acquisitions over Norway spanning the period from 2011 to 2021, specifically collected between June and August.

The data are derived from the first and second global DEM campaigns and downsampled to a Ground Sampling Distance (GSD) of 20 metres. The feature stack, illustrated in Figure 4.2, is generated during the processing described in Section 4.1.1. It includes backscatter intensity (σ^0_{HH}), total coherence (γ_{tot}), ambiguity height ($\mathrm{h_{amb}}$), local incidence angle (θ_{LIA}), volume decorrelation factor (γ_{vol}) and acquisition DEM ($\mathrm{DEM_{ACQ}}$).

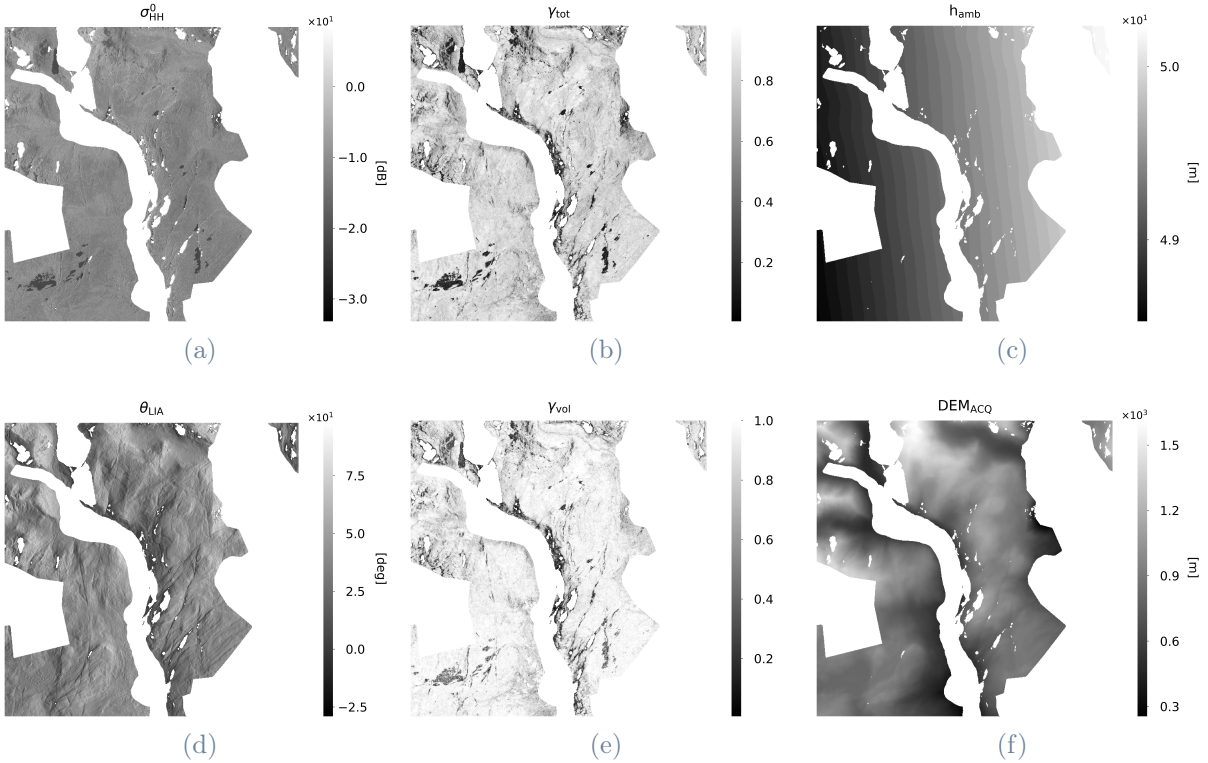


Figure 4.2: Processing output feature collection example, consisting of: (a) backscattering coefficient; (b) interferometric coherence estimate; (c) height of ambiguity; (d) local incidence angle; (e) volume decorrelation factor estimate; (f) acquisition DEM. White areas represent non-forested regions or missing valid satellite data.

4.2. Airborne Laser Scanning Dataset

The Norwegian Institute of Bioeconomy Research (NIBIO) acquired Airborne Laser Scanning (ALS) data as part of a national programme for forest mapping in Norway, conducting nine airborne campaigns between 2009 and 2018. These point cloud datasets cover diverse geographical regions nationwide and provide a robust foundation for canopy height analysis [81]. The extensive spatial coverage, spanning latitudes from 58°N to 69°N and longitudes from 5°E to 18°E, reflects the climatic and ecological diversity of Norway's forests. Furthermore, the temporal distribution of the data, primarily concentrated between July

and November, ensures a comprehensive representation of seasonal forest conditions.

The ALS datasets offer varying point densities, distributed across different density ranges: between 0.1 and 2 points per square metre for 4% of the covered area, 2 to 5 points for 46%, 5 to 10 points for 44%, and 10 to 25 points for the remaining 3%. These high-resolution data facilitate the detailed characterisation of forest structures and provide a reliable basis for integration with satellite SAR images.

To compute canopy height, the raw ALS data — initially referenced to elevation above sea level — are normalised to ground height by subtracting terrain elevation values from the DEM, ensuring accurate measurement of vegetation heights relative to the ground surface. Canopy height is quantified using the percentile height, which closely approximates the maximum height of the forest canopy while mitigating the influence of noise from outlier returns. The computation aggregates the ALS point cloud data onto a 20-metre raster grid, aligning it with the resolution of the SAR products.

4.3. ESA WorldCover Map

The ESA WorldCover map is a global land-cover product at a 10-metre resolution, derived from data acquired by the Sentinel-1 and Sentinel-2 missions and freely accessible [82]. The product classifies the scene into 11 categories: tree cover, shrubland, grassland, cropland, built-up areas, bare/sparse vegetation, snow/ice, permanent water bodies, herbaceous wetland, mangroves and moss/lichen.

This study uses the *WorldCover 2021 v200* product, an improved version based on 2021 acquisitions with an overall classification accuracy of 76.7%. This product is employed as a pre-processing step to mask the scene over undesired land-cover types: built-up areas, snow or ice, permanent water bodies and moss/lichen.

4.4. CORINE Land Cover Map

The CORINE Land Cover (CLC) map is a comprehensive pan-European land cover inventory initiated in 1985 by the European Commission and currently managed by the European Environment Agency (EEA) [83]. It categorises land cover into 44 distinct classes, covering a broad spectrum of natural and anthropogenic landscapes.

The CLC dataset is updated every six years, offering both status layers, representing land cover at specific reference years, and change layers, which detail transitions in land cover between successive periods. The map resolution is 100 metres.

In this study, the *CORINE Land Cover map - 2018 version* is used to incorporate land cover classifications during the performance evaluation phase, enabling performance discrimination across forest types.

4.5. Dataset Pre-processing Pipeline

4.5.1. Dataset Creation Tool

Integrating multiple Earth Observation (EO) data sources necessitates their alignment within a common Geographic Information System (GIS). This process is supported by the internally developed Dataset Creation Tool (DSCT) [17] at DLR, which efficiently manages heterogeneous datasets. This tool enables the processing of hundreds of EO products, amounting to terabytes of data, within a matter of hours.

The *Aggregation* module is designed to associate each dataset with its corresponding intersecting Regions of Interest (ROIs), processing essential metadata, including geographic footprints, projection systems and acquisition dates, while defining the nominal ROIs. In this study, the ROIs are derived from NIBIO's reference data.

Subsequently, the data are *reprojected* and *resampled* onto a unified reference grid using the *Common Grid Interpolation* (CGI) module. This module also validates data by identifying invalid points and generating auxiliary features. Each processed dataset is assigned to a unique identifier (ID), ensuring consistent and traceable data handling throughout the experiments.

4.5.2. Alignment

A significant portion of this thesis has been dedicated to developing a flexible add-on to the DSCT that builds on CGI's output. The first step in this extended framework is the *Alignment* module, which generates unique data entries composed of a single product from each source. This process results in an inflated version of the input data frame, where each entry corresponds to a unique set of EO products. The resulting data frame serves as input for subsequent processing modules. Table 4.2 shows an example of this operation.

4.5.3. Sampling

This step converts the data into formats suitable for deep learning applications. Two main formats are employed: *datacubes*, which stack features across ROIs, and *patches*,

equally sized crops of the former. In this implementation, only datacubes are physically stored, while patches are accessed on demand using memory-mapped files (mmap). This approach minimises disk usage by avoiding redundant data storage and optimises runtime performance.

ROI-ID	NIBIO	TANDEM_X	CORINE LAND COVER	WORLD COVER MAP
33-103-107T2017	33-103-107T2017	TDM1_SAR_COS_BIST_SM_S_SRA_20180713T055855-20180713T055903 TDM1_SAR_COS_BIST_SM_S_SRA_20190608T055901_20190608T055909	CLC_33-103-107T2017	WCM_33-103-107T2017

↓

ROI-ID	NIBIO	TANDEM_X	CORINE LAND COVER	WORLD COVER MAP
33-103-107T2017_0	33-103-107T2017	TDM1_SAR_COS_BIST_SM_S_SRA_20180713T055855-20180713T055903	CLC_33-103-107T2017	WCM_33-103-107T2017
33-103-107T2017_1	33-103-107T2017	TDM1_SAR_COS_BIST_SM_S_SRA_20190608T055901_20190608T055909	CLC_33-103-107T2017	WCM_33-103-107T2017

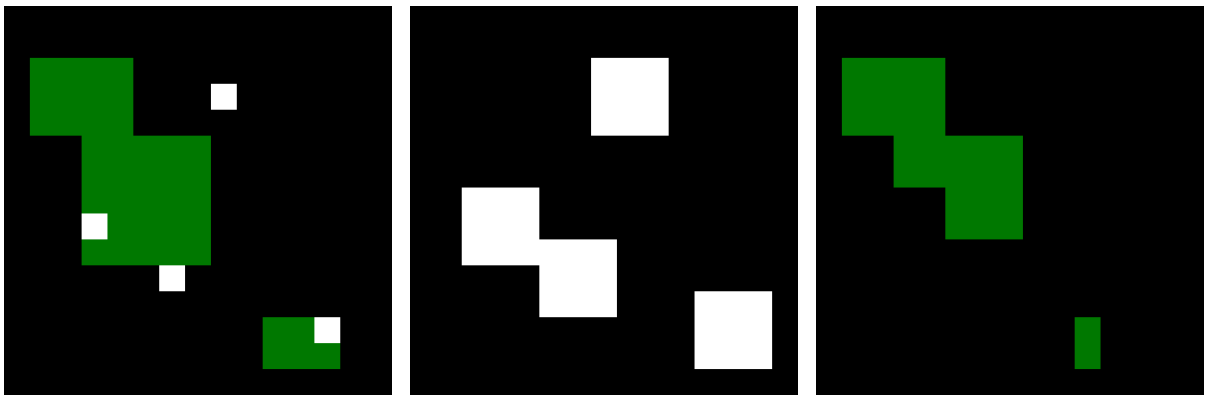
Table 4.2: The input entry (top) consists of one or more products per source. Following inflation (bottom), a collection of unique entries is produced.

Generating datacubes involves iterating through each ROI and its associated sources, reading the associated features, stacking them into a datacube, and saving it as a numpy file. Patch generation builds upon the datacube creation process. Given a nominal patch size, valid patch centres are identified based on specific criteria, such as the absence of *Not-A-Number* (NaN) values and the presence of forested pixels[1]. These criteria are enforced using the ESA WorldCover map (Section 4.3), alongside input and label data, and refined using morphological filtering techniques, such as binary dilation. The locations of valid patches are stored for subsequent use during training. Figure 4.3 shows an example of this operation.

4.6. The Norwegian Context

Norway's forest ecosystem covers approximately 38% of the country's land area, primarily consisting of boreal forests. These forests are characterised by tree heights up to 40 metres,

[1]This approach is employed to exclude pixel classes not directly relevant to the regression task, thereby minimising complexity in the dataset's design and validation process.



(a) (b) (c)

Figure 4.3: (a) Scene (white for NaN areas, green for vegetation); (b) dilated NaN mask; (c) valid patch-centre sampling mask.

with Picea abies (Norway spruce) being a dominant species. The geographical distribution varies significantly, with dense forest coverage concentrated in the central and southern regions, while the northern areas exhibit sparser forestation, transitioning gradually into tundra landscapes [81].

4.6.1. Height Distribution

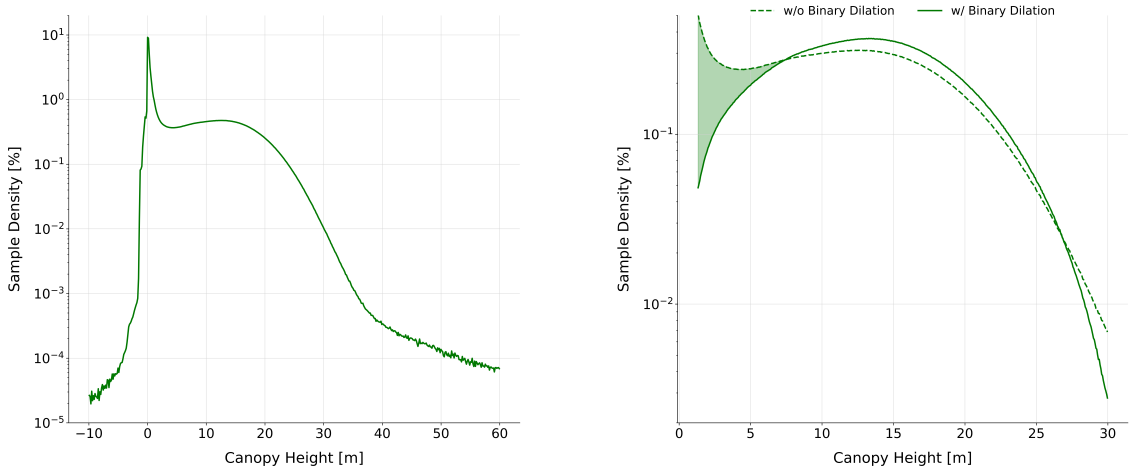
The unique forest conditions in Norway are well captured by the ALS-derived 20-metre gridded height products generated from NIBIO's ALS data. Following the data pre-processing pipeline outlined in Section 4.5, the height distributions within the valid patches exhibit a bimodal structure, as shown in Figure 4.4a. The first peak, centred at approximately 0 metres, reflects a substantial proportion of land with sparse or absent vegetation, consistent with the expected characteristics of the region.

Based on the methodology proposed in [13], this study focuses exclusively on areas with meaningful vegetation heights by masking out measurements below 1.3 metres and above 30 metres. While some trees in Norway exhibit maximum heights up to 40 metres, these measurements represent the average height within a 20 x 20 metres grid (i.e., 400 m^2). The averaging process justifies the upper threshold of 30 metres, as it effectively excludes potential outliers and noisy data. Similarly, excluding vegetation below 1.3 metres ensures the removal of non-vegetated areas and noise below 0 metres.

Acknowledging the potential implications of this masking-out strategy on the network's performance is important. Without applying it, the network might achieve artificially improved performance by exploiting the dominance of low-vegetation areas and predicting

near-zero values in ambiguous cases. However, by masking out heights outside the defined range, the model focuses on scientifically meaningful vegetation heights, potentially at the cost of slightly reduced overall performance but ensuring a stronger alignment with the research's objective.

Initially, a binary dilation operation with a 3×3 kernel was applied around the masked pixels to ensure a conservative margin of confidence for threshold-based exclusion. This approach was intended to mitigate boundary artefacts and ensure robust masking of noisy data and outliers. The resulting configuration, shown in Figure 4.4b, revealed an unintended consequence: the binary dilation accidentally removed vegetation between low and high vegetation classes. Such transitional regions are critical for the analysis of the flora, prompting a revision of the masking approach. As a result, the final configuration excludes binary dilation, as illustrated in the dashed representation of Figure 4.4b, thereby preserving the overall vegetation distribution and spatial patterns across the country.



(a) (b)

Figure 4.4: (a) Norway canopy height histogram derived from NIBIO's ALS data; (b) comparison of binary dilation effects on masking out-of-interest height values.

4.6.2. Geographic Subsetting

The geographic intersection of the available products defines a valid ground area for each ROI. The ROIs are then assigned to the training, validation and test datasets, forming three geographically distinct and non-overlapping subsets. This approach prevents *data leakage*, ensures fair test performance, promotes robust generalisation, and reduces overfitting to specific geographical features. Two partitioning approaches were evaluated: (i) dividing the data into north-south or east-west stripes, where each stripe alternated between datasets, and (ii) a random allocation method. Since the latter did not yield measurable benefits, the random allocation method was ultimately considered more suitable due to the large volume of data, the near-complete coverage of the country (except for a central-southern mountainous region), and its scalability to other case studies under similar conditions. The geographical distribution of the subsets is shown in Figure 4.5, where the colour coding represents the training (green), validation (red), and testing (blue) datasets.

Figure 4.5: Geographical division of the NIBIO's campaign sites.

Figure 4.6 compares the histograms of the reference canopy height obtained for the training, validation and test subsets. Notably, all three subsets exhibit similar bimodal distributions, with the first peak representing extensive areas of very sparse vegetation, and the second peak, around 15 metres, corresponding to regions with moderate canopy height. These distributions suggest that the subsets are equally representative of the regional variability.

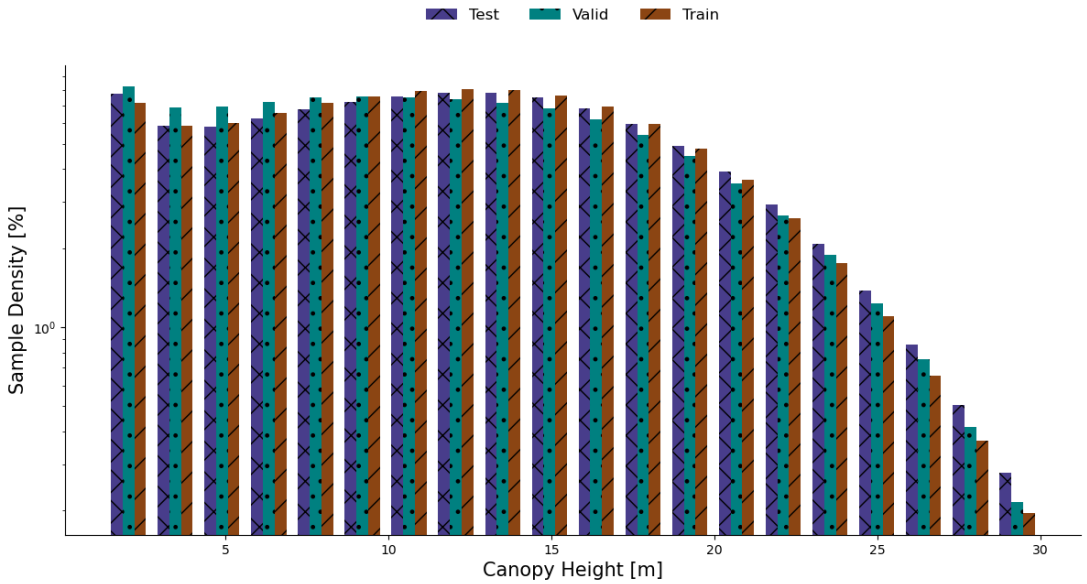


Figure 4.6: The distributions of reference canopy height values for the training, validation and test subsets.

5 | Deep Learning Framework — Definition and Generalisation

This work aims to move beyond single-country studies, enabling detailed and timely monitoring of global forest dynamics at reduced costs by generating reliable, high-resolution, time-tagged predictions derived from a single satellite image input.

As discussed in Chapter 3, deep learning has demonstrated remarkable potential in capturing complex relationships between forest properties and satellite observations. Moreover, its adaptability to diverse environmental conditions and boundary constraints makes it a promising alternative to traditional physics-based models. In this context, Interferometric Synthetic Aperture Radar (InSAR) data serve as a powerful input, as demonstrated in [17], which explored the use of TanDEM-X single-pass InSAR products combined with a purely data-driven deep learning approach for forest height estimation in Gabon.

Building upon that work, a primary objective of this research is the generalisation of the proposed framework. Validation is undertaken in Norway, a region characterised by substantially distinct forest types and land cover compared to Gabon. This shift represents a key step towards addressing generalisation challenges beyond single-country studies.

The initial section of this chapter focuses on a detailed analysis of the dataset at a national scale, supporting the development of appropriate data handling strategies informed by the dataset composition and the operational principles of the deep learning architecture. The chapter then transitions to an in-depth discussion of the implementation details. A baseline scenario and its corresponding performance metrics are first introduced to establish a reference for subsequent analyses. Given the richness of the forest landscape, the research then investigates whether different forest species influence the network's learning process and proposes a more robust set of features to distinguish between vegetated and non-vegetated areas.

As highlighted in Chapter 4, the study relies on 322 TanDEM-X acquisitions that do not perfectly align with the NIBIO's Airborne Laser Scanning (ALS) data due to temporal inconsistencies in the acquisition periods for the same regions of interest. Accordingly,

the final section of this chapter explores the impact of temporal shifts, which may extend up to 11 years, on the model's performance. These temporal inconsistencies introduce natural variability into the problem, potentially acting as a beneficial source of noise from which the network could derive valuable insights during the training phase.

5.1. The SILVA Framework

As discussed in [17], Convolutional Neural Networks (CNNs) are suitable for efficiently processing structured, multi-modal images with large spatial dimensions at a relatively low computational cost. Under the reasonable assumption of a natural imaging process, neighbouring pixels in remotely sensed data exhibit a strong correlation. CNNs leverage this spatial autocorrelation by employing shared kernel functions, efficiently capturing these local dependencies.

The CNN framework — hereinafter referred to as *SILVA* (from the Latin *silva*, meaning "forest") — is directly derived from the comprehensive analysis presented in [17]. The reader is referred to the original work for an in-depth discussion of the design choices and their underlying rationale.

5.1.1. The Model Architecture

The architecture of the model, illustrated in Figure 5.1, is specifically designed to generate predictions while preserving input resolution and minimising computational complexity. The model processes inputs through a series of incremental transformations (layers) and is organised into three main functional components: an *input head*, a series of *hidden blocks* acting as feature extractors, and a *regression head*. However, these components are integrated seamlessly, without explicit boundaries between them.

Input Head The entry block consists of two 1×1 two-dimensional convolutional layers configured with 64 and 128 kernel filters, respectively. These layers are designed to pre-process the input data, integrating channel-wise information while preserving the spatial dimensions. Each convolutional layer is followed by a batch normalisation layer and a ReLU activation to ensure training stability and introduce non-linearities.

Hidden Blocks The feature extraction phase is implemented through ten hidden blocks, each composed of two 3×3 convolutional layers with 128 kernel filters. Together, the hidden blocks extract detailed spatial features from the input data. Each convolution is followed by a batch normalisation layer and a ReLU activation, enabling the network to

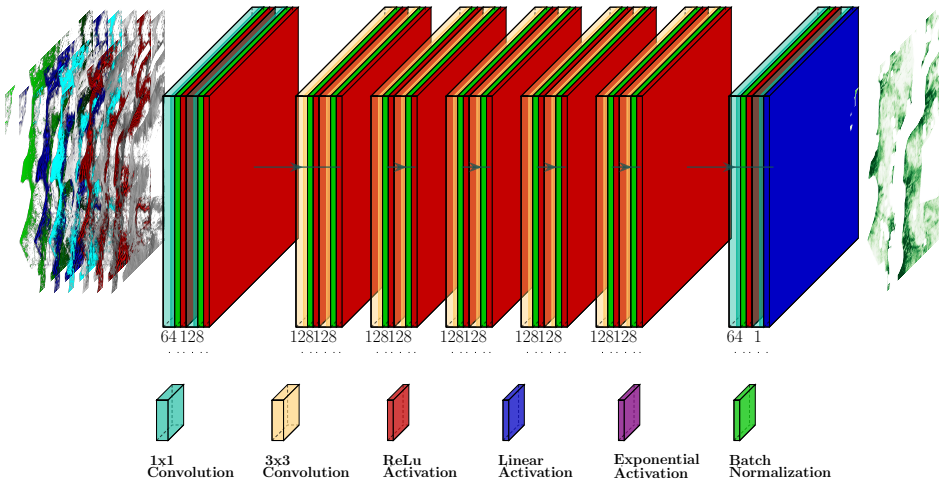


Figure 5.1: The fully convolutional deep learning model, with subscript numbers denoting the number of kernel filters used in each layer.

effectively model complex, non-linear relationships. This stage forms the backbone of the model, acting as the primary feature extractor.

Regression Head The regression head consists of two 1×1 convolutional layers designed to progressively project the high-dimensional feature space into the output space. The first layer contains 64 kernels and is followed by a ReLU activation, while the second layer, with a single filter, is paired with a linear activation function (i.e., a direct bypass), enabling the generation of pixel-wise predictions through linear regression.

The model incorporates approximately 1.4 million trainable parameters and achieves a receptive field of 21×21 pixels, corresponding to 0.1764 km^2.

5.1.2. Training Implementation Details

During training, the model processes mini-batches of 256 randomly sampled patches, each measuring 15×15 pixels, to estimate the corresponding canopy height values. The sampling strategy ensures diverse training inputs, with each patch fully covered by both input and reference features and being labelled as forested. The input dimensions for a single minibatch are $256 \times 6 \times 15 \times 15$, producing outputs of size $256 \times 1 \times 15 \times 15$.

The model is trained using a mean squared error loss function that evaluates the deviation

between predicted outputs and ALS reference data. The loss function is defined as follows:

$$\mathcal{L}_k = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \lambda \sum_{j=1}^{m} w_j^2 \tag{5.1}$$

where $\hat{y}_i$ denotes the predicted value, y_i represents the corresponding reference value, w_j are the trainable weights, and $\lambda = 10^{-4}$ controls the impact of the $\mathcal{L}_2$-regularization term. The loss is computed at the central pixel of each patch, ensuring that the network has access to the surrounding contextual information necessary for making a meaningful prediction.

The backpropagation algorithm, combined with the Adaptive Moment Estimation optimiser (ADAM), updates the model's weights based on the gradients of the prediction loss. Each epoch, consisting of one thousand batch iterations, concludes with an evaluation of the model on the validation set using the same loss function as during training. The training starts with the model initialised using the He method and an initial learning rate of 1×10^{-4}, which is reduced by a factor of 10 after 30 consecutive epochs without improvement in validation loss. To mitigate overfitting, an early stopping mechanism terminates the training if the validation loss fails to improve for 35 consecutive epochs.

The model is implemented in TensorFlow and runs on an NVIDIA H100 GPU, completing the training in approximately twelve hours.

5.2. Baseline Performance

A baseline scenario is introduced as a reference for subsequent experiments, providing a standard for comparison.

The performance metrics presented in Table 5.1 reveal that the baseline model achieves an overall bias of -0.08 metres, a root mean square error of 3.26 metres, and an R^2 score of 0.73. These metrics indicate a reasonable predictive capability of the model.

MAE	**MAE%**	**RMSE**	**RMSE%**	**ME**	**ME%**	**R^2**
[m]	[%]	[m]	[%]	[m]	[%]	[-]
2.47	20	3.26	26	-0.08	-1	0.73

Table 5.1: Performance for the baseline scenario.

The scatter-plot in Figure 5.2a visualises the model's predictions against the reference measurements across Norway. The marginal distributions show the challenge the model faces in capturing the low-height vegetation peak, likely corresponding to transitional regions between forested and non-forested areas. Nonetheless, the model successfully identifies the second mode of the distribution, effectively capturing the predominant vegetation cover.

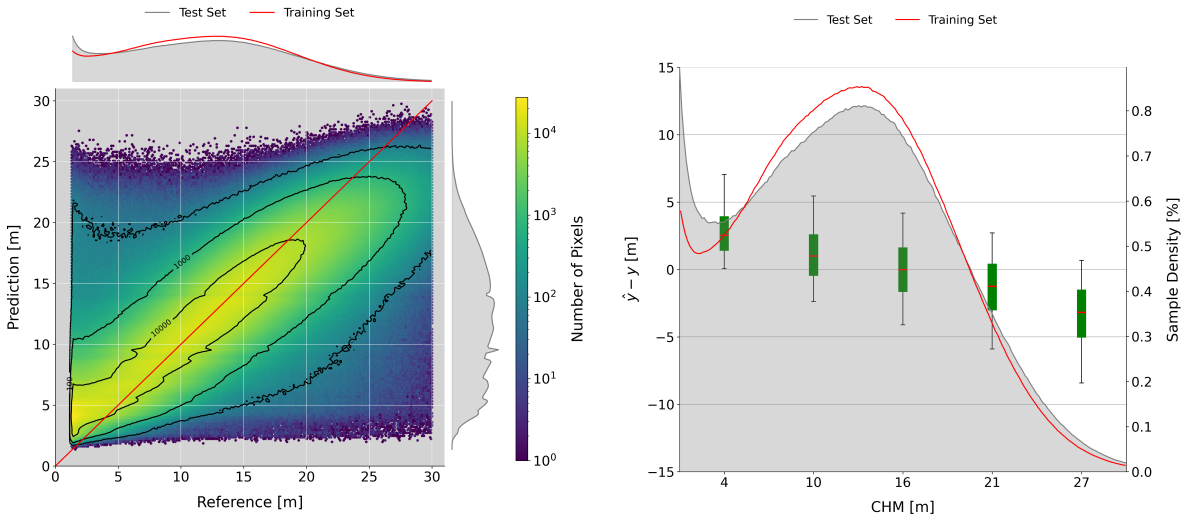


Figure 5.2: Visual representations of prediction performance for the baseline scenario. (a) Scatter-plot in logarithmic scale showing predicted versus reference canopy height values, with marginal distributions of reference, predicted and training samples; (b) mean errors, representing the regression bias, across Canopy Height Model (CHM) ranges. Green box plots represent the interquartile range (25th–75th percentiles), black lines extend to the 5th–95th percentile range, and red lines indicate the mean bias. Grey and red distribution represent sample counts to convey result reliability.

Figure 5.2b further explores the relationship between estimation bias and canopy height subranges, revealing a consistent overestimation for shorter canopies, transitioning to a progressive underestimation as canopy height increases. This pattern likely arises from an insufficient representation of shorter and taller trees in the training dataset.

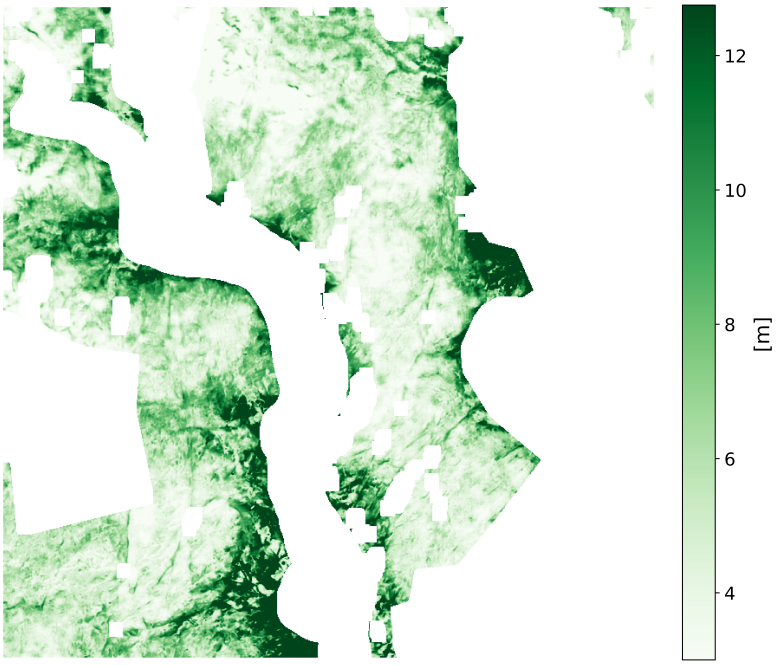


Figure 5.3: Estimated canopy height map, white areas represent non-forested regions or missing valid satellite data.

Finally, Figure 5.3 provides an example of an estimated canopy height map for a selected region of interest in the southwest of Norway, showcasing the model's ability to produce spatially detailed and consistent predictions. The corresponding input data are presented in Figure 4.2. To ensure consistency, all subsequent experiments utilise the same dataset subsetting, minimising variability in training and testing conditions.

5.3. The Impact of Forest Type

This section examines whether the model's performance depends on the inferred forest type. Despite its relatively low resolution, the CORINE Land Cover (CLC) map, described in Section 4.4, is leveraged, as it is deemed acceptable for the scope. The classification it provides is sufficiently detailed to distinguish forest types of interest (e.g., broadleaf, coniferous, mixed forests and others), and it was used to classify the test samples, allowing for separate statistical evaluations.

Figure 5.4 illustrates that no specific forest type consistently underperforms relative to others, highlighting the model's ability to generalise effectively across diverse vegetation classes.

Valuable insights emerge from the analysis within each category, where prediction errors are examined as a function of CHM within each CLC class. For reference, two examples are presented in Figure 5.5. Notably, in each category, the minimal prediction bias occurs for the tree heights with the highest sample density in the training data. This alignment suggests that the model effectively captures type-specific InSAR signatures during training and leverages this knowledge during testing to enhance regression accuracy. These findings demonstrate that the model extends beyond performing generalised regression by adapting

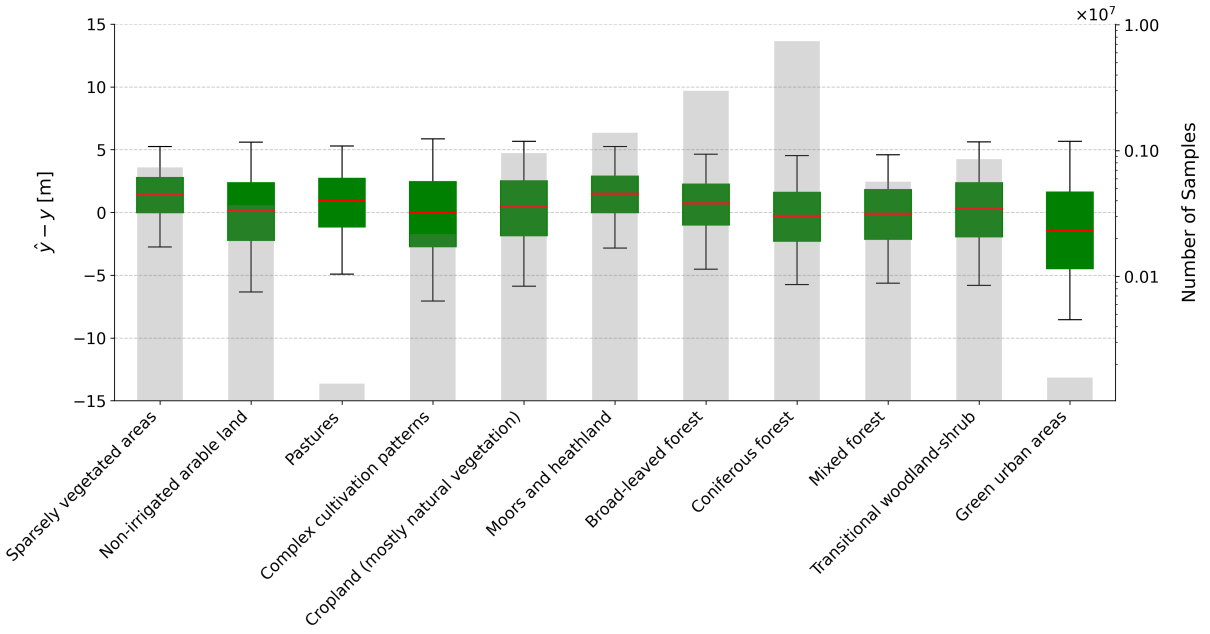


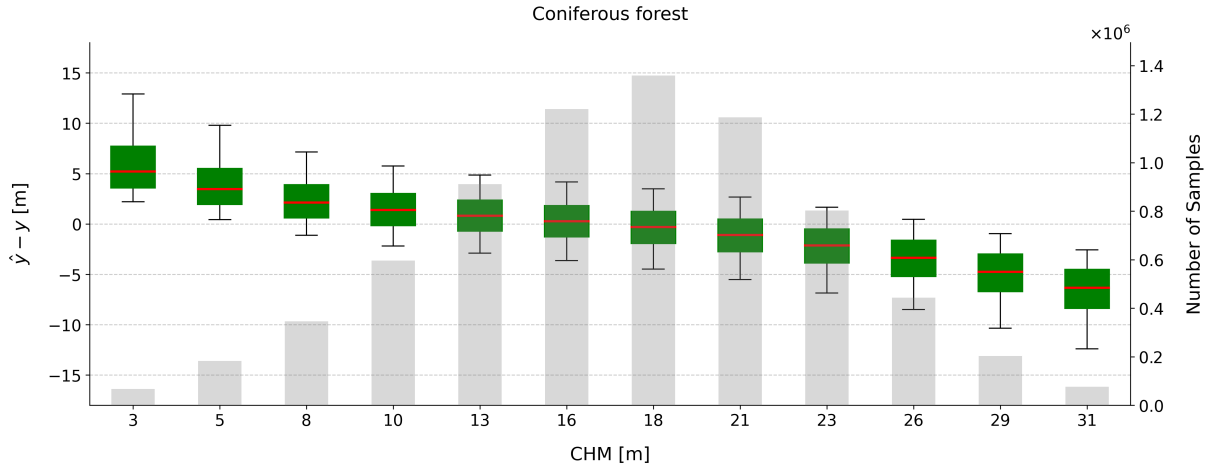
Figure 5.4: Mean errors and sample counts per CLC category. The sample count refers to the test dataset. $\hat{y}$ denotes the prediction and y the reference value.

its predictions based on the unique characteristics of each forest type. Furthermore, the model's ability to distinguish tree types through their InSAR signatures could be further explored and leveraged in future research for classification tasks.

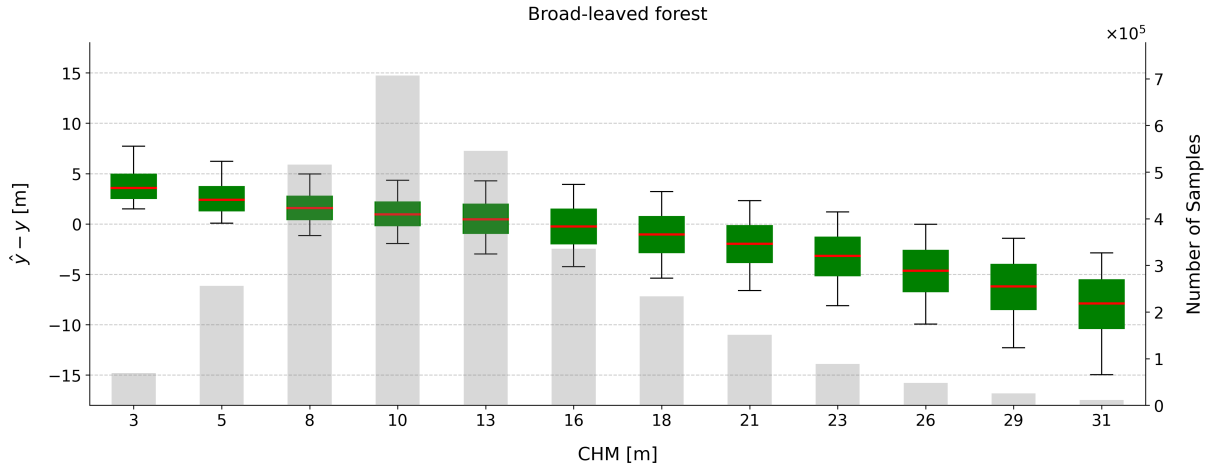
5.4. Inferring Vegetation Presence

It is well established that TanDEM-X DEM measurements (DEM_{ACQ}) provide absolute height information, representing the location of the mean phase centre within the canopy [17]. However, their relationship with actual canopy height is non-linear, as the signal penetration depth into the vegetation layer varies across resolution cells over vegetated areas. The network effectively leverages this variability as an implicit indicator of vegetation presence.

This experiment evaluates whether the network can extract equivalent meta-information directly from raw phase data (ϕ_{ACQ}). If successful, this approach would bypass preprocessing issues such as phase unwrapping errors, which often introduce discontinuities in topographic data, while also removing dependencies on absolute elevation, which have been shown to hinder generalisation at the national scale [17]. By using the sine and cosine of the phase rather than the directly acquired phase values, the network autonomously resolves ambiguities, further enhancing its adaptability.



(a)



(b)

Figure 5.5: Mean error distribution across CHM ranges for: (a) coniferous forest; (b) broad-leaved forest. The sample count refers to the test dataset.

The results, summarised in Table 5.2[1], show that both $\sin\phi_{\mathrm{ACQ}}$ and $\cos\phi_{\mathrm{ACQ}}$, as well as $\mathrm{DEM}_{\mathrm{ACQ}}$, yield equivalent outcomes, indicating no advantage in selecting one representation over another in terms of predictive capability.

[1]The results shown already benefit from the insights gained in Section 5.5.

ID	MAE	MAE%	RMSE	RMSE%	ME	ME%	R^2
	[m]	[%]	[m]	[%]	[m]	[%]	[-]
$\sin\phi_{ACQ}$, $\cos\phi_{ACQ}$	2.32	19	3.08	25	0.05	0	0.75
DEM_{ACQ}	2.36	20	3.12	26	0.17	1	0.75

Table 5.2: Performance comparison between phase-based features and acquisition DEM as input to the model.

5.5. Temporal Consistency and Forest Dynamics

As highlighted in Chapter 4, this investigation relies on 322 TanDEM-X (TDX) acquisitions that do not perfectly overlap with NIBIO's ALS reference datasets, owing to temporal inconsistencies in acquisition periods for the same regions of interest. Building upon the insights presented in Section 5.4, the following analysis examines how the temporal proximity between input and label data during the training process affects the accuracy of vegetation height predictions. Under the assumption of *primary forest*, where the average canopy height remains constant over time due to a balance between tree growth and mortality, the investigation explores whether temporal mismatches — by introducing natural variability into the training data — hinder the learning process, making it more challenging for the network to extract meaningful patterns.

The empirical results, summarised in Table 5.3, illustrate the performance across short and long temporal intervals obtained by systematically adjusting the temporal proximity between input and label data in the training, validation and test datasets. The training and test data collected within one year[1] from the ALS reference data yield the best performance. Conversely, larger temporal gaps in the training data ($|\Delta T_{TDX,ALS}|^{max}_{Training}$ = 11 years) are associated with worse performance, as the model fails to reconcile outdated input data with more recent target conditions, and vice versa. To provide a comprehensive evaluation, two complementary scenarios are also examined, revealing that when both the training and test datasets exhibit similar noise levels, the model's performance is better than in cases where the test data are noisy and the training ones are temporally consistent. This outcome is likely attributable to the network's prior exposure to noisy patterns.

[1]Based on empirical estimates, a one-year period is chosen as it ensures sufficient training and test data to allow for reliable estimates while maintaining reasonable spatial coverage.

Training	Test	MAE	MAE%	RMSE	RMSE%	ME	ME%	R^2
$\lvert\Delta T_{TDX,ALS}\rvert^{max}$		[m]	[%]	[m]	[%]	[m]	[%]	[-]
1	1	2.26	19	2.98	25	0.07	1	0.77
11	1	2.35	21	3.04	27	0.46	4	0.73
11	11	2.43	21	3.19	27	0.35	3	0.73
1	11	2.51	28	3.34	28	-0.19	2	0.72

Table 5.3: Performance comparison across different training and test datasets configurations. The validation dataset follows the same configuration as the training dataset. $\lvert\Delta T_{TDX,ALS}\rvert^{max}$ is computed in years.

These findings underscore the need to tackle temporal inconsistencies during training, thereby reducing ambiguity arising from mismatched temporal associations. This consideration is particularly critical in environments undergoing rapid ecological or environmental transformations. Based on the previous results, only data with a maximum one-year gap are used for training from this point onward unless otherwise stated.

The analysis is further extended by examining model predictions on increasingly mismatched test sets. Specifically, the results are grouped by the number of years elapsed between the TanDEM-X and ALS acquisitions, where negative values indicate that the TanDEM-X data were acquired before the ALS reference ones. To enhance readability and identify clear trends, the results are aggregated in three-year intervals, as shown in Figure 5.6, illustrating the prediction error as a function of the time elapsed between the acquisitions. The box-plot reveals a clear pattern: larger positive temporal gaps (i.e., more recent satellite acquisitions relative to the ALS reference data) tend to result in overestimated predictions, particularly when the gap exceeds five years. On the other hand, when the TDX acquisitions are older than the reference data, there is an average underestimation of canopy height. It is, therefore, essential, when assessing potential over- or underestimations of canopy height, to interpret such deviations within the context of ongoing ecological changes, as they also reflect real-world forest dynamics rather than solely arising from modelling biases.

The theoretical assumption of *primary forest* does not hold in the Norwegian context, where large-scale reforestation initiatives and ecological changes over the past 80 years have significantly contributed to forest growth. Since the 1960s, nationwide afforestation programmes have introduced approximately 100 million spruce saplings annually, increasing forest coverage [84]. Concurrently, rising temperatures and elevated CO_2 con-

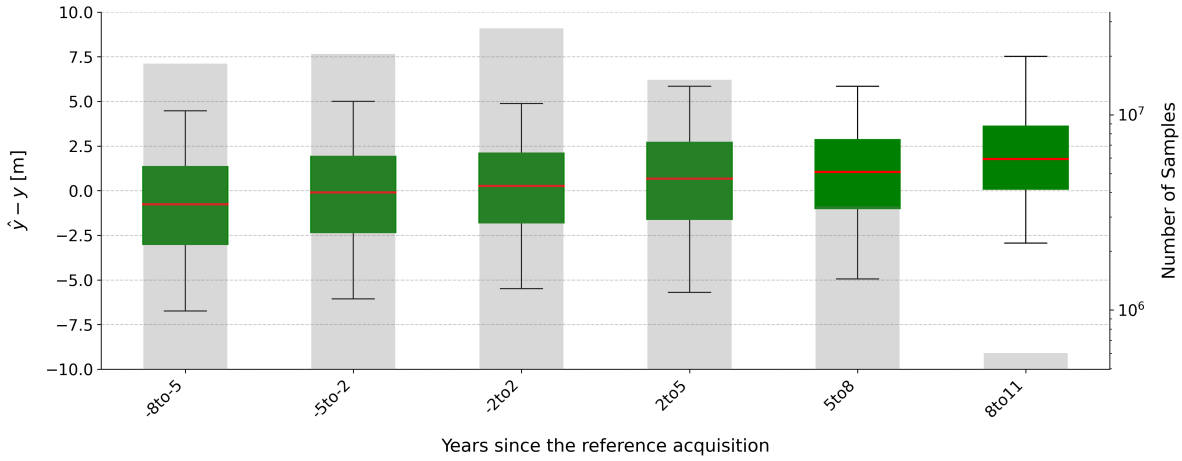


Figure 5.6: Mean errors based on time elapsed between TanDEM-X and NIBIO ALS reference data acquisition.

centrations have extended growing seasons, fostering favourable conditions for tree growth [85]. Additionally, reduced grazing pressure has facilitated natural regeneration and improved the survival rates of young saplings [85]. Modern Norwegian forests, therefore, differ significantly from primary forests, which could explain the observed patterns in the network's canopy height estimates over time.

These results highlight the potential of the current framework for forest monitoring, yet it is still far from its operational deployability. The observed trends are based on pointwise homoscedastic estimates with uncertainty values that have not been evaluated. This limitation presents significant challenges, as monitoring forest evolution would require addressing the propagation of these unquantified prediction uncertainties. A more structured analysis of prediction reliability, moving beyond the homoscedastic assumption, is essential for achieving reliable temporal change monitoring. The following chapter explores heteroscedastic approaches, marking a step forward towards improving the framework's effectiveness in monitoring forest dynamics.

6 | Bayesian Framework Extension

The end of Chapter 5 provides insights into estimation uncertainty — previously overlooked — and its growing impact when shifting from mapping canopy height to monitoring its temporal variations. By comparing multiple predictions over time, uncertainties accumulate, such that only changes exceeding the shrinking confidence threshold can be deemed reliable. Moreover, accounting for uncertainty differentiates between high-confidence and low-confidence estimates, which is essential for downstream applications. In this context, assessing the prediction-associated uncertainty becomes as important as the prediction itself. This chapter adopts a Bayesian perspective to explicitly estimate well-calibrated uncertainties alongside tree height predictions.

Chapter 2 presented a theoretical overview of Bayesian Deep Learning, detailing its mathematical formulation and distinguishing between aleatoric and epistemic uncertainty, before introducing Bayesian Model Averaging as a tool to derive tangible estimations. Given the absence of a closed-form solution, the focus now shifts to approximation strategies and their respective trade-offs.

The discussion begins by investigating viable approaches for capturing aleatoric uncertainty and reviewing their architectural implications. Since conventional regression metrics do not assess the quality of uncertainty estimates, *uncertainty calibration* is introduced as a key benchmark to measure how accurately predicted uncertainties reflect actual errors.

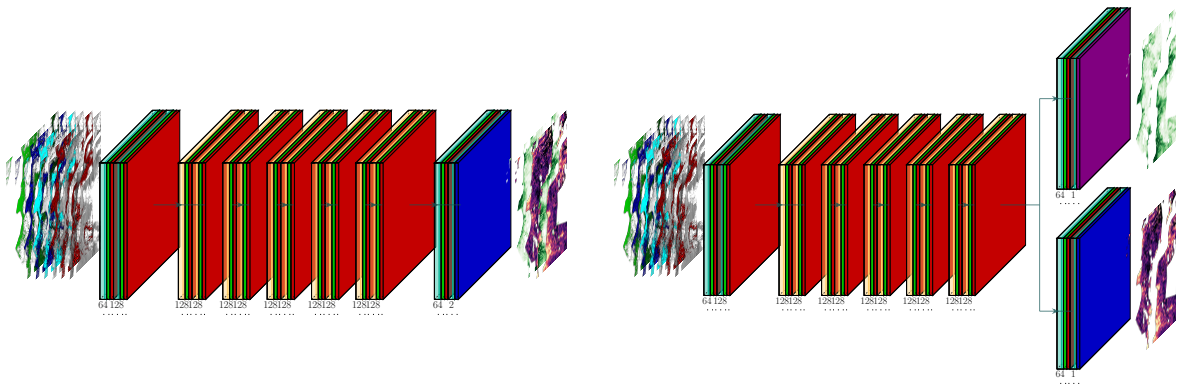
Nevertheless, aleatoric uncertainty alone does not provide a complete uncertainty-aware framework. Consequently, the discussion extends to epistemic uncertainty estimation, offering a comparative analysis of the techniques introduced in Chapter 2. The analysis covers uncertainty calibration performance and evaluates the computational costs associated with each approach — critical factors for deployability. Finally, an Out-Of-Distribution (OOD) analysis is conducted to assess the model's trustworthiness in scenarios beyond the learned distribution.

6.1. Quantifying the Unknown

6.1.1. Intrinsic Noise

The focus is first placed on modelling the uncertainty inherent in the observations — i.e., the aleatoric component introduced in Section 2.3.1. Following [13], predictions are reframed as *Gaussian posterior distributions*, where the network regresses their two defining parameters: the mean, corresponding to the *Maximum A Posteriori* (MAP) estimate, and the variance, which explicitly captures data-related uncertainty.

Consequently, the SILVA model (introduced in Section 5.1) and the loss function are adapted accordingly. Two primary architectural configurations are investigated: a *Dual-Head* (DH) design where each parameter is regressed independently [13], and a *Combined-Head* (CH) design that outputs the mean and the variance as a unified tensor. The dual-head design offers greater flexibility, allowing for tailored architectural choices within each branch. Conversely, the combined-head approach enforces a shared representation, potentially capturing common patterns more efficiently but remaining more generalised and sensitive. Given the absence of a definitive theoretical preference, both configurations are explored across multiple experimental settings. The two possible architectures are illustrated in Figure 6.1.



(a) (b)

Figure 6.1: Bayesian SILVA architectures: (a) combined-head; (b) double-head. The colour legend is depicted in Figure 5.1.

The loss function follows the formulation proposed in [13], maximising the posterior probability of the network parameters given the training data. Its derivation is rooted in Bayesian principles, where a zero-mean isotropic Gaussian prior is imposed over the network parameters (corresponding to $\mathcal{L}_2$ regularisation). For the i-th sample, the Gaussian

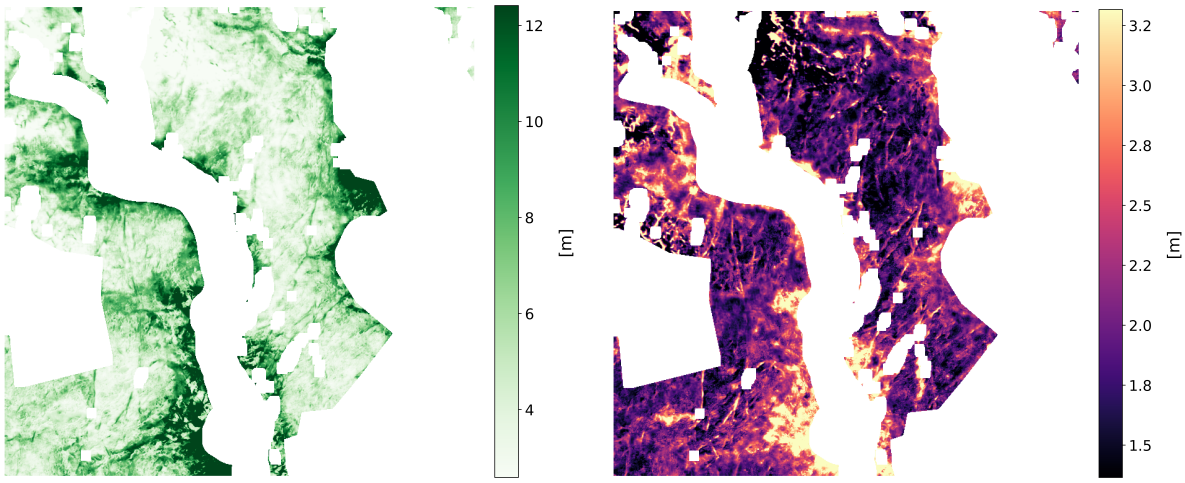
posterior is characterised by a mean $\hat{\mu}_i := \hat{\mu}(\boldsymbol{x}_i; \boldsymbol{\theta})$ and a variance $\hat{\sigma}_i^2(\boldsymbol{x}_i; \boldsymbol{\theta})$. The latter is parametrised through its logarithm as $\hat{s}_i := \log \hat{\sigma^2}_i(\boldsymbol{x}_i; \boldsymbol{\theta})$, ensuring numerical stability while enforcing strictly positive values. The resulting loss function is formulated as follows:

$$\mathcal{L}(\mathcal{D}; \boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_2^2 + \sum_i \left[\hat{s}_i + \exp(-\hat{s}_i)(\hat{\mu}_i - y_i)^2\right] \tag{6.1}$$

where y_i is the target canopy height value and $\lambda = 10^{-4}$ is a regularisation parameter controlling the impact of the $\mathcal{L}_2$ penalty on the network weights. This formulation allows noisy observations to be explained through larger predicted variances (i.e., uncertainties), giving the model greater flexibility in interpreting input data and enhancing its robustness against noisy labels. A detailed derivation and further discussion of Equation (6.1) are provided in Appendix A.

The training remains fully supervised using ALS-derived height values, with further implementation details, including optimiser settings, the learning rate schedule and stopping criteria, provided in Section 5.1.2.

To provide a tangible understanding of the enhanced Bayesian SILVA model's behaviour, an estimated canopy height map ($\hat{\mu}$) is presented alongside the corresponding uncertainty map ($\hat{\sigma}$) in Figure 6.2. The analysis is conducted in the southwest of Norway, covering the same region as in Section 5.2. The corresponding input data are shown in Figure 4.2.



(a) (b)

Figure 6.2: Estimated canopy height map in (a) and uncertainty map in (b), where white areas indicate non-forested regions or missing valid satellite data.

The Logarithmic Space

A careful analysis of Figure 6.1b reveals that, unlike the original SILVA model, the dual-head architecture employs an exponential activation function in the canopy height regression head, while the variance estimation branch retains a linear activation (i.e., a bypass). The rationale behind this design choice is rooted in the mathematical considerations discussed in Section 6.1.1: since the variance head explicitly predicts the log-variance, the network operates in a logarithmic space, diverging from SILVA's original linear approach for height regression. To maintain consistency across all network pathways, an exponential activation function is applied exclusively to the canopy height head, ensuring alignment within the logarithmic domain while preserving the correct scale of height estimates.

While this design choice offers theoretical advantages, it cannot be applied to the combined-head architecture. Therefore, to assess its practical impact and distinguish whether the observed effects stem from the logarithmic domain or the activation function properties, both linear and exponential activation functions are tested in the canopy height regression branch of both the dual-head and Single-Head (SH) architectures, with the latter representing the original SILVA model.

ID	Activation	MAE	RMSE	ME	R^2	TPE	ECE
		[m]	[m]	[m]	[-]	[%]	[-]
DH	**EXP**	2.41	3.24	-0.26	0.74	92	0.24
DH	**LIN**	2.42	3.22	-0.03	0.74	92	0.24
SH	**EXP**	2.41	3.20	0.03	0.75	-	-
SH	**LIN**	2.42	3.20	0.11	0.75	-	-

Table 6.1: Performance comparison between linear and exponential activation function for canopy height regression.

The results, presented in Table 6.1, show that the network does not exhibit significant benefits from the exponential activation function in the single-head configuration nor from enforcing domain consistency in the dual-head architecture. While more pronounced fluctuations are observed in the mean error, no clear trend emerges. These variations are likely attributed to different runs, resulting in distinct yet equally valid parameter sets for the network. It is worth noting that performance slightly deteriorates compared to the original SILVA framework (e.g., Root Mean Squared Error (RMSE) increases from 3.12 metres[1] to 3.22 metres), as the model is now optimising two outputs simultaneously.

[1]Refer to Section 5.4

For consistency, all subsequent experiments operate on the same dataset subsetting, minimising variability in training and testing conditions.

Uncertainty Calibration

Assessing the reliability of the uncertainty estimates involves evaluating their alignment with actual prediction errors. In a perfectly calibrated model, the predicted variance should correspond, on average, to the squared deviation between the predicted and actual values, as dictated by the variance definition.

To empirically evaluate calibration, predictions are grouped into bins according to their predicted variance. The mean squared error of the predictions within each bin is compared to the mean predicted variance. To align with the measurement scale, both quantities are expressed in their root form, yielding the RMSE and the Root Mean Variance (RMV). A well-calibrated model is expected to exhibit good agreement between RMSE and RMV across uncertainty bins, indicating that the predicted uncertainty accurately reflects the actual prediction error. Two key metrics are introduced to quantify calibration performance. The first is the *Total Predicted Error* (TPE) percentage, defined as:

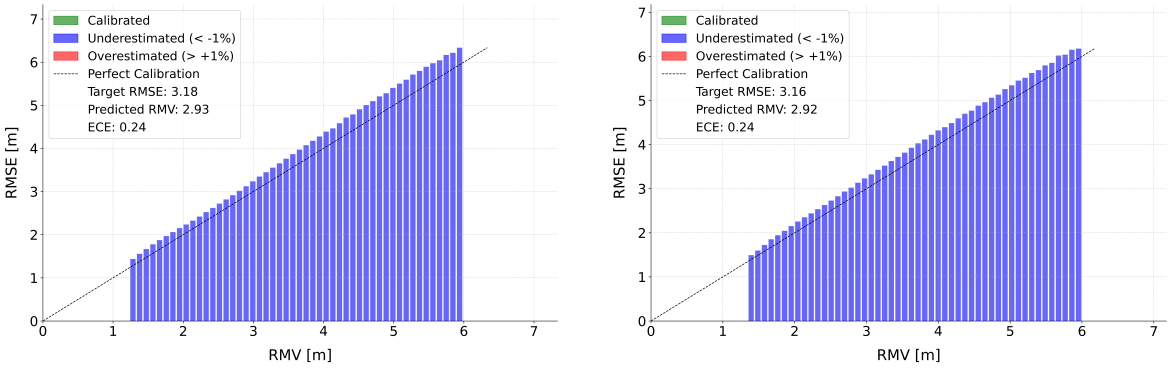
$$\text{TPE} = \frac{\text{RMV}}{\text{RMSE}} \times 100 \tag{6.2}$$

which assesses whether the model is generally over or under-calibrated. The second metric is the *Expected Calibration Error* (ECE), which quantifies the uniformity of calibration errors across uncertainty levels. It is computed based on RMSE across the uncertainty bins as:

$$\text{ECE} = \sum_{m=1}^{M} w_m \left| \frac{\text{RMV}_m}{\text{RMSE}_m} - 1 \right| \tag{6.3}$$

where M represents the number of bins, w_m denotes the proportion of samples in each bin, and $\frac{\text{RMV}_m}{\text{RMSE}_m}$ measures the relative calibration error within each bin. An ideal calibration would be characterised by a TPE of 100% and an ECE of 0.

Figure 6.3 presents the calibration results for the dual-head models: (a) employing the exponential activation function and (b) the linear one, both using 50 uncertainty bins. The results exhibit a modest underestimation of uncertainty, indicative of overconfident behaviour across all uncertainty levels, which is consistent with the existing literature [12] [13] [86]. Nevertheless, the results reported in Table 6.1 demonstrate a satisfactory calibration performance, with both configurations achieving a TPE of 92% and an ECE of 0.24.



(a) (b)

Figure 6.3: Calibration plots for dual-head models with (a) an exponential activation function and (b) a linear activation function on the canopy height regression branch.

6.1.2. The Limits of Knowledge

Uncertainty in predictive models extends beyond the intrinsic noise in the data. While aleatoric uncertainty can be estimated directly from observations, epistemic uncertainty arises from the model's inherent lack of knowledge, stemming from both insufficient training samples and structural assumptions that constrain its ability to generalise beyond observed examples. — Yet, can a model recognise the limits of its own knowledge? — Expecting a deterministic network to infer its epistemic uncertainty is analogous to asking a cognitive system to evaluate the biases embedded in its reasoning: an assessment that is impossible without external feedback. Just as a cognitive system cannot inherently delineate the boundaries of its own knowledge, a neural network lacks the intrinsic capacity to quantify the uncertainty arising from its structural limitations and inductive biases.

Bayesian-based techniques provide a theoretically grounded framework for the robust estimation of epistemic uncertainty by exploring prediction variations across multiple plausible model configurations. A key tool is *Bayesian Model Averaging* (BMA), as detailed in Section 2.3, which aims to compute the posterior predictive distribution by integrating over all possible parameter configurations rather than relying on a single parameter set. As an analytical solution to this integral is intractable, it is typically approximated through *Monte Carlo integration*, where multiple parameter realisations — sampled from an approximation of the posterior distribution over the model weights — are aggregated to derive an approximation of the posterior predictive distribution. This formulation enables the generation of MAP estimates alongside well-calibrated measures of predictive uncertainty. Given M sampled models, the MAP estimates ($\hat{\boldsymbol{\mu}}_*$) for the image $\boldsymbol{x}_*$ are

aggregated by computing their mean ($\bar{\boldsymbol{\mu}}_*$):

$$\mathbb{E}[\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D}] \approx \frac{1}{M}\sum_{k=1}^{M} \hat{\boldsymbol{\mu}}_{*,k} =: \bar{\boldsymbol{\mu}}_* \tag{6.4}$$

while the variances ($\hat{\boldsymbol{\sigma}}_*^2$) are combined using *the law of total variance*:

$$\mathrm{Var}[\boldsymbol{y}_*|\boldsymbol{x}_*, \mathcal{D}] \approx \frac{1}{M}\sum_{k=1}^{M} \left[\hat{\boldsymbol{\sigma}}_{*,k}^2 + (\hat{\boldsymbol{\mu}}_{*,k} - \bar{\boldsymbol{\mu}}_*)^2\right] =: \bar{\boldsymbol{\sigma}}_*^2 \tag{6.5}$$

The following experiments assess Stochastic Weight Averaging-Gaussian (SWAG)[1], Deep Ensemble and Multi-SWAG approximations, evaluating both regression performance and uncertainty calibration. Table 6.2 shows the results, extending the comparison to the MAP and the Stochastic Weight Averaging (SWA) solutions, which serve as baselines for regression performance assessment.

The SWA solution slightly improves regression performance over the MAP one by averaging across multiple well-performing models, reducing specific parameter fluctuations, and yielding a more stable and generalisable solution. However, its impact on calibration remains marginal, with no clear improvement trend across metrics. Progressively increasing the representational complexity of the solution landscape improves calibration while slightly enhancing regression performance. This establishes Multi-SWAG as the best-performing approach in both aspects. However, this comes at a higher computational cost, with a fivefold increase in training time (five models versus one) and inference scaling from 1 to 25. While inference overhead is less critical in experimental setups due to the additional cost of computing statistics, it becomes a key factor in deployment scenarios. For the sake of brevity, Table 6.2 focuses on the linear activation case, as both linear and exponential activation functions exhibit similar behaviour. This is illustrated in Figure 6.4, which provides a side-by-side graphical comparison for clarity and completeness, confirming comparable trends across different methodologies.

As discussed earlier, the computational cost is the primary bottleneck of BMA, increasing training and inference times. Graphical analyses, such as those in Figure 6.4, illustrate performance variations with respect to the number of sampled models and/or basins, aiding in balancing performance against computational constraints. The figure presents results for the linear and exponential activation functions, showing that while they follow similar trends across methodologies, the linear one converges more quickly and achieves

[1]The simplified *SWAG-Diagonal* variant [38] is used, assuming a diagonal covariance matrix.

ID	MAP	SWA	Ensemble	SWAG	M.SWAG	MAE	RMSE	ME	R^2	TPE	ECE
						[m]	[m]	[m]	[-]	[%]	[-]
SH	○	○	○	○	○	2.42	3.20	0.11	0.75	-	-
SH	○	●	○	○	○	2.40	3.19	0.09	0.75	-	-
SH	○	○	○	●	○	2.40	3.19	0.10	0.75	-	-
DH	●	○	○	○	○	2.42	3.22	-0.03	0.74	92	0.24
DH	○	●	○	○	○	2.41	3.21	0.03	0.75	97	0.22
DH	○	○	○	●	○	2.42	3.22	-0.09	0.74	94	0.19
DH	○	○	○	○	●	2.37	3.15	0.02	0.75	99	0.04
DH	○	○	●	○	○	2.39	3.17	0.07	0.75	98	0.04
CH	●	○	○	○	○	2.39	3.20	-0.16	0.75	93	0.22
CH	○	●	○	○	○	2.39	3.19	-0.15	0.75	92	0.26
CH	○	○	○	●	○	2.39	3.18	-0.02	0.75	94	0.19
CH	○	○	○	○	●	2.36	3.15	-0.07	0.76	98	0.07
CH	○	○	●	○	○	2.37	3.16	0.02	0.75	97	0.07

Table 6.2: Comparison of single-head, double-head, and combined-head architectures with the linear activation function. *MAP* – Maximum a posteriori estimate. *SWA* – Stochastic Weight Averaging solution. *Ensemble* – Deep Ensemble of five independently trained models. *SWAG* – derived from a single Deep Ensemble run, with five samples extracted. *M.SWAG* – Multi-SWAG, obtained by sampling five models per basin from the five Deep Ensemble runs.

slightly better overall performance, aligning with the findings of Section 6.1.1.

Optimizer

The original SWA [39] and SWAG [38] papers derive their methodology under the assumption that Stochastic Gradient Descent (SGD) is the optimiser, as their mathematical formulation relies on its properties. Up to this point, all models have been trained using Adaptive Moment Estimation (ADAM), a variant of SGD detailed in Section 2.2.2. This comparative evaluation examines the impact of this design choice on the case study.

Table 6.3 shows that SGD improves uncertainty calibration, yielding higher TPE scores and lower ECE values. Its noisier convergence leads to broader optimisation trajectories, increasing weight variability and enabling a higher-variance Gaussian approximation in the SWAG methodology, thus better capturing epistemic uncertainty. However, SWAG

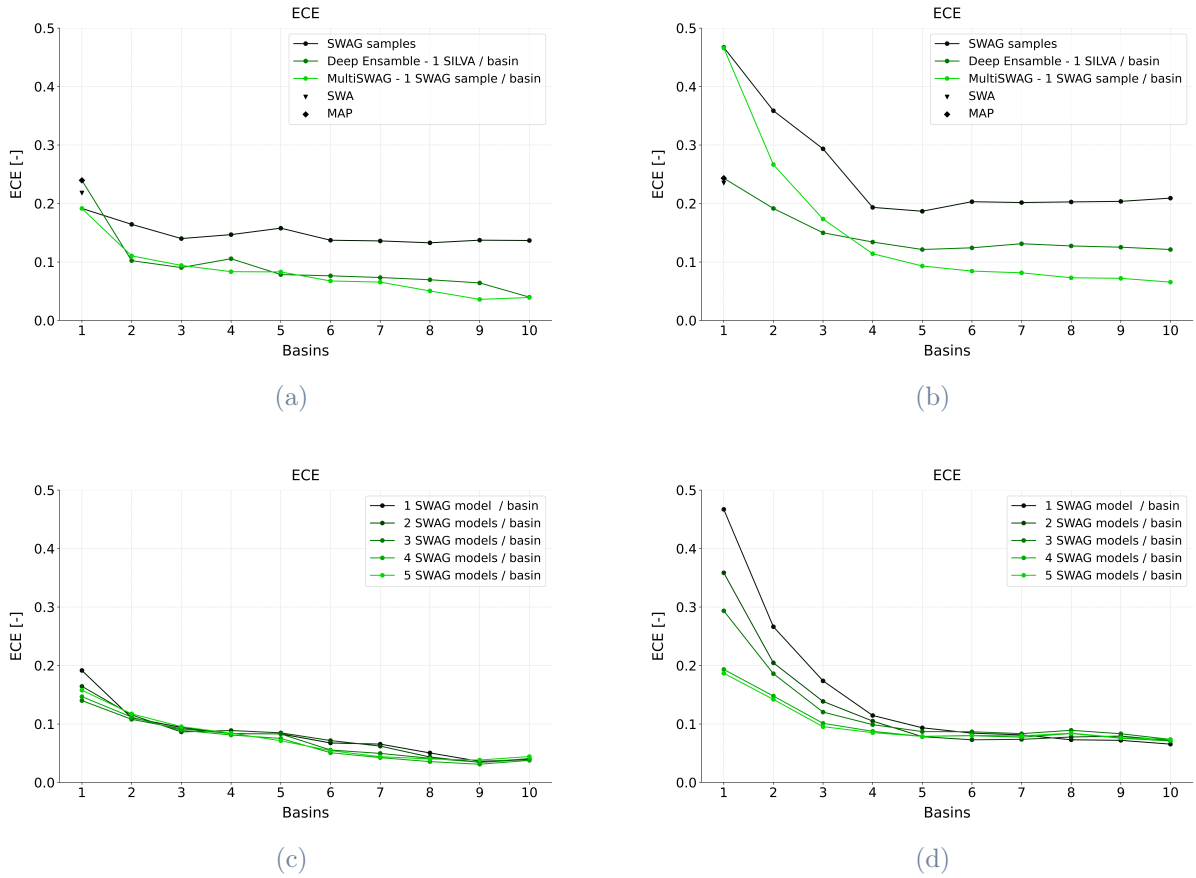


Figure 6.4: Graphical representations of: (top row) comparison between SWAG, Deep Ensemble, and Multi-SWAG approximations alongside SWA and MAP solutions; (bottom row) detailed analysis of Multi-SWAG. Performance variations are shown as a function of the number of sampled models and/or basins, with ECE reported as an example. Figures on the left (a), (c) correspond to the linear case, while those on the right (b), (d) refer to the exponential one.

provides little additional benefit when combined with SGD, as it samples from the same explored region, limiting its ability to refine uncertainty estimation. Improved calibration with SGD, however, comes at the cost of higher RMSE. In contrast, ADAM converges to sharper minima, enhancing regression accuracy but leading to greater overconfidence in uncertainty estimates.

These findings highlight the crucial role of the optimiser in shaping predictive confidence beyond the influence of post-processing techniques. While SGD enhances uncertainty estimation, ADAM remains preferable for optimising regression performance, particularly when paired with an explicit epistemic uncertainty modelling approach.

ID	Optimizer	SWAG	MAE	RMSE	ME	R^2	$\frac{RMV}{RMSE}$	ECE
			[m]	[m]	[m]	[-]	[%]	[-]
SH	ADAM	○	2.42	3.20	0.11	0.75	-	-
SH	ADAM	•	2.40	3.19	0.10	0.75	-	-
SH	SGD	○	2.86	3.78	0.30	0.65	-	-
SH	SGD	•	2.85	3.77	0.06	0.65	-	-
DH	ADAM	○	2.42	3.22	-0.03	0.74	92	0.24
DH	ADAM	•	2.42	3.22	-0.09	0.74	94	0.19
DH	SGD	○	2.62	3.50	0.09	0.70	95	0.16
DH	SGD	•	2.62	3.49	0.11	0.70	95	0.15

Table 6.3: Performance comparison between SGD and ADAM in single-head and double-head architectures. *SWAG* – comprising five extracted samples.

Batch Normalization Recalibration

The original SWA [39] and SWAG [38] papers propose a Batch Normalization (BN) recalibration step as part of their methodologies. This step is deem necessary because the averaging process in SWA and the posterior sampling in SWAG disrupt the BN statistics accumulated during the training. The recalibration procedure involves running the entire training dataset through the model in training mode, without backpropagation, to update the BN layers' running statistics.

Experimental results indicate that while BN recalibration provides marginal improvements in regression performance, it degrades uncertainty calibration, reducing its metrics by a few percentage points. Given its negative impact on uncertainty calibration and substantial computational overhead — adding approximately one-third to one-half of the training time per sampled model — this step was deliberately omitted from the proposed framework.

6.2. Temporal Consistency

In this section, an additional experiment is conducted to assess whether the temporal misalignment between the input and the reference data — previously detailed in Section 4 and experimentally examined in Section 5.5 — can be leveraged as a structured source of noise. The objective is to expose the network to temporal inconsistencies during the training phase (i.e., noisy patterns), stimulating it to learn more robust data representations and better account for potential noise in real-world applications.

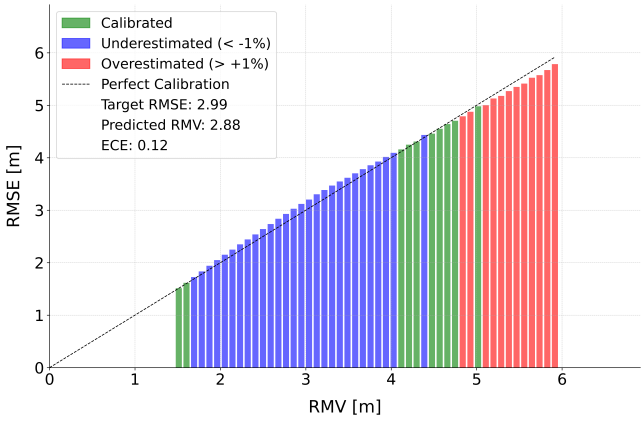


Figure 6.5: Calibration plot for the noisy training experiment.

ID	Noise	MAE	RMSE	ME	R^2	TPE	ECE
		[m]	[m]	[m]	[-]	[%]	[-]
DH	○	2.41	3.24	-0.26	0.74	92	0.24
DH	•	2.32	3.05	-0.07	0.75	96	0.12

Table 6.4: Performance comparison between noisy and standard training dataset.

The experiment is conducted on the dual-head linear model, which demonstrated the most promising performance in capturing aleatoric uncertainty in Section 6.1.2. Training is conducted on all available samples without applying the year-matching filter, introducing temporal gaps of up to 11 years. Conversely, validation and testing are limited to samples with a maximum temporal gap of one year, following the rationale outlined in Section 5.5.

The results, summarised in Table 6.4, suggest that exposure to structured noise during training enhances the regression performance. Regarding calibration, an improvement is observed; however, the calibration graph in Figure 6.5 reveals a more complex pattern: rather than a straightforward over/underestimation trend, large uncertainties tend to be overestimated, while smaller ones are underestimated — reflecting a more complex interaction between the introduced noise and the calibration performance, which remains

unexplained within the current scope of this research.

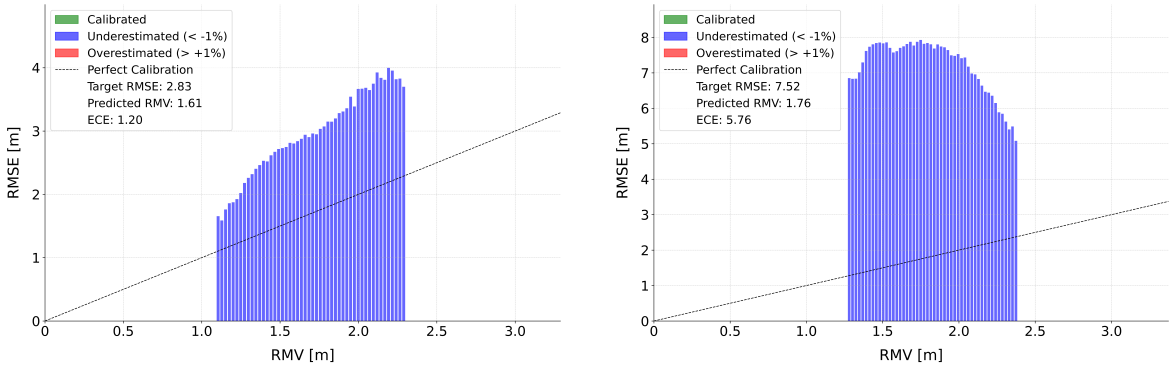
6.3. Out-Of-Distribution Analysis

This final experiment investigates how predicted uncertainty scales in Out-Of-Distribution (OOD) scenarios (i.e., when the model encounters data that significantly deviate from the training distribution) and whether it can help identify such instances.

The study is structured as follows: samples whose height of ambiguity (HOA) exceeds the median value of its distribution across the dataset (approximately 52 metres) are excluded from the training and validation subsets. Testing is then performed on in- and out-of-distribution samples, with uncertainty calibration assessed across different HOA and canopy height ranges. This distinction is necessary as uncertainty estimates typically scale with canopy height. Ideally, the model should assign significantly higher uncertainty to OOD samples, clearly distinguishing them from in-distribution ones.

Figure 6.6 compares two HOA ranges: in-distribution samples (31–46 metres) and OOD ones (61–76 metres). The results show that the model assigns comparable uncertainty levels to both scenarios. In the OOD case, this results in a low TPE and a strong deviation from the identity line in the calibration plots, further supported by elevated ECE values, indicating a lack of OOD detection capability. Calibration performance is affected by the constrained training setup: removing high HOA samples alters the overall dataset distribution, limiting the model's generalisation ability, while the small number of samples per HOA and canopy height ranges reduces the statistical robustness of the calibration estimates. This is reflected in poorer performance in the in-distribution case.

At the current stage, the model does not inherently recognise when making predictions outside its learned distribution, reinforcing the need for dedicated OOD detection strategies for reliable deployment.



(a) (b)

Figure 6.6: Calibration plots for dual-head models with the exponential activation function: (a) shows in-distribution samples (HOA 31–46 metres), while (b) presents OOD cases (HOA 61–76 metres). Results refer to the canopy height range of 10.9-20.4 metres.

7 | Conclusion and Outlook

This Master's thesis presented a novel Bayesian-based approach for forest height estimation from TanDEM-X single-pass Interferometric Synthetic Aperture Radar (InSAR) data.

After introducing the Norwegian forest ecosystem in Chapter 4, along with the datasets used and the respective processing pipelines, the thesis explored in Chapter 5 its first central theme: the challenge of generalising beyond single-nation studies. The SILVA framework demonstrated robust generalisation capabilities across diverse Norwegian ecosystems, suggesting its potential for broader geographical application. A key finding of this research was the neural network's ability to differentiate between tree types based on their InSAR signatures, paving the way for the integration of high-resolution classification tasks within the SILVA framework. However, scaling to continental and global scenarios requires implementing ad hoc strategies for cross-ecosystem generalisation. In this context, transfer learning, carefully designed data sampling strategies and architectural expansion remain promising approaches. Then, this study turned to the challenge of long-term forest monitoring, demonstrating that periodic satellite acquisitions can effectively track forest height changes over time, yet emphasising that their reliability depends on the uncertainty underlying the estimation process. Therefore, the investigation into methods that not only provide accurate predictions but also quantify their associated uncertainty became the central focus of the final chapter of this thesis.

Chapter 6 presented a Bayesian interpretation of the estimation process, with Multi-SWAG proving to be the most effective approximation method for capturing the shape of the solution landscape, improving confidence estimation and refining regression performance. Still, the Out-Of-Distribution (OOD) detection study revealed significant shortcomings in identifying OOD scenarios, highlighting the need for more advanced detection mechanisms. This thesis identified computational costs as a significant barrier to large-scale implementation, as ensemble-based Bayesian methods incur substantial overhead. Therefore, future research should prioritise solutions that enhance real-world deployability, with knowledge distillation, teacher-student paradigms and noise injection emerging

as promising approaches.

While this Master's thesis laid the foundation for uncertainty-aware forest monitoring, it represents only an initial exploration of Bayesian Deep Learning applications in the InSAR domain. Overcoming critical challenges, such as cross-ecosystem generalisation, lightweight uncertainty estimation and OOD detection, will be essential for developing operationally viable products capable of large-scale forest monitoring, ultimately supporting evidence-based forestry management and conservation policy decisions.

Bibliography

[1] Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment 2020*. FAO, Rome, 2020.

[2] Kim Calders, Jennifer Adams, John Armston, Harm Bartholomeus, Sebastien Bauwens, Lisa Patrick Bentley, Jerome Chave, F. Mark Danson, Miro Demol, Mathias Disney, Rachel Gaulton, Sruthi M. Krishna Moorthy, Shaun R. Levick, Ninni Saarinen, Crystal Schaaf, Atticus Stovall, Louise Terryn, Phil Wilkes, and Hans Verbeeck. Terrestrial laser scanning in forest ecology: Expanding the horizon. *Remote Sensing of Environment*, 251:112102, December 2020.

[3] K. M. Bergen, S. J. Goetz, R. O. Dubayah, G. M. Henebry, C. T. Hunsaker, M. L. Imhoff, R. F. Nelson, G. G. Parker, and V. C. Radeloff. Remote sensing of vegetation 3-d structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions. *Journal of Geophysical Research: Biogeosciences*, 114(G2), June 2009.

[4] Ralph Dubayah, James Bryan Blair, Scott Goetz, Lola Fatoyinbo, Matthew Hansen, Sean Healey, Michelle Hofton, George Hurtt, James Kellner, Scott Luthcke, John Armston, Hao Tang, Laura Duncanson, Steven Hancock, Patrick Jantz, Suzanne Marselis, Paul L. Patterson, Wenlu Qi, and Carlos Silva. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. *Science of Remote Sensing*, 1:100002, June 2020.

[5] World Wide Fund for Nature (WWF). Deforestation fronts: Drivers and responses in a changing world. `https://files.worldwildlife.org/wwfcmsprod/files/Publication/file/9bsfj8aq5v_deforestation_fronts___drivers_and_responses_in_a_changing_world___summary_english.pdf`, 2021.

[6] Intergovernmental Panel on Climate Change (IPCC). Climate change and land: An ipcc special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. `https://www.ipcc.ch/srccl/`, 2019.

[7] Nico Lang, Konrad Schindler, and Jan Dirk Wegner. Country-wide high-resolution vegetation height mapping with sentinel-2. *Remote Sensing of Environment*, 233:111347, November 2019.

[8] Mats Nilsson. Estimation of tree heights and stand volume using an airborne lidar system. *Remote Sensing of Environment*, 56(1):1–7, April 1996.

[9] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, March 2013.

[10] R. N. Treuhaft, B. D. Chapman, J. R. dos Santos, F. G. Gonçalves, L. V. Dutra, P. M. L. A. Graça, and J. B. Drake. Vegetation profiles in tropical forests from multibaseline interferometric synthetic aperture radar, field, and lidar measurements. *Journal of Geophysical Research: Atmospheres*, 114(D23), December 2009.

[11] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 3369–3378. IEEE, June 2018.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[13] Alexander Becker, Stefania Russo, Stefano Puliti, Nico Lang, Konrad Schindler, and Jan Dirk Wegner. Country-wide retrieval of forest structure from optical and sar satellite imagery with deep ensembles. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:269–286, January 2023.

[14] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[16] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *ArXiv*, abs/2002.08791, 2020.

[17] Daniel Carcereri. *A Deep Learning Study on the Retrieval of Forest Parameters from Spaceborne Earth Observation Sensors*. PhD thesis, Università degli studi di Trento, 2024.

[18] C. Oliver and S. Quegan. *Understanding Synthetic Aperture Radar Images*. SciTech Publ., 2004.

[19] I.G. Cumming and F.H. Wong. *Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation*. Artech House, 2005.

[20] Richard Bamler and Philipp Hartl. Synthetic aperture radar interferometry. *Inverse Problems*, 14(4):R1–R54, August 1998.

[21] S. Cloude. *Polarisation: Applications in Remote Sensing*. OUP Oxford, 2010.

[22] Richard M. Goldstein, Howard A. Zebker, and Charles L. Werner. Satellite radar interferometry: Two-dimensional phase unwrapping. *Radio Science*, 23:713–720, 1988.

[23] Carolina González, Markus Bachmann, José-Luis Bueso-Bello, Paola Rizzoli, and Manfred Zink. A fully automatic algorithm for editing the tandem-x global dem. *Remote Sensing*, 12(23):3961, December 2020.

[24] H.A. Zebker and J. Villasenor. Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, 30(5):950–959, 1992.

[25] Ridha Touzi, Armand Lopes, Jérôme Bruniquel, and Paris W. Vachon. Coherence estimation for sar imagery. *IEEE Trans. Geosci. Remote. Sens.*, 37:135–149, 1999.

[26] Gerhard Krieger, Alberto Moreira, Hauke Fiedler, Irena Hajnsek, Marian Werner, Marwan Younis, and Manfred Zink. Tandem-x: A satellite formation for high-resolution sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 45(11):3317–3341, 2007.

[27] Paola Rizzoli, Luca Dell'Amore, José-Luis Bueso-Bello, Nicola Gollin, Daniel Carcereri, and Michele Martone. On the derivation of volume decorrelation from tandem-x bistatic coherence. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3504–3518, 2022.

[28] K.P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning series. MIT Press, 2023.

[29] S.J.D. Prince. *Understanding Deep Learning*. MIT Press, 2023.

[30] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, December 1943.

[31] Marvin Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, September 2017.

[32] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989.

[33] D.E. RUMELHART, G.E. HINTON, and R.J. WILLIAMS. *Learning Internal Representations by Error Propagation*, page 399–421. Elsevier, 1988.

[34] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[35] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 315–323, 2011.

[36] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[37] Yan Wang, Xiaofu Wu, Yuanyuan Chang, Suofei Zhang, Quan Zhou, and Jun Yan. Batch normalization: Is learning an adaptive gain and bias necessary? In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, ICMLC 2018, page 36–40. ACM, February 2018.

[38] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019.

[39] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018.

[40] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016.

[42] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2020.

[43] Temilola Fatoyinbo, John Armston, Marc Simard, Sassan Saatchi, Michael Denbina, Marco Lavalle, Michelle Hofton, Hao Tang, Suzanne Marselis, Naiara Pinto, Steven Hancock, Brian Hawkins, Laura Duncanson, Bryan Blair, Christy Hansen, Yunling Lou, Ralph Dubayah, Scott Hensley, Carlos Silva, John R. Poulsen, Nicolas Labrière, Nicolas Barbier, Kathryn Jeffery, David Kenfack, Memiaghe Herve, Pulchérie Bissiengou, Alfonso Alonso, Ghislain Moussavou, Lee T.J. White, Simon Lewis, and Kathleen Hibbard. The nasa afrisar campaign: Airborne sar and lidar measurements of tropical forest structure and biomass in support of current and future space missions. *Remote Sensing of Environment*, 264:112533, October 2021.

[44] Tommaso Jucker, John Caspersen, Jérôme Chave, Cécile Antin, Nicolas Barbier, Frans Bongers, Michele Dalponte, Karin Y. van Ewijk, David I. Forrester, Matthias Haeni, Steven I. Higgins, Robert J. Holdaway, Yoshiko Iida, Craig Lorimer, Peter L. Marshall, Stéphane Momo, Glenn R. Moncrieff, Pierre Ploton, Lourens Poorter, Kassim Abd Rahman, Michael Schlund, Bonaventure Sonké, Frank J. Sterck, Anna T. Trugman, Vladimir A. Usoltsev, Mark C. Vanderwel, Peter Waldner, Beatrice M. M. Wedeux, Christian Wirth, Hannsjörg Wöll, Murray Woods, Wenhua Xiang, Niklaus E. Zimmermann, and David A. Coomes. Allometric equations for integrating remote sensing imagery into forest monitoring programmes. *Global Change Biology*, 23(1):177–190, July 2016.

[45] J.Bryan Blair, David L Rabine, and Michelle A Hofton. The laser vegetation imaging sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54(2–3):115–122, July 1999.

[46] S.S. SAATCHI, J. CHAVE, N. LABRIERE, N. BARBIER, M RÉJOU-MÉCHAIN, A. FERRAZ, and S. TAO. Afrisar: Aboveground biomass for lope, mabounie, mondah, and rabi sites, gabon, 2019.

[47] Thorsten Markus, Tom Neumann, Anthony Martino, Waleed Abdalati, Kelly Brunt, Beata Csatho, Sinead Farrell, Helen Fricker, Alex Gardner, David Harding, Michael Jasinski, Ron Kwok, Lori Magruder, Dan Lubin, Scott Luthcke, James Morison, Ross Nelson, Amy Neuenschwander, Stephen Palm, Sorin Popescu, CK Shum, Bob E.

Schutz, Benjamin Smith, Yuekui Yang, and Jay Zwally. The ice, cloud, and land elevation satellite-2 (icesat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*, 190:260–273, March 2017.

[48] Ibrahim Fayad, Nicolas Baghdadi, and Kamel Lahssini. An assessment of the gedi lasers' capabilities in detecting canopy tops and their penetration in a densely vegetated, tropical area. *Remote Sensing*, 14(13):2969, June 2022.

[49] E. P. W. Attema and Fawwaz T. Ulaby. Vegetation modeled as a water cloud. *Radio Science*, 13(2):357–364, March 1978.

[50] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, 2013.

[51] K.P. Papathanassiou and S.R. Cloude. Single-baseline polarimetric sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 39(11):2352–2363, 2001.

[52] A. Reigber and A. Moreira. First demonstration of airborne sar tomography using multibaseline l-band data. *IEEE Transactions on Geoscience and Remote Sensing*, 38(5):2142–2152, 2000.

[53] Matteo Pardini, Marivi Tello, Victor Cazcarra-Bes, Konstantinos P. Papathanassiou, and Irena Hajnsek. L- and p-band 3-d sar reflectivity profiles versus lidar waveforms: The afrisar case. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(10):3386–3401, 2018.

[54] Mauro Mariotti D'Alessandro and Stefano Tebaldini. Digital terrain model retrieval in tropical forests through p-band sar tomography. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6774–6781, 2019.

[55] Matteo Nannini, Michele Martone, Paola Rizzoli, Pau Prats-Iraola, Marc Rodriguez-Cassola, Andreas Reigber, and Alberto Moreira. Coherence-based sar tomography for spaceborne applications. *Remote Sensing of Environment*, 225:107–114, 2019.

[56] Shaun Quegan, Thuy Le Toan, Jerome Chave, Jorgen Dall, Jean-François Exbrayat, Dinh Ho Tong Minh, Mark Lomas, Mauro Mariotti D'Alessandro, Philippe Paillou, Kostas Papathanassiou, Fabio Rocca, Sassan Saatchi, Klaus Scipal, Hank Shugart, T. Luke Smallman, Maciej J. Soja, Stefano Tebaldini, Lars Ulander, Ludovic Villard, and Mathew Williams. The european space agency biomass mission: Measuring forest above-ground biomass from space. *Remote Sensing of Environment*, 227:44–60, 2019.

[57] Catherine Torres de Almeida, Jéssica Gerente, Jamerson Rodrigo dos Prazeres Cam-

pos, Francisco Caruso Gomes Junior, Lucas Antonio Providelo, Guilherme Marchiori, and Xinjian Chen. Canopy height mapping by sentinel 1 and 2 satellite images, airborne lidar data, and machine learning. *Remote Sensing*, 14(16):4112, August 2022.

[58] Zhanmang Liao, Albert I.J.M. Van Dijk, Binbin He, Pablo Rozas Larraondo, and Peter F. Scarth. Woody vegetation cover, height and biomass at 25-m resolution across australia derived from multiple site, airborne and satellite observations. *International Journal of Applied Earth Observation and Geoinformation*, 93:102209, December 2020.

[59] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[61] Luca Bergamasco, Francesca Bovolo, and Lorenzo Bruzzone. A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2147–2162, 2023.

[62] Elena Donini, Mattia Amico, Lorenzo Bruzzone, and Francesca Bovolo. Unsupervised semantic segmentation of radar sounder data using contrastive learning. In Nazzareno Pierdicca, Lorenzo Bruzzone, and Francesca Bovolo, editors, *Image and Signal Processing for Remote Sensing XXVIII*, page 23. SPIE, October 2022.

[63] Xiao Wang and Haipeng Wang. Forest height mapping using complex-valued convolutional neural network. *IEEE Access*, 7:126334–126343, 2019.

[64] Daniel Carcereri, Paola Rizzoli, Dino Ienco, and Lorenzo Bruzzone. A deep learning framework for the estimation of forest height from bistatic tandem-x data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:8334–8352, 2023.

[65] Daniel Carcereri, Paola Rizzoli, Luca Dell'Amore, José-Luis Bueso-Bello, Dino Ienco, and Lorenzo Bruzzone. Generation of country-scale canopy height maps over gabon using deep learning and tandem-x insar data. *Remote Sensing of Environment*, 311:114270, September 2024.

[66] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 300:113888, January 2024.

[67] Andrés C. Rodríguez, Stefano D'Aronco, Konrad Schindler, and Jan D. Wegner. Mapping oil palm density at country scale: An active learning approach. *Remote Sensing of Environment*, 261:112479, August 2021.

[68] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology and Evolution*, 7(11):1778–1789, September 2023.

[69] Gerhard Krieger, Alberto Moreira, Hauke Fiedler, Irena Hajnsek, Marian Werner, Marwan Younis, and Manfred Zink. Tandem-x: A satellite formation for high-resolution sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 45(11):3317–3341, November 2007.

[70] R. Werninghaus and S. Buckreuss. The terrasar-x mission and system design. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):606–614, February 2010.

[71] Gerhard Krieger, Manfred Zink, Markus Bachmann, Benjamin Bräutigam, Daniel Schulze, Michele Martone, Paola Rizzoli, Ulrich Steinbrecher, John Walter Antony, Francesco De Zan, Irena Hajnsek, Kostas Papathanassiou, Florian Kugler, Marc Rodriguez Cassola, Marwan Younis, Stefan Baumgartner, Paco López-Dekker, Pau Prats, and Alberto Moreira. Tandem-x: A radar interferometer with two formation-flying satellites. *Acta Astronautica*, 89:83–98, August 2013.

[72] Manfred Zink, Alberto Moreira, Irena Hajnsek, Paola Rizzoli, Markus Bachmann, Ralph Kahle, Thomas Fritz, Martin Huber, Gerhard Krieger, Marie Lachaise, Michele Martone, Edith Maurer, and Birgit Wessel. Tandem-x: 10 years of formation flying bistatic sar interferometry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3546–3565, 2021.

[73] Paola Rizzoli, Michele Martone, Carolina Gonzalez, Christopher Wecklich, Daniela Borla Tridon, Benjamin Bräutigam, Markus Bachmann, Daniel Schulze, Thomas Fritz, Martin Huber, Birgit Wessel, Gerhard Krieger, Manfred Zink, and Alberto Moreira. Generation and performance assessment of the global tandem-x digital ele-

vation model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 132:119–139, October 2017.

[74] B. E. Schutz, H. J. Zwally, C. A. Shuman, D. Hancock, and J. P. DiMarzio. Overview of the icesat mission. *Geophysical Research Letters*, 32(21), November 2005.

[75] Marie Lachaise, Carolina Gonzalez, Paola Rizzoli, Barbara Schweibhelm, and Manfred Zink. The new tandem-x dem change maps product. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, page 5432–5435. IEEE, July 2022.

[76] Michele Martone, Paola Rizzoli, and Gerhard Krieger. Volume decorrelation effects in tandem-x interferometric sar data. *IEEE Geoscience and Remote Sensing Letters*, 13(12):1812–1816, December 2016.

[77] Thomas Fritz, Helko Breit, Cristian Rossi, Ulrich Balss, Marie Lachaise, and Sergio Duque. Interferometric processing and products of the tandem-x mission. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, page 1904–1907. IEEE, July 2012.

[78] Pau Prats, Marc Rodriguez-Cassola, Luca Marotti, Matteo Naninni, Steffen Wollstadt, Daniel Schulze, Nuria Tous-Ramon, Marwan Younis, Gerhard Krieger, and Andreas Reigber. Taxi: A versatile processing chain for experimental tandem-x product evaluation. In *2010 IEEE International Geoscience and Remote Sensing Symposium*, page 4059–4062. IEEE, July 2010.

[79] Paola Rizzoli, Luca Dell'Amore, Jose-Luis Bueso-Bello, Nicola Gollin, Daniel Carcereri, and Michele Martone. On the derivation of volume decorrelation from tandem-x bistatic coherence. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3504–3518, 2022.

[80] W.G. Kropatsch and D. Strobl. The generation of sar layover and shadow maps from digital elevation models. *IEEE Transactions on Geoscience and Remote Sensing*, 28(1):98–107, 1990.

[81] NIBIO. National programme for airborne laser scanning in norway. Technical report, Norwegian Institute of Bioeconomy Research, 2015–2018.

[82] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino. Esa worldcover 10 m 2021 v200, 2022.

[83] European Environment Agency. Corine land cover. https://land.copernicus.eu/en/products/corine-land-cover.

[84] Norwegian Institute of Bioeconomy Research (NIBIO). Sustainable forest management: Impacts on water quality and climate change mitigation. *Environmental Science and Policy*, 114:1–10, 2020.

[85] Ketil Haugland and Hans-Petter Fjeldstad. Forest expansion in western norway: Climatic and land use drivers. *Norwegian Journal of Geography*, 69(4):1–13, 2015.

[86] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016.

A | Loss Function Derivation

The process of training the network involves identifying the most likely set of parameters ($\boldsymbol{\theta}$) given the training dataset ($\mathcal{D}$). By applying the Bayes' theorem in the logarithmic domain, the posterior distribution of $\boldsymbol{\theta}$ given $\mathcal{D}$ (Equation A.1) is decomposed into three distinct components: the prior, the data likelihood and the evidence (Equation A.2). The evidence term is omitted in Equation A.3, as it does not depend on $\boldsymbol{\theta}$. Under the standard assumption that each data point is *independent and identically distributed*, the overall likelihood factorises across the dataset, thereby yielding a sum over per-sample likelihood terms (Equation A.4).

As outlined in [13], Equation A.5 defines the model's predictions as a *Gaussian probability density function* and applies an isotropic *Gaussian prior* to the model's parameters, inherently enforcing $\mathcal{L}_2$ regularisation. The strength of the regularisation is determined by the prior's variance (σ_p^2).

$$\arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid D) \tag{A.1}$$

$$= \arg\max_{\boldsymbol{\theta}} \underbrace{\log p(\boldsymbol{\theta})}_{\text{prior}} + \underbrace{\log p(D \mid \boldsymbol{\theta})}_{\text{likelihood}} - \underbrace{\log p(D)}_{\text{evidence}} \tag{A.2}$$

$$= \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \log p(D \mid \boldsymbol{\theta}) \tag{A.3}$$

$$= \arg\min_{\boldsymbol{\theta}} - \log p(\boldsymbol{\theta}) - \sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \tag{A.4}$$

$$= \arg\min_{\boldsymbol{\theta}} - \log \mathcal{N}(\boldsymbol{\theta};\, \mathbf{0}, \sigma_p^2 \boldsymbol{I}) - \sum_{i=1}^{N} \log \mathcal{N}(y_i;\, \hat{\mu}_i, \hat{\sigma}_i^2) \tag{A.5}$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\sum_{i=1}^{N}\left[\log \sigma_i^2 + \frac{(\hat{\mu}_i - y_i)^2}{\hat{\sigma}_i^2} + \log 2\pi\right] \quad \text{with} \quad \lambda \propto \frac{1}{\sigma_p^2} \tag{A.6}$$

$$= \arg\min_{\boldsymbol{\theta}} \underbrace{\lambda\|\boldsymbol{\theta}\|_2^2 + \sum_i \left[s_i + \exp(-\hat{s}_i)(\hat{\mu}_i - y_i)^2\right]}_{\mathcal{L}(D;\boldsymbol{\theta})} \quad \text{with} \quad \hat{s}_i = \log \hat{\sigma}_i^2 \tag{A.7}$$

Equation A.7 represents the final formulation, consisting of two principal terms: a penalty on the parameter norm (quadratic, induced by $\mathcal{L}_2$ regularisation) and a discrepancy term that measures the agreement between the predicted canopy height and the ground-truth observations. The optimisation problem inherently involves two degrees of freedom: the network can refine its predictions either by adjusting the mean — interpreted as the *Maximum A Posteriori* (MAP) estimate — or by increasing the variance, thereby modulating the uncertainty and influencing the plausibility of a given solution within the posterior distribution. This formulation can be interpreted as a variant of the Negative Log-Likelihood (NLL).

List of Figures

List of Tables