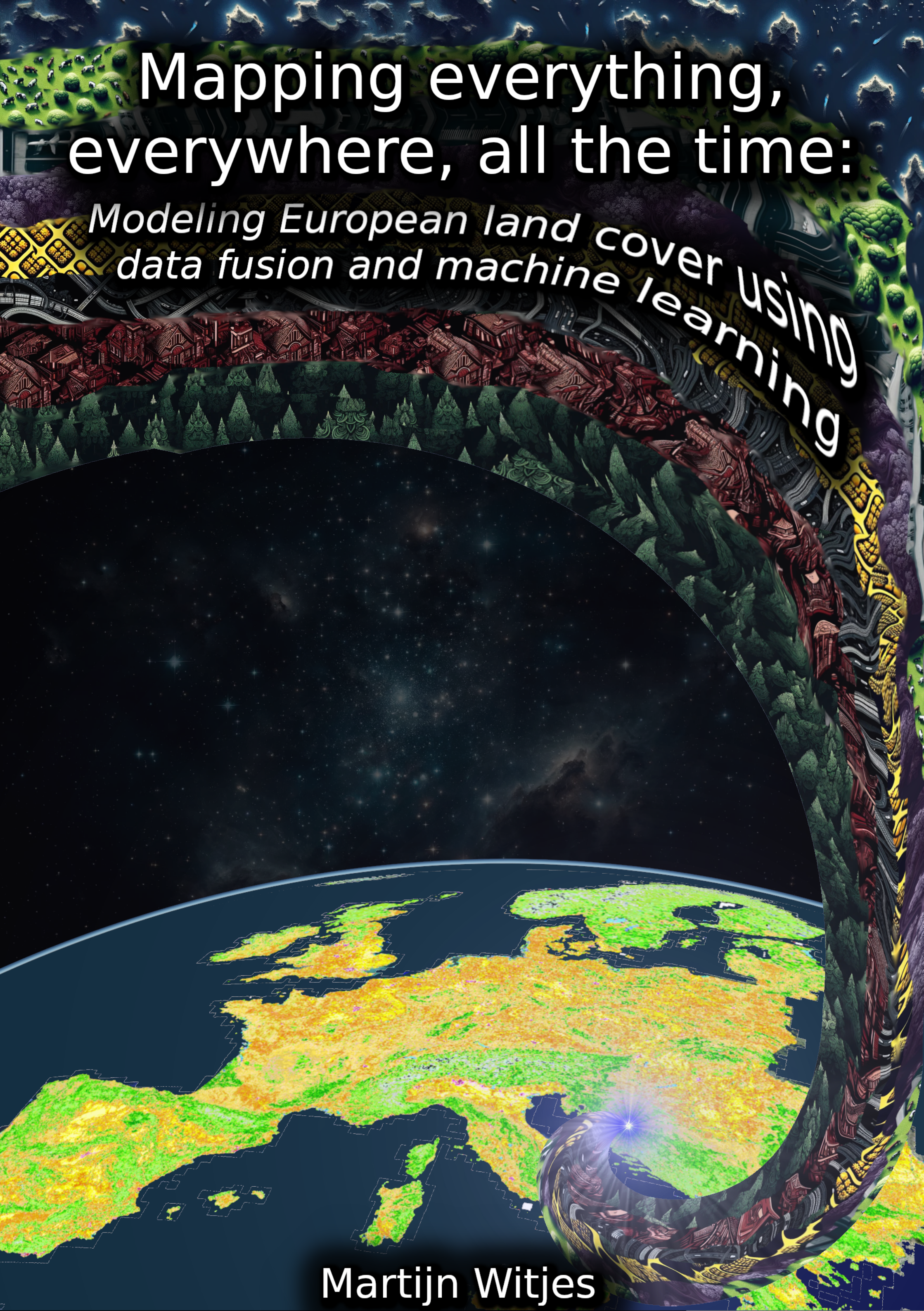
Mapping everything, everywhere, all the time:

Modeling European land cover using data fusion and machine learning



Martijn Witjes

Propositions

1. Open data are sufficient for making accurate, detailed maps. (This thesis)
2. Adjusting for model bias does not imply sacrificing model accuracy. (This thesis)
3. The mental health and career decisions of PhD students would benefit greatly from exchanging their experiences as much and as openly as possible.
4. Poor acronym design is harmful to scientific communication.
5. Hierarchy — while beneficial for land cover classification — is disastrous to society.
6. For-profit corporations are intrinsically unfit as the stewards of any kind of truth.

Propositions belonging to the thesis, entitled

Mapping everything, everywhere, all the time: Modeling European land cover using data fusion and machine learning

Martijn Witjes

Wageningen, 2 July, 2024

Mapping everything, everywhere, all the time

Modeling European land cover using data fusion and machine learning

Martijn Witjes

Thesis committee

Promotor:
Prof. Dr M. Herold
Professor of Geo-information Science and Remote Sensing
Wageningen University

Co-promotors:
Dr S. de Bruin
Associate Professor, Laboratory of Geo-information Science and Remote Sensing
Wageningen University

Other members:
Prof.dr R da Silva Torres, Wageningen University
Dr Raymond Sluiter, Netherlands Space Office
Dr Ruben vd Kerchove, VITO
Dr Miriam Machwitz, Luxembourg Institute of Science and Technology

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Mapping everything, everywhere, all the time

Modeling European land cover using data fusion and machine learning

Martijn Witjes

Thesis
submitted in fulfilment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr C. Kroeze,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 2 July 2024
at 16:00. in the Omnia Auditorium.

Martijn Witjes
Mapping everything, everywhere, all the time
Modeling European land cover using data fusion and machine learning
220 pages.

PhD thesis, Wageningen University, Wageningen, NL (2024)
With references, with summary in English

ISBN 978-90-834311-9-2
DOI 10.18174/659369

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Land cover

Land cover—the physical and biological material found on the surface of the earth—encompasses natural environments like forests, grasslands, wetlands, and things made by humans, like urban areas, croplands, and infrastructure [73].

Its study and monitoring are crucial for a multitude of reasons. Land cover changes significantly influence regional and global climate patterns and are a major driver of biodiversity loss, emphasizing the necessity of monitoring for climate change mitigation and biodiversity preservation [116, 202, 229, 234].

Additionally, these changes have direct implications for the quality and availability of natural resources, such as water and air [75]. Land cover assessment is vital for evaluating the livability of human environments [141, 145] and has key applications in informing policy, analyzing land-based emissions, and estimating local climate extremes [59, 112, 256].

Therefore, understanding land cover dynamics is crucial for effective policy-making at both regional and global levels [153, 240, 268]. As different entities with different objectives require maps with different characteristics and aspects, the choice of which type of land cover to map is a central one.

1.1.2 General land cover classes

Crops and Grasslands

Croplands and grasslands are essential for food production, and together cover over 40% of the European Union's territory [64]. European cropland is a diverse land cover type; over 100 crops are produced at scale in Europe, ranging from cereals such as wheat and rice to sunflowers, tobacco, and herbs such as basil and valerian [65]. Grasslands, on the other hand, cover wide roles beyond dairy and meat production: water supply and flow regulation, carbon storage, erosion control, climate mitigation, pollination, and cultural significance [12].

The agricultural sector is not only a cornerstone of the EU's economy, but also a significant employer, with approximately 4.2% of the EU's population engaged in agriculture. Reflecting its importance, the European Union's Common Agricultural Policy (CAP) has earmarked €264 billion as agricultural subsidy for the 2023–2027 period. The economic magnitude of crop and livestock production makes tracking and understanding their dynamics in the EU a priority for many stakeholders.

Forests

Forests cover nearly a third of the planet's land area [7, 67] and more than 41% of the EU [64]. They are important for biodiversity [34], carbon sequestration [126], and have been supplying a plethora of resources used by human civilization since the palaeolithic. Given recent trends of globalization and deforestation, monitoring forest cover change and its drivers is crucial [165, 257] for a great number of actors, including the UN and national governments, the industrial sector, local communities, and Indigenous peoples. National forest inventories, statisticians, and mapping organizations often disagree about the quantity of forest in any given area, which can lead to fierce scientific debates [139, 190, 201, 225]

Wetlands

Wetlands, recognized as one of the most biologically productive and important natural ecosystems, play a pivotal role in regulating the global climate, maintaining the hydrological cycle, and supporting diverse species [117, 215]. The unique hydrology of these ecosystems influences the distribution of sediment, nutrients, flora, and fauna. They provide essential services such as improving water quality, flood mitigation, recharging aquifers, and supporting an array of wildlife [43, 243]. Despite these benefits, wetlands have historically been undervalued, often seen as wastelands or disease sources, leading to significant losses due to reclamation and degradation [50, 58, 87, 91, 186, 189, 216]. The world has witnessed a dramatic decline in wetland areas, with an estimated 87% degradation since 1700, particularly in the 20th and early 21st centuries, resulting in substantial economic losses and reduced ecosystem service value [44, 88]. While they only cover 1.7% of the EU [64], their vital role and the alarming rate of loss make it crucial to preserve their ecological integrity and ensure the continued provision of their invaluable ecosystem services.

Land Use

There is another way besides land *cover* to describe land: With land *use*. Land Use does not refer to what's there, but how it is being used by humans. A grassland can be an intensively grazed pasture or a protected nature reserve. A 'forest' can be a palm oil plantation, a recreational area, or a rare old-growth forest with religious significance. The term 'Land Use' is sometimes used interchangeably with Land Cover [73], and different authors make different distinctions. For example, Hansen et al. [97] group inland water and wetlands as land cover, and built-up area, cropland and tree cover change as land use. Sometimes, a legend contains classes that are combinations of Land Use / Land Cover (LULC), such as pastures and natural grasslands in CORINE land cover, or even the grass fields at airports in the LUISA Basemap of Europe [203]. Because such classes can be ambiguous and hard to map by computer programs, they are not often reproduced at scale with remote sensing techniques. The LUCAS survey makes a clear distinction between land use and land cover: each observation has a separate listing for both types, with some combinations occurring frequently (like *agricultural* use and *grassland* cover to represent *pastures*), and others being extremely rare (*residential* use and *peat bog* cover).

What should be on a map?

Broad categories such as 'Forest' or 'Water' are relatively simple for people and computers to differentiate. However, adding more detail to the map by splitting up these big categories into more specific classes, or improving the *thematic resolution*, brings many challenges. First of all, maps with many classes are hard to read by humans. Reading a complex map is not always necessary, as users can derive simpler maps from rich datasets to tailor them to their use case [271] or use the mapped values for a different type of analysis. Secondly, the process of making detailed maps is more costly and difficult. You need to collect more examples from more different categories, and not all categories are equally easy to distinguish by surveyors. Someone interpreting aerial imagery might be able to distinguish grassland from forest, but what if they want to map different tree species, or distinguish pastures from natural grassland? Which brings us to the next question: How do you get those maps in spite of those challenges? People used to draw maps based on field surveys. Later maps were based on aerial imagery. In recent decades, we have been increasingly using remote sensing and machine learning.

Land cover and land use mapping and monitoring in EU

Especially in Europe, long-running datasets, such as the Coordination of Information on the Environment (CORINE) Land Cover Project and the Land Use / Cover Area frame Survey (LUCAS) [46], have provided detailed and consistent land cover information for almost two decades [81]. CORINE is a good example of a detailed land cover dataset experiencing th . CORINE is an initiative by the European Environment Agency (EEA) that aims to collect information on the land cover of Europe to support environmental policy development. The project started in 1985 as part of the European Commission's CORINE program, designed to gather and harmonize data related to the environment across the member states. In the subsequent decades, it was used to assess land cover changes in Europe [69, 178].

However, as the project evolved and standards for environmental data became more stringent, its limitations became increasingly apparent. The nature of its legend, which often combines land use and land cover within the same category, and the presence of mixed classes such as *airport* (grass, buildings, roads). Furthermore, the project's reliance on a 25 hectare minimum mapping unit for most LULC types means that it only contains relatively large patches (see Fig. 1.3). This means it under-represents LULC types and is of limited accuracy at fine spatial scales [5, 29, 199]. Regardless, CORINE Land Cover is among the most popular and widely used European land cover datasets and should not be discarded. As using human cartographers at a finer spatial and temporal scale would be prohibitively difficult, slow, and costly, attempts have been made to automate the production of CLC [29], but usually at a lower thematic resolution, for example for 12 [199] or 14 [15] classes. There have also been other projects that accurately map European land cover at high spatial resolution using different legends that are more optimized for a remote sensing context, such as S2GLC [161], and attempts to specifically map crop types with specialized approaches [48, 158]

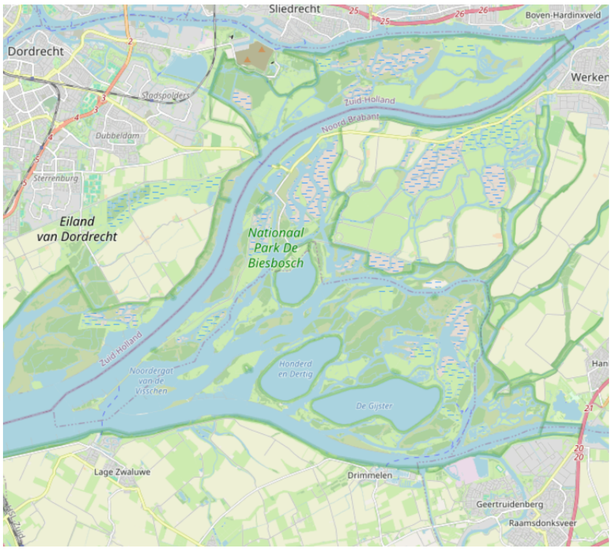


Figure 1.1: The Biesbosch wetlands area in the Netherlands, as represented on OpenStreetMap.

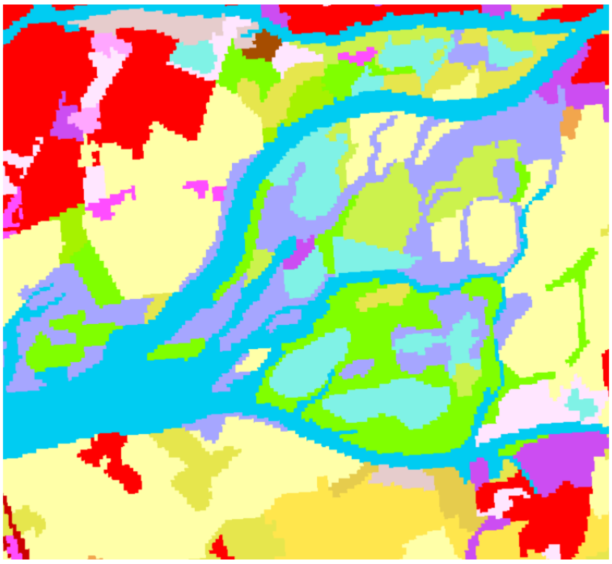


Figure 1.2: The Biesbosch as represented by CORINE Land Cover.

🟥	Continuous Urban Fabric
🟥	Discontinuous Urban Fabric
🟪	Industrial or Commercial Units
🟥	Road and rail networks
⬜	Port areas
⬜	Airports
🟪	Mineral extraction sites
🟫	Dump sites
🟪	Construction sites
🟪	Green urban areas
⬜	Sport and leisure facilities
🟨	Non-irrigated arable land
🟨	Permanently irrigated land
🟨	Rice fields
🟧	Vineyards
🟧	Fruit trees and berry plantations
🟧	Olive groves
🟨	Pastures
🟨	Annual crops associated with permanent crops
🟨	Complex cultivation patterns
🟨	Land principally occupied by agriculture
🟧	Agro-forestry areas
🟩	Broad-leaved forest
🟩	Coniferous forest
🟩	Mixed forest
🟩	Natural grasslands
🟩	Moors and heathland
🟩	Sclerophyllous vegetation
🟩	Transitional woodland-shrub
⬜	Beaches, dunes, sands
⬜	Bare rocks
🟩	Sparsely vegetated areas
⬛	Burnt areas
🟩	Glaciers and perpetual snow
🟦	Inland marshes
🟦	Peat bogs
🟦	Salt marshes
⬜	Salines
🟦	Intertidal flats
🟦	Water courses
🟩	Water bodies
🟩	Coastal lagoons
🟩	Estuaries
⬜	Sea and ocean

Figure 1.3: The 43-class CORINE Land Cover legend.

1.1.3 Land cover classification with machine learning and earth observation data

Automated mapping methods, such as those used by Pflugmacher et al. [199], d'Andrimont et al. [48], Luo et al. [158] and Malinowski et al. [162] are called Land Cover Classification. It typically involves showing many good examples of the things you want to detect on a map to a computer program, which then learns how to recognize them by itself when shown new locations. For this, the following three core components are needed:

- **Examples:** Confirmed and annotated observations of different types of land cover, with a known time and place. We often call these observations **reference data**, *labels*, or *samples*.
- **Earth observation (EO) data:** satellite imagery and other geographic data that cover the area you want to map. Together, these **covariates** form the **feature space**. Spatial and temporal metadata allow aspiring mappers to combine several existing, unrelated datasets into one feature space; a rarity in the machine learning domain.
- A **model:** Some kind of decision-making algorithm that learns to recognize which examples often occur at which combination of the covariates. If this goes well, you can use it to **predict**, or *classify*, which land cover occurs at locations that are covered by the feature space, but where reference data are not available.

It is common to use different sets of reference data: one set to train your model, and one independent set to validate its predictions, namely the land cover maps it has made. It is good if these two datasets were collected separately from each other, because this helps to make sure your model **generalizes** well: that its predictions are accurate in new situations, not just on the data it was trained on.

Training data

Training data is a set of labeled observations that are used to teach a machine learning model to recognize similar things in new situations. To make sure your model generalizes well, you have to properly represent the feature space of the area you want to map in your training data [170]. This does not directly relate to geographical space; large homogeneous areas such as deserts tend to be quite similar in the feature space, while small complex landscapes such as cities with parks, gardens and canals will show great local variation.

It is important to have enough training examples [214, 222, 320], but what is ‘enough’ differs per model [181] and task [140]. Models often learn to estimate classes in the same proportions as their training data. If these proportions do not match the real-world situation of the area they need to map, the model may become **biased**. To minimize this bias and predict the right relative quantities for each class, you need training data that match the desired proportions [100].

Validation data

All maps are wrong [175], and maps produced with machine learning and earth observation data are no exception. Errors can be defined as *quantity disagreement* and *allocation* *disagreement*: are there enough pixels of every class, and are they in the right place [204, 205]? Ideally, both types of errors should be low. Errors can come from various sources. Some classes might be hard to distinguish in the available feature space, confusing the model [239]. Some classes might be overrepresented in the training data, leading the model to predict it more frequently at the expense of other classes and resulting in low quantity agreement.

In general, it is essential that any map is validated with a set of observations. Such validation datasets need to be of high quality: their accuracy should surpass the target accuracy of the map and they need to be representative of the classes in the mapped area.

Biased training data can lead to a biased model that overpredicts certain classes. In this case, the proportions of the classes on the map will deviate from the proportions in the real world. It is very difficult to quantify the bias of a model before it is used [249], and the current best way of fixing the proportions on a map is to measure the model bias with additional sampling after the mapping is complete [250].

Classification errors become even more troublesome when land cover change is measured by comparing maps from two time periods, as any misclassified pixel will be interpreted as a change [188]. This can be counteracted by careful sampling design [187, 248], although some work has been done to circumvent this and derive area estimates directly from the model predictions [3, 138, 230]. All in all, it is important to validate annual large-scale maps with up-to date validation data to ensure there is no dataset drift and to properly quantify accuracy and uncertainty [270].

Earth Observation Data

Earth observation data refers to measurements taken by sensors at various ranges, from ground measurements [238] to drone [262] and airplane [166] imagery, to satellite scans [200]. The most useful for recurring large-scale mapping are satellite scans, as satellites revisit the same area at a regular pace. This allows them to provide a consistent stream of data, which, in some cases, can continue for decades [304, 305]. The activity of remotely collecting Earth observation data is generally referred to as **remote sensing**.

Earth observation data only started becoming openly available after Brazil published its LANDSAT archives as open data in 2008 [164]. Afterwards, more and more Earth observation data and land cover information has become openly available. While this rapidly growing landscape of increasingly large datasets on diverging platforms and systems can be hard to navigate [293], it has fuelled an exponential increase in the amount of mapping initiatives and environmental awareness [304], with applications ranging from deforestation [96] to soil mapping [104]. For further reading on the role of remote sensing in quantifying essential characteristics of the biosphere, cryosphere, and hydrosphere, readers are directed to Radeloff et al. [213], which provides an up-to-date review of the state of the art and current needs.

Spatial machine learning is unique in the sense that it allows modelers to easily combine variables if they can somehow acquire a representation of them at the time and place of the ground truth observation, such as with a spatial or spatiotemporal overlay. Combining different data sources can lead to much higher performance, for instance with crop stress detection [13], urban heat [238], and land cover classification [115, 121, 307, 320]. Having data of the same place from different moments [155] and different spatial resolutions or scales [55] can also improve mapping accuracy, especially of classes that have very similar spectral profiles and/or cyclical temporal dynamics, such as crops and grassland [61].

The current land cover mapping arms race

In the past decades, machine learning has been used to make land use and land cover maps at global scales. Due to limitations in the resolution of the available satellite data, such maps were at resolutions of 5 km for GLASS-GLC [148], 300 m for ESA CCI [52] or 100 m for Copernicus Global Land Cover [27]. This made them useful for large scale analysis, but of limited value for local applications and analyses, such as investigating small-scale land use as drivers of deforestation [165]. However, in the last few years, the increased availability and analysis-readiness of high-resolution from NASA's Landsat and ESA's Sentinel-2 programmes have enabled the creation of global- and continental-scale high-resolution land cover maps by several organizations. Notable examples are ESA's WorldCover [280] and WorldCereal [284], Google's DynamicWorld [24], ESRI's Land Cover [134], and GLAD's Global Land Cover [207]. These maps often have high resolution and high accuracy, but are usually limited in multiple ways: reproducibility, time coverage, and detail:

Reproducibility

Most of the current big mapping initiatives only share their maps, not the reference data that they used to produce them. This not only makes it difficult or impossible to reproduce, challenge, or improve their work, but also prevents it from being used for new projects [272]. If they don't share their validation data, independent validations are dependent on open validation datasets [286], which might not perfectly match their legend, making a fair comparison more difficult.

Time coverage

Most of the current generation of maps use 10 m Sentinel-2 data, which only became available in 2016. Mapping initiatives that use Landsat data can go back much further, but are then limited to mapping at 30m resolution. Examples are the Australian land cover time series made by [32] and the work of UMD GLAD, which produced annual land cover maps from the year 2000 onwards using Landsat data [97]. García-Álvarez et al. [86] provides an thorough overview of LULC datasets with long-running time series.

Detail

Often, these maps depict at most ten land cover categories (see Table 1.1). Mapping more classes is more difficult because there is more chance of confusion between similar classes. This also applies to the creation of training data: The current most common way to create it is by visual inspection of high resolution aerial or satellite imagery, and this becomes more difficult with more specific classes. Just like for a machine learning model, it is much easier for an annotator to differentiate between a forest and cropland than between different cereal crop types. Classifying more different classes is also more difficult for machine learning models and requires more training data, and of higher quality, which is an expensive and complex task [146], especially for large diverse areas (e.g., continents or the whole planet) [247, 270].

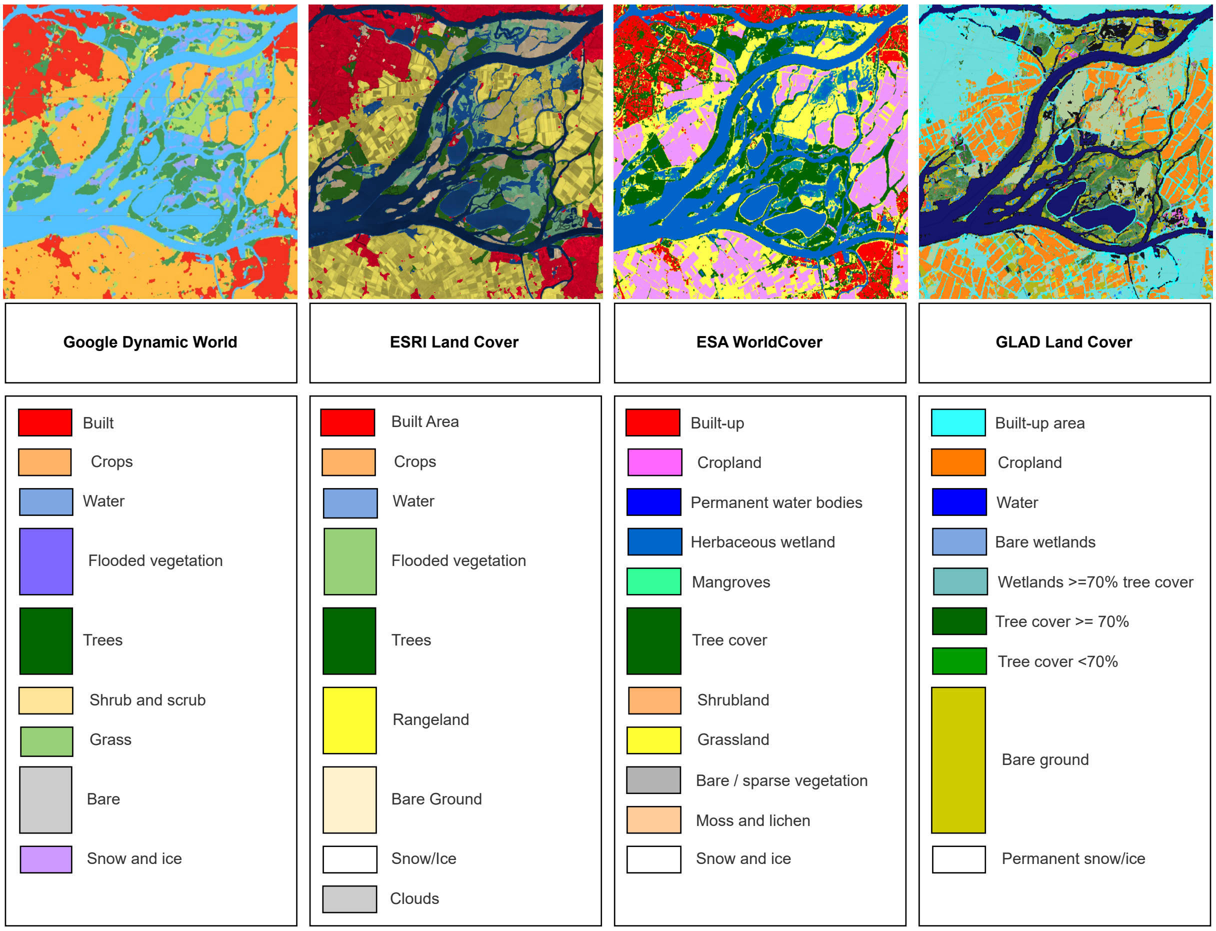


Figure 1.4: Comparison of four global mapping projects in the Biesbosch wetlands area in the Netherlands: Google Dynamic World [24], ESRI Land Cover [134], ESA WorldCover [313], and GLAD Land Cover [207]. Note that each legend is more specific in some thematic types: GLAD differentiates more strongly based on tree cover, while ESA WorldCover differentiates more between different types of non-forested dry land.

Although there are exceptions to each of these issues, in general, all of the current mapping initiatives have at least one of these limitations (see Table 1.1). There is value in mapping as far back as possible, something that the previous generation of global maps did do [86]. An extreme example is the History database of the Global Environment (HYDE), which contains historic estimated maps at 10 km resolution of main land use categories going back more than 10 thousand years. (Klein Goldewijk et al. 2017). HYDE is also quite detailed, and includes irrigated areas, rice, intensive pasture, extensive rangelands and similar classes. This makes it useful despite its low spatial resolution. Regardless, there is a need for modern maps that do not sacrifice detail to maintain useful spatial resolution and accuracy, and which are made transparently and reproducibly, so that they can be challenged, improved, and trusted.

Table 1.1: Summary of Mapping Products and Data Availability

Mapping Product	**Classes**	**Training Data**	**Validation Data**	**Spatial Res.**	**Time Res.**
ESA WorldCover [313]	11	No	No	10 m	2020, 2021
ESA WorldCereal [284]	10	Yes [303]	Yes [144]	10 m	2021
Google Dynamic-World [24]	9	Yes [260]	No	10 m	2016+, monthly
ESRI Land Cover [134]	9	No	No	10 m	2017–2021, annual
GLAD Global Land Cover and Land Use Change [207]	7	No	No	30 m	2000–2020, annual
GLC_FC30 [315]	30	No	Yes [147]	30 m	1982–2021, annual*
S2GLC Maps of Europe [162]	13	No	Yes [129]	10 m	2017
ELC10 [287]	8	Yes [49]	Yes (cross-validation)	10 m	2020
Pan-European land cover [199]	12	Yes [49]	Yes (cross-validation)	30 m	One year
Australian Land Cover [32]	6	No	No	30 m	1985-2015

1.2 Research Gaps

There is a reproducibility crisis in science, and environmental monitoring is no exception [209, 267]. How can decision-makers trust datasets whose creation does not only involve complex workflows, but can not be reproduced or improved? It turns out that they often don't, and keep relying on older methods that might be slow and expensive, but reproducible and reliable. The current boom in for-profit remote sensing has allowed for excellent analysis using high-resolution data —e.g., drivers of deforestation in Africa [165] using 3 m resolution Planet data— but limitations on sharing this data, and the profit and marketing motivations of corporations make sharing, trusting, and using their maps difficult.

There are no reproducible large-scale maps with more than 13 classes, not even for Europe, for which an unprecedented wealth of land cover information is publicly available [49].

Pflugmacher et al. [199] demonstrated that the LUCAS points can be used to train and validate land cover models, but the LUCAS legend is a hierarchical system with over 70 land cover and land use classes. To retain its use as a validation set, training data should be extracted from other human-annotated datasets such as CORINE, EuroCrops [231] and OpenStreetMap [232]. In order to use data with the large minimal mapping unit of CORINE or the potential errors in volunteered geographical information [183], filtering techniques are required to remove spurious training samples.

Furthermore, while some long-term annual mapping projects exist, they either have a low spatial resolution like ESA CCI [98], cover only a select set of classes [207], or did not publish their training data [315].

Lastly, to evaluate the impacts of human activities on the environment and to enable and justify localized interventions, it is essential that trends from area estimates and periodic maps are linked [187, 258, 298]. This requires a mapping framework that minimizes both quantity and allocation error [205]. While many mapping initiatives report class proportions [199] or try to approximate correct area estimation with direct remote sensing measures [138], only a few explicitly investigate methods to force maps to match area estimates [113, 253].

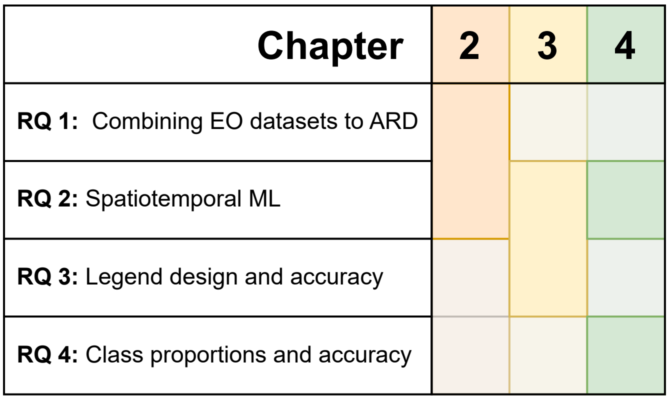
1.3 Objectives

The overall objective of this PhD thesis is to take advantage of the availability of various open European land use / land cover datasets and statistics, and to combine these with Earth observation data and machine learning to make accurate, detailed maps that are compatible with the needs of European policy makers and researchers. More specifically, this thesis aims to answer the following research questions:

1. What are the benefits and challenges of combining multiple large time-series and static EO datasets into analysis-ready data for the purpose of land cover classification?
2. To what extent does training data from multiple times and places improve the accuracy and generalization of land cover classification?
3. How does the number and type of classes in a legend affect the accuracy of land cover classification?
4. What is the effect of enforcing design-based class proportions on map accuracy?

1.4 Thesis Outline

The research questions posed in the previous section are answered through four research papers, which are presented as chapters in this thesis. Fig. 1.5 provides a graphical overview of how each chapter relates to each research question framework. The first two chapters focus on combining earth observation and land cover data from multiple sources to train models that can generalize well to unknown years. Chapter 4 applies a novel algorithm that uses a-priori class proportions from area estimates to optimize the accuracy and quantity of predictions by such models.



Chapter	2	3	4
RQ 1: Combining EO datasets to ARD	●		
RQ 2: Spatiotemporal ML	●	●	●
RQ 3: Legend design and accuracy		●	
RQ 4: Class proportions and accuracy			●

Figure 1.5: Overview of which chapters discuss which research questions.

Chapter 2 focuses on the benefits and challenges of harmonizing and combining large-scale spatiotemporal datasets for land cover mapping, most of which were used in the following chapters. The chapter details the work that went into creating, harmonizing, and imputing multiple Earth observation datasets (Landsat, Sentinel-2, and a new 30m resolution DTM) covering Europe. It introduces and describes the imputation algorithm TMWM that was used to impute the Landsat data, and validates its accuracy in a spatiotemporally explicit way. It then explores how combining the different datasets improves the accuracy of land cover classification models. Lastly, it shows that models trained on samples from a longer time range can generalize better to years that they have not been trained on.

Chapter 3 focuses on the production of annual land use / land cover maps of Europe for 2000-2020. It details the steps taken to harmonize and clean the training data from multiple openly available sources (CORINE, LUCAS) into a legend with 43 classes. A thorough accuracy assessment using cross-validation and an independent set of S2GLC validation points describes how well the model generalizes across space and time, and quantifies the trade-off imposed by having a legend with high thematic resolution. Results show that the maps have similar accuracy as other current continental-scale maps at low thematic resolution, and that a more detailed legend introduces more errors.

Chapter 4 introduces IMP, an algorithm that uses land cover area estimates to iteratively classify land cover from existing probabilities, producing maps whose class proportions match the input estimates. It details the algorithm and showcases its use by mapping five European countries in five years. The accuracy of the maps is compared with maps created using highest likelihood classification. Results show that the proportional maps do not only have more accurate class proportions, but equal or better accuracy than highest likelihood maps. We also compare the accuracy and proportions of maps based on probabilities predicted by models trained on data representative of the area of interest, and probabilities predicted by a general model trained on large parts of Europe. Results show that maps based on general model predictions reach more accurate class proportions, while maps based on local model predictions are slightly more accurate. Finally, it presents an unintentional finding that the iterations at which the algorithm classifies certain pixels is related to the accuracy of those pixels, suggesting that it can be used to generate pixel-level accuracy estimates.

Chapter 2

An Analysis-Ready Data cube for European-scale land cover mapping

This chapter is based on:

M. Witjes, L. Parente, J. Križan, T. Hengl, and L. Antonić. "Ecodatacube.eu: Analysis-ready open environmental data cube for Europe". *PeerJ* 11 (2023), e15478. DOI: `https://doi.org/10.7717/peerj.15478`

Abstract

The paper describes the production steps and accuracy assessment of an analysis-ready, open-access European data cube consisting of 2000–2020+ Landsat data, 2017–2021+ Sentinel-2 data and a 30 m resolution Digital Terrain Model (DTM). The main purpose of the data cube is to make annual continental-scale spatiotemporal machine learning tasks accessible to a wider user base by providing a spatially and temporally consistent multidimensional feature space. This has required systematic spatiotemporal harmonization, efficient compression, and imputation of missing values. Sentinel-2 and Landsat reflectance values were aggregated into four quarterly averages approximating the four seasons common in Europe (winter, spring, summer and autumn), as well as the 25th and 75th percentile, in order to retain intra-seasonal variance. Remaining missing data in the Landsat time-series was imputed with a temporal moving window median (TMWM) approach. An accuracy assessment shows TMWM performs relatively better in Southern Europe and lower in mountainous regions such as the Scandinavian Mountains, the Alps, and the Pyrenees. We quantify the usability of the different component data sets for spatiotemporal machine learning tasks with a series of land cover classification experiments, which show that models utilizing the full feature space (30 m DTM, 30 m Landsat, 30 m and 10 m Sentinel-2) yield the highest land cover classification accuracy, with different data sets improving the results for different land cover classes. The data sets presented in the paper are part of the EcoDataCube platform, which also hosts open vegetation, soil, and land use / land cover (LULC) maps created. All data sets are available under CC-BY license as Cloud-Optimized GeoTIFFs (ca. 12 TB in size) through SpatioTemporal Asset Catalog (STAC) and the EcoDataCube data portal.

2.1 Introduction

Over recent decades, the world has experienced rapid growth in Earth Observation (EO) technology. This has brought many benefits to various applied fields, however, it also brings new challenges to aspiring users: massive data volumes produced by EO sensors and *in-situ* monitoring networks require new specialized expertise and extensive computing capacity. Wagemann et al. [293] lists the following five key challenges to finding, accessing, and combining big environmental data: (1) limited processing capacity on user side, (2) growing data volumes, (3) non-standardized data formats and dissemination workflows, (4) too many data portals, and (5) difficult data discovery. Environmental data needs to be as accessible and useful as possible, while all its limitations, caveats and uncertainties need to be clearly documented to minimize the risk of error propagation. In addition, decision-makers require easy access to open environmental data and critical assessment tools in order to dynamically synthesize the information needed to address many critical environmental and economic challenges [92]. The European Green Deal specifically [241] requires a diversity of environmental information to reach its ambitious project goals, especially those focused on reaching climate neutrality, preservation of natural capital, modernization and simplification of the Common Agricultural Policy (CAP), and connecting farms to forks, all whilst enabling the socio-economic transformation of rural and agricultural areas.

To enhance environmental data use for decision-making, several groups in different areas around the world have been putting effort in building *EO data cubes*: spatially aligned time-series of calibrated multi-dimensional observations [92]; also see Liu et al. [149], Lu et al. [156], and Mirmazloumi et al. [174]. Some prominent examples of EO data cubes include the Earth System Data Cube [160], Digital Earth Australia [157], Digital Earth Africa [312], and the Swiss Data Cube [36]. Infrastructures such as the openEO Platform (`https://openeo.cloud/`), and Google Earth Engine (`https://developers.google.com/earth-engine`) can also be considered EO data cubes [93] due to the ease with which users can combine the various data sets hosted on these platforms.

Two important EO data sources in this context are Sentinel-2 and Landsat. Sentinel-2 has provided global coverage every five days since the launch of its second satellite (Sentinel-2B) in 2017, available freely from multiple sources such as `https://scihub.copernicus.eu` and `https://earthexplorer.usgs.gov/`. In recent years it has served as input data for various global and continental land cover mapping initiatives, such as ESA's Worldcover [280], Google's DynamicWorld [24], and Sentinel-2 Global Land Cover (S2GLC) [162]. The spatial resolution of Sentinel-2 sensors varies; the red, green, blue, and near-infrared (NIR) bands are available at 10 m resolution, while the two shortwave infrared (SWIR) bands are only available at 20 m.

The Landsat program is the world's longest continuously running EO mission [304]. It is *de facto* the only option for assessing long-term dynamics as it provides an uninterrupted supply of satellite imagery since 1972. The entire archive was made available to the public in 2008, leading to widespread use, including refinement into data sets closer to analysis-ready status. The University of Maryland (UMD) Global Land Analysis and Discovery (GLAD) laboratory's Landsat ARD product is another representative example of long-term EO data due to its free availability, global coverage, and its inclusion and harmonization of a succession of Landsat satellite sources [206]. The original data is available in 23× 16–day scenes per year in scaled long format [206]. While this high temporal resolution and numerical precision provide a large amount of information for subsequent modeling and has been successfully utilized as such by teams with access to large computational resources [97], the added benefit for Land Use Land Cover (LULC) classification compared with a compressed, more accessible form, has not yet been quantified. Furthermore, although the data set is nominally analysis-ready, we encountered the following limitations of using this data set for actual Machine Learning for the purpose of vegetation / land cover mapping:

- The data volume of GLAD's original archive (23 4-byte values, or 92 bytes per band per pixel per year) may exclude users without advanced computational capacity from performing country- or continental-scale analysis;
- While multiple Sentinel-2 data sets are now available from 2015 (Copernicus Sentinel program), a harmonized, cloud-optimized product that is freely accessible regardless of institutional membership and computational resources could greatly increase global usage, especially among marginalized users.
- While the aggregation to 23 16-day reflectance values increases coverage, gaps remain in the archive due to e.g. snow cover.

To maximize the usability of the produced data sets and facilitate future work in annual mapping, we have built a data cube and a data portal available on `https://EcoDataCube.eu`. It integrates various layers into a single seamless expandable and open access system. In this paper, we describe the key processing steps used to produce data cubes. We first explain the process of obtaining, gap-filling, artifact removal, and harmonization of EO images: gap-filled Landsat time series from 2000–2020, two Sentinel-2 time-series from 2018–2021 at varying temporal and spatial resolutions, and an optimized Digital Terrain Model created with an ensemble machine learning approach. Finally, we provide examples of EcoDataCube usage and demonstrate case studies for which this data cube provided the main feature space, such as annual LULC maps [301] and potential and realized tree distribution [19].

Methods and Data

In this work we detail the processing and validation workflows of the following four data sets:

1. Quarterly spatiotemporal Landsat aggregates (median, 75th, and 25th percentile) of blue, green, red, NIR, SWIR1, SWIR2, and thermal bands at 30 m resolution between 2000 and 2020.
2. Two spatiotemporal aggregates (median, 75th and 25th percentiles) of Sentinel-2 between 2018 and 2021:
   (a) Annual blue, green, red, and NIR at 10 m resolution;
   (b) Quarterly blue, green, red, NIR, SWIR1 and SWIR2 bands at 30 m resolution;
3. An optimized Digital Terrain Model (DTM) for Europe, created with an ensemble machine learning data fusion approach.

In order to quantify the extent to which these data sets complement each other as a single feature space for annual mapping with machine learning, we include a series of land cover classification experiments where we compare model performance when trained on different combinations of EcoDataCube layers. Fig. 2.1 provides an overview of the general workflow and the resulting output and findings. These data sets all cover the exact same area, which is defined by all member and partner states of the European Environment Agency (EEA) in 2019, with the exception of Turkey (See Fig. 2.2-A).

Landsat

For this work, we used the Landsat Analysis-Ready Data (ARD) product developed by the UMD's GLAD lab, a globally consistent analysis-ready data set for multi-decadal LULC monitoring [206]. It consists of 16-day time-series composites (23 per year, see Table 2.1) from Landsat 5, 7 and 8 which have been calibrated using MODIS surface reflectance.

Table 2.2 shows the bands, their spectral range in the different Landsat sources, and their spatial resolution. These time-series are freely available in 1–degree tiles (see Fig. 2.2-B) and has been screened to flag pixels that likely contain clouds and their shadows in a quality assessment (QA) layer. While this data set is already a level 3 remote sensing product (i.e. temporal composites of gridded data), we aim to make it both more analysis-ready and easier to use by compressing it and imputing any missing values in a computationally efficient way that yields values suitable for classification tasks.

Landsat Temporal Composites

In order to balance the trade-off between computation time of large areas while retaining as much temporal variability as possible, we aggregated the 23 annual GLAD Landsat ARD values into four annual quarterly period medians based on the astronomical seasons described by Trenberth [266], which allows the four periods to act as a proxy for the four typical seasons in large parts of Europe: winter, spring, summer, and autumn. This

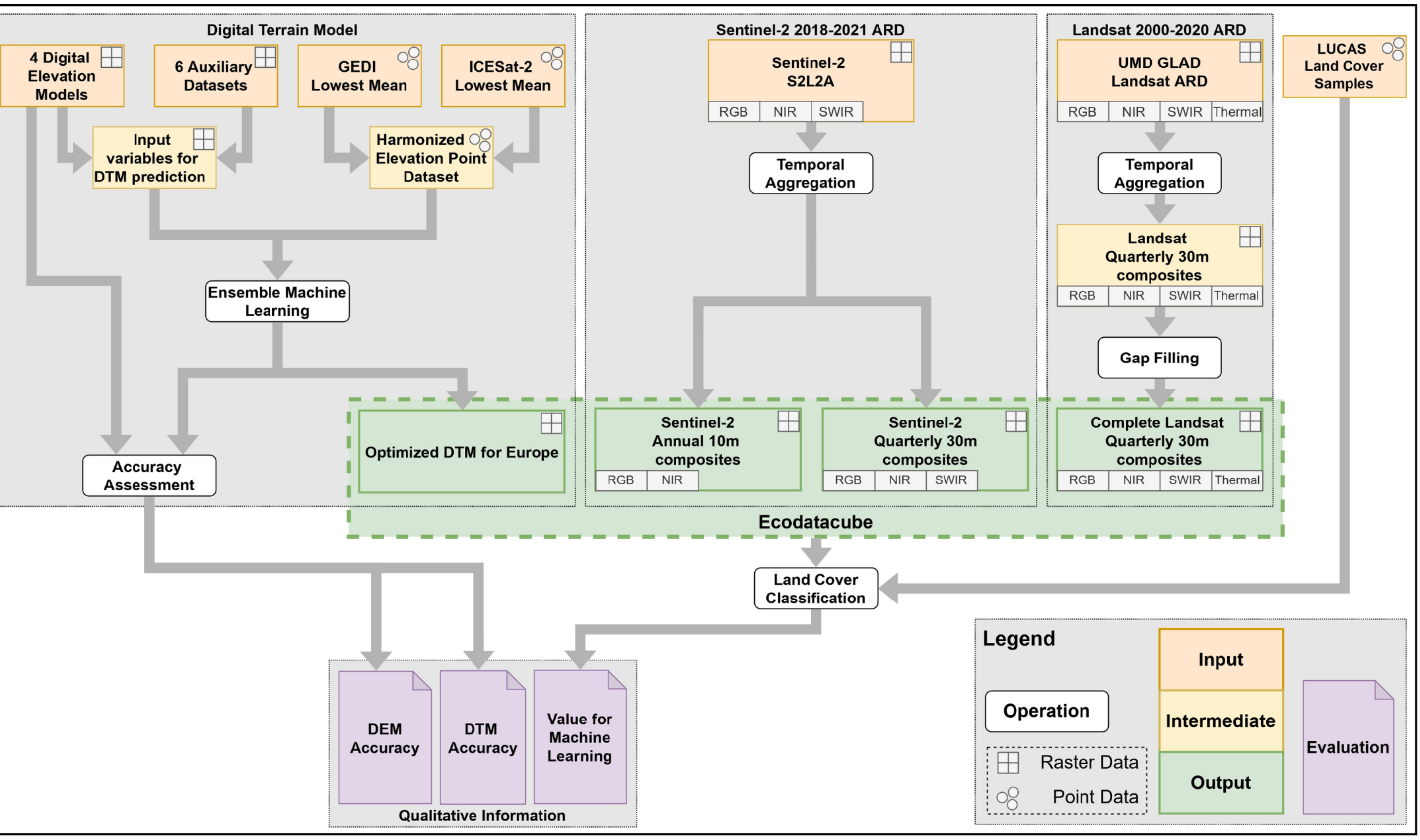
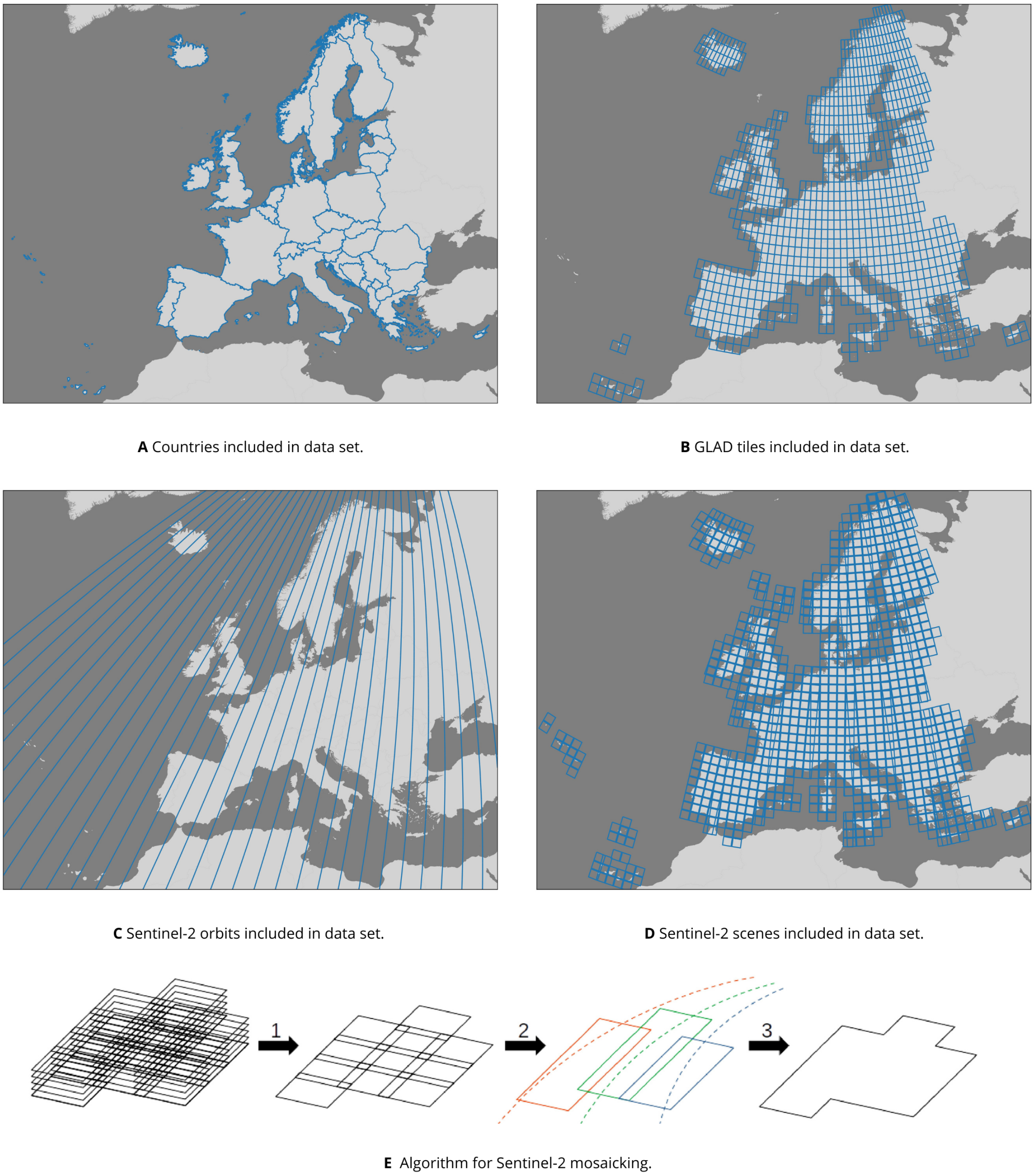


Figure 2.1: Overview of the general workflow with input, intermediate, and output data sets, as well as operations and evaluations of data set quality.

allows us to match the beginning and end of each period with the 16-day intervals used by Potapov et al. [206] (see Table. 2.1 and Fig.2.3-B). We also calculated the 25th and 75th percentile of these aggregated values per pixel in order to maintain a measure of variance within each period that might be useful to recognize intra-annual dynamics. This yields 84 layers for each year (4 quarterly periods × 3 percentiles × 7 Landsat bands, see Fig. 2.3-A) with varying amounts of no-data values based on cloud and snow cover.

Gap-filling

While the temporal composite aggregation reduces the number of no-data pixels in the time series at the cost of temporal resolution, gaps remain. While many gap-filling methods have been proposed, they are only well established for specific purposes (e.g. DINEOF [2] for ocean modeling, Geostatistical Neighborhood Similar Pixel Interpolator [319] for Landsat 7 Scan Line Corrector-off images), too computationally intensive to process multi-decade, continental-scale data (e.g. STAIR 2.0 [159] and linear temporal interpolation), and/or not available as maintained open source software.



A Countries included in data set.

B GLAD tiles included in data set.

C Sentinel-2 orbits included in data set.

D Sentinel-2 scenes included in data set.

E Algorithm for Sentinel-2 mosaicking.

Figure 2.2: Overview of A: the area of interest, B: GLAD Landsat ARD tiles , C-D: Sentinel-2 orbits and scenes used as input sources, and E: the mosaicking algorithm that 1) computes quarterly composites, 2) mosaics the quarterly composites along orbital tracks, and 3) stitches the orbital track mosaics into a single data set.

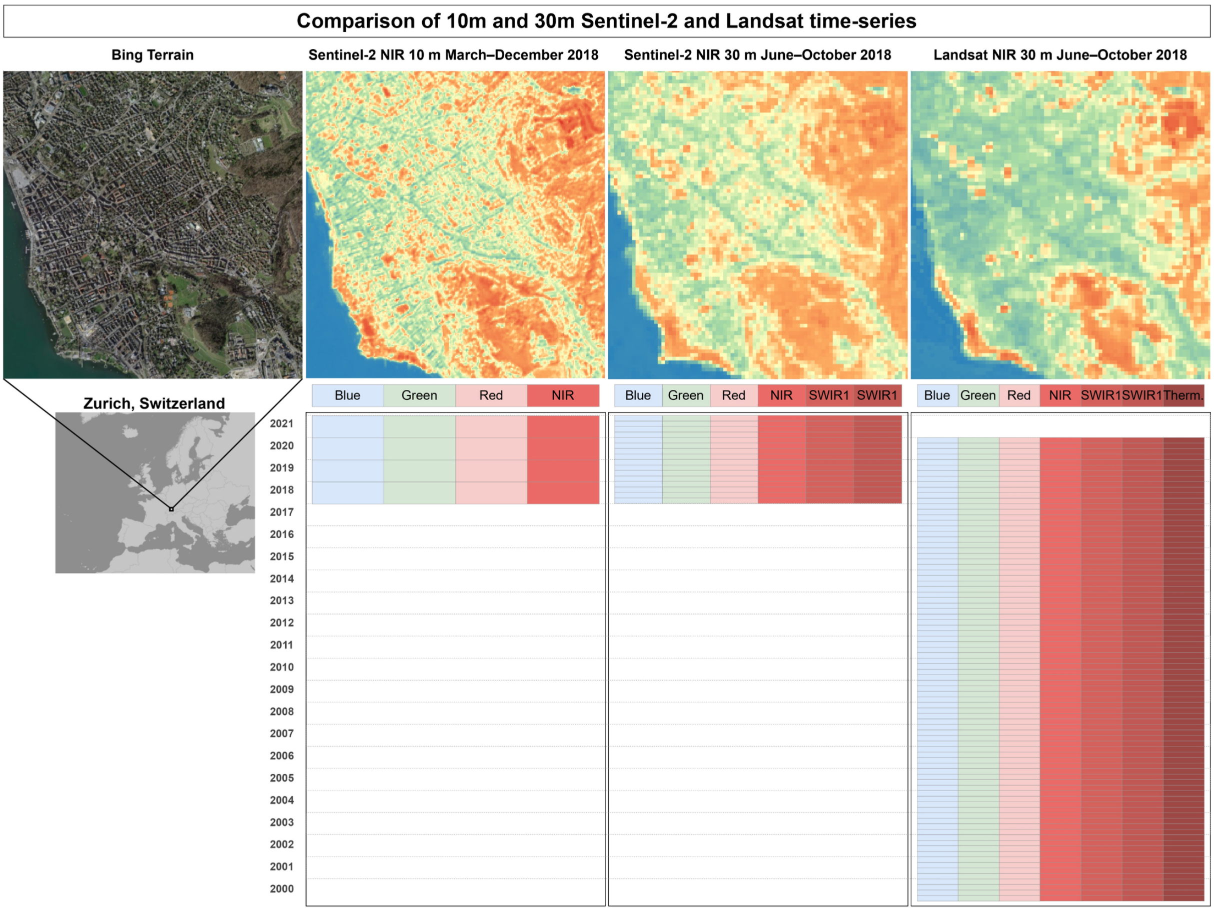
Table 2.1: Overview of the start and end dates of the four temporal composite periods (quartiles) and of which GLAD interval IDs they were composed.

	Date		GLAD interval ID	
Quartile	Start	End	Start	End
1	December 2nd of previous year	March 20nd	22 of previous year	5
2	March 21st	June 24th	6	11
3	June 25th	September 12th	12	16
4	September 13th	December 1st	17	21

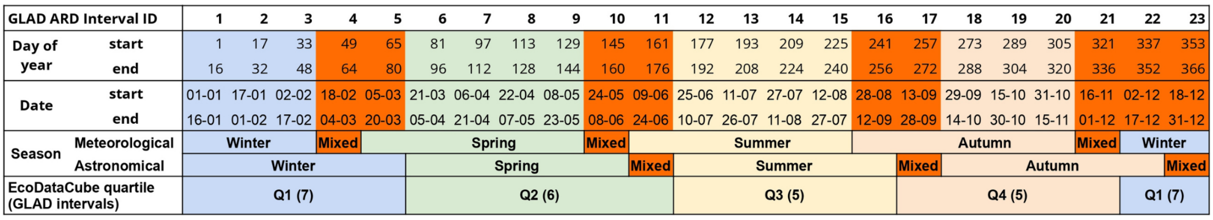
Table 2.2: Spectral bands used by Potapov et al. [206] to create the GLAD Landsat ARD data set from multiple Landsat sensors.

Sensor	Landsat 5	Landsat 7	Landsat 8
Time range	2000–2011	2000–2021	2013–2021
Band		Wavelength (nm)	
Blue	450-–520	441–514	452—512
Green	520-–600	519—601	533—590
Red	630-–690	631—692	636—673
NIR	760-–900	772—898	851-–879
SWIR1	1,550-–1,750	1,547—1,749	1,566—1,651
SWIR2	2,080-–2,350	2,064—2,345	2,107-–2,294
Thermal	10,410-–12,500	10,310—12,360	10,600—11,190

In order to impute these remaining missing values in this multi-decade, continental-scale data set, we developed and implemented a custom gap-filling method: Temporal Moving Window Median (TMWM). The algorithm is designed to be computationally fast and suitable to gap-fill data for annual mapping for machine learning. It therefore only uses existing values in the data set instead of estimated values like averages or linear inter- and extrapolations, which makes sure that any imputed values are from the same feature space subsequent models are trained on. It fills gaps in a pixel by deriving median pixel values from its ‘temporal neighbours’. If the same pixel has a value for the same period in the next and/or previous year, TMWM takes the median of that period in the two ‘adjacent’ years. If the pixel had no value in the same period of the previous or next year, the ‘window’ expands to include values for that period in increasingly earlier and later years. If no value exists for the specified period in any year, TMWM will derive the pixel’s median value in the previous and next period of the same year. If that fails, the ‘window’ will again expand to include the previous and next period of increasingly earlier and later years. If no value can be found in these ways, the window encompasses all values in the entire time series of the pixel. If the pixel lacks data throughout the entire time series, the value is imputed with a local spatial average and assigned a QA value of 100.


Comparison of 10m and 30m Sentinel-2 and Landsat time-series
Bing Terrain
Sentinel-2 NIR 10 m March–December 2018
Sentinel-2 NIR 30 m June–October 2018
Landsat NIR 30 m June–October 2018
Zurich, Switzerland
Blue
Green
Red
NIR
SWIR1
SWIR1
Therm.
2021
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000

A Comparison of spatial and temporal resolution, as well as covered time range, of Sentinel-2 and Landsat data.



GLAD ARD Interval ID		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Day of year	start	1	17	33	49	65	81	97	113	129	145	161	177	193	209	225	241	257	273	289	305	321	337	353
	end	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240	256	272	288	304	320	336	352	366
Date	start	01-01	17-01	02-02	18-02	05-03	21-03	06-04	22-04	08-05	24-05	09-06	25-06	11-07	27-07	12-08	28-08	13-09	29-09	15-10	31-10	16-11	02-12	18-12
	end	16-01	01-02	17-02	04-03	20-03	05-04	21-04	07-05	23-05	08-06	24-06	10-07	26-07	11-08	27-07	12-09	28-09	14-10	30-10	15-11	01-12	17-12	31-12
Season	Meteorological	Winter			Mixed	Spring					Mixed	Summer					Autumn					Mixed	Winter	
	Astronomical	Winter					Spring					Mixed	Summer					Mixed	Autumn					Mixed
EcoDataCube quartile (GLAD intervals)		Q1 (7)					Q2 (6)						Q3 (5)					Q4 (5)					Q1 (7)	

B Comparison of the 23 16-day Landsat GLAD ARD intervals, the meteorological and astronomical seasons in Europe and the four quartile periods used in this work.

Figure 2.3: Overview of the spatial and temporal resolution and covered time range of the presented Landsat and Sentinel-2 data. Fig. A shows a comparison of the spatial and temporal resolution of the satellite imagery data sets included in EcoDataCube, as well as the available bands and time range covered per data set. Microsoft Bing Terrain screenshot © Microsoft Corporation. Fig. B shows which GLAD ARD intervals were used to generate which quartiles, and how this compares to the two commonly defined seasons in Europe. GLAD ARD intervals that would be in different seasons depending on the definition are marked in bright orange.

TMWM attempts to derive a value for missing pixels in three phases. Within each phase, it will first try to use only eligible values within the last X and subsequent X years, where X is half the window size. If that fails, it will try to use values within double the search range. Then, it will try to use eligible values from any year in the time series. If that yields no result, it will move to the next phase, increasing the number of eligible values by including more intra-annual time periods in its search. These phases differ in which metric they calculate and which periods they use in the following way:

1. The median of the same period from different years;
2. An average of the medians of the previous and next periods;
3. The median of all periods.

We validated TMWM's performance on the temporal composite Landsat data from 2000 to 2020. This was done by sampling 100 pixels from each 6,750 30 km tiles and extracting the time series for the 50th percentile of each band and each period. If tiles were not completely covered by land, the number of sampled pixels was reduced in proportion. We then created a boolean mask based on whether the pixel had a value in that year. This 'missing value mask' was then inverted in order to introduce additional simulated gaps, reproducing existing patterns in which missing values occur in the data. We then used the TMWM method to fill the simulated gaps, and compare the imputed values to the original values. We quantified the performance of TMWM by deriving the root mean square error (RMSE) for each band and quartile. As different bands have different value ranges, we normalized the RMSE per band by dividing it by the range (maximum value minus minimum value) of that band's values in the entire data set. This normalized RMSE (NRMSE) allows for a standardized comparison of performance across bands and years.

Sentinel-2

We created two Sentinel-2 2018–2021 time series of the study area: one series with annual values at 10 m resolution limited to the red, green, blue, and NIR bands, and one quartile/seasonal series at 30 m resolution which also includes the SWIR1 and SWIR2 bands. The data was processed in four steps (Fig. 2.2-E):

1. Computing temporal composites (annual and quarterly) for each Sentinel-2 tile;

2. Reprojecting and resampling the tiled composites to `EPSG 3035` (`https://epsg.io/3035`) at 10 m and 30 m resolution;

3. Mosaicking the resampled composites over their respective orbital tracks;

4. Stitching the orbital mosaics together.

The Sentinel-2 mosaics were built from Sentinel-2 Level 2A (S2L2A) imagery (BOA reflectance generated with scene classification and atmospheric correction algorithms), for six bands (see Table 2.3). The mosaics span 1,028 tiles. Each scene over these tiles, imaged over the time period from winter 2017/2018 (2.12.2017) to winter 2020/2021 (1.12.2021), was collected from the AWS Sentinel-2 repository, which is hosted as a Requester Pays S3 bucket (accessible without charge from AWS instances). The mosaicking and temporal aggregation was performed with the *s2mosaic* functionality of the *eumap* python package [1]

Table 2.3: Sentinel-2 bands included in the data cube with their respective wavelengths and original imaging resolutions

Band number	Band name	Wavelength (nm)	Resolution (m)
B02	Blue	496.6–492.1	10 m
B03	Green	560–559	10 m
B04	Red	664.5–665	10 m
B08	NIR	835.1–833	10 m
B11	SWIR1	1613.7–1610.4	20 m
B12	SWIR2	2202.4–2185.7	20 m

Sentinel Temporal Composites

We created annual composites of the four 10 m resolution Sentinel-2 bands: green, blue, red, and NIR. We aggregated all scenes captured between March 21st and December 1st to three percentiles: 25th, 50th (median) and 75th, excluding all pixels flagged in the S2L2A cloud and cloud shadow masks. These were then resampled to `EPSG:3035`. We also created quarterly percentile composites of all bands (green, blue, red, NIR, SWIR1, and SWIR2) with date ranges matching the Landsat quartile periods. The tile-wise composites were resampled to 30 m resolution in `EPSG:3035` and aggregated to the same quartiles as the Landsat ARD data. Since scenes that are acquired along the same orbital track are very likely to be imaged in equivalent conditions, seamless mosaicking is possible by averaging the overlapping pixels. A total of 32 orbital tracks were used (Fig. 2.2-D). The final mosaicking was done by stitching together the orbital mosaics with weighted averaging of overlapping pixels. For each pair of overlapping pixels, the relative distance from their respective orbital track was calculated (from 0 to 1, with 1 being the distance to the neighboring track). Overlapping pixels with a relative distance within the range of 0.4 to 0.6 (inclusive) were averaged by using their relative distances as weights, while pixels with a relative distance below 0.4 were designated the correct value, regardless of overlap, as artifacts were often observed at relative distances above 0.6.

[1] Code is available at `https://gitlab.com/geoharmonizer_inea/eumap/-/tree/master/eumap/datasets/eo/s2mosaic`, documentation at `https://eumap.readthedocs.io/en/latest/notebooks/09_sentinel2_mosaicking.html`

Digital Terrain Model

Although a continental-scale DTM called *"EU-DEM"* already exists, it is based on SRTM and ASTER GDEM [131]; we have built a DTM for the study area using more detailed and more up-to-date elevation products: MERIT DEM [308], ALOS AW3D [261] and GLO-30 (`https://doi.org/10.5270/ESA-c5d3d65`). To generate the best estimate of the land surface/terrain elevation, we used 10 input variables, obtained by overlaying the training points on 5 elevation and 5 auxiliary raster data sets (Table 2.4), and an ensemble machine learning approach trained on a random sample of 7 million (randomly sampled from the 28 million points available) Global Ecosystem Dynamics Investigation (GEDI) points and 2 million ICESat-2 points. We specifically used GEDI level 2B points `elev_lowestmode` column and ICESat-2 (ATL08) `h_te_mean` column, which in both cases represent the *lowest* elevation observed i.e. the most likely bare ground height. We combined these two ground-truth data sets in a total of 9 million data points, and then built a machine learning model that we used to predict height without canopy and report predictions errors.

The ensemble model was composed of a random forest, a cubist model, and a generalized linear model in the `mlr` R package. These models make separate predictions using all input variables. Their estimates are then used as input for the `makeStackedLearner` meta-learner function [16]; in this case a linear regression model that makes the final elevation estimate for each pixel. This approach can be compared to the approach of Hawker et al. [99] who produced a global map of elevation with forests and buildings removed, however, in our approach we also use continental data set EU-DEM and numerous additional layers, at the cost of higher complexity and computational effort. We also validated the four DEMs used as input for the ensemble by comparing their values with the GEDI/ICESat-2 training data set, which can be considered ground-truth data.

Note that the predicted elevations are based on the GEDI data, hence the reference water surface (WGS84 ellipsoid) is about 43 m higher than the seawater surface for a specific EU country. Before modeling, we corrected the reference elevations to the Earth Gravitational Model 2008 (EGM2008) by using the 5-arcdegree resolution correction surface [194].

We assessed the accuracy of the ensemble model and the resulting DTM by performing k-fold spatial cross-validation [154]. We divide the study area into square blocks of 30×30 km, which are grouped to provide folds for the cross-validation approach. Because the publications describing the input DSMs do not report accuracy with comparable metrics or criteria, we also validated each input data set with our training/cross-validation data set i.e. by using the GEDI and ICESat-2 points. This provides an objective comparison between each input DTM, and the cross-validation predictions of the ensemble. The produced Ensemble DTM of Europe and prediction errors were provided as GeoTIFFs using Integer format (elevations rounded to 1 dm) and have been converted to Cloud Optimized GeoTIFFs using GDAL 3.1.4.

Table 2.4: Overview of data sets used as input for the ensemble that produced our DTM

Dataset	Producer	Source
MERIT DEM	University of Tokyo Global Hydrodynamics lab	[309]
EU-DEM	European Environmental Agency (EEA)	[177]
ALOS AW3D0	JAXA Earth Observation Research Center (EORC)	[259]
GLO-30	European Space Agency (ESA)	[60]
Canopy Height	UMD GLAD	[208]
Surface Water Probability	European Commission Joint Research Centre (JRC)	[198]
Tree Cover	Global Forest Watch	[96]
Bare ground cover	UMD GLAD	[96]
Pan-European Land cover	Humboldt University of Berlin	[199]

Land Cover Classification Experiments

Because the intended purpose of the presented data sets is to facilitate annual mapping with machine learning, we compare the usefulness of 30 m Landsat, 30 m Sentinel-2, and 10 m Sentinel-2 for land cover classification. We do this by training several Random Forest (RF) models on 300,543 observations from the European Land Use and Land Cover Survey (LUCAS) data set harmonized by d'Andrimont et al. [47] to predict the 8 LUCAS level-1 land cover classes; each of these RF models uses a different combination of the feature space provided by the data cube (Sentinel-2 10 m and 30 m, Landsat, and DTM). Additionally, to investigate the added value of the multi-decade harmonized Landsat time series, we trained RF models on all 1.4 million available LUCAS observations from 2000–2020. For each classification task we used a RF classifier with 100 trees, 1 minimum sample per leaf, and 2 minimum samples per node, implemented in Python 3.8.6 using Scikit-Learn . The maximum number of features per tree was set to the square root of the amount of total features. We assess the performance of each model through both 5-fold cross-validation on its training set, and by validating each model on one randomly sampled left-out test data set of 33,394 LUCAS observations from 2018 and 2019. The different combinations of time range, data set usage, and train/test points are presented in Table 2.5.

Results

Landsat

Downloading the 2000–2020 time-series of 16-day composites for 1,149 1–degree GLAD geotiffs of 72.3GB each amounted to approximately 81 TB of data, or 1 TB per band and year. Compressing the scaled long integer values to Byte format and aggregation into temporal composites reduced this to 29 GB per band per year, constituting a size reduction of about 97.1%. Removing all pixels not labeled as having *"clear sky"* in the GLAD metadata, resulting in an average of 5.83% empty pixels in spring, 19.7% in

Table 2.5: Overview of land cover classification experiments that were performed to quantify the added value of each data set to the data cube.

Satellite	Resolution	Time range	DTM used	Training points
Landsat	30 m	2000–2020	Yes	1,443,227
Landsat	30 m	2000–2020	No	1,443,227
Landsat	30 m	2018–2021	Yes	300,543
Landsat	30 m	2018–2021	No	300,543
Sentinel-2	30 m	2018–2021	Yes	300,543
Sentinel-2	30 m	2018–2021	No	300,543
Sentinel-2	10 m	2018–2021	Yes	300,543
Sentinel-2	10 m	2018–2021	No	300,543
Sentinel-2	10 m + 30 m	2018–2021	Yes	300,543
Sentinel-2	10 m + 30 m	2018–2021	No	300,543
Landsat + Sentinel-2	30 m	2018–2021	Yes	300,543
Landsat + Sentinel-2	30 m	2018–2021	No	300,543
Landsat + Sentinel-2	10 m + 30 m	2018–2021	Yes	300,543
Landsat + Sentinel-2	10 m + 30 m	2018–2021	No	300,543

summer, 11.73% in autumn, and 54.29% in winter, when aggregated to quarterly temporal resolution. Fig. 2.4 shows that, on average across all years, more gaps occur in Scandinavia and the northern Baltic countries. Fig. 2.4-E shows that the winter quartile of each year consistently had the highest number of gaps, followed by summer in all years except 2003. It also shows a clear reduction in gaps after the winter of 2012–2013 and the inclusion of Landsat-8 data in the archive.

Sampling 100 gap-filling validation pixels for each 30 km tile in proportion to its land area resulted in 566,454 pixels from 6,750 tiles. Table 2.6 shows the average gap-filling NRMSE respectively per band and quartile. Our validation shows that the lowest gap-filling performance was on the NIR band in Q2 (Spring) with a NRMSE of 4.41%, while the thermal band in Q2 was filled the most accurately with a NRMSE of 0.66%. More generally, Q3 (Summer) was gap-filled the most accurately with an average NRMSE of 2.09% , while Q4 (Winter) was gap-filled the least accurately with an average NRMSE of 2.36%. Fig. 2.5 shows the spatial variability of gap-filling NRMSE per quartile, indicating a consistent higher error rate in mountainous areas, especially the Alps and the Scandinavian mountains.

Sentinel-2

A total of 190,884 Sentinel-2 scenes were processed, ranging from 13 to 119 scenes per tile and quartile (see Table 2.7), amounting to a data set size of roughly 15.5 TB per quartile, or 62 TB per year. The input data for each annual composite of Sentinel-2 10 m

Table 2.6: gap-filling validation NRMSE (in percentage) per band and quartile, as well as band & quartile averages.

Band	Q1 (Winter)	Q2 (Spring)	Q3 (Summer)	Q4 (Autumn)	Average
Blue	2.10	1.79	1.84	2.15	1.97
Green	1.64	2.03	1.65	1.99	1.83
Red	2.37	2.25	2.30	2.03	2.24
NIR	3.74	4.41	3.69	4.34	4.05
SWIR1	2.46	2.88	2.43	2.92	2.67
SWIR2	2.20	2.18	2.08	2.37	2.21
Thermal	0.68	0.66	0.67	0.71	0.68
Average	2.17	2.31	2.09	2.36	2.23

RGB+NIR was an average of 62 TB per year. Aggregating them to annual composites reduced the size to roughly 0.3 TB per band, or 1.2 TB for the four bands, resulting in 98.1% compression. The 10 m resolution product had an average of 0.397% gaps per year, with the median among 30 km tiles being 0%. Fig. 2.6 shows that most 30 km tiles with gap percentages above 0.1% occur on tiles next to water.

Table 2.7: Number of Sentinel-2 scenes processed per Sentinel-2 tile and quarter

Quartile	Max	Min	Mean	Standard Deviation
Q1 (Winter)	84	16	47.95	11.82
Q2 (Spring)	119	19	53.07	17.49
Q3 (Summer)	100	16	43.87	14.65
Q4 (Autumn)	81	13	40.07	10.38

The input data for each annual set of Sentinel-2 quarterly composites at 30 m RGB+NIR+SWIR were identical to those used for the annual 10 m composites, amounting to an average of 62 TB per year. Aggregation to quarterly composites reduced the size to 0.2 TB per quarter, or 0.8 TB yearly, resulting in a compression of 98.7%.

Fig. 2.6-E shows that most gaps occur each year in the first quarter, especially in 2018. Except in 2018, the second highest number of gaps occurred in quarter 4 (September-December). Figs. 2.6-A–D shows that across all years, Northern Scandinavia has the least gaps in quarter 2 (March-June), and that relatively more gaps occurred in a stripe pattern across Europe in quarter 1 (December-March).

Digital Terrain Model

Results of modeling terrain with 5 million GEDI and 2 million ICESat-2 elevation measurements, using 4 existing DEMs and 6 auxiliary data sets as input variables, show a maximum cross-validation RMSE of 6.54 m for absolute accuracy of predicting terrain (bare-earth) height with majority of errors between 2–3 m (25% and 75% quartiles). MERIT DEM [308] is the most correlated DEM with GEDI and ICESat-2 points, most likely because it has been systematically post-processed and the majority of canopy problems have been removed (Fig. 2.7).

Fig. 2.8 shows the average RMSE of the four input DEMs per 30 km tile and the standard deviation of these four values. Our results suggest that MERIT and EUDEM are less accurate in large parts of Sweden, while GLO30 and AW3D are less accurate in central Europe. Furthermore, red lines are visible on these maps that match GEDI orbits and no natural features. Table 2.8 compares the RMSE of each DEM and the Ensemble DTM. This shows that MERIT DEM had both the lowest RMSE (8.45 m) and the highest variable importance in the model. In summary, our results show that the DTM produced by the ensemble is approximately 2 meters more accurate than MERIT in the area of interest. A copy of the all inputs, regression-matrix and outputs produced, including the code used to fit models and produce predictions, is available via `https://doi.org/10.5281/zenodo.4056634`.

Table 2.8: RMSE of the four input DEMs and the output DTM produced with ensemble machine learning. The RMSE of the input DEMs was acquired by comparing them to the values of 7 million GEDI/ICESat-2 points. The RMSE of the DTM was acquired through spatial cross-validation of the ensemble model. Variable importance from the random forest in the ensemble is included for the input DEMs.

Dataset	RMSE	Variable importance
MERIT	8.451	430 B
AW3D	9.858	291 B
EUDEM	9.806	132 B
GLO30	9.900	201 B
EcoDataCube DTM	6.544	NA

Land Cover Classification Tests

The random forest utilizing all four data sets (DTM, Landsat, Sentinel-2 10 m and 30 m) achieved the highest cross-validation scores (0.761) and second-highest test score (0.767). In general, models with multiple data sets in their feature space (DTM, Landsat, Sentinel-2 at 10 m spatial resolution, and/or Sentinel-2 at 30 m) outperformed models using a single data set. However, the highest test score of 0.774 was achieved by the Landsat model trained on 300,543 LUCAS observations spread out across 2000–2020, while this model had a relatively low cross-validation score of 0.715 (see Table 2.9). It appears that models utilizing DTM variables achieved higher cross-validation and test F1–scores in every experiment except for the Sentinel-2 10 m model. This model also achieved the lowest cross-validation and test scores.

Fig. 2.9 shows that models trained on data sets with multiple satellite sources and resolutions generally outperformed single-source or single-resolution models. Fig. 2.9-A shows that *“Shrubs“* and *“Wetlands“* were more accurately classified by models only using 30 m data sets. Fig. 2.9-B shows that Sentinel-2 data was more useful for classifying *“Artificial”* and *“Water areas”*, while being less useful for classifying *“Shrubs”* and *“Wetlands”*. For these classes, Landsat data yielded higher F1-scores. Using the DTM and its derived variables generally yielded slight performance increases, except for *“Crops“* and *“Bare Ground“*. Fig 2.9-A shows that including the DTM in the feature space mainly lead to higher accuracy when classifying *“Shrubs”* and *“Wetlands“*.

Table 2.9: Overview of land cover classification tasks that were performed to compare the usefulness of the different products, with weighted averaged F1-scores from cross-validation and from predicting on test set.

Dataset	Resolution	Time range	DTM	Training points	CV F1-score	Test F1-sc
Landsat	30m	2000–2020	Yes	1,273,518	0.710	0.738
Landsat	30m	2000–2020	No	1,273,518	0.698	0.730
Landsat	30m	2000–2020	Yes	300,543	0.715	**0.774**
Landsat	30m	2000–2020	No	300,543	0.707	0.768
Landsat	30m	2018–2021	Yes	300,543	0.731	0.742
Landsat	30m	2018–2021	No	300,543	0.724	0.736
Sentinel-2	30m	2018–2021	Yes	300,543	0.746	0.753
Sentinel-2	30m	2018–2021	No	300,543	0.741	0.748
Sentinel-2	10 m	2018–2021	Yes	300,543	0.705	0.713
Sentinel-2	10 m	2018–2021	No	300,543	0.706	0.715
Sentinel-2	10 m + 30m	2018–2021	Yes	300,543	0.758	0.760
Sentinel-2	10 m + 30m	2018–2021	No	300,543	0.751	0.756
Landsat + Sentinel-2	30m	2018–2021	Yes	300,543	0.750	0.757
Landsat + Sentinel-2	30m	2018–2021	No	300,543	0.747	0.755
Landsat + Sentinel-2	10 m + 30m	2018–2021	Yes	300,543	**0.761**	0.767
Landsat + Sentinel-2	10 m + 30m	2018–2021	No	300,543	0.758	0.765

Discussion

We have demonstrated a full methodological framework for processing various EO data for the purpose of producing an open Data Cube for Europe, evaluated all steps, and investigated the value of combining its component data sets using machine learning applications. Our key findings indicate that:

- Combining all four data sets produced in this work (DTM, Landsat 30 m, Sentinel-2 30 m and Sentinel-2 10 m) yields the highest land cover classification accuracy, with different data sets improving the results for different land cover classes;

- When used separately, the 2000–2020 Landsat data set can be used to model longer time series. In our experiments, models trained on LUCAS observations in this longer time span generalized better than those trained on an equal amount of points, but only sampled from 2018–2019;
- Ensemble machine learning can be used as a data fusion technique to combine global elevation models and create an optimized DTM that is more accurate in the area of interest, based on an independent validation;
- Accuracy and visual assessment of the four input DEMs suggests that DTM could still be much improved if countries would donate their national higher resolution elevation data.

In the next sections we discuss some remaining limitations of the Data Cube and suggest possible strategies to overcome them.

Suitability of Temporal Composite Design

We recognize that the choice for aggregating the 23 × 16–day GLAD measurements into temporal composites that approximate the typical four seasons in central Europe imposes some limitations. Firstly, seasonality differs per region, even within the study area. This can cause differences in performance between regions with a matching seasonality and regions with a different one. This may be a potential explanation for the poorer accuracy of land use / land cover classification along the Mediterranean coast in **Witjes** et al. [301], for which this data set provided a substantial part of the feature space. Secondly, the loss of temporal resolution likely hinders the accurate classification of dynamic classes that are distinguished by intra-seasonal variation, such as different crop types [291] or other modeling tasks involving vegetation phenology [317]. However, Zhao et al. [316] found that choosing an appropriate temporal compositing strategy can reduce the need for a higher temporal resolution.

While a monthly aggregation would likely help solve these issues, but would also pose new challenges: it would be more complicated to derive from the 23 measurements, as they do not perfectly match the 12 months in the Gregorian calendar. In addition, a monthly aggregation would retain more gaps in the data. While the quarterly aggregation has less temporal resolution, the three percentiles quantify some intra-seasonal variability when multiple pixel values are available per quarter. This approach retains information on the variability in the growing seasons, while reducing the lack of data in an efficient way.

Another solution would be to use a non-symmetrical approach, where the growing season is divided into smaller time units, while the non-growing season is aggregated to a higher extent. As there are far fewer gaps in the quartiles roughly covering the growing season (Q2 and Q3, March-September; see Fig. 2.4), this approach would yield a data set that is: (1) more detailed where it matters, (2) requires less gap-filling, and (3) might be more suitable for gap-filling with TMWM. Such a technique, however, should only be used when constructing data cubes of areas with homogenous seasonal dynamics. Further research into the optimal temporal aggregation method for different subsequent modeling tasks would likely improve the usefulness of resulting data sets.

Gains and Limitations of gap-filling with TMWM

The TMWM algorithm is computationally efficient and only imputes sets of existing combinations of pixel values across bands. It does have important drawbacks, however: firstly, in regions where data for a specific period is extremely sparse or non-existent (e.g. Northern Scandinavia in winter), data for this period will be almost completely derived from other periods. This can severely hamper the performance of classification tasks when a model needs intra-annual dynamics to distinguish certain classes. For example, our validation suggests that the Landsat RGB bands are easier to gap-fill with TMWM than infrared bands, especially NIR. This phenomenon is more pronounced in spring, which may be caused by the more dynamic and variable nature of vegetation in that period each year. Filling NIR data with values from other seasons may be exceptionally problematic in this respect.

Secondly, any model trained on data using this method may be less suitable for the timely detection of changes between predicted classes. Because TMWM prioritizes previous and subsequent years when imputing missing values, the resulting feature space may stay constant while the actual situation on the ground has changed. This may make the chosen combination of temporal aggregation and gap-filling less suitable for annual change analysis.

The *qa_f* layers included for every year and quartile at `https://stac.ecodatacube.eu` allow users to programmatically identify all gap-filled pixels and replace them with a gap-filling method that is optimal for their own subsequent modeling.

Finally, no gap-filling was implemented on the Sentinel-2 data sets. Although the total percentage of gaps is very low in most areas (see Fig. 2.6, the Sentinel-2 data is not fully complete and therefore not 100% analysis-ready.

The validation method used to assess the accuracy of the TMWM algorithm was chosen because it reproduces existing nodata patterns as observed in the intra and inter-annual dynamics of our data set, which we expect to yield to a more realistic evaluation of its performance on real-world data. However, mirroring the time-series occurrence of gaps as a validation mask simulates a larger number of missing values than their actual occurrence. Because the algorithm subsequently has less data to derive filling values with, this may lead to an underestimation of the actual accuracy.

Applicability and Limitations of Digital Terrain Model

We validated each of the four input DTMs on the harmonized GEDI/ICESat points that we used to train our model, allowing a comparison between these data sets and with the predictions made by our ensemble. It must be noted that a thorough accuracy assessment of GEDI and ICESat-2 for generating 30 m resolution DTMS on an European scale is not available. This means that it is hard to quantify the highest attainable accuracy when using it as training data. We did notice, however, possible artifacts in the GEDI data. Fig. 2.8 shows that enough GEDI points matched poorly with each input DEM in line-shaped groups for them to cause relatively large average RMSE values in the 30 km spatial tiles. The effect is especially noticeable in Iberia, and does not match any natural features. It does, however, match the GEDI orbit track. This suggests that there might be some orbit issues affecting the local accuracy of the predictions.

For the ensemble DTM we have produced, we noticed that in many places canopy height is still visible on the hillshading images, indicating that even after using the canopy height, the true terrain elevation in forests is overestimated. Additional filtering is needed to remove human built objects in urban areas. Our Ensemble DTM has not been hydrologically corrected and will require additional processing before it can be used for spatial modeling. Furthermore, several EU countries such as Belgium, the Netherlands, and Denmark have high-resolution terrain models built from LIDAR surveys. The ensemble DTM could be further improved by merging such publicly available data. Most importantly, it should be compared with the recently published 30 m global map of elevation with forests and buildings removed [99].

Usefulness of EO Data for Land Cover and Land Use Mapping

The land cover experiments, while limited in scope, clearly show that combining different data sets in the data cube improved modeling performance for classification tasks. This demonstrates the value of spatially, temporally, and spectrally harmonized multivariate data cubes such as the one presented in this work. For some of the 8 land cover classes, the combined models were outperformed by a model using only a specific data set or resolution (e.g. Sentinel-2 models when classifying water areas, and 30 m resolution models when classifying wetlands).

The different Landsat-only experiments also suggest that sacrificing spatial resolution in order to access longer time-series of training data may yield better subsequent modeling results; the 300 K 2000–2021 Landsat model outperformed the other Landsat models on the test data set. This matches findings by **Witjes** et al. [301] and Pflugmacher et al. [199]. However, the lower performance of the 1,273K 2000–2020 Landsat model suggests that using a larger training data set does not necessarily improve model performance.

Shrubland and bare land were consistently the least accurately classified land cover classes in each experiment. This was particularly the case for models trained only on 10 m Sentinel-2 data (see Fig 2.9-A), even when taking into account their lower performance across all classes. This lower performance was likely affected by the lower number of available bands, i.e. the lack of SWIR and NIR. This possible explanation is supported by the fact that models combining 30 m and 10 m Sentinel-2 data outperformed Landsat-only models on bare land, The model trained only on Sentinel-2 data however did outperform other models when classifying water areas, achieving a slightly higher score than even the model combining all data sources. These findings, combined with indications that the Sentinel-2 SWIR bands are highly useful for distinguishing between different tree species [124] suggest that incorporating both 10 m and 30 m Sentinel-2 data sets as part of the data cube enhances subsequent land cover modeling efforts.

The Cost of Accessibility and Analysis-Readiness

The ultimate goal of this work is to present an analysis-ready data cube that is as useful as possible to as many different users as possible, with an emphasis on time-series classification tasks. To this end, we have greatly – sometimes by more than 90% – reduced the size of the input data, reduced the number of gaps to below 1% for both Landsat and Sentinel-2 time-series, and made all products freely accessible through modern technologies and formats such as COG and SpatioTemporal Asset Catalogs (STAC) without requiring any form of user registration. However, this emphasis on accessibility and analysis-readiness comes at the cost of both temporal and spectral detail.

We aggregated a 23–part Landsat time-series and a Sentinel-2 time-series with a highly variable number of observations (13–119) into four quartiles with 3 percentiles each: a standardized total of 12 values. While this low temporal resolution approach requires less gap-filling and makes the data set less susceptible to error propagation, it may not be detailed enough to detect certain classes in some classification tasks such as those relying on vegetation phenology. Our land cover experiments did not try to distinguish between different vegetation types or inform vegetation growth models, which might only be discernible when having a higher number of observations in the growing season [317]. On the other hand, Vuolo et al. [291] present significant increases in Observer's Accuracy (OA) of crop type classification when using multi-date data, and suggest that this method alleviates the issue of finding the optimal temporal window where the highest single-date accuracy can be achieved.

We did not analyze the effects of compressing the Landsat values from long unsigned integer (0–40,000) and unsigned 16-bit integer (0–65,535), respectively, to Byte (0–255). This loss in precision may lead to similar limitations, especially in the case of infrared for vegetation-related modeling. However, Bonannella et al. [19] achieved high accuracy

(0.81–0.89) using the Byte-scale Landsat data when classifying multiple different tree species, suggesting this may not, after all, be a significant limitation.

Future Work

While the EcoDataCube layers presented in this work bridge a large part of the gap between data and users through a consistent design philosophy, they are not completely finished. Notably, gaps remain in both Landsat and Sentinel-2 data. Work continues on a more efficient and effective gap-filling methodology optimized for classification tasks with probability-based post-processing such as the methodology used in **Witjes** et al. [301]. The effects of different temporal resolutions and percentile usage will need to be compared in order to reach the optimal method for any specific task. Purely relying on time-series of the same pixel may be too limiting for areas with frequent and consistent gaps, such as Scandinavia and mountainous regions. A less stringent QA-informed removal of pixels could reduce this issue, but might require additional processing steps. Furthermore, GLAD will discontinue its current ARD Landsat archive, publishing a recomputed version that will be continued in the foreseeable future. While this necessitates a recomputation of all current EcoDataCube Landsat layers as well, it provides an excellent opportunity to compare and implement the next generation of temporal composition and gap-filling techniques.

The EcoDataCube platform already hosts more data sets that follow the same design philosophy of analysis-readiness and accessibility [293], allowing rapid and user-friendly comparisons and synthesis (see Fig. 2.10). Examples are the potential and realized distribution of 16 tree species [19], monthly airborne fine particulate matter levels [122], 43 CORINE land cover classes [301], and daily aerosol optical depth levels [123].

We aim to continuously extend the feature space of the data cube, both by producing new data sets and hosting harmonized products created by third parties. For instance, a backwards estimation of Sentinel-2 values to cover the same time period as the Landsat data set is under consideration. Data fusion approach such as the FORCE [77] provides a systematic solution for this and will likely be used for this purpose. There are thousands of data sets that could potentially be integrated in, and shared on, the EcoDataCube. For instance, with proper documentation and spatiotemporal harmonization, many of the over 1,500 data sets from the European Environmental Agency's archive (`https://www.eea.europa.eu/data-and-maps`) could be added to the feature space and breadth of analyses offered on the platform.

Conclusions

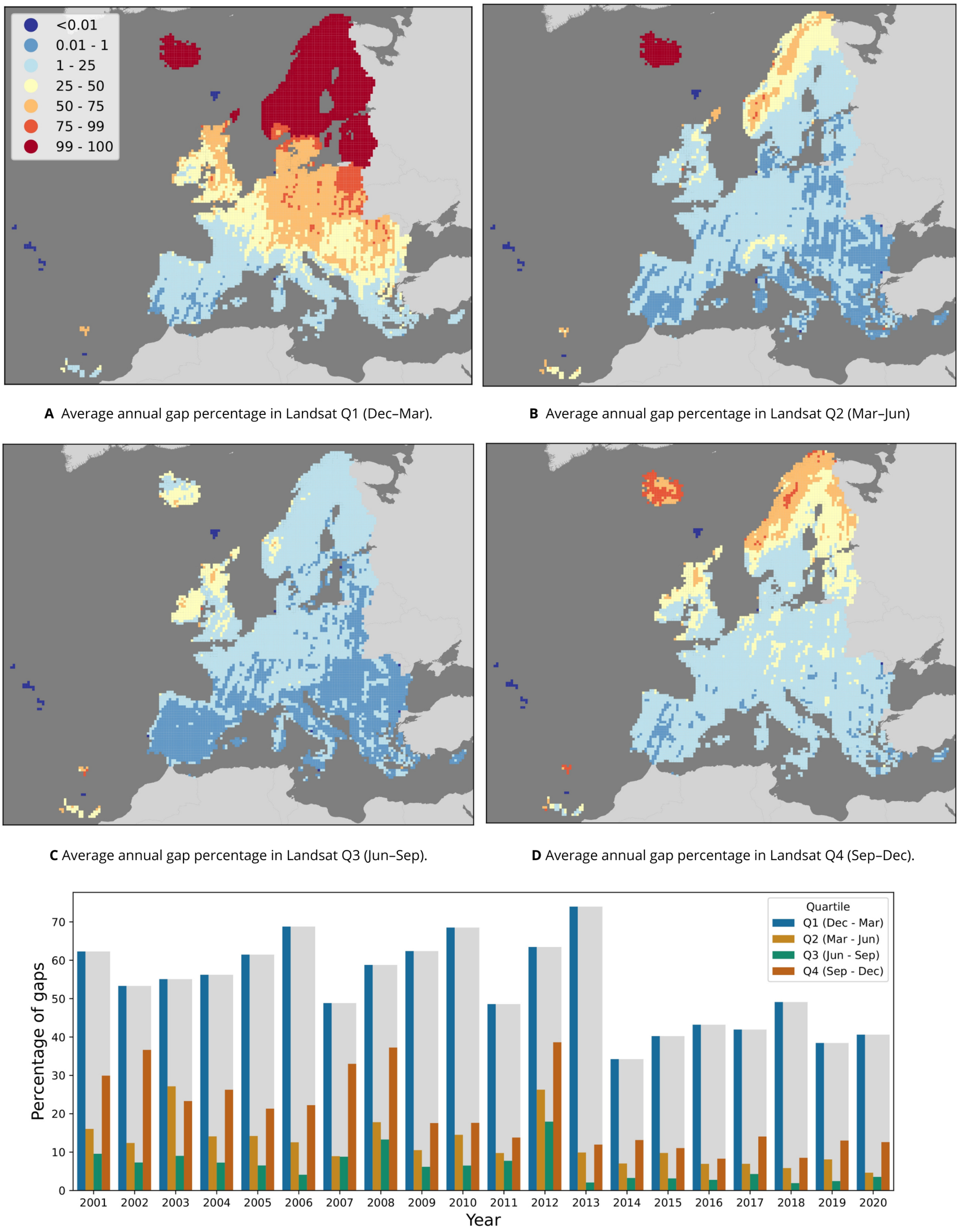
With the EcoDataCube data sets and platform, we present a spatiotemporally consistent, transparently and reproducibly processed continental-scale data set that is hopefully as accessible as possible across platforms. We intend it to reach the widest possible user base and to be put to as many different uses as possible — generating more value for society — while also facilitating collaboration and reproducibility. The intended uses of EcoDataCube include vegetation, soil, land cover and land use mapping projects, environmental monitoring by the EEA, and automated generation of data for statistical offices such as Eurostat (`https://ec.europa.eu/eurostat`).

The EcoDataCube data sets are hosted through modern cloud-based solutions that are both humanly and programmatically accessible without limitation and with minimal effort through STAC and the EcoDataCube platform, and will be continuously updated, maintained, and expanded.

The spatiotemporal harmonization and gap-filling processes described in this work ensure that users can easily combine different data sets to perform their analyses without the need for extensive preprocessing. The included validations and published QA data sets accompanying the TMWM-gap-filled Landsat data ensure that all limitations of the data set can be easily analyzed and communicated. As we continue our work to create and host more open spatiotemporal analysis-ready data, we encourage and invite all interested parties to use these data sets and to provide feedback, especially on inaccuracies or limitations, so that these can be addressed in future versions.

Code and data availability

The Python code used to download the original Sentinel-2 data and GLAD Landsat archives and aggregate them into temporal composites is available at `https://gitlab.com/geoharmonizer_inea/eumap` under MIT license. The code used to perform the landcover classification experiments can be found on `https://gitlab.com/research_m_witjes/ecodatacube`. The codebase primarily uses components of the *“eumap”* python package, which is available from `https://eumap.readthedocs.io/en/latest/`. The accompanying Docker image can run all python code used for this paper. Data from all Landsat and Sentinel bands, quarters, and percentiles is available under a CC-BY license and can be downloaded through STAC at `https://stac.ecodatacube.eu`. It can also be explored and accessed through the EcoDataCube platform. All data sets are presented in the ETRS89-extended / LAEA Europe projection (`EPSG:3035`).



A Average annual gap percentage in Landsat Q1 (Dec–Mar).

B Average annual gap percentage in Landsat Q2 (Mar–Jun)

C Average annual gap percentage in Landsat Q3 (Jun–Sep).

D Average annual gap percentage in Landsat Q4 (Sep–Dec).

E Percentage of gaps per year and quartile in Landsat.

Figure 2.4: Percentage of gaps per pixel in the Landsat 30 m data between 2000–2020. Figs. A, B, C and D show the annual average, calculated per 30 km tile (1 million pixels) for each of the four quartiles that the GLAD Landsat ARD product was aggregated to. Fig. E shows the percentage per year and quartile.

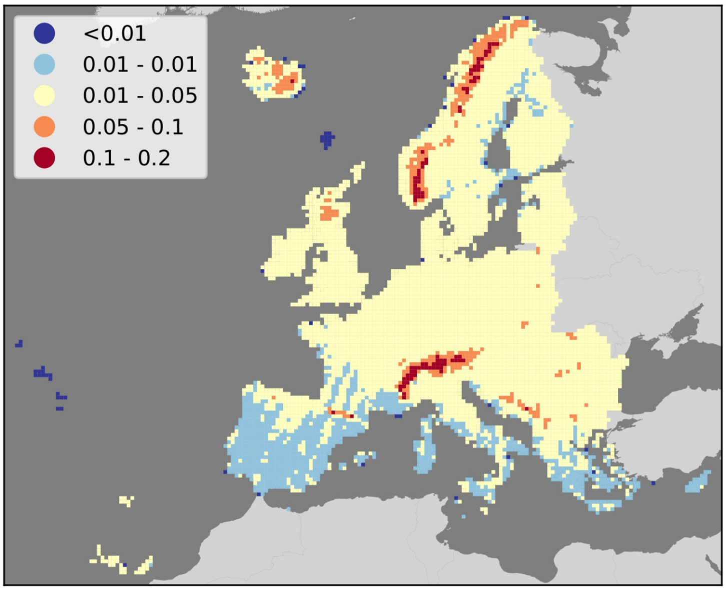
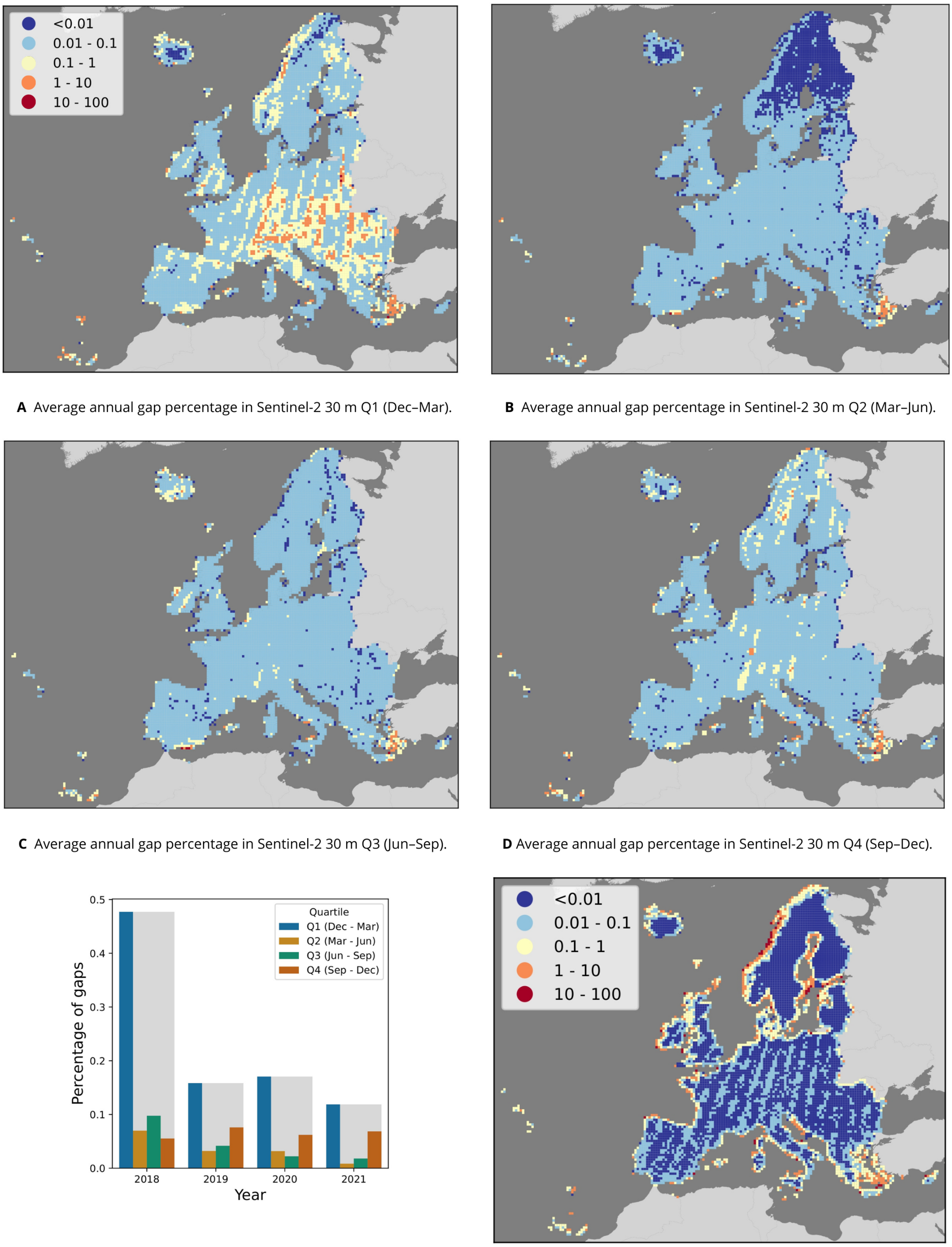


Figure 2.5: Map of average gap-filling NRMSE per 30 km tile in the study area. Results show consistently higher values in mountainous regions, and lower values in Southern Europe.



A Average annual gap percentage in Sentinel-2 30 m Q1 (Dec–Mar).

B Average annual gap percentage in Sentinel-2 30 m Q2 (Mar–Jun).

C Average annual gap percentage in Sentinel-2 30 m Q3 (Jun–Sep).

D Average annual gap percentage in Sentinel-2 30 m Q4 (Sep–Dec).

E Percentage of gaps per year and quartile for Sentinel-2 30 m.

F Average annual gap percentage in Sentinel-2 10 m.

Figure 2.6: Overview of gaps per pixel in the Sentinel-2 time series. Average annual percentage of gaps per quartile and 30 km tile (a, b, c, d) and in total (e) for the 30 m data, as well as average annual percentage per 30 km tile for the 10 m data (f).

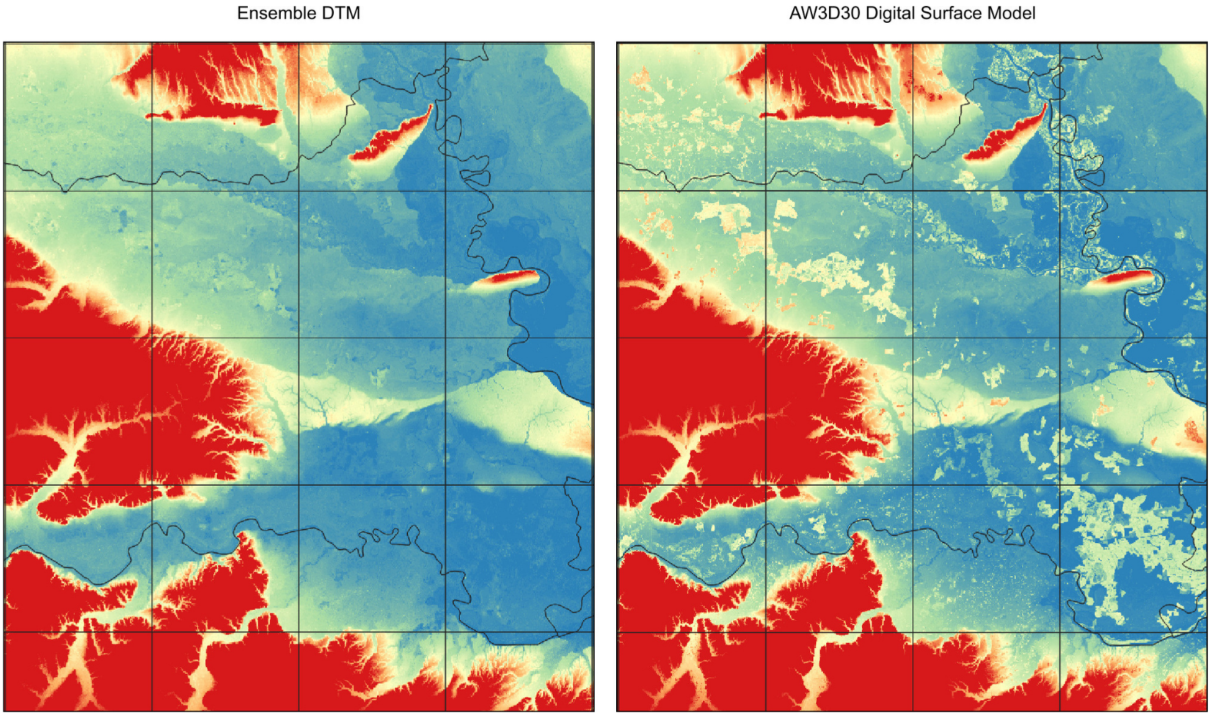
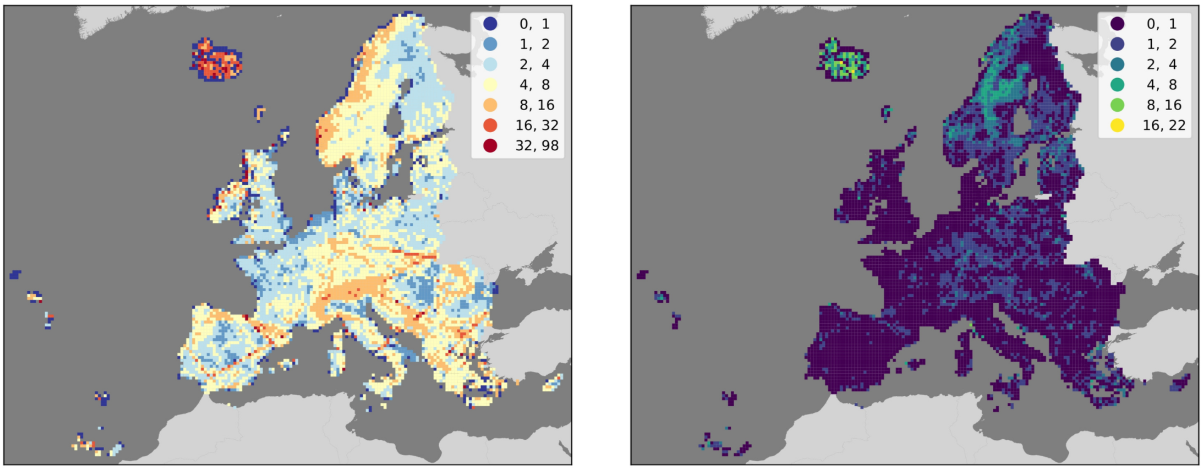


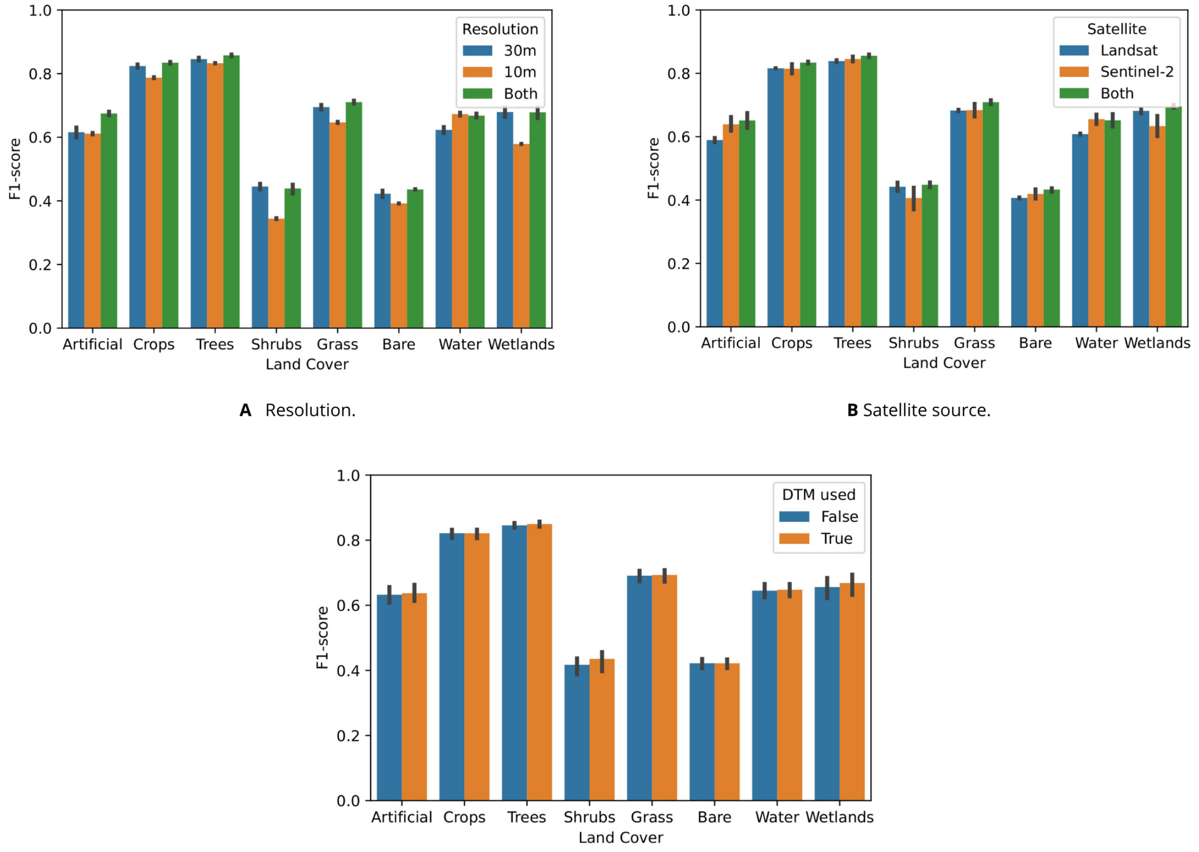
Figure 2.7: Comparison of Ensemble DTM (left) and the AW3D Digital Surface Model for the Pannonian plane in Eastern Croatia (right). Tree height visible on AW3D seems to be systematically removed with the help of machine learning. Grid showing 30 km tiles.



A Average RMSE of input elevation data.

B Standard deviation of RMSE of input elevation data.

Figure 2.8: Results of comparing GEDI and ICESat-2 measurements to AW3D, GLO30 EUDEM and MERIT values in 30 km tiles covering the study area. Fig. A shows the average RMSE across the four data sets, while Fig. B shows the standard deviation among RMSE values, representing the disagreement between the four data sets. Note the straight lines of higher average RMSE values in e.g. Iberia, which Fig. B suggests are consistent across all four data sets. Instead, the data sets disagree most in parts of Northern Europe, especially Iceland.

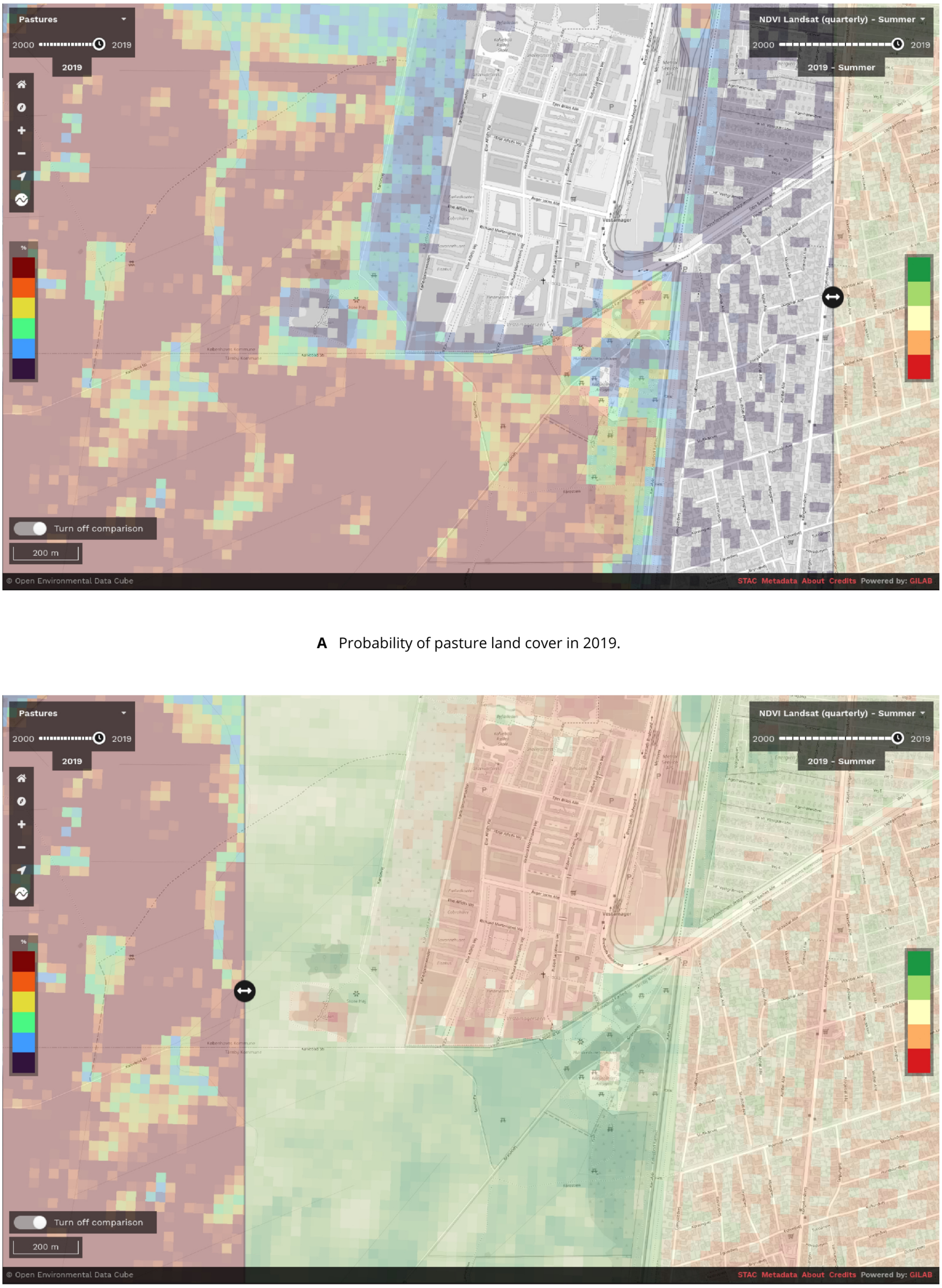


A Resolution.

B Satellite source.

C DTM usage.

Figure 2.9: Test F1-score of random forests trained to classify 8 LUCAS land cover classes, shown per class and aggregated based on which data sets were included in their feature space (A: spatial resolution, B: satellite source, and C: DTM). Fig. A shows that models trained on both 10 m and 30 m resolution data generally achieved the highest classification accuracy, but that models trained solely on 10 m (which is only Sentinel-2) classified water more accurately than models trained either 30m data sets or a combination of 30m and 10m. Especially the *"Shrubs"*, *"Grass"*, and *"Wetlands"* classes were predicted much less accurately by the 10 m models. Combining 10 m and 30 m resolution data lead to the largest performance increase for the *"Artificial"* class. Fig. B shows that models trained on data from both satellite types outperformed other models on every class except *"Water"*. Fig. C shows that models including DTM data in their feature achieved higher accuracy when classifying *"Shrubs"* and *"Wetlands"*.


Pastures
2000
2019
2019
NDVI Landsat (quarterly) - Summer
2000
2019
2019 - Summer
%
Turn off comparison
200 m
© Open Environmental Data Cube
STAC Metadata About Credits Powered by: GILAB

A Probability of pasture land cover in 2019.

B NDVI in the summer of 2019.

Figure 2.10: Screenshot of the EcoDataCube.eu data viewer showing A: the probability of pasture land cover in 2019 near Copenhagen, Denmark, and B: Landsat-derived NDVI in the summer of 2019.

Chapter 3

A Spatiotemporal Ensemble Machine Learning Framework for Generating Land Use / Land Cover Time-series Maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat

This chapter is based on:

M. Witjes, L. Parente, C. J. van Diemen, T. Hengl, M. Landa, L. Brodskỳ, L. Halounova, J. Križan, L. Antonić, C. M. Ilie, V. Craciunescu, M. Kilibarda, O. Antonijević, and L. Glušica. “A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat”. *PeerJ* 10 (2022), e13573. DOI: 10.7717/peerj.13573

Abstract

A spatiotemporal machine learning framework for automated prediction and analysis of long-term LULC dynamics is presented. The framework includes: (1) harmonization and preprocessing of spatial and spatiotemporal input datasets (GLAD Landsat, NPP/VIIRS) including 5 million harmonized LUCAS and CORINE Land Cover-derived training samples, (2) model building based on spatial k-fold cross-validation and hyper-parameter optimization, (3) prediction of the most probable class, class probabilities and model variance of predicted probabilities per pixel, (4) LULC change analysis on time-series of produced maps. The spatiotemporal ensemble model consists of a random forest, gradient boosted tree classifier, and an artificial neural network, with a logistic regressor as meta-learner. The results show that the most important variables for mapping LULC in Europe are: seasonal aggregates of Landsat green and near-infrared bands, multiple Landsat-derived spectral indices, long-term surface water probability, and elevation. Spatial cross-validation of the model indicates consistent performance across multiple years with overall accuracy (a weighted F1-score) of 0.49, 0.63, and 0.83 when predicting 43 (level-3), 14 (level-2), and 5 classes (level-1). Additional experiments show that spatiotemporal models generalize better to unknown years, outperforming single-year models on known-year classification by 2.7% and unknown-year classification by 3.5%. Results of the accuracy assessment using 48,365 independent test samples shows 87% match with the validation points. Results of time-series analysis (time-series of LULC probabilities and NDVI images) suggest forest loss in large parts of Sweden, the Alps, and Scotland. Positive and negative trends in NDVI in general match the land degradation and land restoration classes, with *"urbanization"* showing the most negative NDVI trend. An advantage of using spatiotemporal ML is that the fitted model can be used to predict LULC in years that were not included in its training dataset, allowing generalization to past and future periods, e.g. to predict LULC for years prior to 2000 and beyond 2020. The generated LULC time-series data stack (ODSE-LULC), including the training points, is publicly available via the ODSE Viewer. Functions used to prepare data and run modeling are available via the eumap library for python.

3.1 Introduction

Anthropogenic land cover change has influenced global climate since the Paleolithic [133] and continues to be a major driver of regional [202] and global [116] climate change. Furthermore, it is the single largest cause of global biodiversity loss [229], and has quantifiable consequences for the availability and quality of natural resources, water, and air [75]. Key applications of land cover change maps are to inform policy [59], analyze land-based emissions [112], and help estimate local climate extremes [256]. Quantifying land cover dynamics is often crucial for policy-making at regional and global levels [153, 240, 268].

Land cover mapping was initially done by visual interpretation of aerial photographs and later on with automated classification of multispectral remotely sensed data with semi-supervised or fully-supervised methods [70, 151, 265]. There are currently multiple global [27, 68] and regional [10, 46, 110, 161, 199] land cover products based on using Machine Learning and offering predictions (or their refinements) at high spatial resolutions for the whole of continental Europe (Table 3.1). The increasing number of land cover applications and datasets in Europe can largely be attributed to (1) the extensive LUCAS *in-situ* point data being publicly available for research, and (2) NASA's Landsat and ESA's Sentinel multispectral images being increasingly available for spatial analysis [151, 258].

However, not all land cover prediction systems perform equally. Vilar et al. [290] have done extensive evaluation of accuracy of the CLC products for period 2011–2012 using the LUCAS data and found that agreement with LUCAS was slightly higher for CCI-LC (59%; 18 classes) than for CLC (56%; 43 classes). Gao et al. [83] has evaluated accuracy of the global 30 m resolution products GlobeLand30 with 10 classes [39], and GLFCS30 with 18 classes [**zhang2020glc˙fcs30**] using the LUCAS point data and concluded that the GlobeLand30-2010 product agrees with LUCAS points up to 89%, while GLFCS30-2015 agrees up to 85%. The large difference in the agreement reported by Vilar et al. [290] and Chen et al. [39] can be attributed to the number of classes in the two studies: the absolute accuracy linearly drops with the number of classes [106, 283], and usually the accuracy results for 6–10 classes vs 40 classes can be up to 50% better.

Generally, the accuracy of European land cover mapping projects match those in other parts of the world. For example, Calderón-Loor et al. [32] achieved 90% producer's accuracy when classifying on 6 classes for 7 separate years between 1985 and 2015, using Landsat data of Australia. Tsendbazar et al. [273] reports similar accuracy levels for Africa. Likewise, Liu et al. [148] reports 83% accuracy on 7 classes with 34 years of GLASS data. Finally, the US National Land Cover Database reports accuracy of at least 80% for 16 classes at 30 m in 2001, 2004, 2006, 2008, 2011, 2013, 2016, and 2018 [111].

Inglada et al. [125] report a kappa score of 0.86 for mapping 17 land cover classes for France in 2014. The most-up-to-date land cover products for Europe by Malinowski et al. [161] report a weighted F1-score of 0.86 based on predicting 13 classes with 2017 Sentinel-2 data. The ESA's CCI-LC project classified land cover in three multiyear epochs (see Table 3.1), the last of which achieved an estimated producer's accuracy of 73% [4]. Their new WorldCover project (`https://esa-worldcover.org/`) aims for a consistent accuracy of at least 75% at 10 m spatial resolution. d'Andrimont et al. [46] recently produced a 10 m resolution European crop type map also by combining LUCAS and plot observations and achieved an overall accuracy of 76% for mapping 19 main crop types for year 2018.

Table 3.1: Inventory and comparison of existing land cover data products at finer spatial resolutions (≤300 m) available for the continental Europe.

Product / reference	Time span	Spatial resolution	Mapping accuracy	Number of classes	Uncertainty / Probability
CLC	1990, 2000, 2006, 2012, 2018	100 m (25 ha)	≤85%	44	N / N
ESA CCI-LC	1998-2002, 2003-2007, 2008-2012	300-m	73%	22	N / N
Batista e Silva et al. [10]	2006	100-m	70%	42	N / N
S2GLC [161]	2017	10 m	89%	15	N / N
Pflugmacher et al. [199]	2014-2016	30 m	75%	12	N / N
GLFCS30 [**zhang2020glc˙fcs30**]	2015, 2020	30-m	83%/71%/69%	9/16/24	N / N
Buchhorn et al. [27]	2015, 2016, 2017, 2018	100 m	80%	10	N / **Y**
ESA WorldCover	2020	10 m	≤75%	≤10	N / N
ELC10 [287]	2020	10 m	90%	8	N / N
ODSE-LULC (our product)	2000, 2001, . . . , 2019	30 m		43	**Y** / **Y**

Based on these works, it can be said that the state-of-the-art land cover mapping projects primarily aim at:

(a) Automating the process as much as possible so that land cover maps can be produced almost on monthly or even daily revisit times,

(b) using multi-source Earth Observation data, with especial focus on combining power of the Sentinel-1 and 2 data [287],

(c) producing data of increasingly high spatial and thematic resolution.

Although the modern approaches to land cover mapping listed in Table 3.1 report relatively high levels of accuracy, we recognize several limitations of the general approach:

- Common land cover classification products often only report hard classes, not the underlying probability distributions, limiting the applicability for use cases that would benefit from maximizing either user's or producer's accuracy of specific classes in the legend.
- Per-pixel information on the reliability of predictions is often either not reported or not derived at all.

- Many policy makers require time-series land cover data products compatible with legacy products such as CLC and CCI-LC, while most research produces general land cover maps for recent years only.
- Many continental- or global scale land cover mapping missions employ legends with a low number of classes. While achieving high accuracy, such generalized maps are of limited use to large parts of the policy-making and scientific communities.

Land cover data with higher thematic resolution have shown to help improve the performance of subsequent change detection [31], as well as the performance and level of detail of modeling land cover trends [42] and other environmental phenomena [33, 318]. Increasing thematic resolution while limiting the prediction to one trained classifier, however, poses several challenges: (1) training a single model on multi-year data requires extensive data harmonization efforts, and (2) the exponential increase of possible change types with each additional predicted class complicates the manual creation of post-classification temporal consistency rules.

With an increasing spatial resolution and increasing extent of Earth Observation (EO) images, the gap between historic land cover maps and current 10 m resolution products is growing [46, 283]. This makes it difficult to identify key processes of land cover change over large areas [285, 290]. Hence, a balanced and consistent approach is needed that can take into account both accuracy gains due to spatial resolution, and applicability for time-series analysis / change detection for longer periods of time.

The main objective of this paper is to present a framework for spatiotemporal prediction and analysis of LULC dynamics over the span of 20+ years at high thematic resolution, and to assess its usefulness for reproducing the CLC classification system at an annual basis at 30 m resolution. To properly assess the usefulness of the framework, we investigate whether spatiotemporal models (trained on observations from multiple years) generalize better to earth observation data from unknown years than spatial models (trained on observations from a single year). Furthermore, we investigate whether an ensemble machine learning pipeline provides more accurate LULC classifications than single classifiers. Finally, we provide an in-depth analysis of the feasibility to reproduce the CLC classification system by assessing the performance of our framework at various thematic resolution levels.

To this end, we present results of predicting 43 LULC classes from the CLC classification system for continental Europe using spatiotemporal EML at 30 m spatial resolution. These annual predictions are made by a single ensemble model trained on LULC observations ranging from 2000-2018 and a data cube consisting of harmonized annual multispectral Landsat imagery, derived spectral indices, and multiple auxiliary features.

We include the results of multiple accuracy assessments: Firstly, we use 5–fold spatial cross-validation with refitting [154, 219] to compare the performance of single-year and multi-year models, the performance of the separate component models of our ensemble, and the output of the entire ensemble. Secondly, we test the predictions of our ensemble on the S2GLC validation points, a dataset that was independently collected and published by Malinowski et al. [161].

We use, as much as possible, a consistent methodology, which implies:

1. Using consistent training data based on consistent sampling methodology and sampling intensity over the complete spacetime cube of interest LUCAS; d'Andrimont et al. [47]);
2. Using consistent / harmonized Earth Observation images based on the GLAD ARD Landsat product [206], Night Light images NPP/VIIRS [223] and similar;
3. Providing consistent statistical analysis per every pixel of the space-time cube and per each probability;

Our modeling framework comes at high costs however: the data we have produced is about 50–100 times larger in size than common land cover products with the total size of about 20 TB (Cloud-Optimized GeoTIFFs). A dataset of such volume is more complex to analyze and visualize. To deal with the data size, we ran all processing in a fully automated and fully optimized high performance computing framework. We refer to the dataset we have produced as ODSE-LULC.

In the following section we describe how we prepared data, fitted models, tested spatial vs spatiotemporal models, and fitted pixel-wise space-time regressions for NDVI and probability time-series. We then report the results and discuss advantages and limitations of spatiotemporal EML, and suggest what we consider could be next development directions and challenges.

3.2 Material and methods

Overview

The annual land cover product for continental Europe was generated using spatiotemporal modelling approach. This means that all training points are overlaid with EO variables matching both their location and their survey date, so that classification matrix contains spacetime coordinates (x, y, t); then a spatiotemporal model is fitted using the classification matrix. A detailed overview of the workflow used to fit models and produce predictions of land cover is presented in Fig. 3.1. It was implemented in Python and R programming languages, and is publicly available via the eumap library (`https://eumap.readthedocs.io/`). The eumap library builds upon scikit learn [89, 197]; with `StackingClassifier` as the key function used to produce EML.

All the output predictions were predicted first per tile, then exported as Cloud Optimized Geotiffs (COGs) files and are publicly available through the Open Data Science Europe

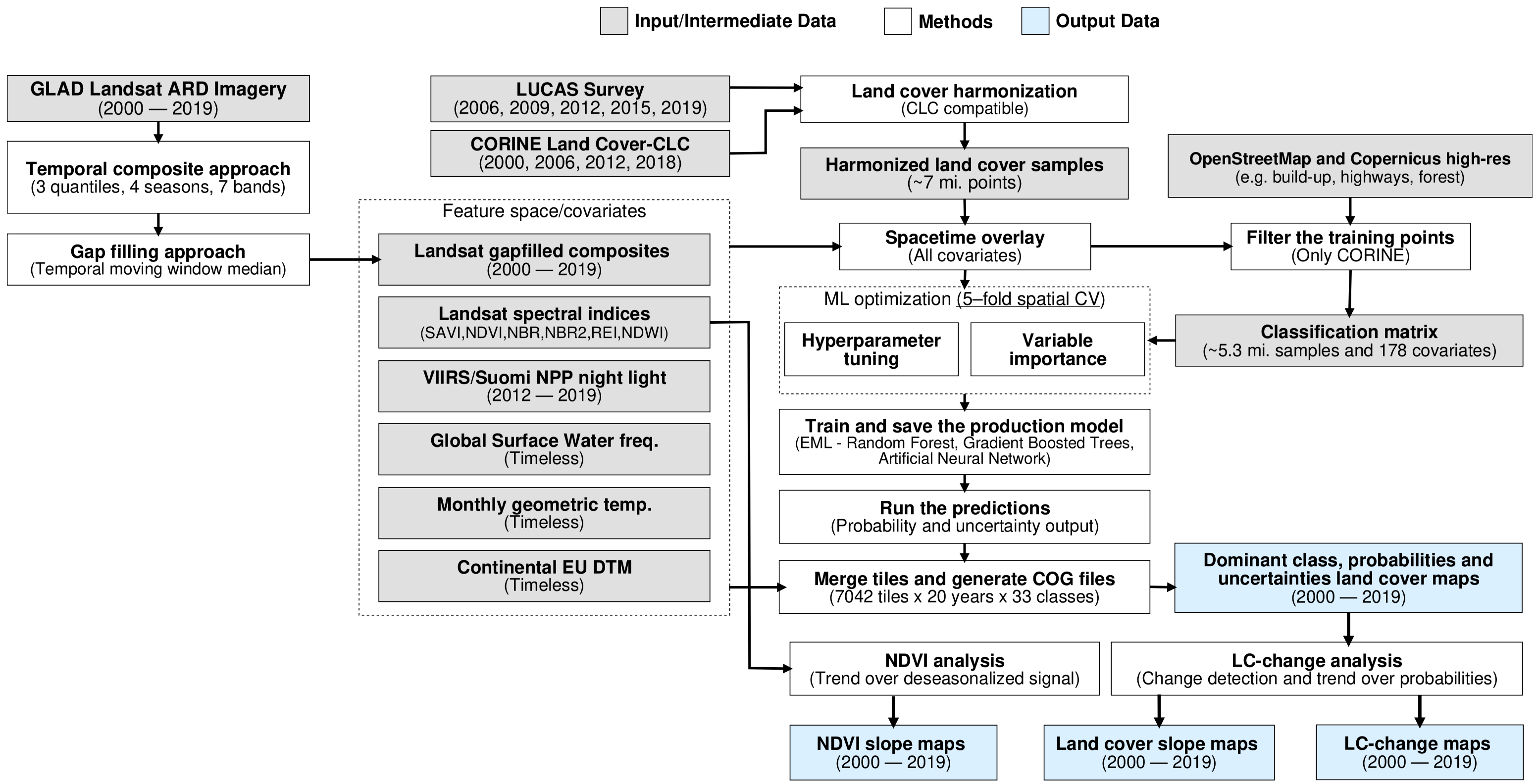

Input/Intermediate Data
Methods
Output Data
GLAD Landsat ARD Imagery (2000 — 2019)
Temporal composite approach (3 quantiles, 4 seasons, 7 bands)
Gap filling approach (Temporal moving window median)
LUCAS Survey (2006, 2009, 2012, 2015, 2019)
CORINE Land Cover-CLC (2000, 2006, 2012, 2018)
Feature space/covariates
Landsat gapfilled composites (2000 — 2019)
Landsat spectral indices (SAVI,NDVI,NBR,NBR2,REI,NDWI)
VIIRS/Suomi NPP night light (2012 — 2019)
Global Surface Water freq. (Timeless)
Monthly geometric temp. (Timeless)
Continental EU DTM (Timeless)
Land cover harmonization (CLC compatible)
Harmonized land cover samples (~7 mi. points)
Spacetime overlay (All covariates)
ML optimization (5–fold spatial CV)
Hyperparameter tuning
Variable importance
Train and save the production model (EML - Random Forest, Gradient Boosted Trees, Artificial Neural Network)
Run the predictions (Probability and uncertainty output)
Merge tiles and generate COG files (7042 tiles x 20 years x 33 classes)
OpenStreetMap and Copernicus high-res (e.g. build-up, highways, forest)
Filter the training points (Only CORINE)
Classification matrix (~5.3 mi. samples and 178 covariates)
Dominant class, probabilities and uncertainties land cover maps (2000 — 2019)
NDVI analysis (Trend over deseasonalized signal)
LC-change analysis (Change detection and trend over probabilities)
NDVI slope maps (2000 — 2019)
Land cover slope maps (2000 — 2019)
LC-change maps (2000 — 2019)

Figure 3.1: General workflow used to prepare point data and variable layers, fit models and generate annual land cover products (2000–2019). Components of the workflows are described in detail via the `eumap` library (`https://eumap.readthedocs.io/`), with technical documentation available via `https://gitlab.com/geoharmonizer_inea/`.

(ODS-Europe) Viewer, the S3 Cloud Object Service, and from `http://doi.org/10.5281/zenodo.4725429`. The classification matrix with all training points and variables is available from `http://doi.org/10.5281/zenodo.4740691`.

Spatiotemporal ensemble modeling

The annual land cover product for continental Europe was generated with an ensemble of three models and a meta-learner. We used a grid search strategy to find the best hyperparameters and used them to train the final model.

Although ensemble training and inference is computationally intensive, it typically achieves higher accuracy than less complex models [237, 314]. Furthermore, when each component learner predicts a probability per class, it is possible to use the standard deviation of the per-class probabilities (also known as *model variance*) as an indicator of the prediction uncertainty (see Fig. 3.2).

We selected three component learners among an initial pool of 10 learners based on their performance on sample data:

1. Random Forest [23];
2. Gradient-boosted trees [40];
3. Artificial Neural Network [167];

Each of these models predicts a probability for each class, resulting in 129 probabilities for 43 classes. These component probabilities are forwarded to the meta-learner, a logistic regression classifier [51], which in turn predicts a single probability per class. The ensemble also outputs the standard deviation of the three component-predicted probabilities per class to generate a class-wise model variance, which can help analyze the data and inform decision-makers where data is more reliable. Because the LUCAS points are based on *in-situ* observations, we considered them as more reliable training data than the CLC centroid points. To prioritize performance on the LUCAS points during model training, we assigned a training weight rating of 100% to the LUCAS points and 85% to the CLC points.

We optimized the hyperparameters of the random forest and gradient boosted trees component learners by minimizing the logistic (log) loss metric [154]:

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k} \tag{3.1}$$

where Y is a binary matrix of expected class labels, N is the total number of observations, K is the number of classes, P is the matrix of probabilities predicted by the model, $y_{i,k}$ indicates whether sample i belongs to class k, and $p_{i,k}$ indicates the probability of sample

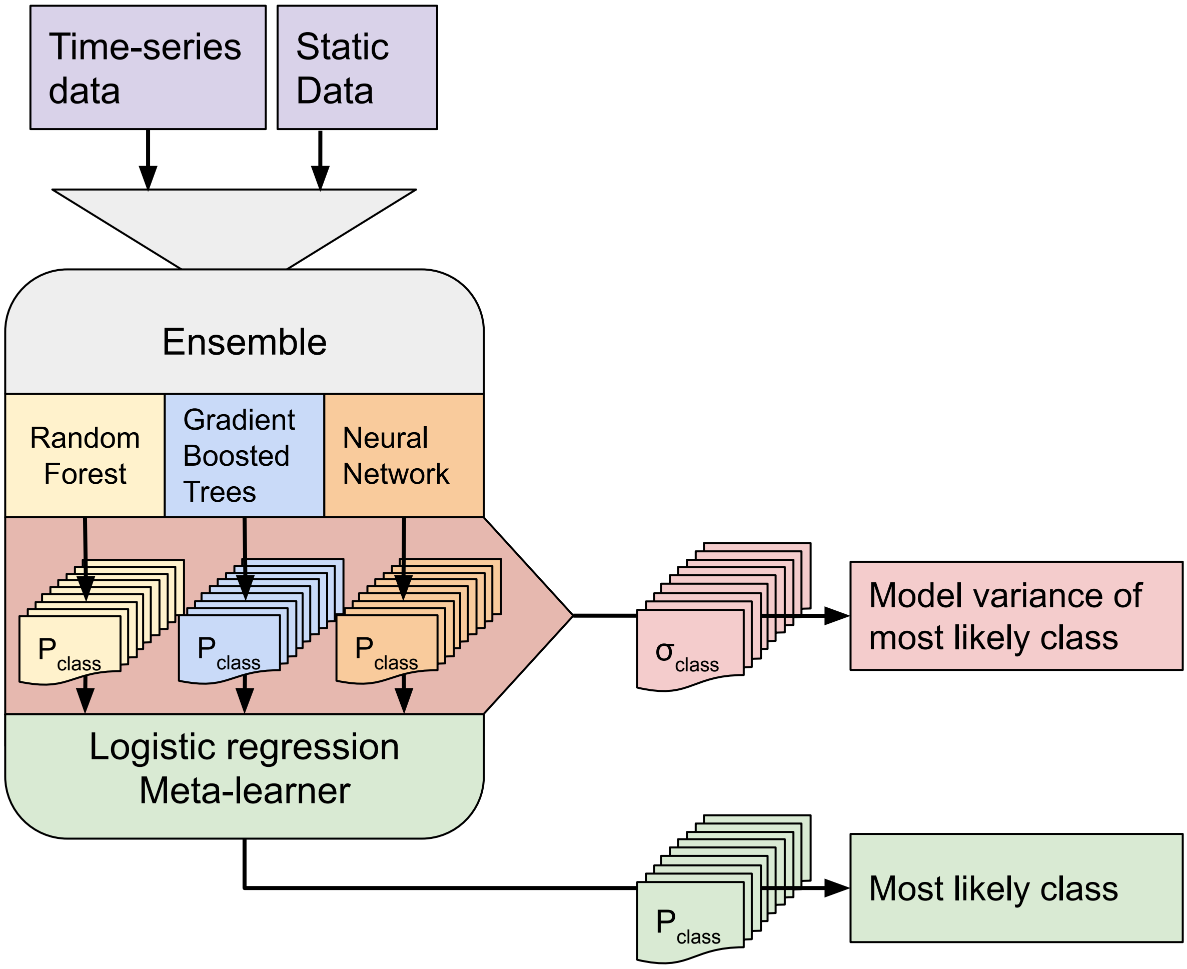


Figure 3.2: Structure of the ensemble. Time-series data and static data are used to train three component models. Each component model predicts 43 probabilities (1 per class). We calculate class-wise model variance as a proxy of prediction uncertainty as a separate output by taking the standard deviation of the three component probabilities per class. The 129 probabilities are used to train the logistic regression meta-learner, which predicts 43 probabilities that are used to map LULC.

i belonging to class j. A log loss value close to 0 indicate high prediction performance, 0 being a perfect match, while values above 0 indicate progressively worse performance.

We performed 5–fold spatial cross-validation for each different hyperparameter combination (see Table 3.2. These combinations were generated per model based on a grid search of 5 steps per hyperparameter. Fig.

We evaluated each set of hyperparameters by performing a spatial 5-fold cross-validation. We did this by creating a Europe-wide grid of 30 km tiles (see Fig. 3.3) and using the tiles' unique identifiers to group their overlapping points into 5 folds.

Table 3.2: Minimum and maximum value of each hyperparameter that was optimized for the random forest and gradient boosted tree learners.

Model	Hyperparameter	Lower value	Upper value
Random Forest	Number of estimators	50	100
	Maximum tree depth	5	50
	Maximum number of features	0	0.9
	Minimum samples per leaf	5	30
Gradient boosted trees	Eta	0.001	0.9
	Gamma	0	12
	Alpha	0	1
	Maximum tree depth	2	10
	Number of estimators	10	50

After hyperparameter optimization we trained the three component learners on the full dataset. The meta-learner was trained on the probabilities predicted by each component model during the cross-validation of their optimal hyperparameters.

Study area and target classification system

The study area covers all countries included in the CLC database, except Turkey (see Fig. 3.3). The spatiotemporal dataset used in this research contains data from the winter of 1999 to the autumn of 2019.

The target land cover nomenclature was designed based on CLC nomenclature [22] and is available in Table 3.3. CLC is probably the most comprehensive and detailed European land cover product to date. The CLC program was established in 1985 by the European Community to provide geographically harmonized information concerning the environment on the continent. The original CLC dataset is mapped in 44 classes with a minimum mapping unit of 25 ha for areal phenomena and 10 ha for changes. CLC mapping relies on harmonized protocol and guidelines that are shared for country-wise visual photo-interpretation.

The ODSE-LULC nomenclature is identical to the CLC legend, excluding class 523: Sea and ocean, as we omitted such areas from our study area to reduce computation time. The CLC classification system has been reported to be unsuitable for pixel-wise classification due to the inclusion of: 1) heterogeneous and mixed classes defined for polygon mapping (e.g. airports, road and rail networks, complex cultivation patterns, agro-forestry, etc.) and 2) classes primarily distinguishable by land use, not land cover (e.g. commercial and industrial units, sports and leisure facilities). We did not remove these classes beforehand to provide objective information about the performance of the CLC level 3 legend for

Table 3.3: The ODSE-LULC land cover legend used based on CLC [22]. Note: To make table formatting easier, we refer to class 243 as *‘Agriculture with significant natural vegetation‘* in all other tables.

Class name	Class description
111: Continuous urban fabric	Surface area covered for more than 80% by urban structures and other impermeable, artificial features.
112: Discontinuous urban fabric	Surface area covered between 30% and 80% by urban structures and other impermeable, artificial features.
121: Industrial or commercial units	Land units that are under industrial or commercial use or serve for public service facilities.
122: Road and rail networks	Motorways and railways, including associated installations.
123: Port areas	Infrastructure of port areas, including quays, dockyards and marinas.
124: Airports	Airports installations: runways, buildings and associated land.
131: Mineral extraction sites	Areas of open-pit extraction of construction materials (sandpits, quarries) or other minerals (open-cast mines).
132: Dump sites	Public, industrial or mine dump sites.
133: Construction sites	Spaces under construction development, soil or bedrock excavations, earthworks.
141: Urban green	Areas with vegetation within urban fabric.
142: Sport and leisure facilities	Areas used for sports, leisure and recreation purposes.
211: Non-irrigated arable land	Cultivated land parcels under rain-fed agricultural use for annually harvested non-permanent crops, normally under a crop rotation system.
212: Permanently irrigated arable land	Cultivated land parcels under agricultural use for arable crops that are permanently or periodically irrigated.
213: Rice fields	Cultivated land parcels prepared for rice production, consisting of periodically flooded flat surfaces with irrigation channels.
221: Vineyards	Areas planted with vines.
222: Fruit trees and berry plantations	Cultivated parcels planted with fruit trees and shrubs, including nuts, intended for fruit production.
223: Olive groves	Cultivated areas planted with olive trees, including mixed occurrence of vines on the same parcel.
231: Pastures	Meadows with dispersed trees and shrubs occupying up to 50% of surface characterized by rich floristic composition.
241: Annual crops associated with permanent crops	Cultivated land parcels with a mixed coverage of non-permanent (e.g. wheat) and permanent crops (e.g. olive trees).
242: Complex cultivation patterns	Mosaic of small cultivated land parcels with different cultivation types (annual and permanent crops, as well as pastures), potentially with scattered houses or gardens.
243: Land principally occupied by agriculture with significant areas of natural vegetation	Areas principally occupied with agriculture, interspersed with significant semi-natural areas in a mosaic pattern.
244: Agro-forestry areas	Annual crops or grazing land under the wooded cover of forestry species.
311: Broad-leaved forest	Vegetation formation composed principally of trees, including shrub and bush understorey, where broad-leaved species predominate.
312: Coniferous forest	Vegetation formation composed principally of trees, including shrub and bush understorey, where coniferous species predominate.
313: Mixed forest	Vegetation formation composed principally of trees, including shrub and bush understory, where neither broad-leaved nor coniferous species predominate.
321: Natural grasslands	Grasslands under no or moderate human influence. Low productivity grasslands. Often in areas of rough, uneven ground, also with rocky areas, or patches of other (semi-)natural vegetation.
322: Moors and heathland	Vegetation with low and closed cover, dominated by bushes, shrubs (heather, briars, broom, gorse, laburnum etc.) and herbaceous plants, forming a climax stage of development.
323: Sclerophyllous vegetation	Bushy sclerophyllous vegetation in a climax stage of development, including maquis, matorral and garrigue.
324: Transitional woodland-shrub	Transitional bushy and herbaceous vegetation with occasional scattered trees. Can represent either woodland degradation or forest regeneration / re-colonization.
331: Beaches, dunes, sands	Natural un-vegetated expanses of sand or pebble/gravel, in coastal or continental locations, like beaches, dunes, gravel pads.
332: Bare rocks	Scree, cliffs, rock outcrops, including areas of active erosion.
333: Sparsely vegetated areas	Areas with sparse vegetation, covering 10-50% of the surface.
334: Burnt areas	Areas affected by recent fires.
335: Glaciers and perpetual snow	Land covered by ice or permanent snowfields.
411 Inland marshes	Low-lying land usually flooded in winter, and with ground more or less saturated by fresh water all year round.
412 Peat bogs	Wetlands with accumulation of considerable amount of decomposed moss (mostly Sphagnum) and vegetation matter. Both natural and exploited peat bogs.
421 Salt marshes	Vegetated low-lying areas in the coastal zone, above the high-tide line, susceptible to flooding by seawater.
422 Salines	Sections of salt marsh exploited for the production of salt by evaporation, active or in process of abandonment, distinguishable from marsh by parcellation or embankment systems.
423 Intertidal flats	Area between the average lowest and highest sea water level at low tide and high tide. Generally non-vegetated expanses of mud, sand or rock lying between high and low water marks.
511: Water courses	Natural or artificial water courses for water drainage channels.
512: Water bodies	Natural or artificial water surfaces covered by standing water most of the year.
521: Coastal lagoons	Stretches of salt or brackish water in coastal areas which are separated from the sea by a tongue of land or other similar topography.
522: Estuaries	The mouth of a river under tidal influence within which the tide ebbs and flows.

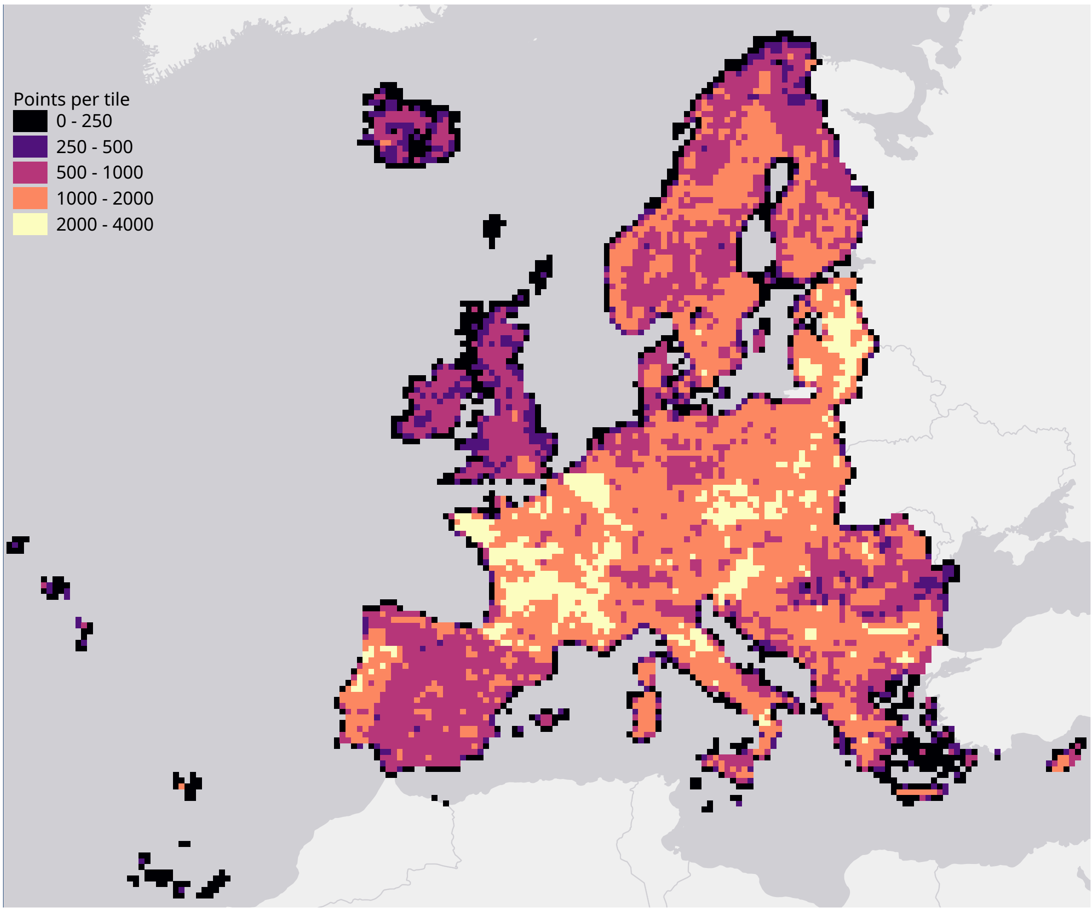


Figure 3.3: Map of the study area, overlaid with a grid of 30 km tiles that was used for spatial 5-fold cross-validation. Grid color indicates the number of training points aggregated per tile.

pixel-wise classification, and to enable a complete comparison to the S2GLC nomenclature, which is more optimized for such pixel-based classification.

Training points

We obtained the training dataset from the geographic location of LUCAS (*in-situ* source) and the centroid of all CLC polygons (as shown in Fig. 3.4), harmonized according to the 43 land cover classes (see Table 3.3) and organized by year, where each unique combination of longitude, latitude and year was considered as an independent sample, resulting in more than 8 million training points.

The LUCAS data from 2006, 2009, 2012, 2015 and 2018, as provided by Eurostat (obtained from: `https://ec.europa.eu/eurostat/web/lucas`) is the largest and most comprehensive *in-situ* land cover dataset for Europe. The survey has evolved since 2000 and requires harmonisation before it can be used for mapping over several years. We imported datasets from individual years and harmonized these before merging it into one common database with an automated workflow implemented in Python and SQL (Fig. 3.1). For the multi-year harmonization procedure we first harmonized attribute names, re-coded variables, harmonized point locations, and aggregated the points based on their location in space and time. After these operations, we translated the LUCAS land cover nomenclature to the ODSE-LULC nomenclature, Table 3.3, according to the method designed by Buck et al. [28]. The distribution of all reference points per CLC class and per survey year is shown in Fig. 3.5.

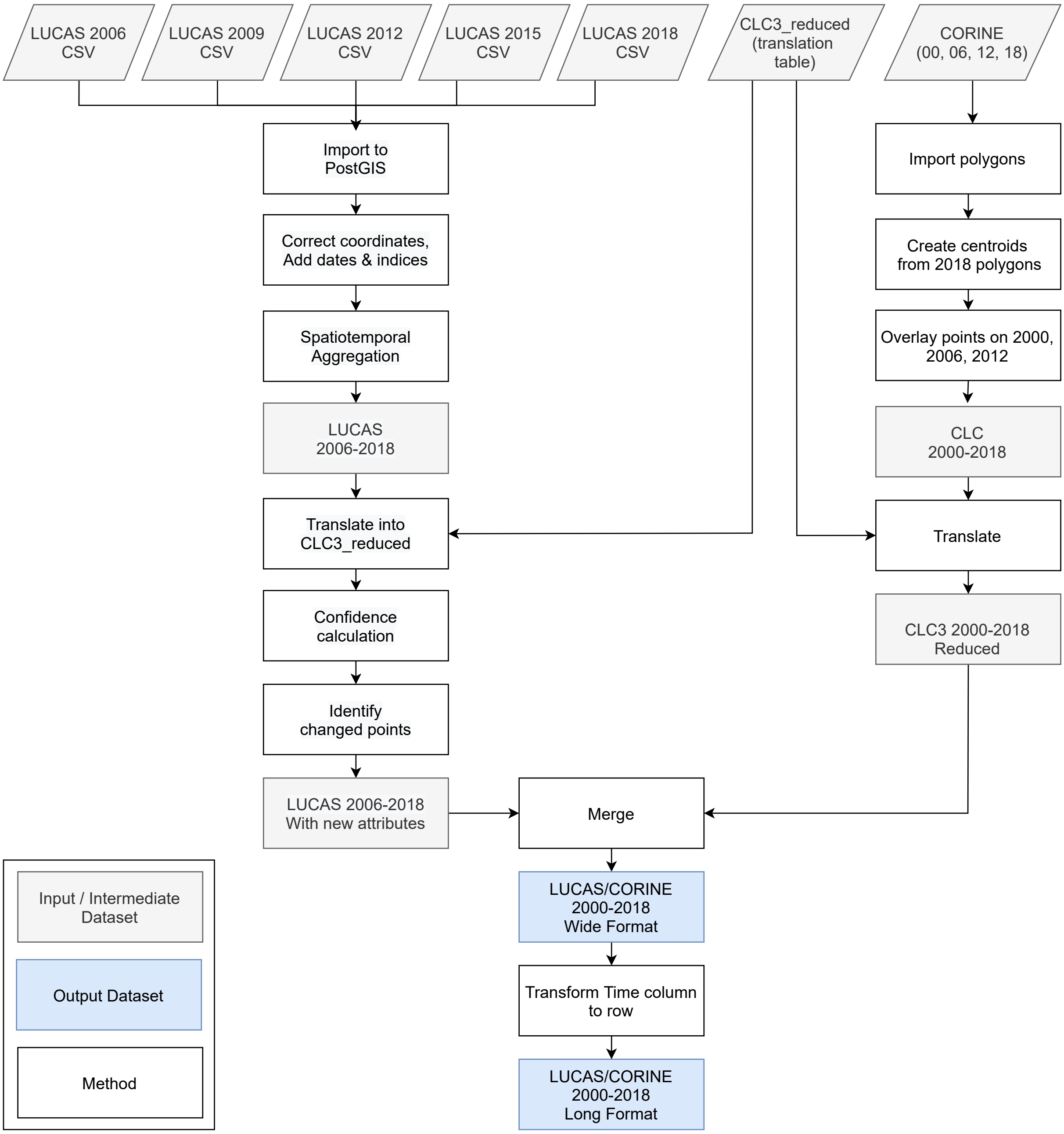

LUCAS 2006 CSV
LUCAS 2009 CSV
LUCAS 2012 CSV
LUCAS 2015 CSV
LUCAS 2018 CSV
CLC3_reduced (translation table)
CORINE (00, 06, 12, 18)
Import to PostGIS
Correct coordinates, Add dates & indices
Spatiotemporal Aggregation
LUCAS 2006-2018
Translate into CLC3_reduced
Confidence calculation
Identify changed points
LUCAS 2006-2018 With new attributes
Import polygons
Create centroids from 2018 polygons
Overlay points on 2000, 2006, 2012
CLC 2000-2018
Translate
CLC3 2000-2018 Reduced
Merge
LUCAS/CORINE 2000-2018 Wide Format
Transform Time column to row
LUCAS/CORINE 2000-2018 Long Format
Input / Intermediate Dataset
Output Dataset
Method

Figure 3.4: General workflow for merging training points obtained from LUCAS and CLC.

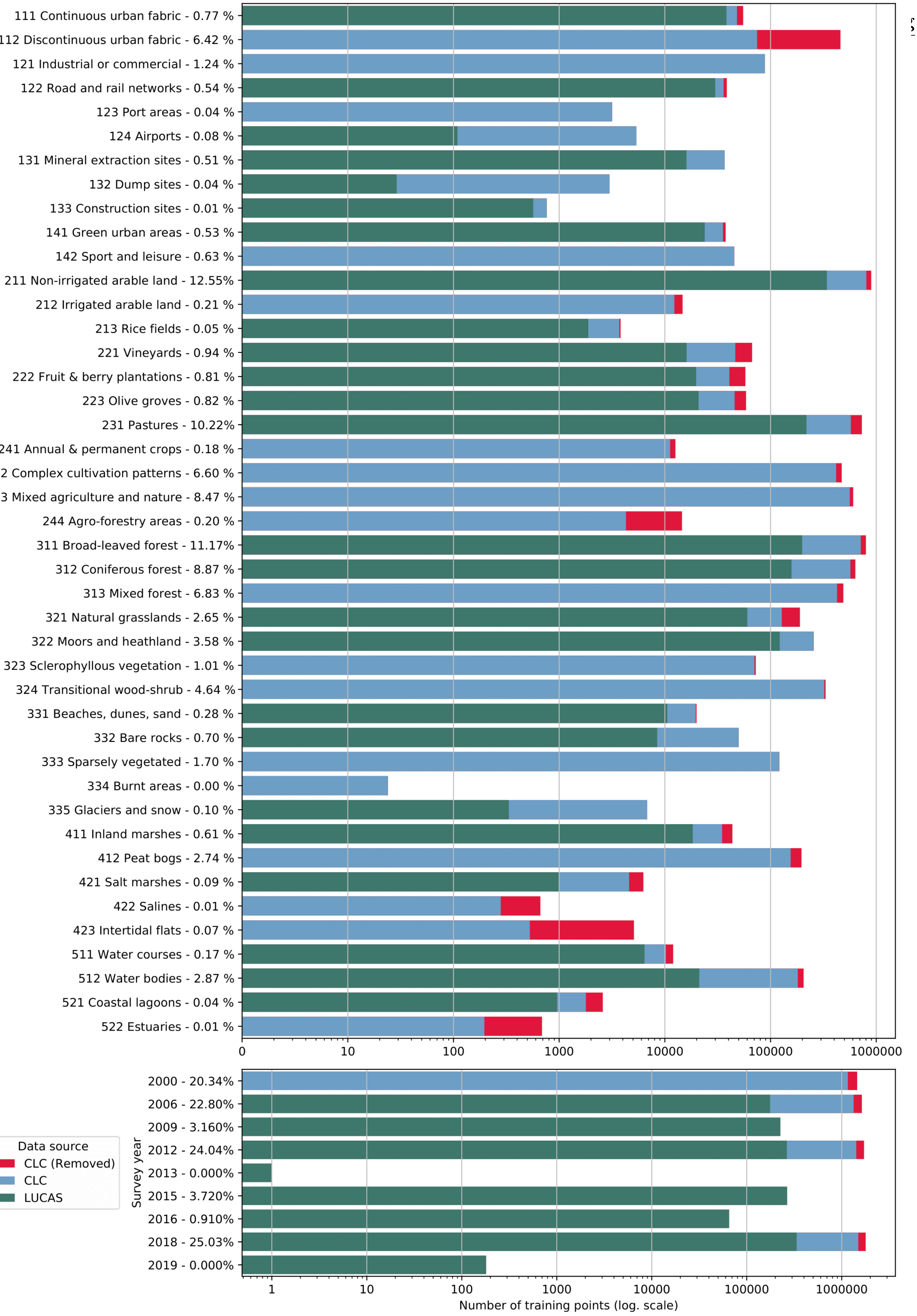


Figure 3.5: Distribution of training points per data source (blue and green), class (top) and

The CLC minimal mapping unit of 25 ha required filtering on the training points before they could be used to represent 30 m resolution LULC, for example, to remove points for *"111: urban fabric"* located in small patches of urban greenery (<25 ha). For this purpose, we extracted vector data from OSM layers for roads, railways, and buildings (obtained from `https://download.geofabrik.de/`). We then created a 30 m density raster for each feature type. This was done by first creating a 10 m raster where each pixel intersecting a vector feature was assigned the value 100. These pixels were then aggregated to 10 m resolution by calculating the average of every 9 adjacent pixels. This resulted in a 0—100 density layer for the three feature types. Although the digitized building data from OSM offers the highest level of detail, its coverage across Europe is inconsistent. To supplement the building density raster in regions where crowd-sourced OSM building data was unavailable, we combined it with Copernicus High Resolution Layers (HRL) (obtained from `https://land.copernicus.eu/pan-european/high-resolution-layers`), filling the non-mapped areas in OSM with the Impervious Built-up 2018 pixel values, which was averaged to 30 m. The probability values produced by the averaged aggregation were integrated in such a way that values between 0—100 refer to OSM (lowest and highest probabilities equal to 0 and 100 respectively), and the values between 101—200 refer to Copernicus HRL (lowest and highest probability equal to 200 and 101 respectively). This resulted in a raster layer where values closer to 100 are more likely to be buildings than values closer to 0 and 200. Structuring the data in this way allows us to select the higher probability building pixels in both products by the single boolean expression: pixel $>$ 50 AND pixel <150.

We also use HRL products to filter other classes: Table 3.4 shows the exact conditions points of specific LULC classes needed to meet in order to be retained in our dataset. This procedure is similar to the one used by Inglada et al. [125]. This filtering process removed about 1.3 million points from our training dataset, resulting in a classification matrix with a total of ca. 8.1 million samples and 232 variables. The classification matrix used to produce ODSE-LULC is available from `http://doi.org/10.5281/zenodo.4740691`.

We assessed the quality of the training dataset by comparing it to a number of existing land cover products:

- GLFCS30–2015 [**zhang2020glc˙fcs30**];
- GLFCS30–2020 [**zhang2020glc˙fcs30**];
- S2GLC [161];
- The European land cover product for 2015 created by Pflugmacher et al. [199];
- ELC10 [287].

For each comparison, we reclassified the training dataset to the nomenclature of the target dataset and overlaid all points from our dataset with survey dates from within one year

Table 3.4: Per-class conditions applied only to CLC points during the filtering step. All the raster layers were upsampled to 30×30 m resolution by average and the points that did not meet the specified condition were omitted from the training dataset.

		Condition	HRL						OSM		HRL+OSM
Code	Class		Tree Cover	Grass	Imp.	Perm. Water	Perm. Wetness	Temp. Wetness	Rails	Roads	Buildings
111	Continuous urban fabric	-									>50 and <150
112	Discontinuous urban fabric										>50 and <150
121	Industrial or commercial units										
122	Road and rail networks and associated land	OR			>30				>30	>30	
123	Port areas										
124	Airports										
131	Mineral extraction sites	AND	= 0	= 0							
132	Dump sites										
133	Construction sites										
141	Green urban areas	(OR) AND	>0	>0							<50 or >150
142	Sport and leisure facilities										
211	Non-irrigated arable land	AND	= 0						= 0	= 0	<50 or >150
212	Permanently irrigated arable land		= 0						= 0	= 0	<50 or >150
213	Rice fields								= 0	= 0	<50 or >150
221	Vineyards	AND		= 0					= 0	= 0	<50 or >150
222	Fruit trees and berry plantations	AND		= 0					= 0	= 0	<50 or >150
223	Olive groves	AND		= 0					= 0	= 0	<50 or >150
231	Pastures	AND	= 0						= 0	= 0	<50 or >150
241	Annual crops associated with permanent crops								= 0	= 0	<50 or >150
242	Complex cultivation patter								= 0	= 0	<50 or >150
243	Agriculture with significant natural vegetation								= 0	= 0	<50 or >150
244	Agro-forestry areas		>0						= 0	= 0	<50 or >150
311	Broad-leaved forest	AND	>0						= 0	= 0	<50 or >150
312	Coniferous forest	AND	>0						= 0	= 0	<50 or >150
313	Mixed forest		>0						= 0	= 0	<50 or >150
321	Natural grasslands	AND	= 0	>0					= 0	= 0	<50 or >150
322	Moors and heathland								= 0	= 0	<50 or >150
323	Sclerophyllous vegetation								= 0	= 0	<50 or >150
324	Transitional woodland-shrub								= 0	= 0	<50 or >150
331	Beaches, dunes, sand								= 0	= 0	<50 or >150
332	Bare rocks								= 0	= 0	<50 or >150
333	Sparsely vegetated areas								= 0	= 0	<50 or >150
334	Burnt areas								= 0	= 0	<50 or >150
335	Glaciers and perpetual snow								= 0	= 0	<50 or >150
411	Inland marshes	OR					>0	>0	= 0	= 0	<50 or >150
412	Peat bogs								= 0	= 0	<50 or >150
421	Salt marshes								= 0	= 0	<50 or >150
422	Salines								= 0	= 0	<50 or >150
423	Intertidal flats								= 0	= 0	<50 or >150
511	Water courses					>50					
512	Water bodies	-				= 100					
521	Coastal lagoons					>50					
522	Estuaries					>50					

of the land cover product. We then calculated the weighted F1-score as if the points represented predictions. Points with classes of the target products that were completely absent in the training point subsets (due to the target nomenclature of the training points) were removed before these assessments, potentially resulting in varying numbers of classes for the same dataset.

The GLFCS30 nomenclature was not suitable for direct translation because some land cover types (such as forests) are separated into several subcategories. We therefore aggregated their thematic resolution to the higher level of abstraction described in **zhang2020glc˙fcs30**. The complete translation scheme is available via the GitLab repository of the GeoHarmonizer project (https://gitlab.com/geoharmonizer_inea/spatial-layers).

Input variables

In this work we combine harmonized time-series data of varying temporal resolution with static datasets. The time-series data consists of the following:

- Seasonal aggregates of Landsat spectral bands (blue, green, red, NIR, SWIR1, SWIR2, thermal), divided into 3 reflectance quantiles per and 4 seasons, resulting in 12 layers per band;
- Spectral indices calculated from the seasonal Landsat data: NDVI, SAVI, MSAVI, NDMI, Landsat NBR, REI, and NDWI derived according to formulas in Table 3.5;
- Terrain Roughness Index (TRI) of the Landsat green band (50th reflectance quantile of summer);
- SUOMI NPP VIIRS night light imagery downscaled from 500 m to 30 m resolution [109];
- Monthly geometric minimum and maximum temperature [137];

Additional static datasets are:

- Probability of surface water occurrence at 30 m resolution [198];
- Continental EU DTM-based elevation and slope in percent [103];

All variables used by our model are derived from remotely sensed EO data from multiple sources, the largest share being derived from Landsat imagery. Although EO data with higher spatial and temporal resolution, as well as actual surface reflection values are available (e.g. Sentinel-2), such sources do not cover the timespan required for the long-term analysis proposed by this framework. The Landsat data used in this work was obtained by downloading the Landsat ARD, provided by GLAD [206], for the years 1999 to 2019 and for the entire extent of continental Europe (see eumap landmask [103]). This imagery archive was screened to remove the cloud and cloud shadow pixels, maintaining only the quality assessment-QA values labeled as clear-sky according to GLAD. Second, we averaged the individual images by season according to three different quantiles (25th, 50th and 75th) and the following calendar dates for all periods:

- Winter: December 2 of previous year until March 20 of current year;
- Spring: March 21 until June 24 of current year;
- Summer: June 25 until September 12 of current year;
- Fall: September 13 until December 1 of current year.

We decided to use the equal length definition provided by Trenberth [266] representing four seasons and matching the beginning and end of each season with the 16-day intervals used by Potapov et al. [206]. From more than 73 TB of input data we produced 84 images (3 quantiles × 4 seasons × 7 Landsat bands) for each year with different occurrences of no-data values due to cloud contamination in all observations of a specific season.

We next impute all missing values in the Landsat temporal composites using the *"Temporal Moving Window Median"* (TMWM) algorithm, implemented in python and publicly available in the `eumap` library (see Fig. 3.1). The algorithm uses the median values derived from temporal neighbours to impute a missing value using pixels from 1) the same season, 2) neighboring seasons and 3 the full year. For example, for a missing value in the spring season, the algorithm first tries to use values from spring seasons of neighbouring years. If no pixel value is available for the entire period (i.e. 2000–2019), the algorithm tries to use values from winter and summer of neighbouring years. If no pixel value is available from data of adjacent seasons from the same year, pixel values from adjacent years are used to derive the median values. Ultimately, a missing value will not receive an impute value only if the pixel lacks data throughout the entire time-series. The median calculation considers different sizes of temporal windows, which expands progressively for each impute attempt (i.e. `time_win_size` parameter); in this work we used a maximum `time_win_size` of 7. We selected the TMWM approach from a set of 4 algorithms through a benchmarking process. To our knowledge, it provides the best combination of gap-filling accuracy and computational costs on the scale of this project.

We include several spectral indices as a form of feature engineering because they are each designed and tested to help identify or distinguish different types of land cover. Table 3.5 provides an overview of how we derived them from the Landsat data. This was done for each quantile and each season, resulting in 4 × 3 = 12 variables per spectral index.

Table 3.5: Spectral indices derived from the Landsat data and used as additional variables in the spatiotemporal EML.

Spectral Index	Equation	Reference
NDVI	$\frac{nir - red}{nir + red}$	[275]
SAVI	$\frac{nir - red}{(nir + red + 0.5) \times 1.5)}$	[120]
MSAVI	$\frac{(2 \times nir + 1) - \sqrt{(2 \times nir + 1)^2 - 8 \times (nir - red)}}{2}$	[212]
NDWI	$\frac{green - swir2}{green + swir2}$	[82]
NBR	$\frac{nir - thermal}{nir + thermal}$	[135]
NDMI	$\frac{nir - swir1}{nir + swir1}$	[130]
NBR2	$\frac{swir1 - thermal}{swir1 + thermal}$	[136]
REI	$\frac{nir - blue}{nir + blue} \times nir$	[238]

The TRI [218] gives an indication of how different pixel values are from those of its neighbors. Is usually calculated from elevation data, but we include it as a derivative of the Landsat green band in order to help the model distinguish between pixels that are part of larger, homogeneous regions from pixels that are located inside more heterogeneous landscapes (e.g. airports, urban green areas, and forest edges).

The Suomi-NPP VIIRS night light imagery [109] was included to introduce a variable that may help the model recognize the built-up environment, but also distinguish different types of land use within that category. This data is originally in 500 m resolution, but we re-sampled them to 30 m using a cubic spline.

The geometric minimum and maximum temperature is a geometric transformation of latitude and the day of the year [137]. We include these variables to improve performance on LULC classes that occur in different situations under distant latitudes e.g. coniferous forest in Greece and Norway. It can be defined anywhere on the globe using Eq.(3.2):

$$t_{min} = 24.2 \cdot \cos\phi - 15.7 \cdot (1 - \cos\theta) \cdot \sin|\phi| - 0.6 \cdot \frac{z}{100} \tag{3.2}$$

$$t_{max} = 37 \cdot \cos\phi - 15.4 \cdot (1 - \cos\theta) \cdot \sin|\phi| - 0.6 \cdot \frac{z}{100} \tag{3.3}$$

where θ is derived as:

$$\theta = (day - 18) \cdot \frac{2\pi}{365} + 2^{1-\mathrm{sgn}(\phi)} \cdot \pi. \tag{3.4}$$

where *day* is the day of year, ϕ is the latitude, the number 18 represents the coldest day in the northern and warmest day in the southern hemisphere, z is the elevation in meter, 0.6 is the vertical temperature gradient per 100 m, and sgn denotes the signum function that extracts the sign of a real number.

We include a long-term (35-year) probability estimate of surface water occurrence [198] based on the expectation that it would improve model performance when classifying LULC classes associated with water, such as wetlands and rice fields.

Accuracy assessment

We evaluate the suitability of the proposed framework with three assessments:

1. Comparison of spatial and spatiotemporal models;
2. 5-fold spatial cross-validation;
3. Validation on S2GLC point data.

we compare the performance of spatial and spatiotemporal models to assess whether training models on data from multiple years can improve their ability to generalize to data from unknown years. We expect models trained on observations from multiple years to generalize better on data from unknown years than models trained on observations from a single year. In order to investigate this, we trained multiple ensemble models on several subsets of our training data that were selected from either one or several years, and validated them on data from years included in their training data and on observations from 2018, the last year of the training dataset, upon which no model was trained.

The validation on the S2GLC point data is included to assess the extent to which the choice of legend affects the classification accuracy of our framework. The S2GLC legend contains less classes and does not

The results produced by the 5-fold spatial cross-validation are used to assess four characteristics of the proposed methodology:

1. The difference in performance between the ensemble model and its component models;
2. classification accuracy of the framework when reproducing the 43-class CLC classification system;
3. consistency of prediction accuracy by the framework through time;
4. consistency of prediction accuracy by the framework through space;

In all comparisons and experiments, we discriminate model performance with the Weighted F1-score metric [281]:

$$\mathrm{WF}_1 = \sum_{c=1}^{n} S_c \cdot \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{3.5}$$

where n is the number of classes, and S_c is the support (the number of training points), P_c the precision (producer's accuracy), and R_c the recall (user's accuracy) of a given class c. We used a weighted version of this metric because it distinguishes classification performance more strictly on imbalanced datasets, such as the one used in this work.

Spatial Cross-validation

Before mapping LULC in continental Europe for all years, we performed spatial 5-fold cross-validation using the hyperparameters of the final EML model to assess its performance. The predictions for the points from each left-out fold were merged into one set of predicted values, which we used to assess the performance of our final model. We did this for each of the three levels in the CLC nomenclature (with 43, 15, and 5 classes) to investigate the effect of legend size. We aggregated predictions to the higher level in the hierarchy by taking the highest probability among subclasses within the same higher level class before selecting the most probable class. Besides this general performance on the total dataset, we also analyzed the performance of the ensemble per class, year, and cross-validation tile.

Analyzing the performance per class and per level in the hierarchy allows us to quantify the performance increase gained from aggregating specific classes. We do this by calculating the weighted average of the F1-score of all sub-classes of a higher-level class (e.g. 311: Broad-leaved forest, 312: Coniferous forest, and 313: Mixed forest, which together comprise the level 2 class 31: Forests and seminatural areas). Finally, we subtract the weighted average F1-score of the subclasses from the F1-score of the higher-level class to quantify the performance gain. This value will tend to be higher when the model frequently confuses sub-classes of a higher-level class, as aggregation then removes more classification errors.

We analyzed the temporal and spatial consistency of our model performance by calculating the weighted F1-scores for the cross-validation predictions on points from each separate year and tile, respectively. We calculated the standard deviation of these scores to assess the consistency of the model.

Finally, we also compare the cross validation log loss score per class, as well as aggregated per CLC level, with a baseline log loss score. This baseline log loss is what a random classifier would score when predicting on a given dataset. A dataset with more classes and a more unequal distribution has a higher baseline log loss score. We also calculate a log loss ratio to give a measure of model performance that is agnostic of the number and distribution of classes, instead only reflecting how well a given model performed given the difficulty of its task. We define this ratio as follows:

$$R(Y, P) = 1 - \frac{L_{\log}(Y, P)}{B_{\log}(Y, P)} \tag{3.6}$$

where L_{log} indicates the log loss score of the prediction and B_{log} indicates the baseline log loss score that would be scored by a randomly predicting model. A ratio of 0 means that the model did not outperform a random predictor, a ratio of 1 means a perfect prediction with a log loss score of 0.

Validation on S2GLC points

After training an ensemble model with the same hyperparameters on all training data, we classified LULC in 2017. This prediction was validated with the S2GLC dataset which Malinowski et al. [161] used to validate their 2017 land cover product. The dataset contains 51,926 points with human-verified land cover classifications which were collected with a stratified random sampling method from 55 proportionally selected regions of Europe.

As the S2GLC points follow a different nomenclature, we translated the ODSE-LULC predicted classes according to Table 3.6. Because any predicted classes outside the S2GLC nomenclature (labeled as 000: None in Table 3.6) would be automatically counted as errors, we performed two validations: (1) a conservative assessment that included points with such predictions, and (2) an optimistic assessment where they were omitted.

Comparison of ensemble and component models

Previous studies have shown that ensemble models can outperform their component models [237, 314]. To investigate if this was the case for our approach, we compared the spatial cross-validation accuracy of the three selected component models with that of the full ensemble. We also compared variable importance of the gradient boosted trees and random forest models in order to discover to what extent the different models used different parts of the available feature space.

Comparison of spatial and spatiotemporal models

We decided to use a spatiotemporal model trained on reference data from multiple years because we expect it to generalize better to data from years that were not included in its training data. We expect this because the EO covariates are more diverse in multi-year datasets, which leads to a larger feature space and likely reduces overfitting.

We also expected better performance from spatiotemporal models because combining data from multiple years allows for larger training datasets, which generally improves the predictive power of a model.

To investigate these two benefits, we trained three types of models:

- Spatial models, trained on 100,000 points from a single year;
- Small spatiotemporal models, trained on 100,000 points sampled from our multi-year dataset;
- Large spatiotemporal models, trained on 100,000 points from each year of our multi-year dataset.

We trained a small and a large spatiotemporal model to gain separate insight into the effects of dataset size and dataset diversity. The years 2000, 2006, 2009 and 2012 had sufficient points for this experiment, resulting in 4 spatial models, 1 small spatiotemporal model, and 1 large spatiotemporal model. We then evaluated each model's classification performance on a dataset sampled from the same years as the model's training data, and a dataset sampled from 2018, which was excluded from the training data selection. Every model's validation dataset was $\frac{1}{3}^{\text{rd}}$ the size of its training dataset. The validation on data from 2018 represents each model's ability to generalize to data from years that it was not trained to classify. We averaged the performance of all spatial models to obtain the performance of one *'spatial model'*.

To investigate the effect of combining the CLC and LUCAS points, we performed this experiment three times by training and validating on only CLC points, only LUCAS points, and a combination of CLC and LUCAS points.

Time-series analysis

After classifying LULC in Europe between 2000–2019, we analyzed the dynamics of land cover predicted by our model in three ways:

- Probability and NDVI trend analysis using logistic regression on NDVI and the probabilities for key classes;
- Change class per year and between 2001–2018;
- Prevalent change mapping;

These LULC change dynamics were not validated and serve as a means of analyzing the output of the presented framework. Furthermore, the GLAD ARD data-set by Potapov et al. [206] is produced for analyzing land cover change but should not be used for land surface reflectance applications directly. Therefore we do not use NDVI trends as an indication of absolute vegetation vigor but only as a relative measure of change. Also, NDVI trends are only applied as a tool to understand the changes and to enhance interpretation.

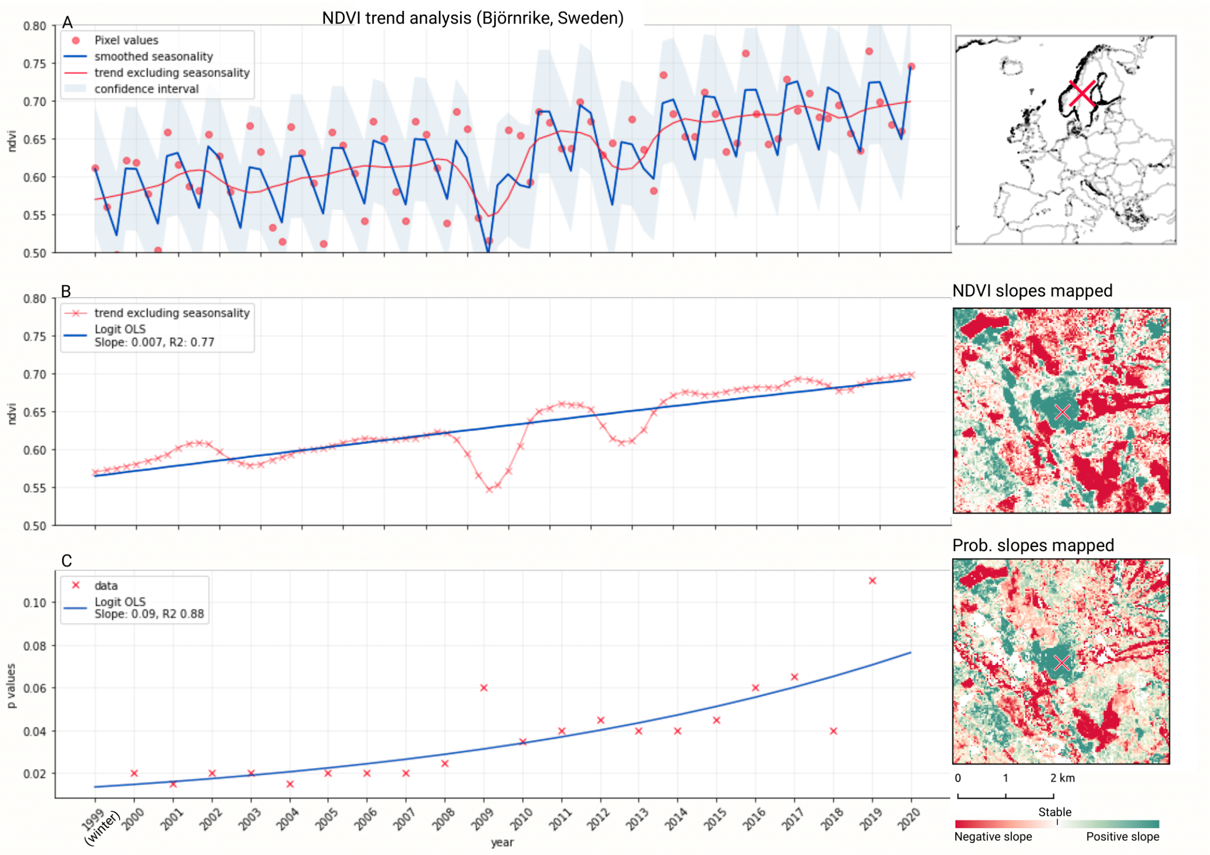


Figure 3.6: Example of deseasonalization [233] and subsequent Logit OLS applied on a single pixel in Sweden (Coordinates: 62°24'43.7"N 13°56'00.3"E): (a) red dots represent pixel values, the blue line represents a local weighted regression smoothed line based on the pixel values plus a light blue area indicating the confidence interval, the red line represents the trend after removing the seasonal signal; (b) red line and crosses represent the trend after removing the seasonal signal, the blue line visualizes the regression model based NDVI values in the logit space; (c) Trend analysis on probability values for non-irrigated arable land. In the case above the gradient value is 0.09 with the model R-square = 0.88

We analyzed the trend over the years between 2000 and 2019 by fitting an Ordinary Least Squares (OLS) regression model on the time-series of probabilities of every pixel. We use the coefficient as a proxy for the gradual change through time. Because probabilities only have meaningful values between 0 and 1 and NDVI are only meaningful for values between -1 and 1, we applied a logit transformation to the input data of the OLS analysis. We applied this trend analysis on the four most prevalent LULC classes: (1) coniferous forest, (2) non-irrigated arable land, (3) broad leaved forest, and (4) pastures. We also applied this method on a deseasonalized [233] NDVI time-series (see Fig. 3.6 and present this trend analysis as an additional tool to qualitatively appraise large-scale, long-term trends.

In order to visualise change implied by our LULC predictions, we first implement a smoothing post-processing strategy before categorizing change processes. The smoothing strategy considers the classification of a pixel in the previous and next years. If a pixel is classified as one class, but as another single class in the year before and after, this classification is considered an error. In such a case, the pixel's class is changed to match the previous and subsequent class. We call this a *"T-3 temporal filter"*.

After this preprocessing step, we categorize LULC change processes by applying the change classes seen in the Copernicus land cover map [27] to our classification scheme. We translated the CLC classes to the land cover classes used by the Copernicus land cover map according to Table 3.7. Some examples of changes include: changing from Dump sites into Urban fabric is classified as *"No change"*, changing from Non-irrigated arable land into Urban fabric to *"Urbanization"*, changing from Airports to Mineral extraction sites to *"Other"* etc. Two notable exceptions are the *"forest loss"* and *"Reforestation"* classes. In this paper we will refer to *"Forest loss"* and *"Forest increase"* instead. We renamed these change classes because we wanted to avoid making assumptions regarding the drivers of the detected trends in forest cover.

In order to identify and visualize the dominant LULC change trends in Europe, we mapped the *"prevalent change"* at two scales of aggregation: 5×5 km and 20×20 km. We created a Europe-covering grid with cells at both scales. Then, we counted the number of 30×30 m pixels of each change class within each grid cell. The predominant change class (see Table 3.7) was then assigned to each grid cell. We also calculated *"change intensity"* by dividing the number of 30×30 m pixels of the prevalent change class, by the sum of all pixels in each grid cell. For example, at a 20×20 km scale, each grid cell contains have $(20,000/30) \cdot (20,000/30) =\ 444,444$ pixels. If the prevalent change class is present in >94,000 pixels this means that it covers >20% of the total area.

Table 3.6: Reclassification key used to validate the predictions of our ensemble model on the S2GLC point dataset collected by Malinowski et al. [161].

S2GLC	ODSE-LULC
111: Artificial Surfaces	111: Continuous urban fabric 112: Discontinuous urban fabric 121: Industrial or commercial units 122: Road and rail networks and associated land 123: Port areas 124: Airports 132: Dump sites 133: Construction sites
311: Broadleaf tree Cover	311: Broad-leaved forest
312: Coniferous Tree Cover	312: Coniferous forest
211: Cultivated Areas	211: Non-irrigated arable land 212: Permanently irrigated arable land 213: Rice fields 241: Annual crops associated with permanent crops 242: Complex cultivation patterns 243: Agriculture with significant natural vegetation 244: Agro-forestry areas
231: Herbaceous Vegetation	231: Pastures 321: Natural grasslands
411: Marshes	411: Inland Marshes 421: Salt Marshes 422: Salines 423: Intertidal Flats
322: Moors and Heathland	322: Moors and heathland
331: Natural Material Surfaces	131: Mineral extraction sites 331: Beaches, dunes, sands 332: Bare rocks
000: None	141: Green urban areas 142: Sport and leisure facilities 222: Fruit trees and berry plantations 223: Olive groves 313: Mixed Forest 324: Transitional woodland-shrub 333: Sparsely vegetated areas 334: Burnt areas
412: Peat Bogs	412: Peat Bogs
335: Permanent Snow	335: Glaciers and perpetual snow
323: Sclerophyllous Vegetation	323: Sclerophyllous vegetation

Table 3.7: Harmonization scheme used to convert ODSE-LULC nomenclature to Copernicus Global Land Cover classes. On the left side, ODSE-LULC classes are converted to Forest, Other Vegetation, Wetland, Bare, Cropland, Urban, and Water classes. Each transition from one Copernicus class to another is then categorized into a change class in the cross-table.

ODSE-LULC class	Copernicus change class	Forest	Other Vegetation	Wetland	Bare	Cropland	Urban	Water
311: Broad-leaved forest 312: Coniferous forest	**Forest**		Forest loss			Deforestation and crop expansion	Deforestation and urbanization	Water expansion
321: Natural grasslands 322: Moors and heathland 324: Transitional woodland-shrub 323: Sclerophyllous vegetation	**Other Vegetation**	Reforestation		Other	Desertification	Crop expansion	Urbanization	
411: Inland wetlands 421: Maritime wetlands	**Wetland**		Wetland degradation		Wetland degradation and desertification	Wetland degradation and crop expansion	Wetland degradation and urbanization	
332: Bare rocks 333: Sparsely vegetated areas 334: Burnt areas 335: Glaciers and perpetual snow 335: Beaches, dunes, and sands	**Bare**		Other			Crop expansion	Urbanization	
211: Non-irrigated arable land 212: Permanently irrigated arable land 213: Rice fields 221: Vineyards 222: Fruit trees and berry plantations 223: Olive groves 231: Pastures	**Cropland**		Land abandonment		Land abandonment and desertification			
111: Urban fabric 122: Road and rail networks and associated land 123: Port areas 124: Airports 131: Mineral extraction sites 132: Dump sites 133: Construction sites 141: Green urban areas	**Urban**		Other					
511: Water courses 512: Water bodies 523: Sea and ocean 522: Estuaries 521: Coastal lagoons	**Water**	Water reduction						

3.3 Results

Quality of reference data

Table 3.8 shows how well each compared land cover product matched ODSE-LULC training data. The comparison with S2GLC with our points from 2016 and 2018 resulted in the highest F1-scores, while the land cover product made by Pflugmacher et al. [199] fits more closely to the 2015 subset (0.657). The 2019 point subset was considered too small to perform any meaningful comparison between ELC10 and GLFCS30. The number of classes can vary per dataset per year because we excluded all classes from the translated dataset that do not appear in the target land cover product.

Table 3.8: Weighted F1-score of other land cover products when validated with the ODSE-LULC training dataset.

Land cover product	Validation year	Data source	Samples	Weighted F1-Score	Number of classes	Res. (m)
S2GLC	2016	LUCAS	756	0.724	8	10
Pflugmacher et al. [199]	2016	LUCAS	719	0.719	10	30
GLFCS30–2015	2016	LUCAS	724	0.677	10	30
Pflugmacher et al. [199]	2015	LUCAS	144,027	0.657	11	30
S2GLC	2018	LUCAS	295,152	0.653	11	10
S2GLC	2018	CLC	1,000,063	0.604	12	10
ELC10	2018	LUCAS	42,629	0.596	8	10
GLFCS30–2015	2015	LUCAS	138,342	0.503	12	30
ELC10	2018	CLC	172,382	0.456	8	10
GLFCS30–2020	2018	LUCAS	308,838	0.424	12	30
GLFCS30–2020	2018	CLC	1,026,914	0.420	12	30

Spatiotemporal ensemble modelling results

The EML model optimization resulted in the following hyperparameters and architecture:

- Random forest: Number of trees equal to 85, maximum depth per tree equal to 25, number of variables to find the best split equal to 89, and 20 as minimum number of samples per leaf.
- Gradient boosted trees: Number of boosting rounds equal to 28, maximum depth per tree equal to 7, minimum loss reduction necessary to split a leaf node equal to 1, L1 regularization term on weights equal to 0.483, learning rate equal to 0.281, greedy histogram algorithm to construct the trees, and softmax as objective function.

- Artificial Neural Network: Four fully connected hidden layers with 64 artificial neurons each; ReLU as activation function, dropout rate equal to 0.15 and batch normalization in all the layers; softmax as activation function for output layer; batch size and number of epochs equal to 64 and 50, respectively; and Adam with Nesterov momentum as optimizer considering 5e-4 as learning rate.
- Logistic Regression: SAGA solver and multinomial function to minimize the loss.

The variable importance, generated by the two tree-based learners and presented in Fig. 3.7, shows that the 50th quantile for summer and winter of the Landsat green band were most important to the random forest and gradient boosted tree models, respectively. In addition to spectral bands, several Landsat-derived spectral indices (NBR2, SAVI, NDVI, REI, NDWI, MSAVI appear amongst the 40 most important variables. Global surface water frequency was the third most important for the random forest. Fig. 3.7 also shows that the summer aggregates of Landsat green (25th quantile) and NDVI are the two most important variables where the highest importance among the two models is less than double the importance of the other model. Except for Landsat green and NDVI, most variables were found important by only one model. For instance, the geometric temperatures and nighttime land surface temperatures were only important for the random forest. The differences in variable importance indicate that the component models use different parts of the feature space before their predictions are combined by the meta-learner, suggesting that ensembles can utilize a wider proportion of the feature space than single models.

Accuracy assessment results

Spatial cross-validation

We performed 5-fold spatial cross-validation with the final hyperparameters for our ensemble. The predictions on the left-out folds were aggregated to assess model performance on the entire dataset. Table 3.9 shows that the model achieved higher weighted user and producer accuracy, as well as F1-score and log loss ratio, when predictions were aggregated to their next level in the CLC hierarchy. Table 3.10 shows that the model only achieved an F1-score over 0.5 for 10 out of 43 classes (112, 121,211,213,311,312,332,335,412,512). The model performed best when predicting 512: Water bodies (0.924), 335: Glaciers and perpetual snow (0.834), and 412: Peat bogs (0.707). It achieved the lowest F1-scores for 334: Burnt areas (0.011), 132: Dump sites (0.026) and 133: Construction sites (0.065). However, log loss ratios for each class and each CLC level overall were higher than 0, indicating that the model assigned probabilities more accurately than a random classifier even for the most difficult classes.

When the predictions were aggregated to 14 level 2 classes (see Table 3.11), the model performed best when classifying 51: Inland waters (0.924), 31: Forests and seminatural areas (0.813) and 41: Inland wetlands (0.708). The biggest increase in performance

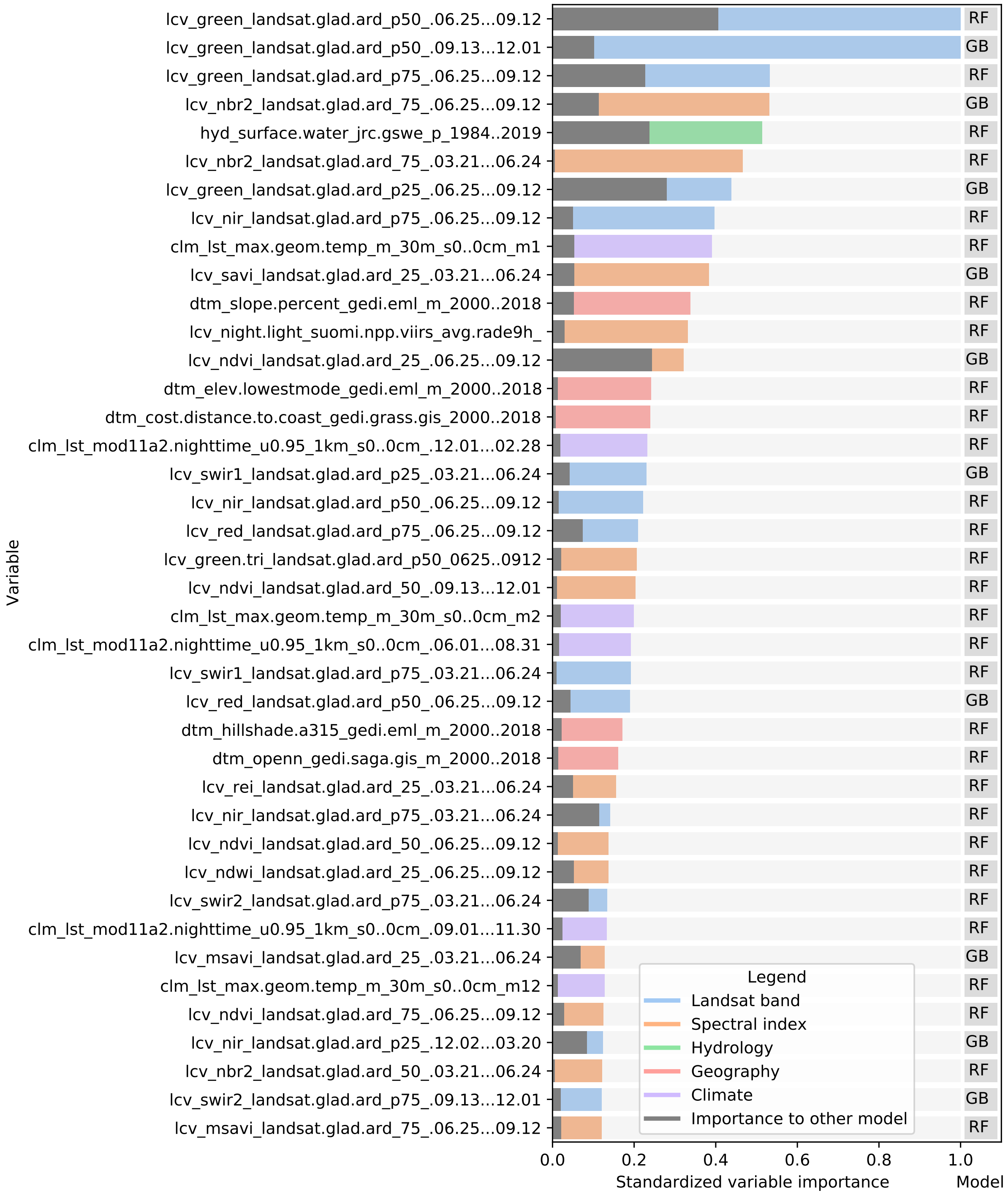


Figure 3.7: Standardized importance of the top-40 most important variables to the random forest and gradient boosted tree models. The colored bar indicates the highest importance of the variable among the two models. This model is indicated to the right of each bar. The corresponding grey bar indicates the importance to the other model. The color of each bar indicates the data type. Each variable name is prefixed with either LCV (either part of a Landsat band or a landsat-derived spectral index), HYD (Hydrological data), CLM (climatic data), or DTM (digital terrain model). This prefix is followed by the specific data source, e.g. *[color or index]_landsat* indicates a Landsat band or derived spectral index. The last part of each name indicates the timespan over which the data was aggregated.

through aggregation to level 2 was in 31: Forests, as the weighted average F1-score of its subclasses (311,312,313) was 0.553. The least accurately predicted classes were 14: Artificial, non-agricultural vegetated areas (0.308), 13: Mine, dump and construction sites (0.370) and 22: Permanent crops (0.412).

Table 3.12 shows that at the highest level of aggregation with 5 general classes, the model classified 5: Water bodies most accurately (0.926) and 1: Artificial surfaces the least (0.688). The best performance improvement from aggregation was for 2: Agricultural areas, as the weighted average F1-score of its subclasses (21, 22, 23, 24) was 0.546, but increased with 0.279 upon aggregation.

Table 3.9: Producer's and user's accuracy, Weighted F1-score, and Log loss of the ensemble predictions during spatial cross-validation.

Corine level	Number of classes	Prod acc.	User acc.	Weighted F1	Log Loss	Baseline Log Loss	Log Loss Ratio
1	5	0.835	0.835	0.834	0.456	2.018	0.774
2	14	0.636	0.639	0.509	1.033	3.596	0.713
3	43	0.494	0.502	0.491	1.544	5.142	0.700

We calculated a separate weighted F1-score for each tile that was used for spatial cross-validation to investigate spatial patterns in classification performance. The average weighted F1-score per tile was 0.463, with a standard deviation of 0.150. Fig. 3.8 shows a disparity in performance between northern and southern Europe. Fig. 3.9 shows that there is a significant correlation (0.125, p=0.000) between the number of reference points and the weighted F1 score of a tile.

We calculated a separate weighted F1-score for all cross-validation predictions from each separate year. Table 3.13 shows that the average weighted F1-score per year was 0.489 with a standard deviation of 0.135. It only scored higher than 0.5 on years with less than 1 million points.

Validation on S2GLC points

We validated the ensemble on S2GLC dataset. We overlaid 49,897 S2GLC points with our input variables for 2017 and classified 43 LULC classes with our model. These 43-class predictions were reclassified to the S2GLC nomenclature. 3,484 points had a predicted class that was not in the S2GLC nomenclature (see Table 3.6). The *'conservative'* assessment (on all 49,897 points) including the non-S2GLC classes resulted in a weighted F1-score of 0.854 and a kappa score of 0.794 (see Table 3.14). The *'optimistic'* assessment excluding non-S2GLC predictions resulted in a weighted F1-score of 0.889 and a kappa score of 0.867 (see Table 3.15).

Taking into account possible noise from the translation process, these results are similar to those reported by Malinowski et al. [161]. Weighted average user and producer accuracy

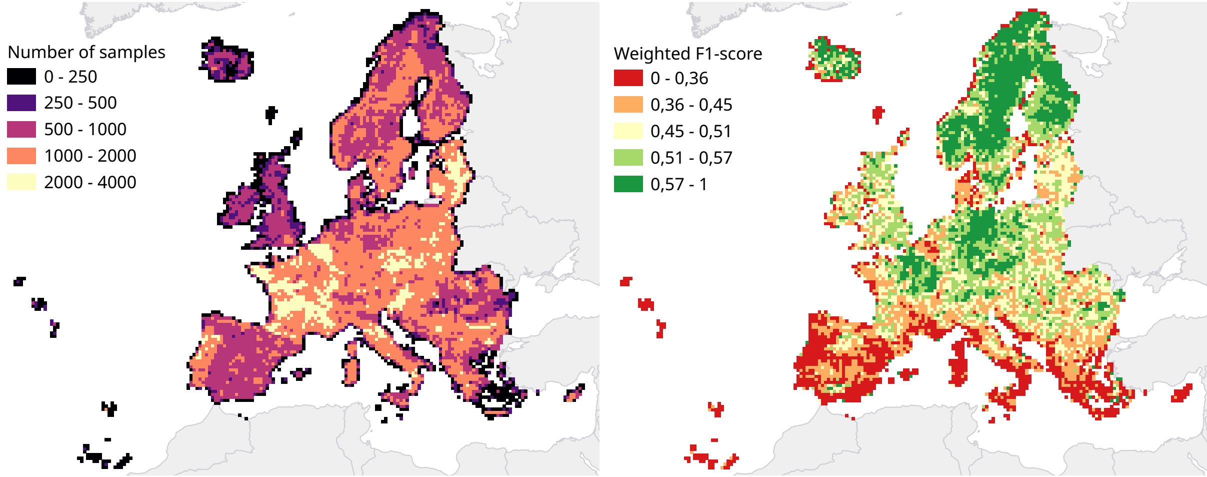


Figure 3.8: 30 km tiling system used for spatial cross-validation, showing the number of samples per tile (left) and the cross-validation weighted F1-score per tile (right).

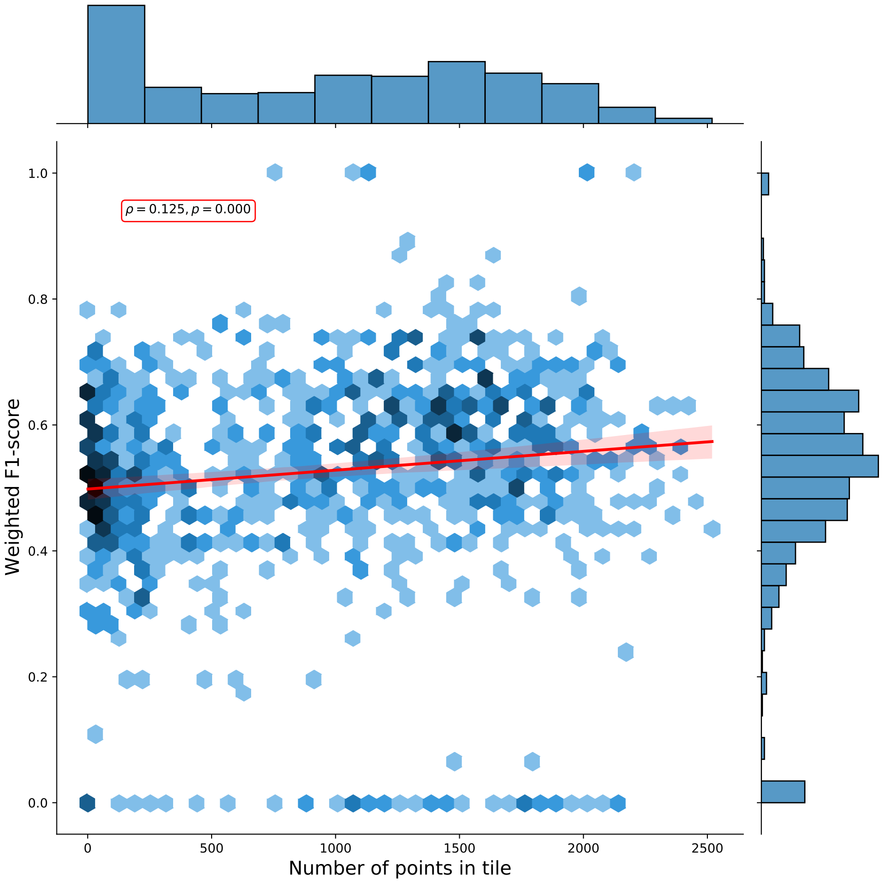


Figure 3.9: Hexbin plot of the weighted F1-score and number of overlapping points per tile. The Pearson correlation coefficient of 0.125 (p: 0.000) indicates there is a weak positive correlation between the number of points in a tile and the cross-validation weighted F1-score.

Table 3.10: Classification report for 43 CLC level 3 classes, based on the predictions made with 5-fold spatial cross-validation.

CLC code (level 3)	Producer Acc.	User Acc.	F1-score	Support	Log loss	Baseline Log Loss	Log Loss Ratio
111: Continuous urban fabric	0.523	0.166	0.252	51,989	0.0230	0.0388	0.409
112: Discontinuous urban fabric	0.509	0.572	0.539	92,151	0.0256	0.0623	0.590
121: Industrial or commercial units	0.496	0.623	0.552	129,661	0.0382	0.0821	0.535
122: Road and rail networks and associated land	0.294	0.068	0.111	39,832	0.0244	0.0311	0.213
123: Port areas	0.543	0.321	0.403	3,994	0.0018	0.0042	0.578
124: Airports	0.300	0.023	0.043	6,702	0.0049	0.0067	0.265
131: Mineral extraction sites	0.482	0.307	0.375	53,447	0.0264	0.0397	0.335
132: Dump sites	0.375	0.013	0.026	6,509	0.0048	0.0065	0.267
133: Construction sites	0.217	0.038	0.065	6,728	0.0047	0.0067	0.299
141: Green urban areas	0.312	0.125	0.179	15,717	0.0091	0.0141	0.350
142: Sport and leisure facilities	0.407	0.200	0.268	64,308	0.0326	0.0463	0.297
211: Non-irrigated arable land	0.604	0.733	0.662	998,381	0.1892	0.3735	0.493
212: Permanently irrigated arable land	0.447	0.146	0.221	29,786	0.0139	0.0243	0.428
213: Rice fields	0.762	0.496	0.601	4,839	0.0020	0.0050	0.596
221: Vineyards	0.506	0.308	0.383	66,213	0.0287	0.0474	0.394
222: Fruit trees and berry plantations	0.411	0.131	0.199	63,659	0.0344	0.0459	0.251
223: Olive groves	0.432	0.355	0.390	63,578	0.0244	0.0459	0.469
231: Pastures	0.455	0.529	0.489	529,466	0.1509	0.2415	0.375
241: Annual crops associated with permanent crops	0.269	0.067	0.107	16,883	0.0101	0.0150	0.326
242: Complex cultivation patter	0.348	0.351	0.349	594,648	0.1942	0.2624	0.260
243: Agriculture with significant natural vegetation	0.355	0.373	0.363	782,237	0.2558	0.3176	0.194
244: Agro-forestry areas	0.276	0.052	0.087	10,497	0.0060	0.0099	0.396
311: Broad-leaved forest	0.537	0.660	0.592	855,499	0.1971	0.3373	0.416
312: Coniferous forest	0.596	0.646	0.620	759,215	0.1644	0.3112	0.472
313: Mixed forest	0.461	0.377	0.414	612,430	0.1707	0.2680	0.363
321: Natural grasslands	0.406	0.314	0.354	400,875	0.1431	0.1971	0.274
322: Moors and heathland	0.493	0.350	0.409	301,693	0.1100	0.1591	0.309
323: Sclerophyllous vegetation	0.311	0.372	0.339	143,521	0.0532	0.0890	0.403
324: Transitional woodland-shrub	0.472	0.431	0.450	724,404	0.2117	0.3013	0.297
331: Beaches, dunes, sand	0.551	0.207	0.301	25,688	0.0147	0.0214	0.312
332: Bare rocks	0.664	0.495	0.567	58,234	0.0162	0.0427	0.621
333: Sparsely vegetated areas	0.522	0.471	0.495	152,571	0.0457	0.0935	0.511
334: Burnt areas	0.224	0.006	0.011	2,263	0.0021	0.0026	0.177
335: Glaciers and perpetual snow	0.852	0.818	0.834	7,250	0.0008	0.0072	0.883
411: Inland marshes	0.425	0.228	0.297	39,784	0.0192	0.0310	0.382
412: Peat bogs	0.684	0.731	0.707	174,314	0.0333	0.1039	0.680
421: Salt marshes	0.505	0.441	0.471	5,598	0.0023	0.0057	0.600
422: Salines	0.481	0.081	0.139	320	0.0002	0.0004	0.577
423: Intertidal flats	0.497	0.209	0.295	788	0.0004	0.0010	0.570
511: Water courses	0.360	0.108	0.166	11,214	0.0068	0.0105	0.353
512: Water bodies	0.895	0.956	0.924	187,981	0.0108	0.1103	0.902
521: Coastal lagoons	0.594	0.429	0.498	1,904	0.0006	0.0022	0.708
522: Estuaries	0.382	0.082	0.135	353	0.0002	0.0005	0.566
Macro average	0.460	0.327	0.356	8097140	0.083	0.137	0.452
Weighted average	0.494	0.502	0.491		0.157	0.253	0.389
Accuracy	0.502						
Kappa score	0.459						
Log Loss (baseline)	1.544 (5.142)						

and F1-scores are also higher than our cross-validation scores at all thematic resolution levels (see Table 3.9). They are also higher than what we obtained when we transformed our cross-validation predictions to the S2GLC nomenclature, which yielded a weighted F1-score 0.611 and a kappa score of 0.535.

Fig. 3.10 shows a normalized confusion matrix of our validation on the S2GLC dataset. It shows the rate at which each true class (rows) was predicted as each other class (columns). The diagonal cells report the true positive rate of each class. Class 000 represents classes

Table 3.11: Classification report for 14 CLC level 2 classes, based on the predictions made with 5-fold spatial cross-validation.

CLC code (level 2)	Producer Acc.	User Acc.	f1-score	Support	Log Loss	Baseline Log Loss	Log Loss Ratio
11: Urban Fabric	0.643	0.535	0.584	144,140	0.039	0.089	0.564
12: Industrial, commercial and transport units	0.568	0.551	0.559	180,189	0.057	0.107	0.469
13: Mine, dump and construction sites	0.533	0.283	0.370	66,684	0.032	0.048	0.331
14: Artificial, non-agricultural vegetated areas	0.479	0.227	0.308	80,025	0.038	0.055	0.315
21: Arable land	0.622	0.738	0.675	1,033,006	0.191	0.382	0.500
22: Permanent crops	0.558	0.326	0.412	193,450	0.072	0.113	0.363
23: Pastures	0.455	0.529	0.489	529,466	0.151	0.242	0.375
24: Heterogeneous agricultural areas	0.488	0.496	0.492	1,404,265	0.364	0.461	0.212
31: Forests and seminatural areas	0.788	0.840	0.813	2,227,144	0.302	0.588	0.487
32: Shrub and/or herbaceous vegetation associations	0.592	0.511	0.548	1,570,493	0.384	0.492	0.218
33: Open spaces with little or no vegetation	0.736	0.591	0.656	246,006	0.061	0.136	0.555
41: Inland wetlands	0.719	0.697	0.708	214,098	0.044	0.122	0.643
42: Coastal wetlands	0.591	0.465	0.520	6,706	0.003	0.007	0.618
51: Inland waters	0.913	0.936	0.924	199,195	0.013	0.115	0.884
52: Marine waters	0.614	0.392	0.479	2,273	0.001	0.003	0.699
Macro average	0.620	0.541	0.569	8,097,140	0.117	0.197	0.482
Weighted average	0.636	0.639	0.634		0.262	0.420	0.393
Accuracy	0.639						
Kappa score	0.565						
Log Loss (baseline)	1.033 (3.596)						

Table 3.12: Classification report for 5 CLC level 1 classes, based on the predictions made with 5-fold spatial cross-validation.

CLC code (level 1)	Producer Acc.	User Acc.	F1-score	Support	Log Loss	Baseline Log Loss	Log Loss Ratio
1: Artificial surfaces	0.784	0.613	0.688	471,038	0.123	0.222	0.445
2: Agricultural areas	0.798	0.854	0.825	3,160,187	0.457	0.669	0.317
3: Forest and seminatural areas	0.872	0.848	0.860	4,043,643	0.526	0.693	0.241
4: Wetlands	0.722	0.696	0.708	220,804	0.045	0.125	0.639
5: Water bodies	0.917	0.936	0.926	201,468	0.013	0.116	0.884
Macro average	0.819	0.789	0.802	8,097,140	0.233	0.365	0.505
Weighted average	0.835	0.835	0.834		0.450	0.626	0.309
Accuracy	0.835						
Kappa score	0.720						
Log Loss (baseline)	0.456 (2.018)						

not present in the S2GLC dataset; as there were no ground truth points in the dataset with these classes, the top row of the matrix is empty. The matrix shows that, when normalized for support, the biggest sources of error were the incorrect classification of classes 323: Sclerophyllous vegetation and 322: Moors and Heathland as classes not in the S2GLC dataset with 29.9% and 27.0% of all errors for these classes, respectively, and of 411: Marshes as 231: Herbaceous vegetation (28.4%). We include a similar confusion matrix of our cross-validation predictions (Fig. 3.11, transformed to the S2GLC nomenclature, to allow a comparison between our cross-validation and independent validation. It shows that many classes have a higher true positive rate in the independent validation on S2GLC points than in our cross-validation results, except for 211: Cultivated areas, 335: Permanent snow cover, and 412: Peatbogs.

Table 3.13: Cross-validation performance of our ensemble model per year.

Year	Weighted F1-score	Support
2000	0.497	1,658,715
2006	0.491	1,852,645
2009	0.558	225,416
2012	0.487	1,971,812
2015	0.588	265,830
2016	0.632	65,235
2018	0.481	2,057,306
2019	0.535	180
Average	0.489	1,012,142
Standard deviation	0.135	882,783

Table 3.14: Conservative classification report of our 2017 LULC prediction on 49,897 S2GLC points that counts 3484 points with predicted classes without an equivalent S2GLC class as errors (141: Green urban areas, 142: Sport and leisure facilities, 222: Fruit trees and berry plantations, 223: Olive groves, 313: Mixed forest, 324: Transitional woodland-shrub, 333: Sparsely vegetated areas, and 334: Burnt areas).

S2GLC Class	Producer Acc.	User Acc.	F1-score	Support
111: Artificial surfaces	0.933	0.933	0.933	1,826
211: Cultivated areas	0.849	0.965	0.903	13,470
221: Vineyards	0.826	0.694	0.754	500
231: Herbaceous vegetation	0.861	0.686	0.764	6,776
311: Broadleaf tree cover	0.967	0.814	0.884	10,944
312: Coniferous tree cover	0.975	0.914	0.943	8,626
322: Moors and heathland	0.641	0.491	0.556	2,070
323: Sclerophyllous vegetation	0.780	0.265	0.396	815
331: Natural material surfaces	0.915	0.751	0.825	2,110
335: Permanent snow cover	0.624	0.800	0.701	85
411: Marshes	0.331	0.327	0.329	324
412: Peatbogs	0.629	0.482	0.546	745
511: Water bodies	0.992	0.974	0.983	1,606
Macro average	0.737	0.650	0.680	49,897
Weighted average	0.892	0.830	0.854	
Accuracy	0.830			
Kappa score	0.794			

Table 3.15: Optimistic classification report of our 2017 LULC prediction on 49,897 S2GLC points where all 3484 points with predicted classes without an equivalent S2GLC class were removed before calculating accuracy metrics (141: Green urban areas, 142: Sport and leisure facilities, 222: Fruit trees and berry plantations, 223: Olive groves, 313: Mixed forest, 324: Transitional woodland-shrub, 333: Sparsely vegetated areas, and 334: Burnt areas).

S2GLC Class	Producer Acc.	User Acc.	F1-score	Support
111 : Artificial surfaces	0.933	0.935	0.934	1,823
211 : Cultivated areas	0.849	0.967	0.905	13,429
221 : Vineyards	0.826	0.720	0.769	482
231 : Herbaceous vegetation	0.861	0.722	0.785	6,441
311 : Broadleaf tree cover	0.967	0.937	0.952	9,512
312 : Coniferous tree cover	0.975	0.973	0.974	8,098
322 : Moors and heathland	0.641	0.672	0.656	1,511
323 : Sclerophyllous vegetation	0.780	0.378	0.509	571
331 : Natural material surfaces	0.915	0.866	0.889	1,831
335 : Permanent snow cover	0.624	0.819	0.708	83
411 : Marshes	0.331	0.351	0.341	302
412 : Peatbogs	0.629	0.494	0.554	726
511 : Water bodies	0.992	0.975	0.984	1,604
Macro average	0.794	0.755	0.766	46,413
Weighted average	0.893	0.892	0.889	
Accuracy	0.892			
Kappa score	0.867			

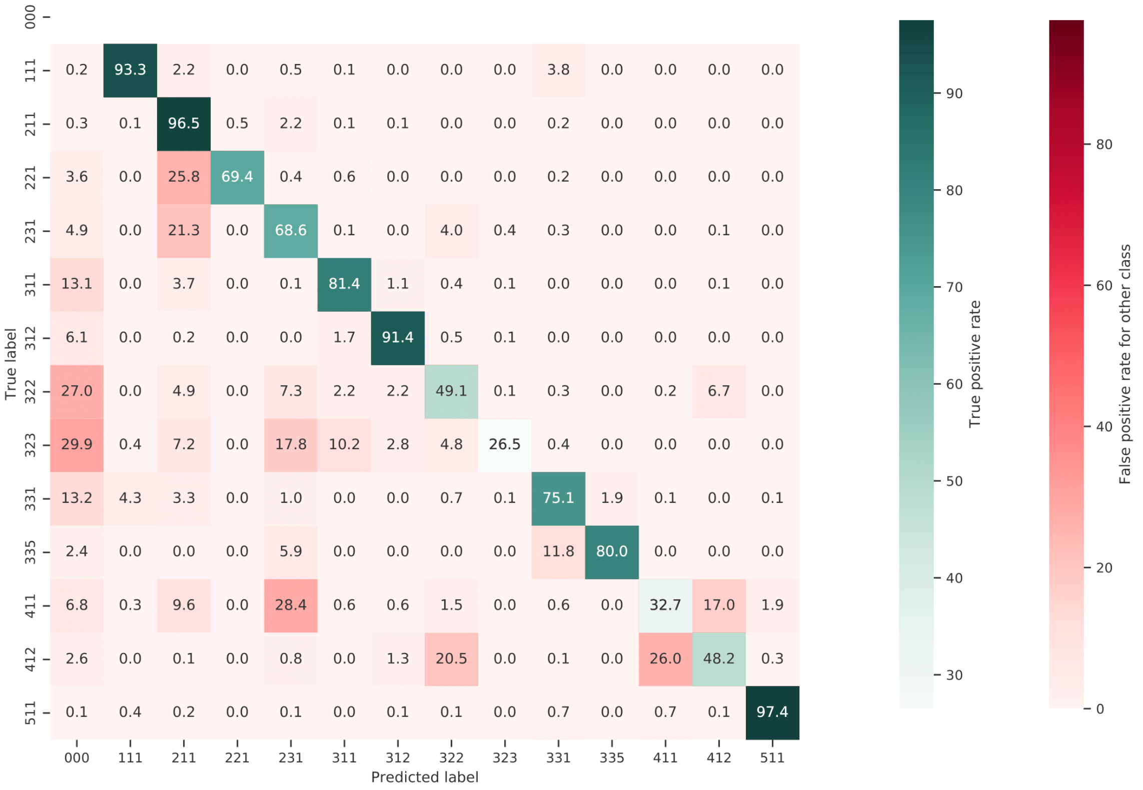


Figure 3.10: Normalized confusion matrix of our prediction on the independently collected S2GLC validation points. Each cell shows the percentage of the true label predicted as the predicted label.

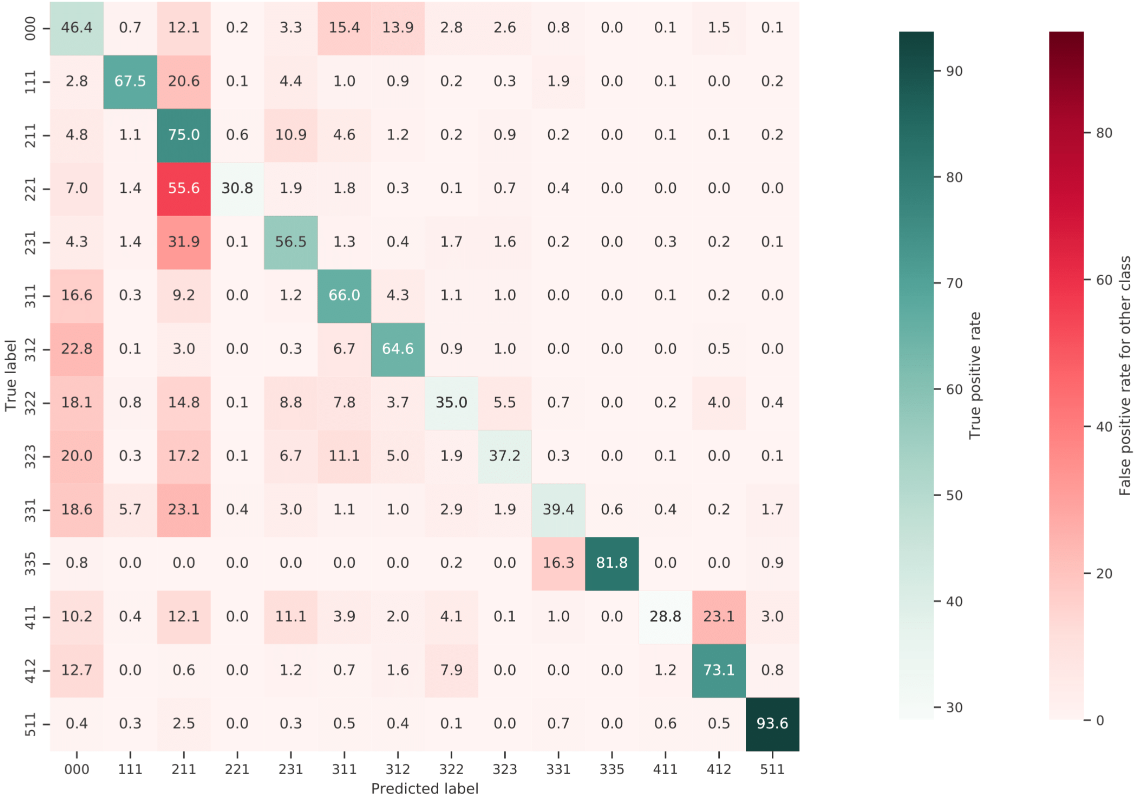


Figure 3.11: Normalized confusion matrix of the predictions made by our model during spatial cross-validation on our own dataset, reclassed to the S2GLC nomenclature. Each cell shows the percentage of the true label predicted as the predicted label.

Comparison of spatial and spatiotemporal models

We trained two types of models and compared their performance: Spatial models, which were trained on 100,000 points sampled from one year, and spatiotemporal models, which were trained on 100,000 points equally distributed across multiple years. Table 3.16 shows the weighted F1-scores obtained through validating each model on 33,333 points from the same year(s) as its training data, and on 33,333 points from the year 2018, which was left out of all training datasets.

The results show that all models performed better when validated on points from the same year as their training data, regardless of data source. However, spatial models achieved higher F1-scores on average when trained and validated on only LUCAS points, while the spatiotemporal models performed better when trained and validated on only CLC points.

The spatiotemporal model trained on only CLC points achieved the highest F1-scores for both known-year and unknown-year classification. This model outperformed spatial models on known-year classification by 2.7% and unknown-year classification by 3.5% as seen in Table 3.16.

Table 3.16: Weighted F1-scores obtained by validating spatial and spatiotemporal models on data from known years and an unknown year (2018).trained on CLC points, LUCAS points, and a combination of both.

Model	Training Year	Points	Trained on CLC		Trained on LUCAS		Trained on CLC and LUCAS	
			Tested on raining year(s)	Tested on 2018	Tested on training year(s)	Tested on 2018	Tested on training year(s)	Tested on 2018
Spatial	2000	100,000			0.610	0.542	0.611	0.515
Spatial	2006	100,000	0.595	0.437	0.604	0.563	0.587	0.534
Spatial	2009	100,000	0.595	0.482			0.602	0.415
Spatial	2012	100,000	0.559	0.476	0.611	0.574	0.565	0.529
Spatial	Average	400,000	0.583	0.465	0.608	0.560	0.591	0.498
Spatiotemporal	All	100,000	0.612	0.576	0.568	0.478	0.574	0.532
Spatiotemporal	All	400,000	0.625	0.579	0.608	0.491	0.595	0.543

Comparison of ensemble and component models

We compared the F1-score of each component model and the meta-learner. The neural network achieved the highest weighted F1-score of 0.514. The meta-learner scored 0.513, the random forest 0.506, the gradient boosted trees 0.471. Fig. 3.12 shows the difference in performance per model per class. When scored per class, the meta-learner achieved the highest F1-score on 36 out of 43 classes, the random forest on 1 class (523), the gradient boosted trees on 6 classes (132,334,422,423,521,522), and the neural network on 1 class (221).

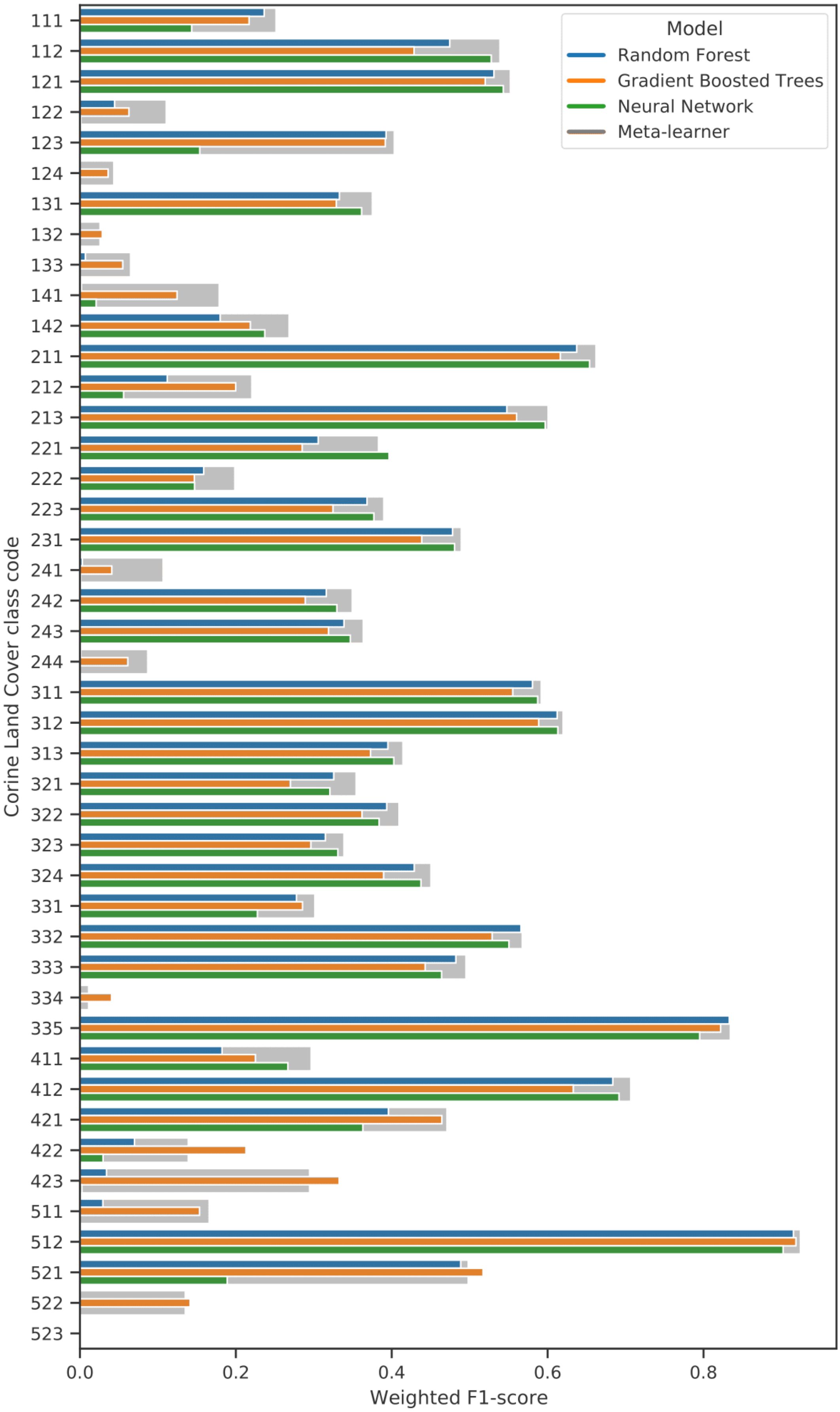


Figure 3.12: Grouped bar plot of the F1-scores CLC class, plotted separately per model of the ensemble. Meta-learner performance is indicated in red on the background of each bar. If the random forest (blue), gradient boosted trees (orange) or neural network (green) outperformed the meta-learner, its bar will exceed the bigger meta-learner bar, indicating that

Time-series analysis results

Our NDVI slope maps show which areas have an increase or decrease in NDVI over time. We selected 19500 LUCAS points that experienced LULC change and overlaid these with our NDVI slope values. Figs. 3.13 and 3.14 show clear differences in NDVI trend between LUCAS points that have undergone different LULC change processes.

We generated annual maps for change classes (see Fig. 3.15 for the maps of 2000 and 2019). Filtered data as well as the removed noise can be viewed from the ODS-Europe viewer.

Fig. 3.16 demonstrates how trend analysis can be used to explore large-scale trends and pixel-level details.

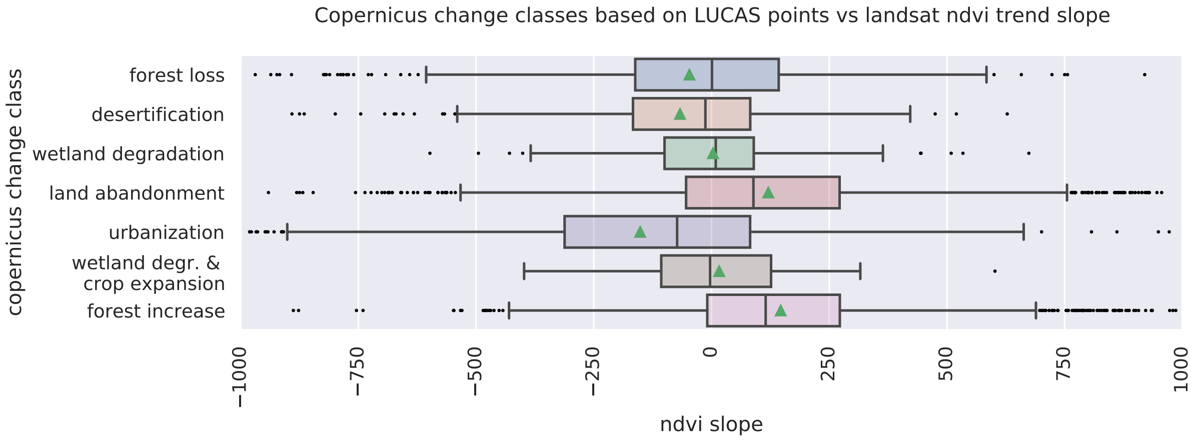


Figure 3.13: NDVI trend slope values of LUCAS points with selected LULC change dynamics, categorized according to the Copernicus change classes. The mean NDVI trend value is indicated with green triangles.

Fig. 3.16-B1 and Fig. 3.16-B2 show areas of negative and positive slope occur adjacent to each other without gradual transitions. Fig. 3.16-B3 and Fig. 3.16-B4 show examples of relatively large areas with homogeneous NDVI slope values. Overall, NDVI slopes in Europe tend to be positive, the largest exceptions being negative slope regions in Northern Scandinavia, Scotland, the Alps, South West France, Spain, Italy and Greece.

The right-most subplots of Fig. 3.16 show examples of where sudden land cover change classes at 30×30 m tend to match relatively large negative slopes, especially for change classes such as forest loss and urbanization.

Fig. 3.17 presents the long-term LULC change processes as suggested by our classification results. Fig. 3.17-A presents the dominant type of LULC change in a 5×5 km grid, while Fig. 3.17-B shows the intensity of change as part of the total area on a separate map using 20×20 km areas. Large parts of mainland Europe are characterized with reforestation as the main change with patches of urbanization scattered in between. Norway, Sweden and

Finland are characterized with forest loss as the main LULC change class. Large areas in Spain have land abandonment and crop expansion as the main land use class. When taking into account the intensity of the changes the central European countries seem to be stable with the Iberian peninsula, Scandinavia and parts of eastern Europe exhibiting more intense changes.

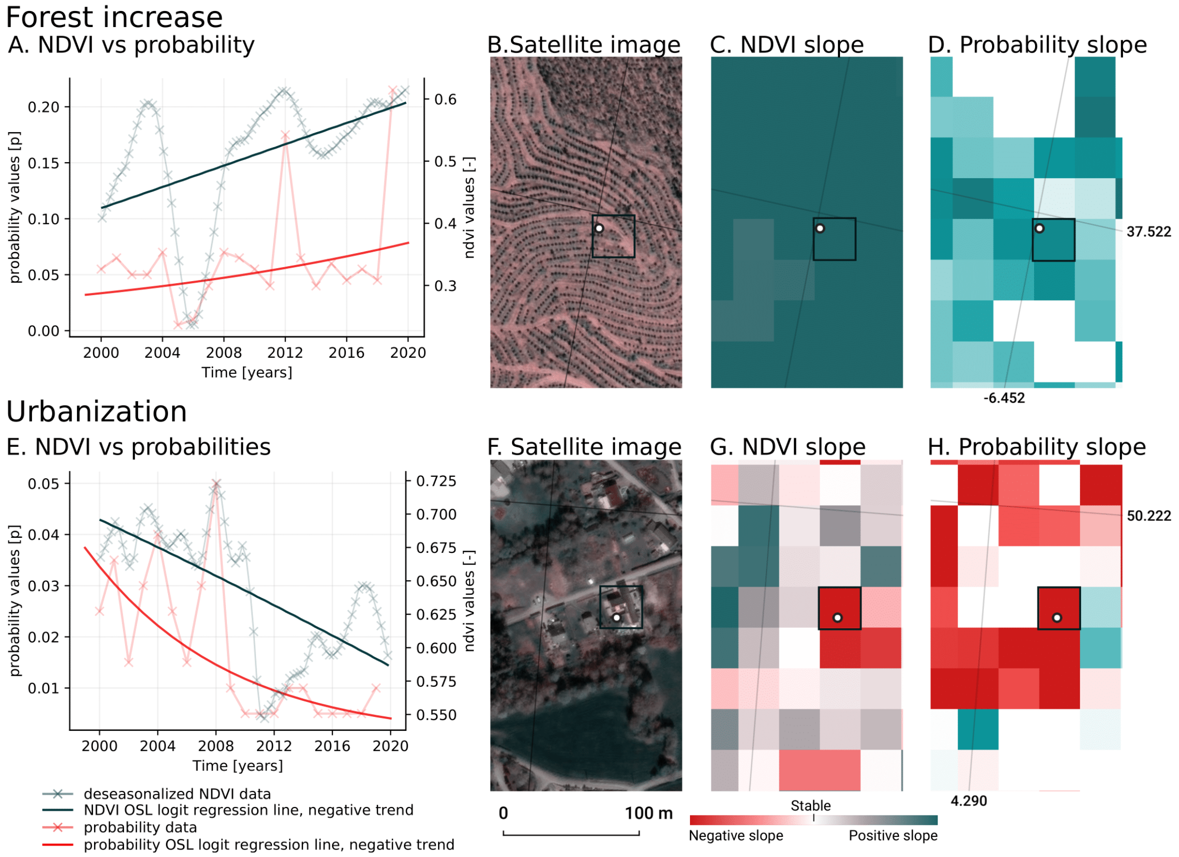


Figure 3.14: Detail plot of NDVI and LULC trends between 2000–2020 for 2 LUCAS points. NDVI trend is compared to forest increase (top) and urbanization (bottom). Left (A and E): A graph comparing the two trends, with green depicting de-seasonalized NDVI data and its trend, as calculated by logit OLS regression. Red depicts the annual probability values and associated trend of the compared LULC change classes (*"312: Coniferous forest"* and *"111: Continuous urban fabric"*, respectively). The maps, from left to right, depict the spatial context of the two points in (B/F) high-resolution satellite RGB, (C/G) slope of Landsat ARD NDVI trends, and (D/H) slope of LULC change class trends as predicted by our ensemble. The *"in-situ"* observations of both points match the dynamic presented in the graph: Point 28681762 (top) experienced forest increase, while point 39143028 (bottom) is located in a recently constructed urban area.

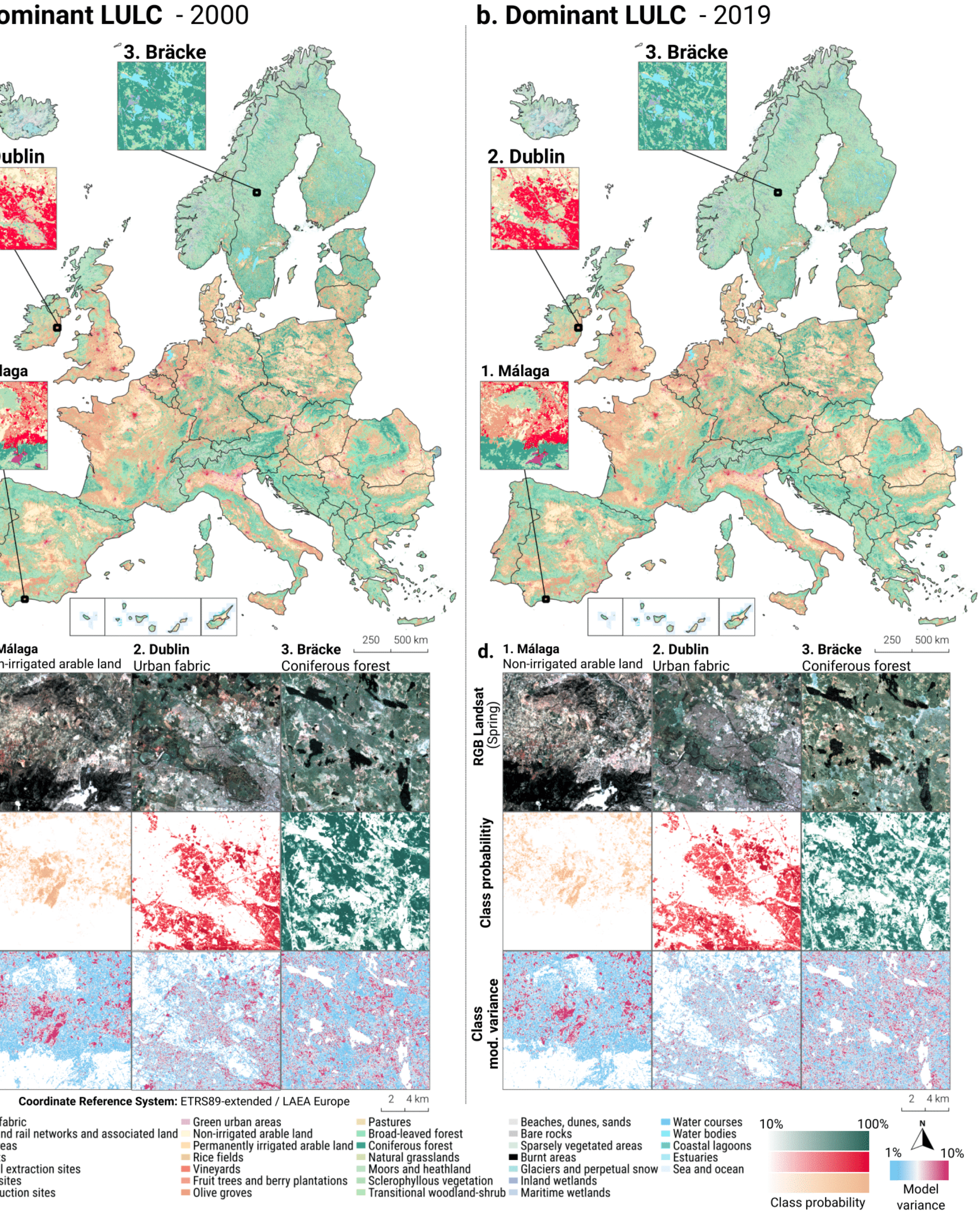


Figure 3.15: Dominant LULC classes, predicted probability and model variance for Non-irrigated arable land, Coniferous forest and Urban Fabric, RGB Landsat temporal composite (Spring season) for the years 2000 and 2019.

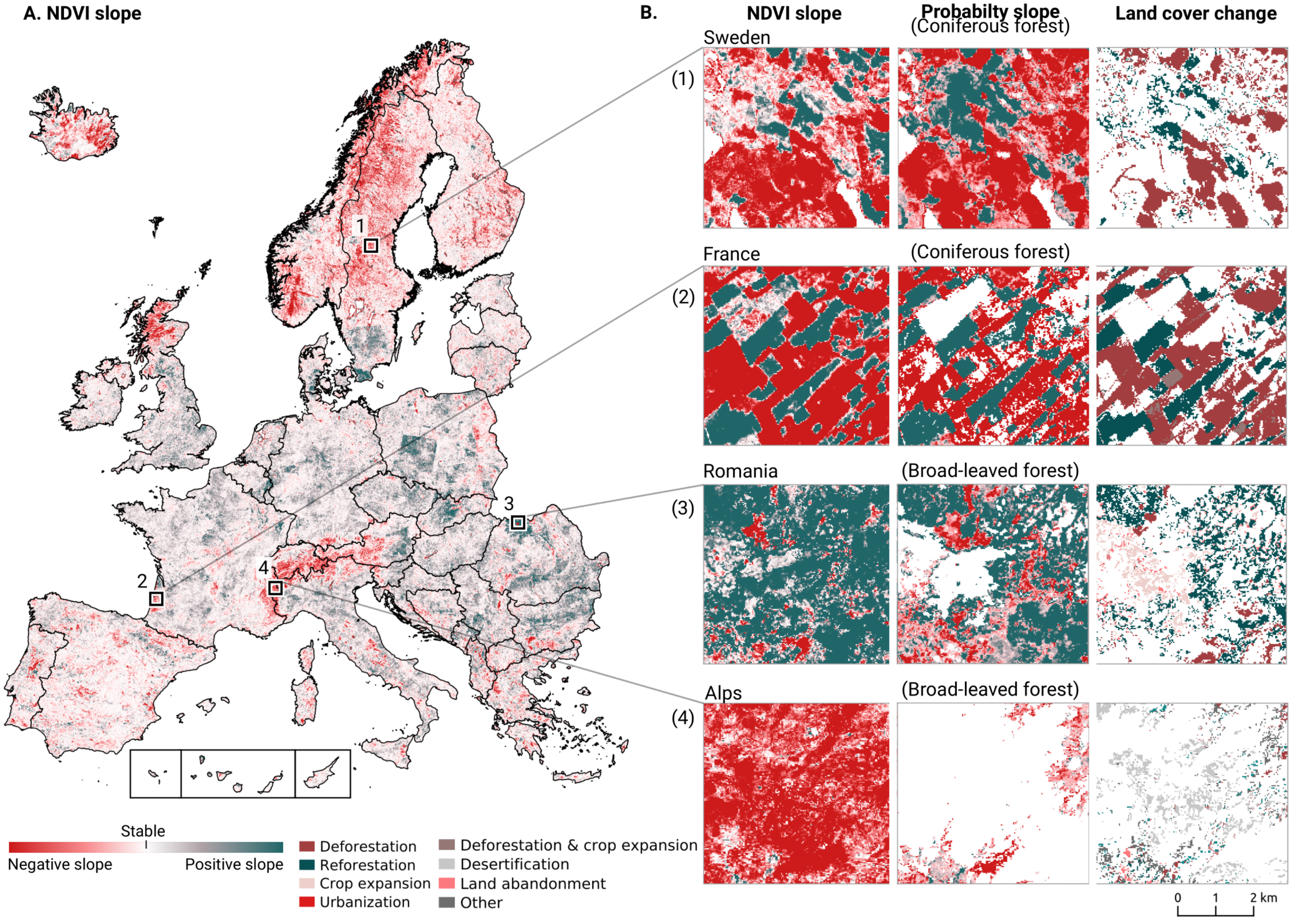


Figure 3.16: Trends in NDVI values between 2000 and 2019 compared to trends in LULC probabilities predicted by our ensemble model, as well as the derived LULC change classes between 2001 and 2018.

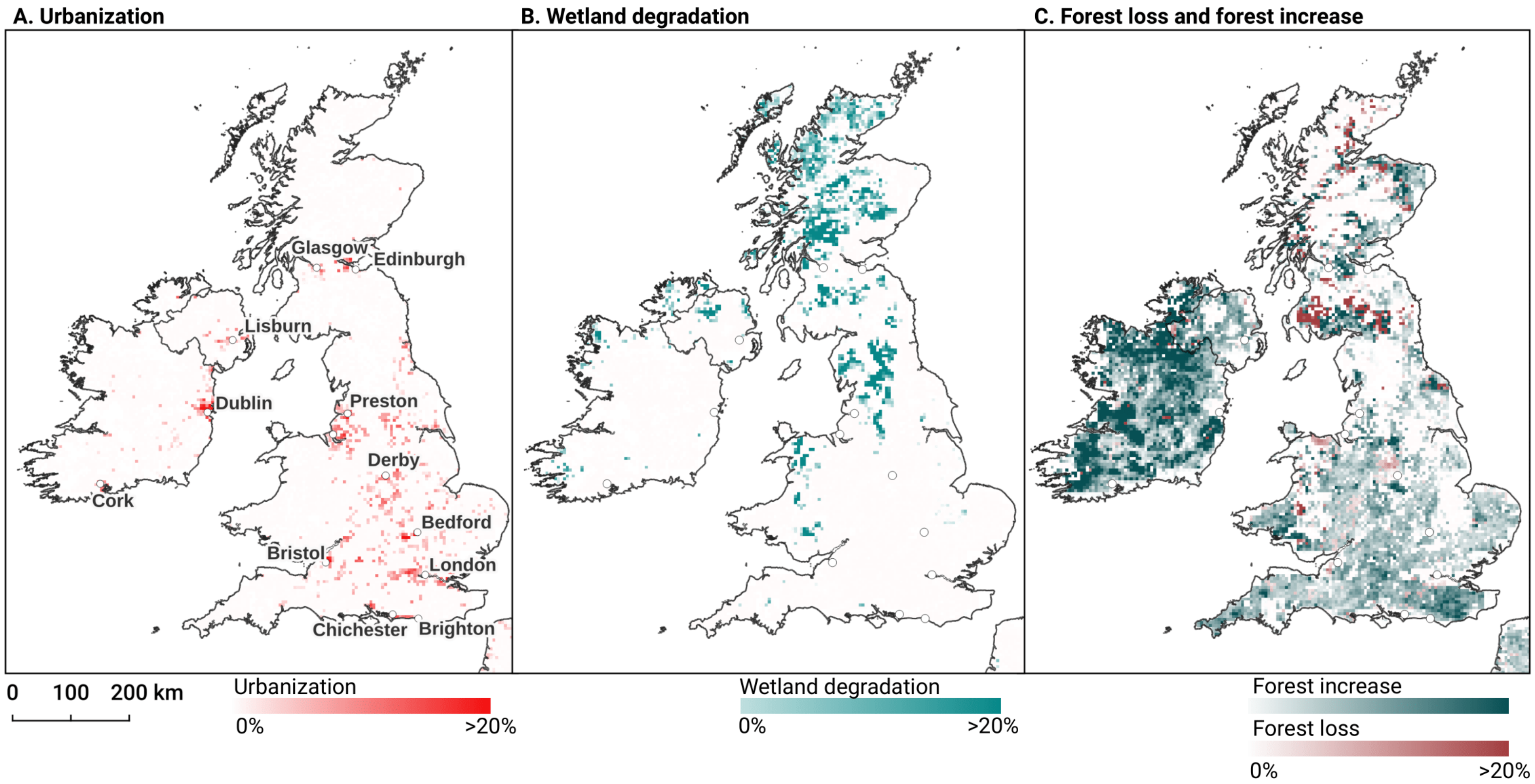


Figure 3.17: Prevalent LULC change and change intensity on the British isles aggregated to 5×5km tiles, for three dynamics: Urbanization (A), Wetland degradation (B), and forest increase/decrease (C).

3.4 Discussion

> *"The appropriateness and adequacy of the 10-class schema used to describe land cover in today's human-dominated world needs a serious rethink. What is the value of a 10 m (resolution) landcover map that cannot capture a grassland being turned into a solar farm?"*
>
> **Mysore Doreswamy Madhusudan**

Summary findings

We have presented a framework for automated prediction of land cover / land use classes and change analysis based on spatiotemporal Ensemble Machine Learning and per-pixel trend analysis. In this framework, we focused not only on predicting the most probable class, but also on mapping each probability and associated model variance. We believe that such detailed information gives a more holistic view of the land cover and land use and allows any future users to derive their own specialized maps of certain classes using probability thresholds and model variance per pixel and class, and/or to incorporate it in further spatial modeling.

We show that in the context of reproducing the CLC legend, models trained on multi-year observations generalize better to unknown years than models trained on single-year observations, and that ensemble machine learning marginally outperforms single classifiers overall. Our accuracy assessment however indicates that several CLC classes remain hard to reproduce in the proposed workflow. The on-par performance on the S2GLC validation points, however, suggests that the framework is capable of generating accurate predictions for relatively detailed legends if they do not contain heterogeneous classes.

We further explained the time-series analysis framework for processing partial probabilities and NDVI values aiming at detection of significant spatiotemporal trends. We provide pixel-wise uncertainty measures (standard deviation of the slope / beta coefficient and R-square), which can also be used in any further spatial modeling. The whole framework, from hyper-parameter optimisation, fine-tuning, prediction and time-series analysis, is fully automated in the (`eumap` python package `https://eumap.readthedocs.io/`) and generates consistent results over time with quantified uncertainty, making it more cost-effective for future updates and additions.

Model performance

Our spatial cross-validation accuracy assessment results indicate limited hard-class accuracy (Weighted F1-score of 0.494) at the highest classification level (43 classes) with several classes such as *124: Airports*", *334: Burnt Areas* performing poorly, likely rendering them unfit for further use. However, a comparison of each class' separate log loss score indicates that the model predicted each class more accurately than the baseline. For example, *522: Estuaries* was one of the least accurately predicted classes in the hard-class classification, but had a log loss ratio of 0.566. This means that probabilities were frequently correctly assigned to validation points in estuaries but overshadowed by other, more numerous classes (e.g. *512: Water Bodies*), allowing a more accurate mapping of estuaries by adjusting the probability threshold for that specific class. Furthermore, our validation on the independent S2GLC dataset collected by Malinowski et al. [161] indicates that the accuracy of our model is comparable to the model used in their publication. Our conservative estimate (counting all points with predicted classes outside the S2GLC legend as errors) resulted in a weighted average F1-score of 0.854 and a kappa score of 0.794 and our optimistic estimate (where those points were removed before calculation) yielded F1: 0.889 and kappa: 0.867, while Malinowski et al. [161] reported 0.86 and 0.83, respectively. While these points were sampled to validate a 10 m resolution map and it is unclear how this affects the accuracy assessment, we could not find a reason to expect overestimated accuracy values in existing literature.

This suggests the nomenclature used by Malinowski et al. [161] is more optimized for remote sensing-based classification than the CLC legend and that the framework presented in this work is capable of achieving accuracy levels comparable to state-of-the-art 10 m resolution land cover products when using a more suitable legend. However, when we transformed our cross-validation results to the S2GLC legend, we obtained an F1-score of 0.611 and a kappa score of 0.535, which is considerably lower. This is unlikely to happen when comparing two datasets that are both sampled in a representative, proportional approach; it is therefore likely that the mismatch is caused by the training points in the ODSE-LULC dataset that were generated from CLC centroids.

The average weighted F1-score per year was 0.489 with a standard deviation of 0.135, while the average weighted F1-score per tile was 0.463, with a standard deviation of 0.150. This means that our model was more consistent through time than through space. A possible explanation is the unequal distribution of training points derived from the CLC data; we did not sample this data based on how much area they cover, but instead on how many separate areas occur in the data. Regions of Europe and classes with smaller CLC polygons may be over-represented in the data. Fig. 3.9 shows that there is a slight but significant correlation between the number of points and cross-validation F1-score. This suggests that improving the CLC sampling strategy may improve the spatial consistency of our model.

Advantages and limitations of combining CLC and LUCAS points

We included LUCAS points in our dataset in order to base our modeling and predictions on a consistent and quality-controlled dataset. However, in this work we found that training spatiotemporal models on LUCAS points lead to lower classification accuracy estimates than when only using CLC points (see Table 3.16). This was unexpected, as LUCAS land cover information stems from actual ground observations, while the CLC points are pseudo-ground truth points from a dataset with a large minimum mapping unit. This suggests that either the LUCAS points are harder to reproduce with remote sensing techniques, or that the harmonization and data filtering process needs to be improved. Further testing is needed to clarify this.

Advantages and limitations of using spatiotemporal models

The results of testing the generalization potential of spatiotemporal models with separate experiments (see methods and results sections about spatial vs spatiotemporal machine learning) show that spatiotemporal models generalize better to data from years they were not trained on. These findings suggest that we can use the existing model to predict land cover for 2020 and 2021 without collecting new training data: Preparing Landsat images for these periods would be likely enough.

Our results also suggests that we can use contemporary reference data to make consistent predictions for periods *prior* to the year 2000, for which very little training data is available. We intend to produce predictions for the years 1995, 1990 and to 1985 in the next phase of our project. We did not do this previously because the Landsat ARD data [206] is only available after 1997. We need to compute and re-calibrate the Landsat 5, 6 and 7 products ourselves, which adds a higher level complexity due to the differences in sensors and acquisition plans.

What further limits us the fact that the long-term spatiotemporal approach aims at 30 m resolution data, while most current land cover products aim at a 10 m resolution. Furthermore, our approach is highly dependent on the availability of quality reference data from multiple years. Many continents except North America and Australia do not have access to datasets similar to LUCAS, which might become real challenge for applying the framework outside Europe, and especially in Africa, Latin America and Asia.

Advantages and limitations of using ensemble models

We implemented ensemble machine learning in our framework for two main reasons. Firstly, to achieve the highest accuracy possible, and secondly, to allow for the inclusion of model variance as a proxy for the uncertainty of its predictions [314]. Our results indicate that using an ensemble approach can indeed increase accuracy. Although the neural network component model scored a slightly higher weighted average F1-score than the meta-learner, the meta-learner achieved the highest F1-score on most classes, suggesting that the meta-learner sacrificed a slight amount of overall performance in order to improve performance on classes that the neural network could not recognize.

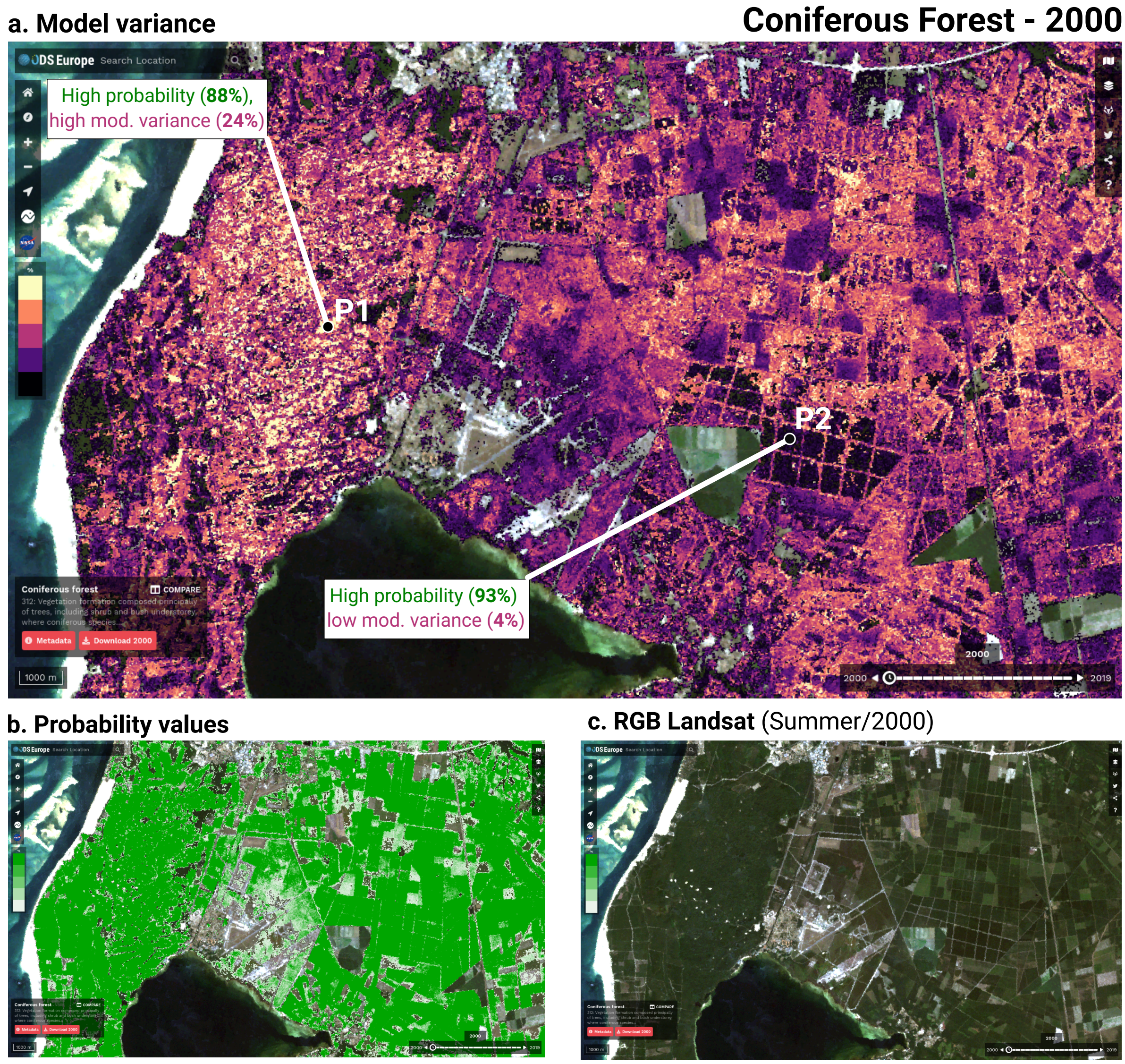


Figure 3.18: Example of model variance (prediction uncertainty) in the city is of La Teste-de-Buch (France) for the class *"Coniferous forest"*, visualized in the odse viewer (`https://maps.opendatascience.eu`): (a) model variance map with examples of two locations (P1 in 44°33'33.6"N 1°10'33.2"W; P2 in 44°32'11.8"N 1°02'38.0"W) with low and high variances, (b) probability values showing relatively high confidence, (c) original Landsat images RGB composite used for classification.

Another advantage of doing ensembles with 5–fold CV with refitting of models and then stacking, is that we can generate maps of model variance (showing where multiple models have difficulties predicting probabilities). This allows users to identify problem areas (see Fig. 3.18), determine where best to collect additional samples, or adjust their classification legend or probability thresholds. To our knowledge, mapping model error of predicted probabilities is a novel area and none of existing landcover datasets for EU provides such information on a per-pixel basis.

Time-series analysis, interpretations and challenges

Palahi et al. [190] found that the transition between Landsat 7 and 8 caused temporal inconsistency in the reflectance data. We tested whether these inconsistencies were propagated into our aggregated and harmonized dataset by calculating the NDVI values of 11 million pixels of our dataset. We then performed a two-sided t test in order to analyze whether there was a difference in NDVI values before and after the launch of Landsat 8 in 2013 (see Fig. 3.19). The t test did not indicate a significant difference (test statistic of 0.0 and p=1.0) between the two distributions, suggesting that the inconsistencies from the transition were not propagated through our preprocessing step.

The results of the probability trend analysis show some interesting patterns. We have focused on four geographic areas: (1) Sweden, as its forest dynamics have already garnered academic attention and it is an exemplary area where remote sensing techniques and on the ground measurements might come to different conclusions (see e.g. Ceccherini et al. [35]). (2) South West France, as it is similar to the Sweden both in our data and is also compared by other authors [236]. (3) Northern Romania because it shows a large region with positive trends for both NDVI and broad-leaved forest land cover, suggesting it is reforesting at high rates. Finally, we found large regions in the Alps (4) that show a strong negative trend for NDVI values that does not seem to correspond to a clear land use change. This signal in our data suggests there may be more artifacts and that further research is needed.

Forest loss in Europe is currently highly debated in academia [35, 190, 201, 235, 236]. Discrepancies between national forest inventories and remote sensing techniques has led to disagreements in Sweden [193], Finland [139], and Norway [225]. For instance, it was found that existing remote sensing products are deemed not fit for these types of analysis [190]. For these reasons, and because we do not validate our trend results, we neither attribute specific causes, nor do we analyze differences between specific time periods.

Further comparison of the most prominent change between 2001–2018 and our results suggest that forest is disappearing more than it is re-appearing in multiple locations. This is corroborated by Global Forest Watch forest gain data; for example, the Jämtland region in Sweden lost 287k ha of tree cover and gained 164k ha between 2001 and 2012 [96]. We present the case of the Landes region in France here as well as it shows a similar pattern to large parts of Sweden and is a known area for large scale forest harvesting [236]. These cases exemplify the usefulness of our maps for finding similar processes all over Europe by using a combination of the data that is presented here. More testing and ground-validation of the land cover changes is needed to assess which changes are over-estimations and which are realistic.

Our data suggests that reforestation is the most prominent land cover change dynamic on a European scale. This change is accompanied by an observed increase of NDVI values.

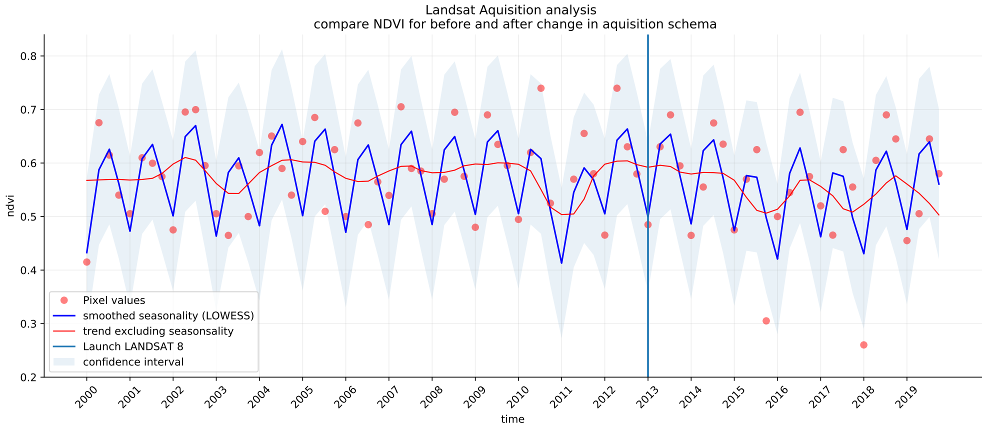


Figure 3.19: NDVI signal for 880 million pixel values in our Landsat data between 2000 and 2019. Red dots indicate the average for each season for 880 million pixels over 11 tiles. The vertical line indicates the launch of Landsat 8, after which the acquisition scheme changed. This sample suggests that the structural difference between the two acquisition schemes in the Landsat ARD product created by Potapov et al. [206] were not propagated into our aggregated and harmonized dataset.

This observation is corroborated by the FAO's State of Europe's Forests report 2020 which states that European forest cover has increased by 9% between 1990 and 2020 [217] and with global estimates that forest cover has increased by 7% between 1982 and 2016 [244]. This increase is consistent with expectations that increased CO2 will enhance plant growth in general. Another concern that is raised is that most of the increase in forest gain is by planted forests [195] that are less valuable in terms of biodiversity and carbon sequestration [152] and less adaptable to climate change. One exemplary area with observed reforestation is found in Northern Romania in all parts of our time-series analysis: we see a change from grassland to forests making reforestation the dominant change class, the broad-leaved forest class probability is increasing, and NDVI values show positive trends.

Finally, our data for the Alps shows unexpected negative NDVI trends for large parts of the Alps. This may be related to changes in snow cover as found by Wang et al. [296] in the Tibetan Plateau and by Buus-Hinkler et al. [30] in the Arctic regions. However, this is not corroborated by the probability slope for class *"Glaciers and perpetual snow"* in our data. It is also possible that this is an artifact from our gap-filling step. Again, further study is necessary before any conclusions can be drawn.

Future work

Even though our framework is comprehensive and has produced predictions of comparable accuracy to the current state-of-the-art on a less complex legend (see results section on

S2GLC), after almost 14 months of processing the data and modeling land cover, we have found that that many aspects of our system could be improved:

- *Improving performance without sacrificing detail*: We consider the poor performance on the 43-class level 3 CLC legend to be the main weakness of our approach. Including such a large and hierarchical legend theoretically makes the resulting data more useful to more potential users, but this will only manifest if the classifications are also reliable for research and policy. To this purpose, we will continue research on methods to improve classification performance while maintaining (or expanding) thematic resolution.
- *Cross-validation of land cover trends*: It was beyond the scope of our project to validate the results of our long-term trend analysis. Independently identifying and quantifying both sudden land cover changes (e.g. due to natural hazards such as fires and floods) and gradual dynamics such as urbanisation and vegetation succession. We have however published all our data online, enabling other research groups to test their usability for land monitoring projects.
- *Combining classification with Object-Based Image Analysis (OBIA) and pattern recognition*: Incorporating spatial context to our workflow could potentially improve performance for several classes that are defined by land use. For instance, class 124: *"Airports"* was frequently misclassified as either urban fabric, non-irrigated arable land, pastures, or Sport and leisure facilities, another complex class that contains buildings and green areas. These predictions likely matched the land cover of the pixel, but missed the spatial patterns that make airports easily recognizable by humans (elongated landing paths). The same issue applies to most other artificial surface LULC classes. The relatively high importance of the TRI of the Landsat green band (see Fig. 3.7) suggests that additional feature engineering or other forms of incorporating the spatial context would improve classification performance on complex classes.

The field of land cover mapping is rapidly evolving. With exciting new global 10 m resolution products such as ESA WorldCover and Google's Dynamic World Map expected in 2021, we expect the LULC mapping bar to be raised quickly to higher resolution and higher accuracy. Venter and Sydenham [287] used low-cost infrastructure to produce land cover map of Europe at 10 m — thanks to ESA and NASA making the majority of multispectral products publicly available, today everyone could potentially map the world's land cover from their laptop. Szantoi et al. [258] show that many land cover products, however, are often ill-suited for practical actions or policy-making. As the quote at the start of this sections says *"The appropriateness and adequacy of the 10-class schema used to describe land cover in today's human-dominated world needs a serious rethink"*, we assert that one should not look for land cover classification legends that are *"low-laying fruits"* for the newest Sentinel imagery, but build people- and policy-oriented datasets that can directly help with spatial planning and land restoration. Our primary focus, thus, will remain on producing harmonised, complete, consistent, current and rapidly-updatable land cover maps that link to the past and allow for the unbiased estimation of long-term trends. We intend for this type of data to facilitate a better understanding of the key drivers of land degradation and restoration, so that we can help stakeholders on the ground make better decisions, and hopefully receive financial support for the ecosystem services our environment provides to us all.

3.5 Conclusions

The spatiotemporal ensemble machine learning framework presented achieved a cross-validation weighted F1-score of 0.49, 0.63, and 0.83 when predicting 43 (level-3), 14 (level-2), and 5 classes (level-1). These values are lower than those reported by other current works that use classification systems with more optimized legends, and less classes. Our validation on an independent test dataset [161] with such an optimized legend yielded accuracy metrics comparable to Malinowski et al. [161]. This indicates that the framework is capable of achieving similar performance to state-of-the-art methods, without any post-processing, and on a coarser spatial resolution, given a less ambitious task.

In our experiments, spatiotemporal models generalized better to EO data from previously unseen years: Spatiotemporal models outperformed spatial models on known-year classification by 2.7% and unknown-year classification by 3.5%. This suggests that spatiotemporal modeling, as incorporated in the presented framework, can be used to predict LULC for years of which no LULC observations exist, even prior to 2000 and beyond 2020.

Other methodological advantages of using spatiotemporal ML are (1) that it helps produce harmonized predictions over the span of years, (2) that the fitted model can be used to predict LULC in years that were not included in its training dataset, allowing generalization to past and future periods, e.g. to predict LULC for years prior to 2000 and beyond 2020. Also, it is an inherently simple system with whole land cover of EU represented basically with a single ensemble ML (a single file). The disadvantages of using spatiotemporal ML is that it requires enough training points spread through time, and EO data needs to be harmonized and gap-filled for the time-period of interest (in this case 2000–2019).

Time-series analysis of predicted LULC probabilities and harmonized NDVI images over continental Europe suggests forest loss in large parts of Sweden, the Alps, and Scotland. The Landsat ARD NDVI trend analysis in general matches the land degradation / reforestation classes with urbanization resulting in the biggest decrease of NDVI in Europe.

Chapter 4

Iterative Mapping of Probabilities: A data fusion framework for generating accurate land cover maps that match area statistics

This chapter is based on:

M. Witjes, M. Herold, and S. de Bruin. *Iterative Mapping of Probabilities: A data fusion framework for generating accurate land cover maps that match area statistics.* 2024

Abstract

Providing land cover estimates with both correct pixel-level class predictions and regional class area estimates is important for many monitoring and accounting purposes but rarely achieved by current land monitoring efforts. We propose a framework that uses class probabilities predicted by machine learning to guarantee that the mapped proportion of each class matches independent area estimates. We used CatBoost models trained on CORINE data to predict probabilities for 8 primary LUCAS land cover classes in five European countries. We then used the proposed algorithm to produce proportional class maps that match Eurostat class area estimates. We validate these proportional class maps and baseline highest likelihood class maps with LUCAS land cover observations and S2GLC validation points. Our results show that the framework and algorithms create maps that match area estimates, and that may also be more accurate than maps created with highest likelihood classification. This is especially the case with general-purpose models trained on data whose class proportions are not representative of the mapped area, which means that this algorithm can be used to localize such models for more accurate mapping of individual countries.

4.1 Introduction

Land cover changes are fundamental to understanding the complex interplay between human activities and the environment [298]. Locating and quantifying this process is essential for several UN Sustainable Development Goals (SDG) [54, 224], predicting and combating climate change [278], and preserving the diversity of life on earth [234]. For this, we currently rely on two main techniques: model-based mapping (pixel- or polygon based predictions) and design-based area estimation for a given region.

Land cover maps enable visualization and analysis of spatial patterns, allowing the identification of drivers of change [257], quantifying carbon emissions [6], and targeted land management [288]. Design-based land area estimates, on the other hand, provide statistically-robust, model-free insights and estimates (typically with confidence intervals) of long-term trends and comparisons necessary for resource allocation, economic assessments, and international accountability [80, 187]. As a result, policy and decision-makers often continue to rely on design-based area estimates, while there also is interest in maps that match the statistical area estimates. Due to the costs involved in performing such sampling surveys, there has long been a large interest in deriving area estimates directly from Earth Observation [79]. Counting the number of pixels per class in the mapped area is the simplest method, but strongly discouraged due to unpredictable biases that may lead to over- and under-prediction of specific classes [79, 187, 295]. These area biases stem from many sources such as imbalanced training data [100, 169, 320], the interaction between spatial resolution and pixel heterogeneity [107, 252], regional accuracy differences [57, 294, 301], and classifier design [53, 90, 294], and are therefore difficult to quantify in order to assess the uncertainty of predictions.

Unbiased area estimation based on pixel counting is possible using the confusion matrix computed from additional statistical reference data that are e.g. stratified according to the mapped classes. The map-based area estimates are then adjusted using commission and omission errors from according to the confusion matrix [187, 188, 249, 250]. While it is not always feasible to obtain additional samples directly from the mapped area, many organizations still adhere to this approach as their standard practice. This is primarily because methods that rely on obtaining (additional) samples directly from the field tend to yield more precise area estimates compared to other methods [3, 62, 72, 80]. Recent developments suggest that unbiased area estimation may be possible without requiring post-classification sampling. For instance [138] used Land Use/Cover Area Frame Survey (LUCAS) data and Copernicus High Resolution Layers (HRL) layers to show that it may be possible when the sampling design of the reference data is appropriate for the mapped phenomenon. Furthermore, Sales et al. [230] derived area estimates from probabilities predicted by a random forest that were more accurate than pixel counting in a binary classification context. Finally, the prediction-powered inference framework recently proposed by Angelopoulos et al. [3] makes it possible to derive quantity estimates with statistically valid confidence intervals that are smaller than those of purely sample-based estimates, without the need for post-classification sampling to adjust for model bias.

While these approaches show promise towards the end goal of deriving unbiased and accurate area estimates without requiring additional sampling, policy and decision makers continue to rely on sample-based area estimates such as the European Commission's LUCAS [80], and require maps that match their class proportions to enable localized interventions [187]. Linking trends from area estimates to periodic maps would also facilitate temporally and spatially explicit assessment of land change, which is crucial for evaluating the impacts of critical human activities on the environment [187, 258, 298]. A key issue is that land cover monitoring from remote sensing data has been producing accurate spatial maps or providing ”best” area estimates but rarely the focus has been on addressing both objectives together: an accurate map whose spatial distribution of classes is an exact match to those from a provided, trusted area estimate. This would effectively minimize both allocation and quantity disagreement [205]; pixel-wise classification errors and map-wide class quantities, respectively. Although this issue has been subject to study since the early years of remote sensing [253], little research on related approaches has been presented so far. Janssen and Middelkoop [128] showed that using ancillary data about class area proportions can be used to improve classification accuracy, especially when there is uncertainty caused by 'mixed pixels' or difficulty separating classes in the feature space. Mingguo et al. [173] further analyzed such methods and showed that using prior probabilities to adjust classification thresholds can be used to change the balance of user's and producers's accuracy (precision and recall, respectively), but that this can cause small classes to disappear from the classified map. There have been few attempts to use ancillary data to go beyond improving per-pixel accuracy and actually making maps that match area estimates. A notable exception is the work done by Tröltzsch et al. [269] and Brus et al. [25], who made 1 km within-pixel tree species proportion maps of Europe, and used an iterative scaling and calibration technique to make them correspond to national forest statistics.

More recently, Horvath et al. [113] transformed predicted probability surfaces for vegetation types [114] into classified maps whose class distributions matched estimates derived from area frame survey data [26], iterating over each species and assigning pixels on the map to that species until the expected prevalence was reached. However, they found that while their proposed methods produced maps with correct area proportions, these maps were less accurate than a map where each pixel was assigned to the class with the highest predicted probability, which suggests a trade-off between allocation and quantity disagreement. Furthermore, their proposed method left approximately 10% of the map unclassified because not every pixel had probabilities for every class. As soon as a pixel is assigned to one class, it can no longer be assigned to another; if pixels with probabilities above 0 for a certain class are rare, this leaves gaps in the classified map that must be filled with other methods. We hypothesize that the amount of remaining unclassified pixels can be reduced by using 'smoother' input probability data. We consider such data 'smooth' when for each class, there are more pixels with a predicted probability value above zero than the number of required pixels on the final map. This might be achieved by improving a model's ability to generalize such as using a bigger training dataset. This is often unfeasible in land cover classification due to the cost involved in collecting reference data, especially when area proportions must be correctly represented to obtain matching proportional predictions [138, 230]. Horvath et al. [113] state that the accuracy of their proportional maps might be improved by using more accurately predicted input probabilities, and models trained on bigger training datasets tend to be more accurate [181, 222]. While there are general recommendations for training dataset size [76, 140] and indications that tree-based methods can attain high accuracy without large training datasets [214] it is important to represent as much of the feature space as possible [170, 292]. Furthermore, **Witjes** et al. [301, 302] showed that models trained on larger portions of CORINE-derived training data generalized better on unseen data, especially when mapping land cover in years not covered by the training dataset.

This paper proposes and demonstrates a framework that employs predicted probabilities for land cover classes to create land cover maps that are both accurate and adhere to the class distribution determined by independent area estimates. We present an expanded version of the approach suggested by Horvath et al. [113] that iterates over each class multiple times to minimize the overlap between classes: Iterative Mapping of Probabilities (IMP). We investigate the effectiveness and potential advantages of the proposed approach by answering the following research questions:

1. How does the quantity disagreement vary between proportional maps produced by IMP compare to that of highest likelihood class maps?

2. How does the allocation disagreement vary between proportional maps and highest likelihood class maps?

3. What is the impact of using machine learning models trained on national or larger area training datasets for producing proportional maps?

We will answer these questions by creating proportional and highest likelihood class maps for five European countries. We will measure quantity disagreement by comparing the proportion of predicted classes to EuroStat area estimates, and allocation disagreement by validating the maps LUCAS land cover samples and S2GLC validation points. To assess the impact of using larger but less proportional training datasets, we do this twice: Once with a model trained on land cover data from the mapped country, and once with a model trained on a data from a group of countries.

4.2 Materials and Methods

In this study, we classify land cover and produce maps that match area estimates by Eurostat across five neighboring European countries: Belgium, Czechia, Germany, Luxembourg, and The Netherlands. We utilized CORINE training points produced by [301], from which we removed potential labeling errors using data from Copernicus HRL and OpenStreetMap. CatBoost classification models were then trained on these filtered points, with distinct models trained on subsets of the CORINE points from each country (referred to as *local models*) and one *general model* trained on CORINE points from all countries that adopted the LUCAS survey in 2006: Belgium, Czechia, France, Germany, Hungary, Italy, Luxembourg, The Netherlands, Poland, and Slovakia. For each of the five mapped countries, we predicted LUCAS land cover probabilities with the country's local model, as well as with the general model for the years 2009, 2012, 2015, 2017, and 2018. Each set of probabilities was used to create two hard-class land cover maps: One *highest likelihood class map* created through highest likelihood classification, and one *proportional class map*, created by the proposed algorithm in an attempt to match land cover class quantities with official LUCAS estimates. We validated our annual maps of 2009, 2012, 2015 and 2018 with the LUCAS points compiled by [49]. The maps for 2017 were validated with a separate dataset that was explicitly created to validate the S2GLC land cover map of 2017 [129]. We used this evaluation to compare the classification performance of local and general models, as well as highest likelihood class maps and proportional maps. An overview of our methodology is provided in Fig. 4.1.

4.2.1 Training and reference data

We used three different land cover point datasets for this study: 1) centroids of CORINE land cover polygons from 2006, 2012 and 2018 extracted by **Witjes** et al. [301] (available on Zenodo [142]), 2) in-situ LUCAS samples of harmonized by d'Andrimont et al. [49], and 3) validation points of S2GLC land cover maps [162] produced by Jenerowicz et al. [129] (see Fig. 4.2). All classifiers were trained using CORINE centroids as reference, and all produced maps were validated on S2GLC and LUCAS samples. The strict sampling design and high accuracy of the LUCAS survey has made it a valuable resource in training and validating land cover models across Europe [11, 199, 245, 289, 301], while S2GLC was specifically designed to validate land cover maps of Europe, using a stratified random sampling design to ensure proportional coverage of all European countries equal or larger in size than Luxembourg (see section 2.4 of Malinowski et al. [162]).

Legend harmonization and filtering

We reclassified all CORINE-derived and S2GLC points to eight LUCAS land cover classes (level-1). In the case of CORINE points, we removed any points that belong to CORINE

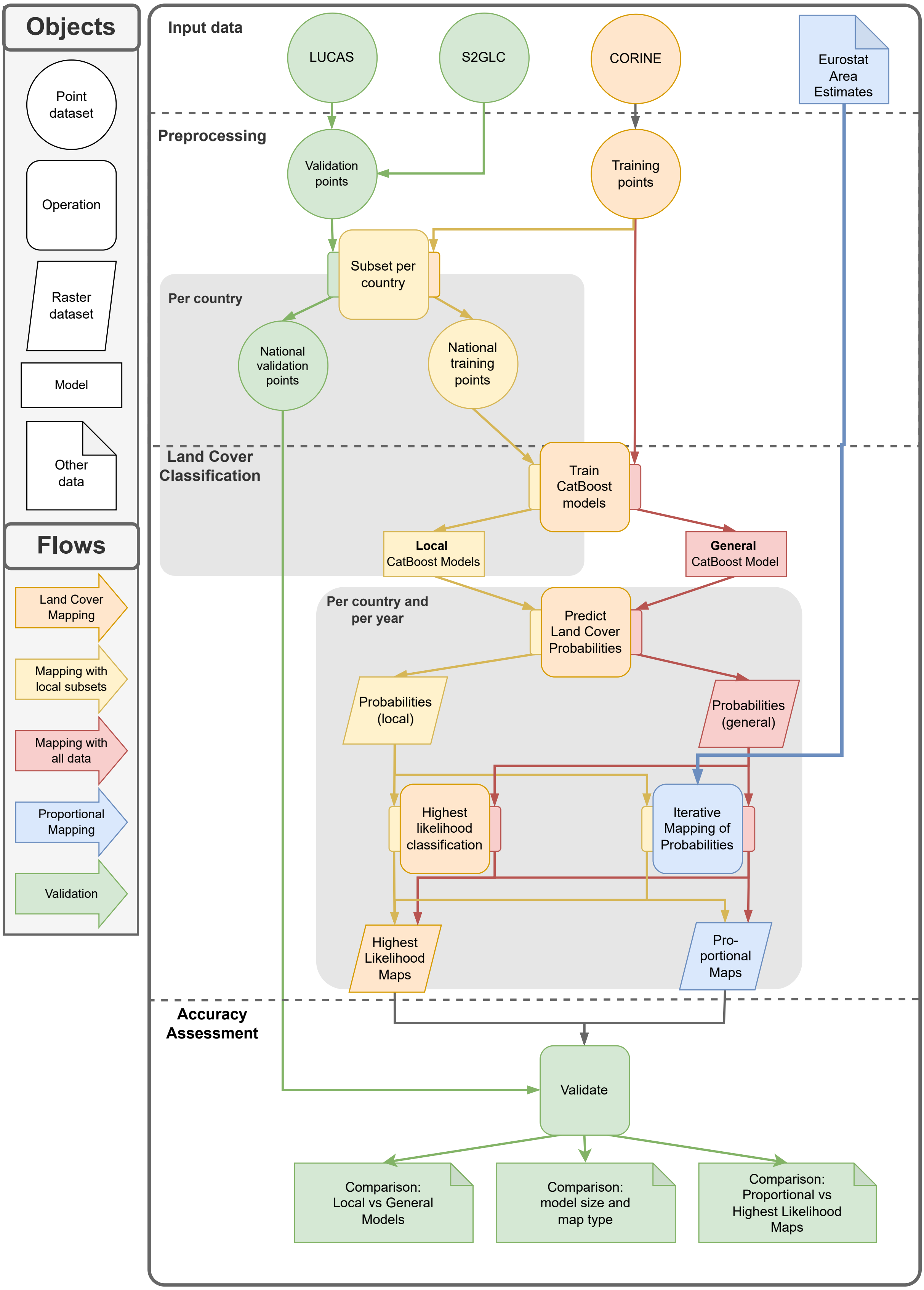


Figure 4.1: Overview of the methodology to produce, validate, and compare highest likelihood and proportional maps. More detailed information on the production of the training data and features can be found in [301] and [302].

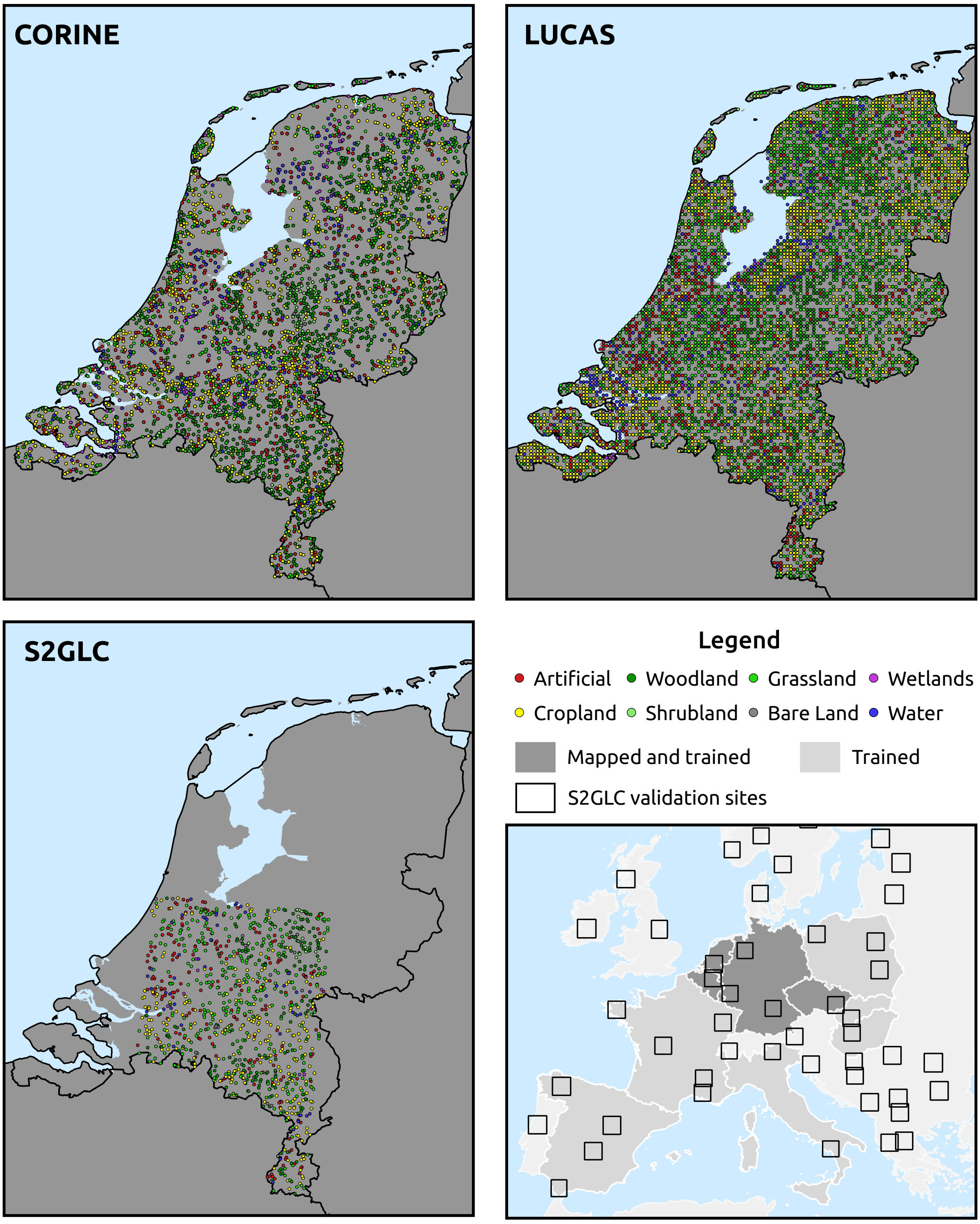


Figure 4.2: Example of the distribution of CORINE training data [301], LUCAS validation data [49], and S2GLC validation data [129], each subset to the Netherlands to visualise their spatial distribution. Bottom right: An overview of countries and S2GLC validation sites surrounding the area of interest. Countries that were mapped and from which CORINE training data was extracted (”*Mapped and trained*”) are marked in dark gray, while countries from which additional CORINE points were extracted for the general model (*”Trained”*)are marked in light gray.

classes with no clear and exclusive match to a LUCAS class. This process was conducted according to the key shown in table 4.1

Considering that the minimum mapping unit of CORINE is 25 hectares and the minimal width of mapped features 100m, CORINE polygons can encompass smaller-scale land cover types that differ from the main category of the polygon, introducing a risk of labeling errors. We counteracted this by screening the CORINE-derived points and removing all points whose land cover class were inconsistent with data from Copernicus HRL layers and OpenStreetMap in a similar way as the one detailed in **Witjes** et al. [301]. For example, grassland training points were removed if Copernicus HRL layers indicated tree cover, or if OpenStreetMap rasters indicated roads or buildings were present at the points' coordinates. 4.7 provides an overview of the data and conditions used to remove potentially faulty training points for each class.

Table 4.1: Reclassification key of CORINE and S2GLC land cover codes to LUCAS level 1 land cover. CORINE centroids of classes in the *Not Used* category were removed from the training set.

LUCAS land cover	CORINE codes	S2GLC codes
Artificial	111, 112, 121, 122, 132, 133	111
Cropland	211, 212, 213, 221, 222, 223, 241	211, 221
Woodland	311, 312, 313	311, 312
Shrubland	322, 323, 324	322
Grassland	231, 321	231
Bare land	331, 332, 333, 334, 335	331, 335
Wetlands	411, 412, 421, 422, 423	411, 412
Water	511, 512, 521, 522, 523	511
Not used	123, 124, 131, 141, 142, 242, 243, 244	

Feature space

All training points were overlaid on 224 covariates: Landsat data, derived spectral indices, a digital terrain model, and monthly minimum and maximum geometric temperature.

The Landsat data were originally published by Potapov et al. [206], aggregated to seasonal composites and gap-filled with a temporal moving window median (TMWM) algorithm by **Witjes** et al. [302], and are openly available for download on `stac.ecodatacube.eu`. From the original bands (Blue, Green, Red, NIR, SWIR1, SWIR2, Thermal), several spectral indices were calculated:

1. Normalized Difference Vegetation Index (NDVI) [226],
2. Soil Adjusted Vegetation Index (SAVI) [119],

3. Modified Soil Adjusted Vegetation Index (MSAVI) [**qi1994modified**],
4. Normalized Difference Water Index (NDWI) [168]
5. Normalized Difference Moisture Index(NDMI) [82],
6. Normalized Burn Ratio (NBR) [85],
7. Normalized Burn Ratio Plus (NBR+) [1],
8. Road Extraction Index (REI) [238],
9. Enhanced Vegetation Index (EVI) [150]

For each Landsat band and spectral index, the highest 25th percentile, median, and 75th percentile was included for each of the 4 seasons typical in Central Europe, resulting in 12 covariates for each of 7 bands and 8 indices, amounting to 192 Landsat-derived covariates.

The digital terrain model was originally published by Hengl et al. [101]. We used 8 derived variables:

1. Slope percent
2. Elevation (Lowest mode)
3. Northness
4. Easterness
5. Positive openness [310]
6. Negative openness [310]
7. Multidirectional hillshade [163]
8. 315 degree sun azimuth hillshade [163]

The minimum and maximum geometric temperature is a geometric transformation of latitude and the day of the year [137]. Aggregated to monthly averages, this amounts to 24 covariates.

Area estimates

The area estimates used as input for the proposed algorithm, and to validate the quantity disagreement of all produced maps, were obtained from Eurostat [63]. This database reports how much of each LUCAS land cover class covers each country in each year that the LUCAS survey was performed: 2006, 2009, 2012, 2015, and 2018. We derived proxy area estimates for 2017 through linear interpolation of the area estimates of 2015 and 2018.

4.2.2 Machine learning

We trained CatBoost classifiers on the filtered and overlaid CORINE-derived points. CatBoost is an implementation of gradient boosting [211] that has seen much use in recent years due to its ability to achieve relatively high accuracy on large datasets in several fields [95], notably being used to produce ESA WorldCover [313] and WorldCereal [284].

We trained six models in total: five *local* models trained exclusively on CORINE training data from each country, and one *general* model trained not only on data from the five countries, but also on CORINE data from all countries that participated in the LUCAS program in 2006. Each model was trained on CORINE data from all available years: 2006, 2012, and 2018. Table 4.2 shows how many training points were used for the local models of each country, with the column *Other* representing the countries from which training data was extracted, but which were not mapped by a local model (see also Fig. 4.2). We included this general model in our analysis to investigate whether a larger feature space compensates for a less balanced class distribution.

Table 4.2: Summary of CORINE points per country and LUCAS level 1 class, used to train the land cover models in this work. Local models were only trained on the available points for that country, while the general model was trained on all points, including those from other countries.

Lucas class	Belgium	Czechia	Germany	Luxembourg	Netherlands	Other	Total
Woodland	13,661	46,555	202,657	1,575	4,934	683,623	953,005
Cropland	12,637	25,506	112,383	1,104	3,608	512,865	668,103
Grassland	10,176	21,923	69,541	689	2,647	255,673	354,649
Artificial	2,028	3,286	19,274	170	2,046	65,012	90,816
Shrubland	494	1,005	2,932	11	614	144,393	150,449
Water	451	1,994	8,415	20	1,222	25,085	37,187
Wetlands	50	272	2,903	3	584	8,385	12,197
Bare land	41	18	833	0	152	31,431	32,475
Total	39,538	100,559	418,938	3,572	15,807	1,726,467	2,250,881

The training points were split up into 2996 30 km tiles. We randomly selected 5% of these tiles as validation data. The remaining points were used to train all models. To prevent overfitting, we validated each model after each iteration on the points from the validation tiles. Training was automatically stopped when validation accuracy had not improved for 10 consecutive epochs, and the model resulting from the epoch where the most recent validation accuracy improvement was recorded was selected as the final model.

4.2.3 Land Cover Classification

For each country, we predicted probabilities for the eight LUCAS land cover classes in the years 2009, 2012, 2015, 2017 and 2018. We did this once with that country's local model, and once with general model. For each set of predicted probabilities, we created a *highest probability class map* (HPC) by assigning each pixel to the class having the highest class probability. This resulted in ten highest probability class maps per country (five years,times two models).

4.2.4 Iterative Mapping of Probabilities

We implemented IMP, a post-processing algorithm, on each set of predicted probabilities (year, country, model). IMP is designed to create the most accurate possible hard-class map with 1) a given set of predicted probabilities and 2) an existing area estimate, based on the following assumptions:

1. **Bias between classes** Models can have unknown biases and may overpredict certain classes by assigning relatively higher probabilities for them on average, which leads to rarer classes being underrepresented by highest likelihood classification [100, 295].
2. **Ranking within classes** Within each class, the pixels with higher predicted values are more likely to correspond with actual occurrence of that class in a given pixel, regardless of the predicted probability values for other classes in the same pixel. Essentially, even if all probabilities for a single class are relatively low, they are at least roughly ranked in the correct order of likelihood for a given class. This is not always guaranteed [184], but can generally can be expected from accurate classifiers.
3. **Overlap between classes:** When selecting pixels based on the level of their within-class relative probability, some pixels may be the best candidates for multiple classes (i.e. being in the top percentile of probabilities), either due to model bias, or due to multiple classes actually occurring inside the same pixel [113].

In general, IMP functions similar to the method proposed by Horvath et al. [113]: It loops over every class, selecting the pixels with the highest predicted probability for that class and assigns the corresponding pixels on the output map to that class. However, to minimize the overlap problem, IMP does not do this once, but several times, each time selecting only the top percentile of available pixels for each class. We set the number of iterations to 20 in our presented experiments. This means that at every iteration, IMP selects the best 5% of the target proportion from the best available pixels for each class. For example, a class which was estimated to cover 20% of a country's surface, at the first iteration, only the pixels with 0.2% of that class' highest predicted probabilities will be assigned to that class. At the second iteration, it would select and the pixels that are within 0.4% of that range, but some of those pixels will have been assigned to other classes. Instead, it will select the pixels with the top 0.2% highest probabilities for that class *amongst the remaining unassigned pixels in the output map.* Figure 4.3 presents a visualisation of how IMP gradually fills a map until the proportions of each class matches those in the target area estimate. A detailed description of IMP is provided in 4.7.

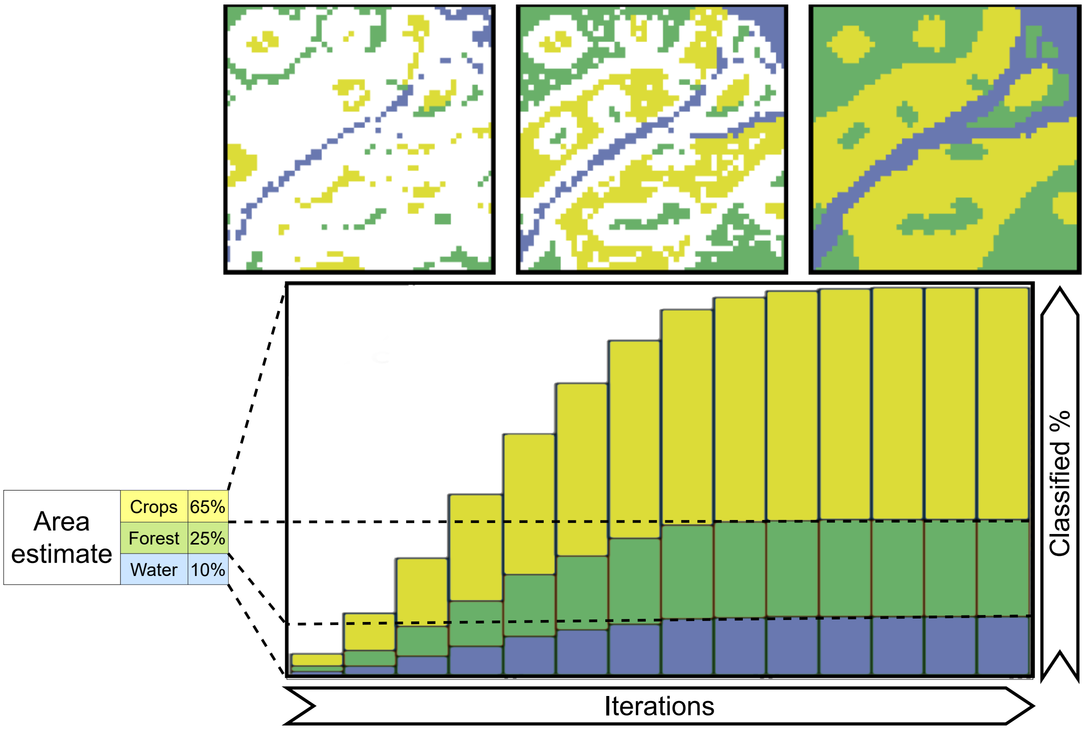


Figure 4.3: Visualisation of how IMP gradually classifies each pixel in the study area, selecting the pixels with the highest available probabilities for each class in each iteration.

4.2.5 Accuracy Assessment

We assess the allocation and quantity disagreement [205] of the produced models and maps by validating the maps of 2009, 2012, 2015 and 2018 with LUCAS points [49] and the maps of 2017 with S2GLC points [129]. Table 4.3 shows the support per class per country for LUCAS points, and table 4.4 for the S2GLC points. We quantify allocation disagreement with the Weighted F1-score metric [282], a harmonic mean of user's and producer's accuracy (precision and recall), because it distinguishes classification performance more strictly on datasets with imbalanced class distributions. This makes it a useful metric to compare the classification performance when both accuracy and proportion of each class are important. This is done both for the *highest likelihood* class maps (based on the highest probability per point or pixel) and the *proportional* class maps produced by IMP. We measure quantity disagreement with the percentage that class-wise proportions of each map deviate from the area assessment of the target country in the target year.

Table 4.3: Summary of LUCAS points per country and LUCAS level 1 class, used to validate the land cover maps of 2009, 2012, 2015, and 2018

Lucas class	Belgium	Czechia	Germany	Luxembourg	Netherlands	Total
Artificial	1,365	1,120	8,911	83	1,742	13,221
Cropland	4,006	10,487	47,926	285	4,147	66,851
Woodland	2,902	8,552	34,813	347	1,845	48,459
Shrubland	139	218	1,042	15	271	1,685
Grassland	4,332	6,199	30,540	402	6,014	47,487
Bare land	236	306	1,332	13	286	2,173
Wetlands	150	283	1,786	8	720	2,947
Water	42	62	542	0	101	747
Total	13,172	27,227	126,892	1,153	15,126	183,570

Table 4.4: Summary of S2GLC points per country and LUCAS level 1 class, used to validate the land cover maps of 2017

Lucas class	Belgium	Czechia	Germany	Luxembourg	Netherlands	Total
Artificial	115	39	183	8	193	538
Cropland	325	439	1,211	16	246	2,237
Woodland	225	258	1,039	23	202	1,747
Shrubland	2	1	8	0	18	29
Grassland	200	60	459	17	329	1,065
Bare land	15	14	42	0	21	92
Wetlands	12	6	10	0	49	77
Water	2	0	23	0	7	32
Total	896	817	2,975	64	1,065	5,817

4.3 Results

4.3.1 Training data preprocessing

The LUCAS/CORINE training dataset used by [301] contained 3,381,460 CORINE centroids with CORINE classes that were compatible to the LUCAS level 1 legend. Filtering the CORINE centroids with Copernicus HRL and OpenStreetMap data removed 1,076,579 points, or 31.84 percent of the data. 4.7 presents a full overview of the amount of training points removed by each filtering rule.

4.3.2 Land cover classification and Proportional Post-Processing

Predicting eight land cover classes for five years with two models (a country-specific local model and a common general model) resulted in 80 predicted probability layers per country. Creating hard-class maps with highest likelihood classification yielded one map with eight classes for each year and each country, resulting in 25 hard-class maps.

Applying the algorithm in 20 iterations on the probabilities predicted by a country-specific local model and the general model produced an equal number of proportional class maps each country, year, and model type. Fig. 4.4 shows the proportional land cover map of the entire study area for 2009, based on probabilities predicted by the LUCAS mode. Fig. 4.5 shows an example of the iterative classification process resulting in a proportional land cover map, as well as a comparison with the highest likelihood map and the iteration at which each pixel was filled. Note that pixels that were classified at later iterations tend to also be marked as differences with the highest likelihood map, and that there are distinct spatial patterns in their occurrence: For instance, the edges of the grassland patches in the northern part of the area, the suburban part in the southern area, and the urban green areas in the city proper.

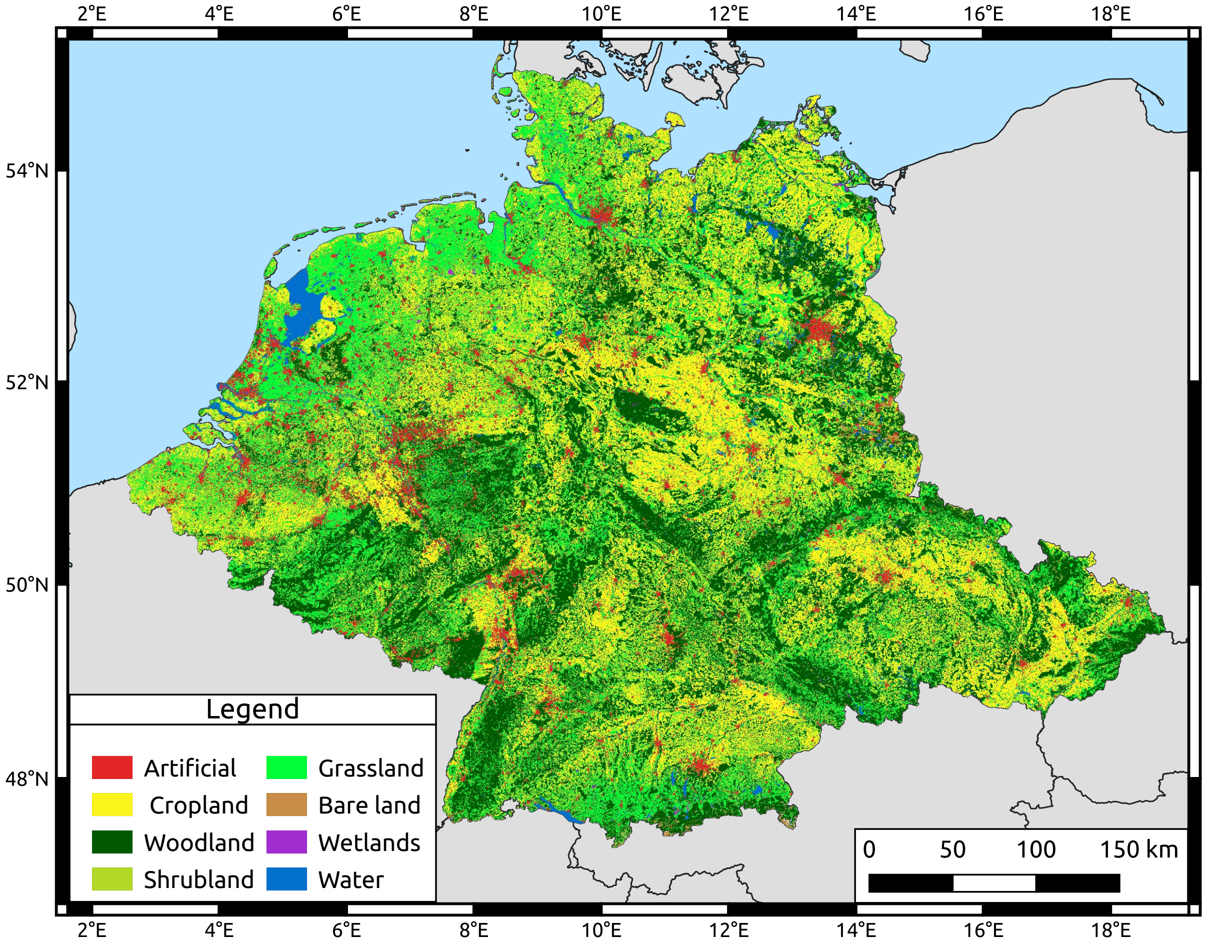


Figure 4.4: LUCAS Land cover maps of Belgium, Czechia, Germany, Luxembourg and the Netherlands of 2009 generated with the proposed algorithm, using Eurostat national area estimates and probabilities predicted by the general CatBoost model trained on CORINE centroid points from all countries that implemented LUCAS in 2006.

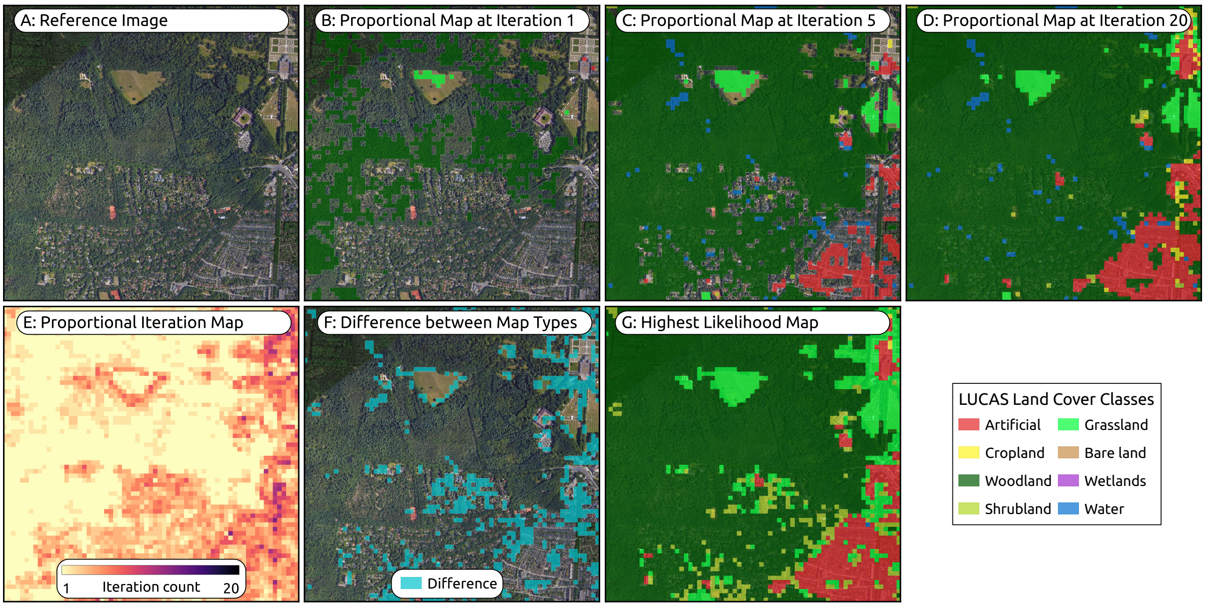

A: Reference Image
B: Proportional Map at Iteration 1
C: Proportional Map at Iteration 5
D: Proportional Map at Iteration 20
E: Proportional Iteration Map
1
Iteration count
20
F: Difference between Map Types
Difference
G: Highest Likelihood Map
LUCAS Land Cover Classes
Artificial
Grassland
Cropland
Bare land
Woodland
Wetlands
Shrubland
Water

Figure 4.5: Example of the IMP iterative classification process and comparison to highest likelihood classification, using the northwestern outskirts of Apeldoorn, the Netherlands. A: high-resolution reference imagery from Google Earth; B-D: classifications over subsequent iterations, identified in E; F: differences between the proportional map and the highest likelihood map; G: the highest likelihood map itself.

A complete overview of the area estimates by Eurostat and mapped area per country, year, class, model type, and map type is included in 4.7. Almost all proportional maps had pixel proportions that fell within the confidence intervals of the Eurostat-derived area estimate. Only some maps derived from probabilities predicted by local models had class proportions that were outside the confidence intervals of the Eurostat area estimates. The only case of underrepresentation was *Bare Land* in the maps derived from probabilities predicted by the local model for Luxembourg, as this class was not represented in the training data. Also in Luxembourg, Woodland was overrepresented by the local model by an average of 0.03 percent above the upper bound of the Eurostat-supplied confidence interval. Both issues are likely due to the fact that Luxembourg's local model was trained on a relatively small dataset compared to the other models, as no under- or overrepresentation occurred in proportional maps derived from probabilities produced by the general model.

The *Artificial* class was overrepresented in the 2012 map of Germany by 0.04 percent of the upper bound, and finally, the *Shrubland* class was overrepresented in the 2009 map of Czechia by 11 percent. This is the biggest percentual error in all proportional maps, and constitutes a representation of 0.87 percent of Czechia's surface instead of the 0.69 percent estimated by Eurostat, with an upper confidence bound of 0.78 percent. This relatively large overrepresentation was caused after the final iteration, where remaining gaps in the map were filled with the highest likelihood class. These gaps were assigned to *Grassland* and *Shrubland*, for which more pixels were available with higher probabilities. Fig. 4.6

shows how the iterative classification, or ’filling’ of the 2009 map of Czechia proceeded. The general behavior of each class is similar to those observed in other countries and years, but not every class approaches the area estimate at the same rate, with *Woodland* reaching the confidence interval at iteration 8, and *Grassland* at iteration 19. We observed a similar pattern in other maps: More gaps remained at the final iteration when there was a relatively small surplus of pixels with predicted probabilities for each class. Generally, classes for which there was a low number of pixels with any probability compared to their proportion according to the Eurostat estimate were classified less accurately, and in proportions that less closely matched the Eurostat estimates. Predictions by the general model tended to have probabilities for more different classes in each pixel, especially for rarer classes. Fig. 4.7 shows this for predictions for the Netherlands in 2009. Both models predicted probabilities above zero for *Cropland* and *Grassland* on a large number of pixels compared to their prevalence as estimated by Eurostat, but the local model barely predicted enough pixels for *Bare Land*, and an insufficient amount for *Water*. While the general model predicted a smaller surplus of pixels with probability above zero for Artificial and Woodland, it predicted surpluses for each class. The comparison of Eurostat estimates and counts based on classification in 4.7 show that proportional maps based on probabilities predicted by the general model matched the Eurostat estimates more closely than those based on probabilities predicted by the local model.

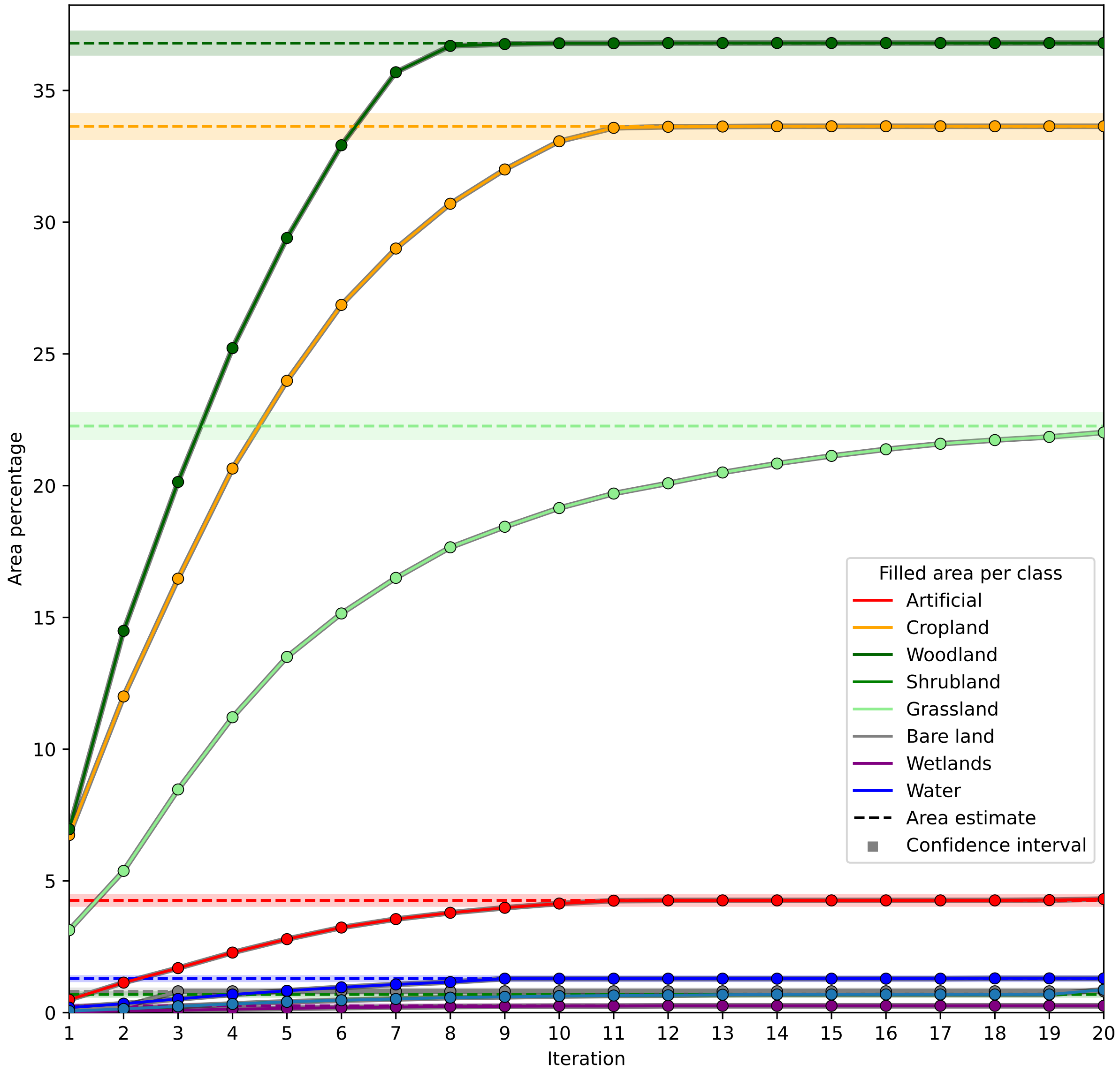
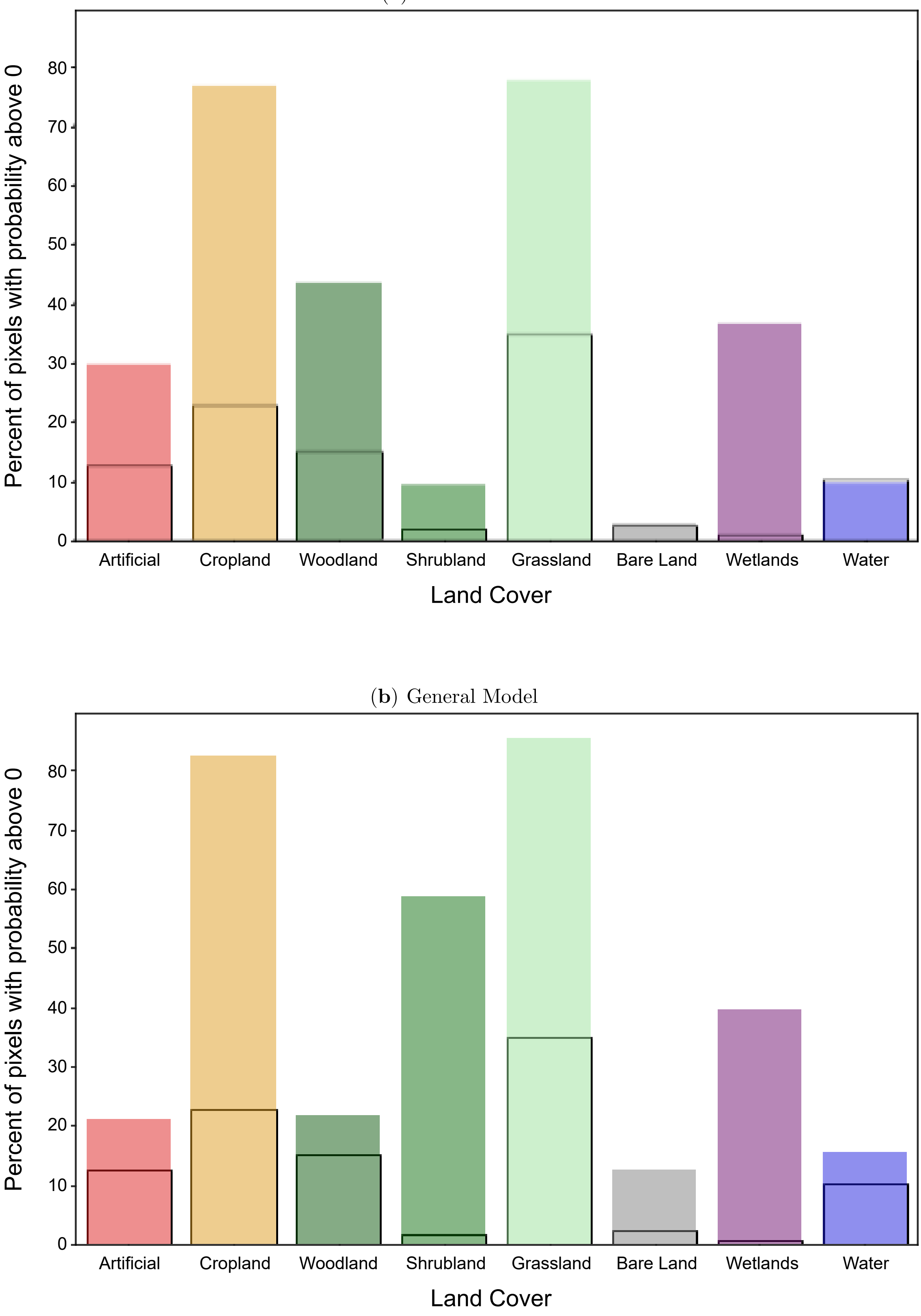


Figure 4.6: Percentage of map area filled per iteration of the IMP algorithm, for the map of Czechia for 2009. Each line with dots indicates the percentage of area assigned to each LUCAS land cover class at each iteration. The dashed lines indicate the mean area estimated by Eurostat, with the lighter-colored area around each dashed line indicating the accompanying confidence interval.



(**b**) General Model

Figure 4.7: Percentage of pixels for which any probability above zero was predicted per class, compared to the proportion of that class according to the Eurostat area estimate used to create proportional maps of the Netherlands in 2009. Note that the general model predicted a surplus of probabilities above zero for each class, while the local model did not predict probabilities above zero in enough pixels for *Water*.

4.3.3 Accuracy assessment

F1-scores calculated by overlaying the highest likelihood and proportional maps of 2009, 2012, 2015 and 2018 on LUCAS points of matching years and the maps of 2017 on S2GLC points show (see Fig. 4.8) the accuracy of proportional maps tended to be higher or equal to the accuracy of highest likelihood maps. The difference between highest likelihood maps and proportional maps created from the same predicted probabilities is more pronounced on maps created by the general model, which was trained on the entire training dataset. On average across all mapped years (see table 4.5, proportional maps were more accurate than highest likelihood maps in most cases. Proportional maps consistently achieved a weighted F1-score above 0.7 on LUCAS points, and above 0.85 on S2GLC points. Fig. 4.9 shows that the highest likelihood maps generally had higher precision (User's accuracy) than proportional maps, while Fig. 4.10 shows that proportional maps generally had higher recall (Producer's accuracy). This sacrifice of precision for gains in recall can explain the noted increase in F1-score. Note that we use the *weighted* precision and recall metrics to give equal importance to the performance of every class.

As shown in Fig. 4.11, the proportional maps had a lower quantity disagreement than the maximum likelihood maps, reducing it to near-zero in most cases. It also shows that maximum likelihood maps based on probabilities predicted by local models tended to be more accurate than those based on probabilities predicted by the general model.

Table 4.5: Comparison of F1-scores for different countries, validation datasets, and model types. The table presents the F1-scores for both highest likelihood and proportional maps across five countries, using two validation datasets (S2GLC and LUCAS) and two model types (local and general). The "Difference" column quantifies the difference in F1-scores between the highest likelihood and proportional maps. Note that the LUCAS values are averages of the 4 years that were mapped and validated: 2009, 2012, 2015 and 2018.

Country	Dataset	Model Type	F1-score		
			Proportional	Highest Likelihood	Difference
NL	S2GLC	local	0.85	0.86	-0.01
		general	0.86	0.80	0.06
	LUCAS	local	0.71	0.72	-0.01
		general	0.73	0.67	0.06
LU	S2GLC	local	0.95	0.98	-0.03
		general	0.97	0.91	0.06
	LUCAS	local	0.75	0.73	0.02
		general	0.73	0.68	0.05
DE	S2GLC	local	0.94	0.93	0.01
		general	0.93	0.83	0.10
	LUCAS	local	0.77	0.77	0.00
		general	0.76	0.70	0.06
BE	S2GLC	local	0.90	0.87	0.03
		general	0.92	0.87	0.05
	LUCAS	local	0.71	0.71	0.00
		general	0.72	0.70	0.02
CZ	S2GLC	local	0.94	0.91	0.03
		general	0.96	0.88	0.08
	LUCAS	local	0.78	0.78	0.00
		general	0.78	0.73	0.05

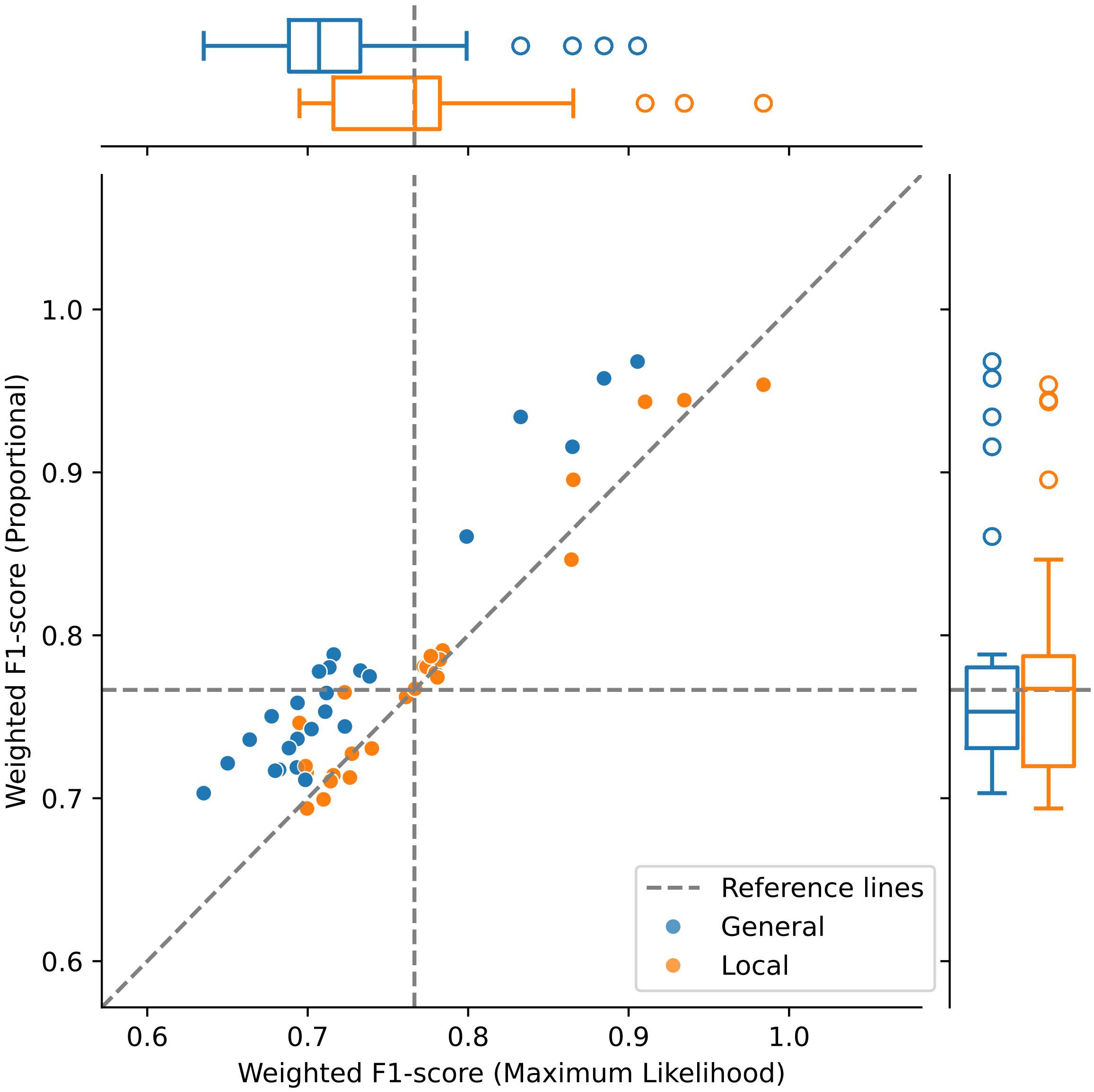


Figure 4.8: Weighted F1-score of the land cover maps classified with maximum likelihood (X-axis) and the iterative mapping (Y-axis), validated on LUCAS and S2GLC reference data. Maps based on probabilities predicted by the general model are shown in blue, while maps based on probabilities predicted by local models are shown in orange. The boxplots at the top and right summarize the F1-scores of maximum likelihood and proportional classification, respectively. The diagonal reference line shows where the F1-score would be equal, while the horizontal and vertical line represent the average across all F1-scores. Points above the diagonal reference line indicate maps where proportional maps were more accurate. In most cases, proportional maps had higher F1-scores than maximum likelihood maps. This difference was generally larger when using probabilities predicted by the general model, although maps based on predictions by local models were more accurate on average.

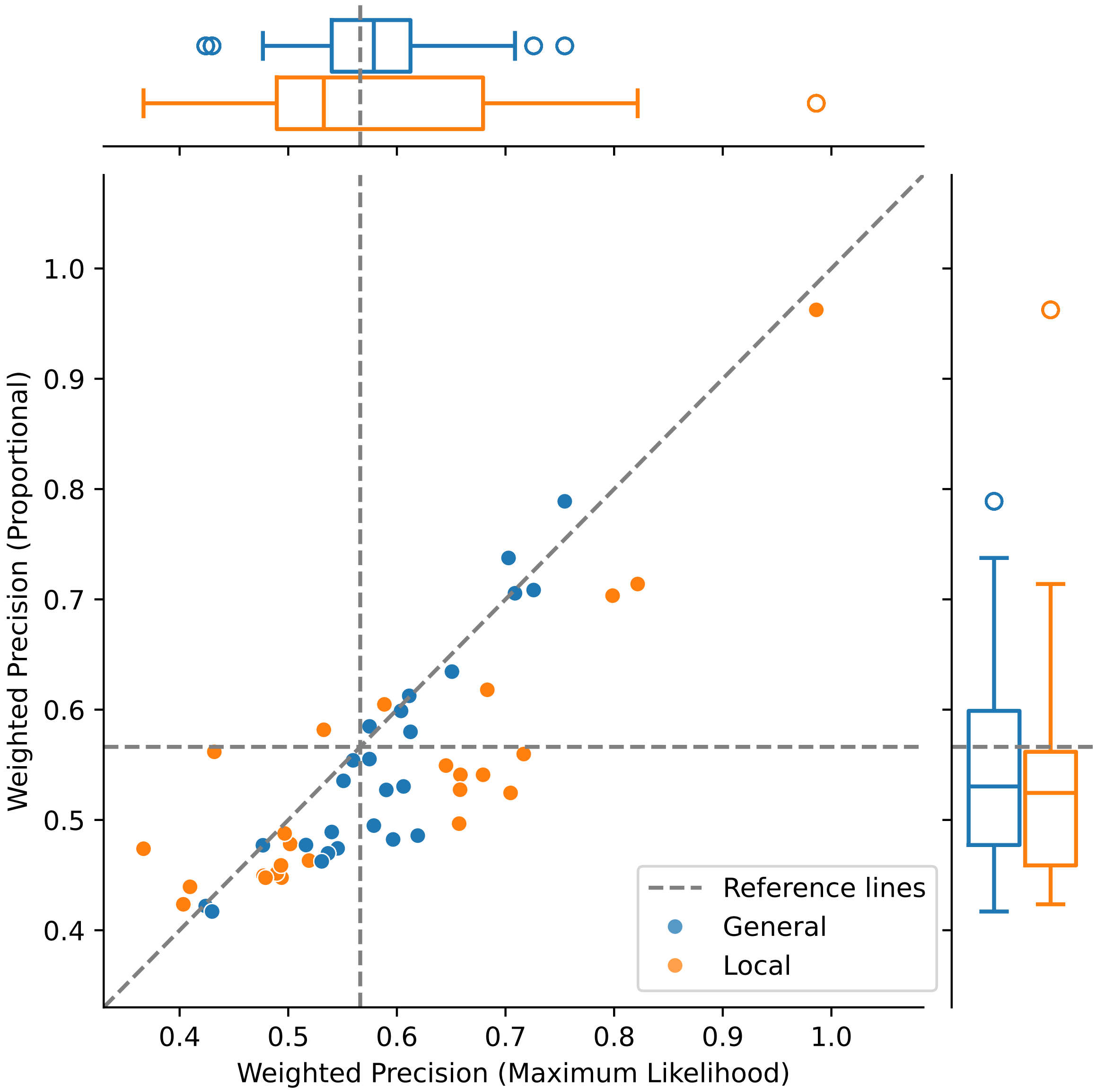


Figure 4.9: Weighted precision (User's accuracy) of the land cover maps classified with maximum likelihood (X-axis) and the iterative mapping (Y-axis), validated on LUCAS and S2GLC reference data. Maps based on probabilities predicted by the general model are shown in blue, while maps based on probabilities predicted by local models are shown in orange. The boxplots at the top and right summarize the precision of maximum likelihood and proportional classification, respectively. The diagonal reference line shows where the F1-score would be equal, while the horizontal and vertical line represent the average across all precision scores. Points above the diagonal reference line indicate maps where proportional maps were more precise. Note that proportional maps generally had a lower precision than maximum likelihood maps, and that this difference was bigger for maps based on probabilities predicted by local models.

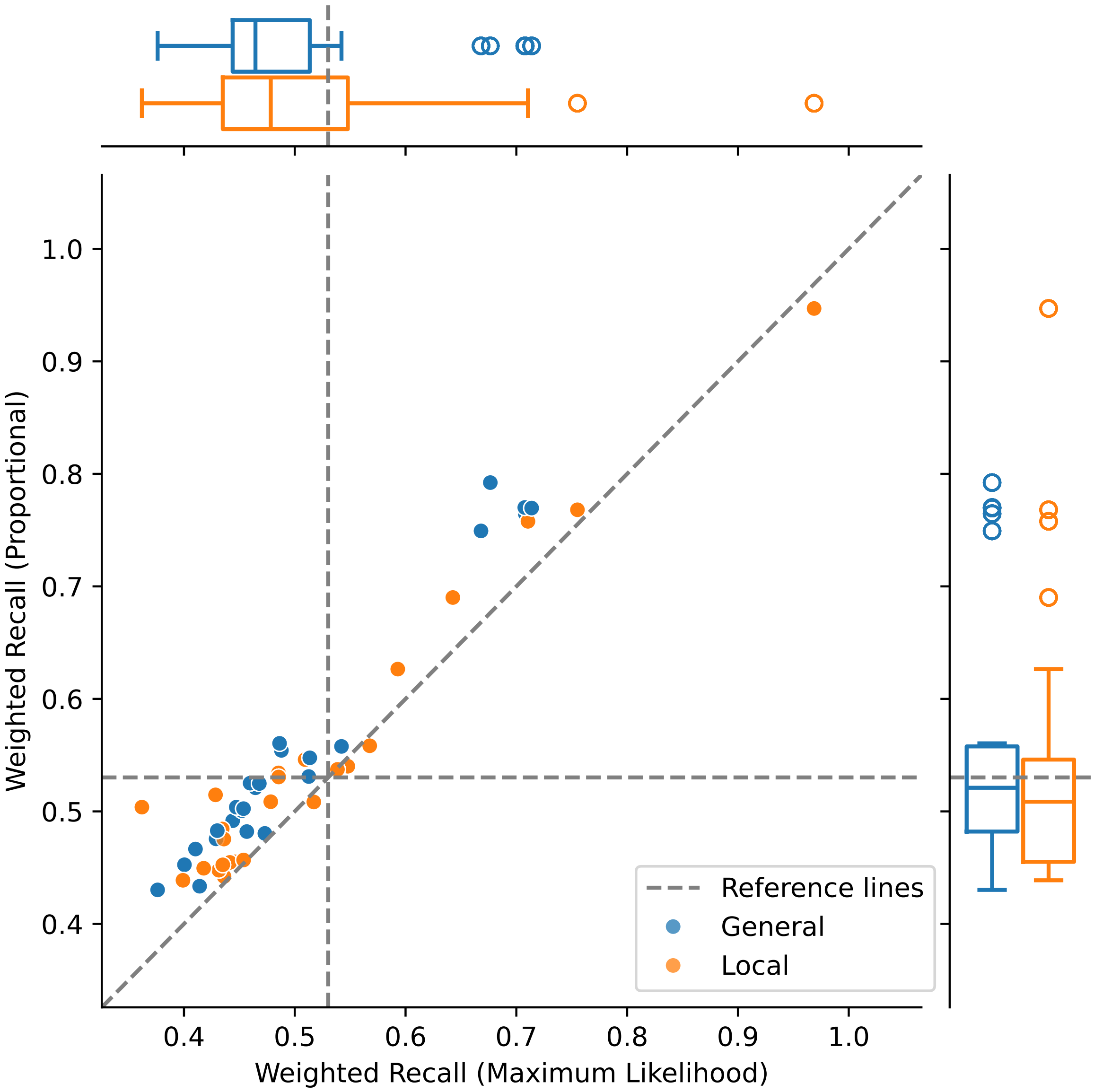


Figure 4.10: Weighted recall (Producer's accuracy) of the land cover maps classified with maximum likelihood (X-axis) and the iterative mapping (Y-axis), validated on LUCAS and S2GLC reference data. Maps based on probabilities predicted by the general model are shown in blue, while maps based on probabilities predicted by local models are shown in orange. The boxplots at the top and right summarize the recall of maximum likelihood and proportional classification, respectively. The diagonal reference line shows where the F1-score would be equal, while the horizontal and vertical line represent the average across all recall scores. Points above the diagonal reference line indicate maps where proportional maps had a higher recall. Observations include: 1) Proportional class maps generally exhibit a marginally higher F1-score, 2) The range of quantity error for proportional class maps narrow compared to highest probability class maps, and 3) Highest probability maps generated by the general model were less accurate in terms of both quantity and allocation disagreement than those generated by local models.

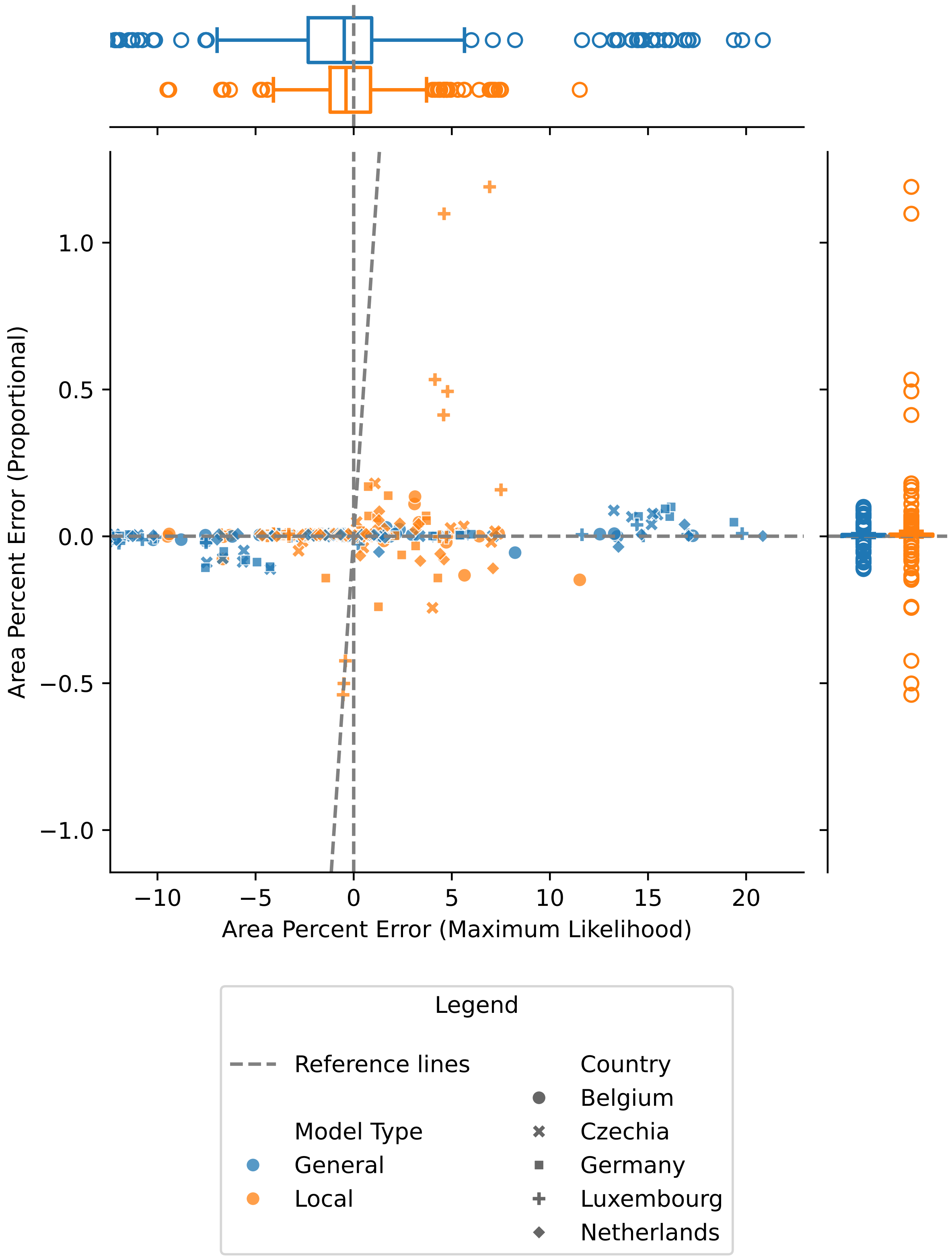


Figure 4.11: Class-wise allocation disagreement of maximum likelihood (X-axis) and proportional (Y-axis) maps based on probabilities predicted by the general model (blue) and local models (orange). Allocation disagreement is quantified as area percent error: the percent point difference between the target area estimate and the amount of pixels classified on the map. Note that proportional maps had errors close to zero, with the exception of Luxembourg (indicated with +), and that highest likelihood maps based on probabilities predicted by local models had lower quantity disagreement than those based on predictions by the general model.

4.4 Discussion

4.4.1 Algorithm design and performance

We found that IMP yielded higher accuracy compared to highest probability class assignment.

Where the use of prior probabilities to adjust classification thresholds such as summarized by Mingguo et al. [173] does not guarantee method proposed by [113] leaves considerable gaps due to overlapping 'best probability' pixels for several classes, our iterative approach with a decaying threshold minimizes this problem by classifying the best pixels for every class first. Any remaining unclassified pixels can be filled in with the highest likelihood class without severely affecting the distribution of classes. The number of iterations can be increased to improve convergence to area estimates, which might be needed if there are relatively few pixels with probabilities for one or more classes. Using models that push probability mass away from 0 and 1 like boosting [184], as was used in this work, may play a role as well.

We did not expect the proportional maps to be consistently either equally or more accurate than those created by highest likelihood classification. A possible explanation for this improvement in accuracy is that by forcing the classification to be stratified according to an accurate area estimate, we reduce the bias of the model. This is supported by the fact that the difference in accuracy between proportional and highest likelihood class maps was higher when using probabilities predicted by the general model: A general model, having been trained on a dataset that is less representative of any given country, is therefore likely to be more biased [100], which gives more space for the iterative algorithm to improve the map. This corresponds to claims by Sales et al. [230] and Kleinewillinghöfer et al. [138] that area estimates are better derived from predictions by models that were trained on datasets representative of the area of interest.

4.4.2 Limitations and potential improvements

While the proposed algorithm correctly mapped to area proportions in most attempts, this is not generally guaranteed. Experiments with different preprocessing workflows, model types, and parameters such as the number of iterations, suggested that the algorithm performs better in both quantity and allocation accuracy when, respectively, more pixels in the area of interest have a predicted probability for multiple classes (see Figs.4.6 and **??**), and when these probabilities are properly ranked within each class. We did not include these initial experiments for the sake of brevity, but this should be taken into account when applying our method to other probability predictions. Based on our omitted findings about the quantity of pixels with predicted probabilities for any given class, we expect techniques that push probability mass away from extreme values, such as label smoothing [180] might help the algorithm converge to the area estimates, as more pixels will be available with probabilities for more classes. This might not lead to more accurate maps by itself, however. It is possible that accuracy of proportional maps could be increased by using probabilities that more closely meet the assumption that probabilities are properly ranked, such as those produced by Platt Scaling [184].

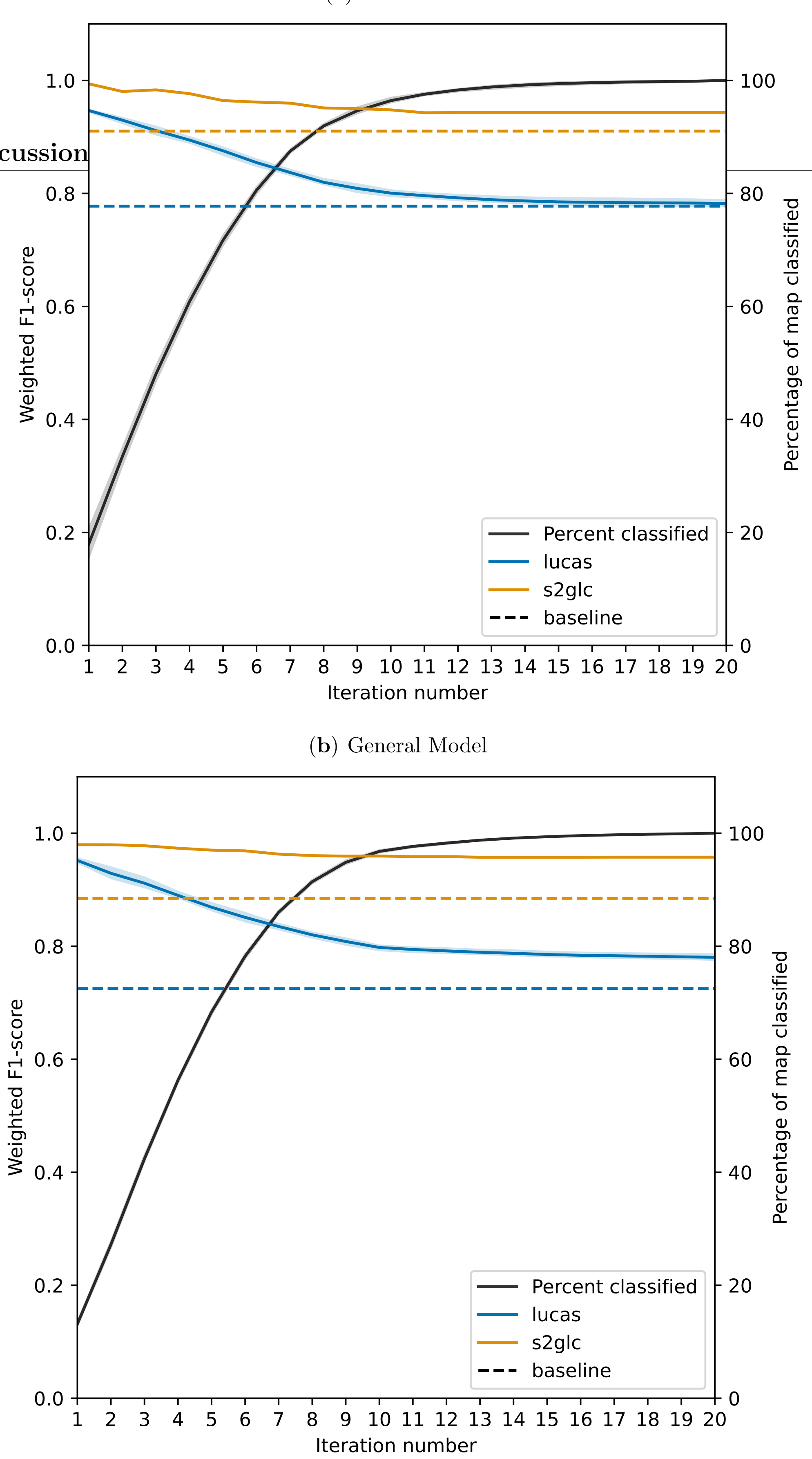
4.4.3 Potential applications

The primary purpose for which this algorithm was designed is to create, or update, land cover maps based on predicted probabilities, using a given area estimate. This area estimate can be derived from an external source, such as a sample-based measurement, like in the case of this study. However, the algorithm might also be used to update a map used to derive area estimates with post-stratification sampling [187], using the corrected area estimates and the original probabilities that were used to make the initial map. Besides this purpose, however, we suggest interested parties consider the following:

Our results show that the single, bigger *general model* was often less accurate when mapping a single area of interest than the *local models* that were trained only on data from their corresponding country. However, this difference in accuracy disappeared in proportional maps, and the general model successfully detected a rare class that was not present in the training data for Luxembourg, whereas the local model simply skipped these classes.

Additionally, the iterative nature of IMP enables the production of spatially explicit uncertainty assessments. Pixel-level uncertainty is useful because it can be propagated in spatial models that use land cover as a covariate [108]. During this study, we found that the pixels classified in earlier iterations have a higher validation allocation disagreement than those classified in later iterations (see Fig. 4.12 for an example). It is therefore possible to calculate separate precision (User's accuracy) scores per class and per iteration, leading to a more fine-grained indicator of each pixel's reliability for decision makers or use in subsequent modeling. The combination these 'iteration maps' (see Fig. 4.5 E) with the user's accuracy of each iteration needs further study.

Furthermore, it is possible that our proposed algorithm might improve the temporal consistency of land cover classification, leading to better land cover change mapping. Due to variations in satellite reflectance per mapped time period, spurious changes are often included in land cover change maps that do not implement proper post-processing techniques. While many such techniques already exist, they often rely on contiguous time series of predictions or are computationally intensive. Our proposed algorithm does not have these limitations, only requiring an up-to-date area estimate of each year, or an interpolation such as the one used to make maps of the year 2017 in this study. If further research indicates that the algorithm can reduce spurious change, it would therefore be a useful addition to the existing methods.



(**b**) General Model

Figure 4.12: Validation performance of proportional maps compared to the highest likelihood baseline. This example of the weighted F1 score of the partially filled-in map during each iteration of IMP, and a comparison with the weighted F1-score of the highest likelihood map when derived from probabilities predicted by the local (a) and general (b) model. The lines represent validations of all maps of Czechia. Note that the pixels classified at early iterations achieve a higher score, and that the score gradually decreases as a greater number of pixels is classified, stabilizing at or above baseline established by highest likelihood maps.

Lastly, during this study it became clear that the method might be suitable for more than only making maps whose proportions match area estimates. The observed improvement in accuracy of proportional maps output by IMP over highest likelihood classification suggests that forcibly improving quantity disagreement with posterior quantity estimates can improve allocation disagreement to a certain extent. This means that IMP could be used to validate area estimates themselves. When provided with 1) a validation dataset that was not used in the area estimation process and suitable for the mapped resolution (such as the S2GLC points in this study), 2) predicted probabilities of sufficient quality, and 3) multiple area estimates, the best area estimate should lead to the most accurate proportional map possible. Exploring this further was beyond the scope of this study but may prove a promising use case.

4.5 Conclusion

We present an algorithm that can transform predicted probabilities by a given machine learning model into maps whose class proportions match area estimates. Our validation on two independent test datasets (LUCAS and S2GLC points) across five years and five countries show that these maps were also equally, if not more, accurate than maps classified according to the maximum probability per pixel. In our experiments, this increase in accuracy was achieved by sacrificing some degree of precision (User's accuracy) for relatively larger gains in recall (Producer's accuracy). Because pixels classified at each iteration have different precision and recall values, the iteration at which a pixel was classified can be used to approximate pixel-wise uncertainty. IMP greatly improves the accuracy of models trained on a bigger, but imbalanced, training dataset. This means that it can be used to optimize predictions by big models to local contexts if reliable area estimates are available. This offers potential for the increasing number of global and continental-scale maps which are trained on a rich feature space and achieve high overall accuracy, but have also been criticized for their relatively low accuracy and/or usefulness at more local scales. By negating the bias from large scale models, IMP can therefore contribute to make accurate and useful maps of smaller regions. This, combined with the potential to generate pixel-based error estimates, suggests that IMP can be a valuable tool for decision makers and other stakeholders.

4.6 Data Availability

The 400 predicted probability layers, 50 highest likelihood class maps, and 50 proportional class maps, are available on zenodo at `https://zenodo.org/records/10641340`. This download is accompanied by the used area estimates, and a script that produces proportional maps.

4.7 Appendices

Training point filtering

Table 4.6: Point filtering rules per class used to remove potential label errors from the CORINE centroids before using them as training data, as well as the number of points that resulted.

LUCAS class	Condition	dataset	Source	Removed
Artificial	> 150	OSM Buildings / Cop Imp.	[301]	68,444
Artificial	< 50	OSM buildings / Cop Imp.	[301]	333,526
Non-Artificial	> 50 < 150	OSM buildings / Cop Imp.	[301]	9,594
Non-Artificial	> 30	Rasterized OSM Roads	[301]	161,576
Non-Artificial	> 30	Rasterized OSM Railroads	[301]	2,759
Woodland	< 1	Cop Tree Cover	[210]	62,731
Shrubland	> 50	Cop Tree Cover	[210]	121,464
Shrubland	> 50	Cop Grassland	[210]	48,734
Grassland	< 1	Cop Grassland	[210]	174,616
Grassland	> 30	Cop Tree Cover	[210]	48,443
Bare Land	> 1	Cop Tree Cover	[210]	16,156
Bare Land	> 1	Cop Grassland	[210]	11,052
Wetlands	< 1	Cop Temp and Perm Wetness	[210]	7,973
Water	< 50	Cop Permanent Water	[210]	9,511

IMP Algorithm in pseudocode

Algorithm 1 Iterative Mapping of Probabilities

```
Input:
• $P$, predicted probabilities for each class $C$
• $E$, the target area estimate
• $I$, the number of iterations
Output:
• $M$, the land cover map whose class proportions should match $E$.

1: Initialize the empty output map $M$
2: for iteration $i = 1$ to $I + 1$ do
    2.1 Set ratio $r_i$ to $\frac{i}{I}$
    2.2 for each $C$ do
        2.2.1 $E_C$ = The target number of pixels of $C$ in the output $M$ according to $E$
        2.2.2 $M_{Ci}$ = Number of pixels in $M$ already classified as $C$ at $i$
        2.2.3 $N_{Ci}$ = Number of pixels that must still be classified as $C$ to match $E_C$
        2.2.4 $P_C$ = All pixels in $P$ with probability for $C$ above 0
        2.2.5 $P_{Ci}$ = All unclassified pixels in $P_C$
        2.2.6 $E_{Ci}$ = Number of pixels to classify in $i = N_C \times r_i$
        2.2.7 $Q$ = the top $\frac{E_{Ci}}{P_{Ci}}$-th percentile of $P_{Ci}$
        2.2.8 $P_Q$ = The lowest probability among pixels in $Q$
        2.2.9 $N_{CQ}$ = Number of pixels in $P$ with probability values $\geq P_Q$ for $C$.
        2.2.10 if $N_{CQ} > N_{Ci}$ do
          • Select all pixels with probability $\geq P_Q + 1$ & update $N_{CQ}$
          • Randomly sample $N_{CQ} - N_{Ci}$ pixels with probability $P_Q$
          • Classify both pixel groups as $C$
        else
          • Classify all pixels in $P_{Ci}$ with probability $\geq P_Q$ as $C$.
3: Classify all remaining unclassified pixels using highest likelihood classification.
```

Full area results

This appendix shows the area per land cover as estimated by Eurostat, predicted by both the country's local model and the general model, and processed into a hard-class map by both highest likelihood classification and iterative mapping of probabilities. The 'Best' columns show which combination of model type and mapping type resulted in a mapped area that most closely matched the mean estimated by Eurostat. The Eurostat 'Min' and 'Max' columns are the estimated mean minus and plus the reported variance of the estimate; these indicate the range that should ideally contain the mapped area.

Belgium

Proportional maps of Belgium consistently matched Eurostat area estimates more closely than highest likelihood maps. Out of 5 years for 8 classes, classes mapped from predictions by general models matched Eurostat estimates more closely than those made from predictions by local models 8 times, while classes mapped from local model predictions had a closer match 6 times.

Table 4.7: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for BE in 2009.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	9.39	9.92	10.45	16.32	11.58	9.92	9.95	Local	Prop.
Cropland	26.05	26.72	27.38	20.41	19.16	26.72	26.72	Tie	Prop.
Woodland	23.88	24.54	25.2	25.0	21.31	24.54	24.54	Tie	Prop.
Shrubland	0.79	0.94	1.09	0.96	3.27	0.94	0.96	Local	Prop.
Grassland	34.37	35.29	36.2	36.82	43.51	35.27	35.23	Local	Prop.
Bare land	0.9	1.13	1.37	0.0	0.13	1.14	1.14	Tie	Prop.
Wetlands	0.05	0.07	0.09	0.0	0.08	0.08	0.08	Tie	Tie
Water	1.23	1.4	1.57	0.49	0.95	1.4	1.4	Tie	Prop.

Table 4.8: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for BE in 2012.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	10.33	10.84	11.35	16.15	11.48	10.85	10.85	Tie	Prop.
Cropland	28.7	29.44	30.18	19.95	16.9	29.44	29.42	Local	Prop.
Woodland	23.75	24.31	24.87	25.23	21.2	24.31	24.31	Tie	Prop.
Shrubland	0.74	0.85	0.97	1.39	3.83	0.86	0.86	Tie	Prop.
Grassland	31.08	31.98	32.88	36.7	45.42	31.96	31.98	General	Prop.
Bare land	0.7	0.81	0.91	0.0	0.11	0.81	0.81	Tie	Prop.
Wetlands	0.28	0.34	0.4	0.0	0.12	0.35	0.35	Tie	Prop.
Water	1.24	1.43	1.62	0.58	0.94	1.43	1.43	Tie	Prop.

Table 4.9: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for BE in 2015.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	10.9	11.39	11.88	14.49	10.59	11.5	11.39	General	Prop.
Cropland	28.2	28.54	28.88	21.9	19.75	28.54	28.53	Local	Prop.
Woodland	24.42	24.67	24.92	25.21	21.94	24.68	24.67	General	Prop.
Shrubland	1.56	1.63	1.7	1.14	2.94	1.63	1.63	Tie	Prop.
Grassland	30.4	31.02	31.64	36.67	43.57	30.89	31.03	General	Prop.
Bare land	0.75	0.83	0.91	0.0	0.11	0.83	0.83	Tie	Prop.
Wetlands	0.27	0.47	0.67	0.0	0.08	0.47	0.47	Tie	Prop.
Water	1.29	1.45	1.61	0.59	1.01	1.46	1.46	Tie	Prop.

Table 4.10: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for BE in 2017.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	11.13	11.62	12.12	14.75	10.74	11.76	11.63	General	Prop.
Cropland	28.53	28.89	29.26	19.49	18.67	28.9	28.88	Local	Prop.
Woodland	25.6	25.86	26.12	23.47	19.67	25.86	25.86	Tie	Prop.
Shrubland	1.33	1.39	1.45	1.13	3.24	1.39	1.39	Tie	Prop.
Grassland	28.51	29.11	29.71	40.63	46.39	28.96	29.11	General	Prop.
Bare land	1.33	1.44	1.55	0.0	0.11	1.45	1.45	Tie	Prop.
Wetlands	0.31	0.46	0.6	0.0	0.16	0.46	0.46	Tie	Prop.
Water	1.1	1.22	1.34	0.54	1.02	1.23	1.23	Tie	Prop.

Table 4.11: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for BE in 2018.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	11.25	11.74	12.24	15.07	10.72	11.79	11.75	General	Prop.
Cropland	28.69	29.07	29.44	22.4	21.55	28.99	29.05	General	Prop.
Woodland	26.19	26.46	26.72	25.02	21.66	26.46	26.46	Tie	Prop.
Shrubland	1.21	1.27	1.33	1.39	3.41	1.28	1.28	Tie	Prop.
Grassland	27.56	28.15	28.74	35.54	41.43	28.16	28.16	Tie	Prop.
Bare land	1.63	1.75	1.87	0.0	0.12	1.76	1.76	Tie	Prop.
Wetlands	0.33	0.45	0.57	0.0	0.09	0.46	0.46	Tie	Prop.
Water	1.01	1.11	1.21	0.58	1.01	1.11	1.11	Tie	Prop.

Czechia

Proportional maps of Czechia consistently matched Eurostat area estimates more closely than highest likelihood maps. Out of 5 years for 8 classes, classes mapped from predictions by general models matched Eurostat estimates more closely than those made from predictions by local models 4 times, while classes mapped from local model predictions had a closer match 9 times.

Table 4.12: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for CZ in 2009.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	4.05	4.26	4.47	4.41	4.0	4.31	4.27	General	Prop.
Cropland	33.16	33.63	34.11	31.7	26.96	33.64	33.56	Local	Prop.
Woodland	36.35	36.8	37.24	34.74	23.82	36.8	36.8	Tie	Prop.
Shrubland	0.6	0.69	0.78	1.77	7.78	0.87	0.69	General	Prop.
Grassland	21.77	22.26	22.75	26.28	36.44	22.02	22.33	General	Prop.
Bare land	0.7	0.81	0.92	0.0	0.05	0.81	0.81	Tie	Prop.
Wetlands	0.19	0.26	0.32	0.28	0.06	0.26	0.26	Tie	Prop.
Water	1.19	1.29	1.39	0.83	0.89	1.3	1.3	Tie	Prop.

Table 4.13: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for CZ in 2012.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	4.21	4.41	4.61	4.42	4.13	4.41	4.41	Tie	Prop.
Cropland	32.21	32.84	33.46	30.03	25.35	32.79	32.75	Local	Prop.
Woodland	37.0	37.72	38.43	35.21	25.57	37.72	37.72	Tie	Prop.
Shrubland	0.67	0.78	0.89	1.63	6.43	0.79	0.79	Tie	Prop.
Grassland	20.95	21.99	23.02	27.6	37.47	22.02	22.06	Local	Prop.
Bare land	0.6	0.7	0.79	0.0	0.06	0.7	0.7	Tie	Prop.
Wetlands	0.14	0.18	0.21	0.25	0.06	0.18	0.18	Tie	Prop.
Water	1.14	1.39	1.65	0.86	0.95	1.4	1.4	Tie	Prop.

Table 4.14: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for CZ in 2015.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	4.42	4.62	4.83	3.82	3.71	4.63	4.63	Tie	Prop.
Cropland	31.77	32.03	32.29	31.72	27.78	32.04	31.92	Local	Prop.
Woodland	37.27	37.53	37.79	34.64	26.23	37.53	37.53	Tie	Prop.
Shrubland	0.93	1.0	1.07	1.49	5.63	0.96	1.0	General	Prop.
Grassland	22.07	22.31	22.56	27.25	35.57	22.34	22.4	Local	Prop.
Bare land	0.81	0.87	0.92	0.0	0.06	0.87	0.87	Tie	Prop.
Wetlands	0.23	0.26	0.29	0.21	0.07	0.26	0.26	Tie	Prop.
Water	1.33	1.38	1.43	0.86	0.95	1.38	1.38	Tie	Prop.

Table 4.15: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for CZ in 2017.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	4.32	4.46	4.59	4.22	3.91	4.46	4.46	Tie	Prop.
Cropland	32.93	33.17	33.41	30.9	27.57	33.17	33.12	Local	Prop.
Woodland	37.82	38.06	38.31	34.82	27.07	38.07	38.07	Tie	Prop.
Shrubland	0.94	1.0	1.06	1.11	4.36	1.0	1.0	Tie	Prop.
Grassland	20.64	20.86	21.08	27.87	36.04	20.84	20.9	Local	Prop.
Bare land	0.82	0.87	0.93	0.0	0.04	0.88	0.88	Tie	Prop.
Wetlands	0.27	0.29	0.32	0.19	0.06	0.3	0.3	Tie	Prop.
Water	1.24	1.29	1.33	0.89	0.96	1.29	1.29	Tie	Prop.

Table 4.16: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for CZ in 2018.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	4.28	4.37	4.47	4.41	4.16	4.38	4.38	Tie	Prop.
Cropland	33.5	33.74	33.97	31.13	28.08	33.72	33.65	Local	Prop.
Woodland	38.1	38.33	38.56	34.67	26.14	38.34	38.34	Tie	Prop.
Shrubland	0.95	1.0	1.06	1.41	5.21	1.0	1.0	Tie	Prop.
Grassland	19.93	20.13	20.33	27.34	35.37	20.15	20.21	Local	Prop.
Bare land	0.82	0.88	0.93	0.0	0.05	0.88	0.88	Tie	Prop.
Wetlands	0.28	0.31	0.33	0.17	0.05	0.31	0.31	Tie	Prop.
Water	1.2	1.24	1.28	0.86	0.93	1.24	1.24	Tie	Prop.

Germany

Proportional maps of Germany consistently matched Eurostat area estimates more closely than highest likelihood maps, with the exception of a tie for Artificial land cover in 2017. Out of 5 years for 8 classes, classes mapped from predictions by general models matched Eurostat estimates more closely than those made from predictions by local models 7 times, while classes mapped from local model predictions had a closer match 6 times.

Table 4.17: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for DE in 2009.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	6.7	6.84	6.98	7.6	5.69	6.91	6.84	General	Prop.
Cropland	32.04	32.27	32.5	35.96	26.77	32.34	32.19	Local	Prop.
Woodland	32.09	32.28	32.47	30.7	20.18	32.28	32.28	Tie	Prop.
Shrubland	0.86	0.91	0.97	0.94	6.29	0.92	0.92	Tie	Prop.
Grassland	24.49	24.74	24.99	23.32	39.26	24.6	24.81	General	Prop.
Bare land	0.76	0.81	0.85	0.11	0.33	0.81	0.81	Tie	Prop.
Wetlands	0.36	0.38	0.41	0.15	0.25	0.39	0.39	Tie	Prop.
Water	1.7	1.76	1.83	1.2	1.23	1.77	1.77	Tie	Prop.

Table 4.18: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for DE in 2012.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	6.93	7.06	7.19	7.8	5.78	7.23	7.06	General	Prop.
Cropland	31.88	32.21	32.53	33.38	24.65	32.27	32.1	Local	Prop.
Woodland	32.51	32.8	33.1	31.43	21.35	32.81	32.81	Tie	Prop.
Shrubland	1.01	1.12	1.22	0.98	6.54	1.12	1.12	Tie	Prop.
Grassland	23.27	23.62	23.97	24.88	39.8	23.38	23.72	General	Prop.
Bare land	0.83	0.89	0.94	0.11	0.34	0.89	0.89	Tie	Prop.
Wetlands	0.48	0.53	0.58	0.16	0.27	0.53	0.53	Tie	Prop.
Water	1.65	1.78	1.9	1.25	1.27	1.78	1.78	Tie	Prop.

Table 4.19: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for DE in 2015.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	7.2	7.38	7.57	7.73	5.75	7.4	7.38	General	Prop.
Cropland	31.72	32.27	32.82	34.37	27.34	32.27	32.18	Local	Prop.
Woodland	32.54	33.79	35.04	30.41	21.87	33.79	33.79	Tie	Prop.
Shrubland	0.92	1.06	1.2	0.92	5.22	1.06	1.06	Tie	Prop.
Grassland	21.14	21.88	22.63	25.04	37.99	21.85	21.95	Local	Prop.
Bare land	1.03	1.23	1.43	0.09	0.29	1.24	1.24	Tie	Prop.
Wetlands	0.46	0.57	0.69	0.17	0.23	0.58	0.58	Tie	Prop.
Water	1.55	1.82	2.08	1.27	1.31	1.82	1.82	Tie	Prop.

Table 4.20: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for DE in 2017.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	7.11	7.5	7.88	7.5	5.65	7.5	7.5	Tie	Tie
Cropland	31.77	32.27	32.78	34.03	25.65	32.41	32.22	General	Prop.
Woodland	33.21	34.34	35.47	30.61	21.22	34.34	34.34	Tie	Prop.
Shrubland	0.94	1.07	1.19	0.87	5.09	1.07	1.07	Tie	Prop.
Grassland	20.47	21.13	21.79	25.42	40.51	20.99	21.18	General	Prop.
Bare land	1.18	1.38	1.57	0.11	0.32	1.38	1.38	Tie	Prop.
Wetlands	0.48	0.57	0.66	0.18	0.28	0.57	0.57	Tie	Prop.
Water	1.53	1.75	1.97	1.27	1.29	1.75	1.75	Tie	Prop.

Table 4.21: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for DE in 2018.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	7.07	7.56	8.04	7.5	5.61	7.56	7.56	Tie	Prop.
Cropland	31.79	32.27	32.76	34.72	28.01	32.21	32.17	Local	Prop.
Woodland	33.54	34.61	35.69	30.95	20.83	34.62	34.62	Tie	Prop.
Shrubland	0.95	1.07	1.19	0.83	7.07	1.08	1.08	Tie	Prop.
Grassland	20.13	20.76	21.38	24.47	36.63	20.81	20.85	Local	Prop.
Bare land	1.25	1.45	1.64	0.1	0.33	1.45	1.45	Tie	Prop.
Wetlands	0.49	0.56	0.64	0.16	0.22	0.57	0.57	Tie	Prop.
Water	1.51	1.72	1.92	1.26	1.3	1.72	1.72	Tie	Prop.

Luxembourg

Proportional maps of the Netherlands consistently matched Eurostat area estimates more closely than highest likelihood maps, with the exception of ties for the Wetlands class, which was estimated at zero percent by Eurostat and not predicted inside Luxembourg by any of the general models. Out of 5 years for 8 classes, classes mapped from predictions by general model matched Eurostat estimates more closely than those made from predictions by local models 12 times, while classes mapped from local model predictions had a closer match 6 times.

Table 4.22: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for LU in 2009.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	7.59	8.86	10.13	5.56	5.65	8.86	8.86	Tie	Prop.
Cropland	16.53	18.45	20.37	25.96	10.93	18.61	18.43	General	Prop.
Woodland	31.34	33.55	35.77	38.16	32.97	34.65	33.56	General	Prop.
Shrubland	0.44	0.69	0.94	0.01	1.58	0.7	0.7	Tie	Prop.
Grassland	33.51	36.86	40.22	30.11	48.5	36.87	36.87	Tie	Prop.
Bare land	-0.08	1.27	2.63	0.0	0.02	0.0	1.28	General	Prop.
Wetlands	0.0	0.0	0.0	0.2	0.0	0.0	0.0	Tie	Tie
Water	-0.0	0.31	0.62	0.0	0.34	0.31	0.31	Tie	Prop.

Table 4.23: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for LU in 2012.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	8.29	9.71	11.12	5.62	5.53	9.71	9.71	Tie	Prop.
Cropland	19.05	20.8	22.55	25.17	10.03	20.8	20.79	Local	Prop.
Woodland	31.09	32.55	34.01	39.48	34.24	33.74	32.55	General	Prop.
Shrubland	0.62	1.23	1.84	0.01	1.18	1.24	1.23	General	Prop.
Grassland	31.53	33.9	36.27	29.5	48.65	33.9	33.9	Tie	Prop.
Bare land	0.28	1.19	2.1	0.0	0.02	0.0	1.2	General	Prop.
Wetlands	0.0	0.0	0.0	0.23	0.0	0.0	0.0	Tie	Tie
Water	0.34	0.62	0.89	0.0	0.34	0.62	0.62	Tie	Prop.

Table 4.24: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for LU in 2015.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	8.74	9.82	10.9	5.74	6.0	9.83	9.83	Tie	Prop.
Cropland	22.4	23.31	24.21	24.89	11.38	23.31	23.29	Local	Prop.
Woodland	33.53	33.94	34.34	38.52	31.49	34.35	33.94	General	Prop.
Shrubland	3.05	3.31	3.57	0.01	2.09	3.32	3.32	Tie	Prop.
Grassland	28.26	28.89	29.53	30.62	48.69	28.89	28.9	Local	Prop.
Bare land	0.36	0.42	0.49	0.0	0.01	0.0	0.43	General	Prop.
Wetlands	0.0	0.0	0.0	0.22	0.0	0.0	0.0	Tie	Tie
Water	0.29	0.31	0.33	0.0	0.33	0.31	0.31	Tie	Prop.

Table 4.25: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for LU in 2017.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	7.37	8.23	9.1	6.07	5.9	8.24	8.24	Tie	Prop.
Cropland	21.47	22.28	23.09	23.93	10.32	22.29	22.26	Local	Prop.
Woodland	34.17	34.54	34.9	39.32	32.79	35.03	34.54	General	Prop.
Shrubland	2.18	2.41	2.64	0.01	2.05	2.42	2.42	Tie	Prop.
Grassland	30.83	31.54	32.26	30.46	48.57	31.55	31.55	Tie	Prop.
Bare land	0.43	0.5	0.57	0.0	0.01	0.0	0.51	General	Prop.
Wetlands	0.0	0.0	0.0	0.2	0.0	0.0	0.0	Tie	Tie
Water	0.46	0.49	0.52	0.0	0.35	0.49	0.49	Tie	Prop.

Table 4.26: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for LU in 2018.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	6.68	7.44	8.2	5.59	5.42	7.44	7.44	Tie	Prop.
Cropland	21.01	21.77	22.53	26.5	11.65	21.77	21.77	Tie	Prop.
Woodland	34.49	34.84	35.18	38.98	33.08	35.37	34.84	General	Prop.
Shrubland	1.75	1.97	2.18	0.0	2.2	1.97	1.94	Local	Prop.
Grassland	32.11	32.87	33.63	28.78	47.31	32.88	32.91	Local	Prop.
Bare land	0.46	0.54	0.62	0.0	0.01	0.0	0.54	General	Prop.
Wetlands	0.0	0.0	0.0	0.14	0.0	0.0	0.0	Tie	Tie
Water	0.54	0.58	0.62	0.0	0.34	0.58	0.58	Tie	Prop.

Netherlands

Proportional maps of the Netherlands consistently matched Eurostat area estimates more closely than highest likelihood maps. Out of 5 years for 8 classes, classes mapped from predictions by general models matched Eurostat estimates more closely than those made from predictions by local models 15 times, while classes mapped from local model predictions had a closer match 4 times.

Table 4.27: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for NL in 2009.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	10.39	10.83	11.27	12.47	10.93	10.83	10.83	Tie	Prop.
Cropland	23.04	23.73	24.41	21.38	13.52	23.73	23.73	Tie	Prop.
Woodland	11.37	11.85	12.32	15.0	7.68	11.87	11.85	General	Prop.
Shrubland	1.75	2.0	2.26	0.82	3.57	2.01	2.01	Tie	Prop.
Grassland	38.8	39.72	40.63	40.05	53.21	39.65	39.68	General	Prop.
Bare land	1.0	1.25	1.51	0.32	0.72	1.26	1.26	Tie	Prop.
Wetlands	0.19	0.24	0.3	1.81	1.25	0.27	0.25	General	Prop.
Water	9.92	10.38	10.84	8.16	9.12	10.39	10.38	General	Prop.

Table 4.28: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for NL in 2012.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	11.04	11.48	11.92	12.0	10.88	11.48	11.48	Tie	Prop.
Cropland	23.56	24.19	24.82	19.43	12.1	24.2	24.18	Local	Prop.
Woodland	11.7	12.13	12.55	15.44	7.97	12.17	12.13	General	Prop.
Shrubland	1.76	1.96	2.17	0.9	3.44	1.97	1.97	Tie	Prop.
Grassland	36.35	37.24	38.13	41.84	54.31	37.16	37.24	General	Prop.
Bare land	1.34	1.5	1.65	0.34	0.8	1.5	1.5	Tie	Prop.
Wetlands	0.48	0.58	0.67	1.73	1.37	0.6	0.58	General	Prop.
Water	10.02	10.92	11.83	8.32	9.12	10.93	10.93	Tie	Prop.

Table 4.29: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for NL in 2015.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	11.67	12.14	12.6	12.19	10.65	12.14	12.14	Tie	Prop.
Cropland	23.98	24.17	24.36	20.25	12.91	24.17	24.17	Tie	Prop.
Woodland	12.91	13.02	13.12	15.37	8.39	13.06	13.02	General	Prop.
Shrubland	1.88	2.02	2.17	0.91	3.31	2.03	1.97	Local	Prop.
Grassland	35.89	36.29	36.69	40.7	53.15	36.23	36.33	General	Prop.
Bare land	0.86	0.94	1.01	0.34	0.72	0.94	0.94	Tie	Prop.
Wetlands	0.92	1.05	1.19	1.75	1.46	1.06	1.06	Tie	Prop.
Water	9.96	10.37	10.79	8.5	9.4	10.38	10.38	Tie	Prop.

Table 4.30: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for NL in 2017.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	12.01	12.45	12.89	12.06	10.41	12.45	12.45	Tie	Prop.
Cropland	23.03	23.22	23.4	18.55	10.64	23.22	23.2	Local	Prop.
Woodland	14.29	14.39	14.5	15.29	7.53	14.4	14.4	Tie	Prop.
Shrubland	1.69	1.81	1.93	0.8	3.17	1.82	1.81	General	Prop.
Grassland	34.95	35.34	35.73	42.44	56.19	35.23	35.34	General	Prop.
Bare land	1.67	1.8	1.92	0.32	0.77	1.8	1.8	Tie	Prop.
Wetlands	0.75	0.83	0.92	2.14	1.92	0.92	0.84	General	Prop.
Water	9.77	10.16	10.55	8.39	9.36	10.17	10.17	Tie	Prop.

Table 4.31: Land cover area estimated by Eurostat and classified by both model (local and general) and map (highest likelihood and proportional) types for NL in 2018.

	Eurostat			Highest Likelihood		Proportional		Best	
	Min	Mean	Max	Local	General	Local	General	Model	Map
Artificial	12.18	12.6	13.03	11.88	10.29	12.61	12.61	Tie	Prop.
Cropland	22.56	22.74	22.92	22.47	15.78	22.75	22.73	Local	Prop.
Woodland	14.98	15.08	15.19	15.71	9.18	15.09	15.09	Tie	Prop.
Shrubland	1.6	1.7	1.81	0.97	3.3	1.71	1.7	General	Prop.
Grassland	34.48	34.86	35.25	38.26	49.53	34.78	34.87	General	Prop.
Bare land	2.07	2.23	2.38	0.34	0.78	2.23	2.23	Tie	Prop.
Wetlands	0.66	0.73	0.79	2.01	1.75	0.78	0.73	General	Prop.
Water	9.68	10.05	10.43	8.35	9.4	10.06	10.06	Tie	Prop.

Chapter 5

Synthesis

This thesis aims to contribute towards efforts made in large-scale land cover mapping, with an emphasis on the benefits of combining several datasets from different sources and of different types. It presents different steps of a methodology to extract training data from multiple rich human-annotated datasets and overlay them on Earth observation data from diverse sources. It furthermore details the challenges and benefits of creating land cover maps that navigate the trade-off between spatial, temporal, and thematic resolution, as well as quantity and allocation accuracy.

This thesis, especially in its first two chapters, describes a relatively applied line of research. Besides the chapters themselves, several datasets were produced, and have been published as open data (CC-BY license):

1. 2000-2020 quarterly Landsat composites at 30 m resolution and 7 bands (on `stac.ecodatacube.eu`);
2. 2016-2019 quarterly/annual Sentinel-2 composites at 10–30m resolution, depending on the band (on `stac.ecodatacube.eu`);
3. Harmonized LUCAS and CORINE training points for LULC classification [142];
4. Input dataset for gap filling and land cover mapping using the eumap library [191];
5. An Ensemble Digital Terrain Model of Europe at 30 m resolution [102]
6. 2000-2019 annual Land Use / Land Cover maps of Europe at 30m resolution and 43 classes [192];
7. Five annual land cover maps of five European countries (Belgium, Czechia, Germany, Luxembourg and The Netherlands) at 30m resolution and 8 classes, whose class proportions match Eurostat area estimates [300].

The third chapter of the thesis describes a relatively more innovative attempt to model land cover in a way that is faithful to design-based area estimates. The proposed algorithm (IMP) may have further uses and implications, these will also be discussed in this synthesis chapter. It is divided into two sections: the first summarizes contributions to the research objectives formulated in Section 1.3 and the second section reflects on these contributions, offering perspectives on future research opportunities.

5.1 Main Findings

The chapters of this thesis all address multiple research questions. Their results will be discussed from the perspective of the research questions; each question is discussed below.

5.1.1 What are the benefits and challenges of combining multiple large time-series and static EO datasets into Analysis-Ready Data for the purpose of land cover classification?

This research question was largely explored in Chapters 2 and 3. Chapter 2 details the construction of an Earth observation data cube from various sources (Landsat, Sentinel-2 and several DTMs), and includes experiments that investigate the performance of land cover models using different combinations of these datasets. Chapter 3 uses most of this data cube to train a single LULC classification model, which is subsequently used to create annual maps of Europe between 2000 and 2019.

Benefits

Our land cover experiments in Chapter 2 show that random forest models trained on the largest combination of datasets achieved a higher classification accuracy than models trained only on data from one satellite program during both cross-validation (0.761 vs 0.741) and on the test set (0.774 vs 0.768). This is in line with other work reporting that including auxiliary variables in the feature space can improve performance [55, 115, 121, 320].

Chapter 2 also details the creation of an Ensemble DTM which was more accurate than its four source datasets, with a RMSE 6.544 vs RMSE of 8.451-9.900. This demonstrates that in the case of continuous variables, different versions or representations of that variable can be used as input for a machine learning algorithm, and that the predictions by this algorithm can be more accurate.

In Chapter 3, the top 15 most important variables of the ensemble land cover model contain data from six sources: several bands from two satellite programs (Landsat and MODIS), long-term probability of surface water occurrence [198], geometric temperature [137], multiple DTM variables, and the cost distance to the nearest coastline (see Fig. 3.7).

Challenges

Experiments in Chapter 2 showed that models using the full feature space (Landsat, Sentinel-2 and DTM) achieved the highest classification accuracy, with different datasets improving the results for different land cover classes. This is supported by variable importance in Chapter 3: data from four different sources (Landsat, surface water frequency, DTM, and distance to coast) were among the top 15 of 200 variables for LULC modeling. While it is clear that combining different datasets into one feature space can improve model performance, there are some challenges:

Spatial resolution: Overlaying different raster datasets onto training samples may yield high accuracy (like in Chapter 2), but can cause artifacts in the shape of low-resolution raster cells when creating maps at the highest resolution among the used datasets. Fortunately, recent findings suggest that certain modeling techniques, such as convolutional neural networks, can counteract this effect [220]. In Chapters 3 and 4, we used low-resolution MODIS data as part of the feature space. To avoid artifacts, we smoothed it while resampling it to a 30 m resolution, but this is a very unsophisticated method and may have led to the loss of a lot of explanatory power. This is also reflected by its low position in the variable importance of the ensemble model in Chapter 3.

Temporal resolution can be subdivided into the temporal resolution of the feature space and that of the mapped units. It is essential to have a high temporal resolution in the feature space to correctly detect and distinguish different vegetation types due to the temporal dynamics of their phenology. The time range represented by a single map has different implications for the accuracy of different classes. In Chapters 3 and 4, we chose to make annual maps, and in Chapter 3, we made annual maps for each consecutive year. This does not work equally well for each class. For example, mapping classes that only periodically cover a certain patch of ground, such as crops, or areas that are prone to different water levels, such as flood plains. At a high thematic resolution, this problem may be minimal: a change of crops due to crop rotation will still be cropland, but when mapping different crops, this will quickly become problematic. Another example from Chapter 3 is *Burnt areas*, which was among the least accurately mapped classes. This is likely due to the fact that these areas don't stay *Burnt* for a long time.

In conclusion, integrating large time-series and static Earth observation datasets into Analysis-Ready Data can enhance land cover classification accuracy. However, challenges such as spatial resolution discrepancies, temporal resolution constraints, and class balance issues present hurdles that must be addressed to make full use of their potential.

5.1.2 To what extent does training data from multiple times and places improve the accuracy and generalization of land cover classification?

The land cover experiments in Chapters 2 and 3 investigate the generalization potential of models trained on data from a single year, and those trained on data from several years. Results from both chapters show that models trained on samples from different years were more accurate; both on years with available training data, and on unseen years.

In the land cover classification experiments of Chapter 2, the model that was trained on a small multi-year Landsat-only data outperformed the model trained on Landsat, sentinel, and DTM data on the test set. This suggests that there is a unique benefit in training a model on data from a larger time range.

In Chapter 3, we combined LUCAS points with samples extracted from CORINE polygons to create a training dataset with samples from 8 years (see table 3.13). On average, models trained on CORINE and LUCAS points from multiple years were more accurate (0.543 vs 0.498) than those trained only on one year when validated on points from 2018, which was left out of the training set in that experiment. We did notice that training on the LUCAS data lead to lower accuracy than on only CORINE (0.579) across all experiments, whether it be in combination with CORINE points (0.543) or stand-alone (0.491). This is likely because the LUCAS points are a stricter validation set, whose class only describes a 10 m square. Leveraging the higher amount of points available from using a multi-year dataset improved generalization accuracy on 2018 by 1% and accuracy on trained years by 2%. Cross-validation of the final model showed that the weighted F1-score of the model was more consistent through time than through space (with standard deviations of 0.135 per year and 0.150 per 30 km tile, respectively). While this is only a small increase, we achieved high accuracy when validating our map of 2017 on the points collected by Jenerowicz et al. [129] to validate the S2GLC maps made by Malinowski et al. [162]: We achieved a similar accuracy without having training data from 2017 and operating on a lower spatial resolution.

Dealing with bias

Combining training data from multiple sources and a wide range of locations and times can introduce bias in the model because the proportion of classes might not match the proportions on the ground in the mapped area in a given time frame. Collecting land cover observations from multiple sources requires consideration of the quality and type of each component dataset. which all have their own problems. For instance, we tried to reproduce the CORINE land cover legend at 30 m resolution, but some classes, such as *urban fabric*, don't directly translate to *buildings* at CORINE's large minimal mapping unit. This can lead to sampled points of an urban area falling within a garden, and therefore being a bad example of a 'buildings' class. Furthermore, some classes with small or narrow patches (e.g., roads) can be relatively rare at the scale of CORINE, which leads to under-representation in the resulting 30 m training set.

In Chapter 3, we did not pay any attention to this issue and selected one point per polygon, which led to some classes being drastically underrepresented and more inaccurately classified than expected, even with a large legend of heterogeneous combinations of land use and land cover classes such as CORINE. In chapters 2 and 4, we only used the LUCAS points because the emphasis was on the benefits of combining EO data sources and the effects of the IMP algorithm, respectively. In future research, a logical improvement to the point sample extraction from polygons will be to sample several points from each polygon, and having the number be dependent on the size and land cover class of the polygon. The scarcity of narrow and small-polygon classes is best dealt with by integrating more different datasets, such as OpenStreetMap for roads, and EUBUCCO for buildings [172]. In volunteered geographical information, errors can come from different sources. For example, when collecting *Buildings* data from OpenStreetMap for Chapter 3, we encountered buildings labeled in many different terms, including *residential*, personal names, *yes*, and, perplexingly, *no*. To correctly capture the right instances of a class, these issues need to be dealt with.

Our results in Chapter 3 show that the ensemble model was much less accurate in Southern Europe than in Central and Northern Europe. We found a weak but significant positive correlation between the number of training samples and the F1-score in 30 km tiles across Europe (see Figs. 3.8 and 3.9). We did not analyze whether this was due to low precision or recall, but it is likely due to differences in proportions of locally abundant classes, such as *Sclerophyllous vegetation*—which are mostly found around the Mediterranean—played a role in lowering the accuracy where they occur more than the European average. On a large scale, the spatial distribution of its probabilities matches its geographical spread quite well (See Fig. 5.1) but this class was hardly predicted in the hard-class map, being frequently overshadowed by *Coniferous forest* and other vegetation types.

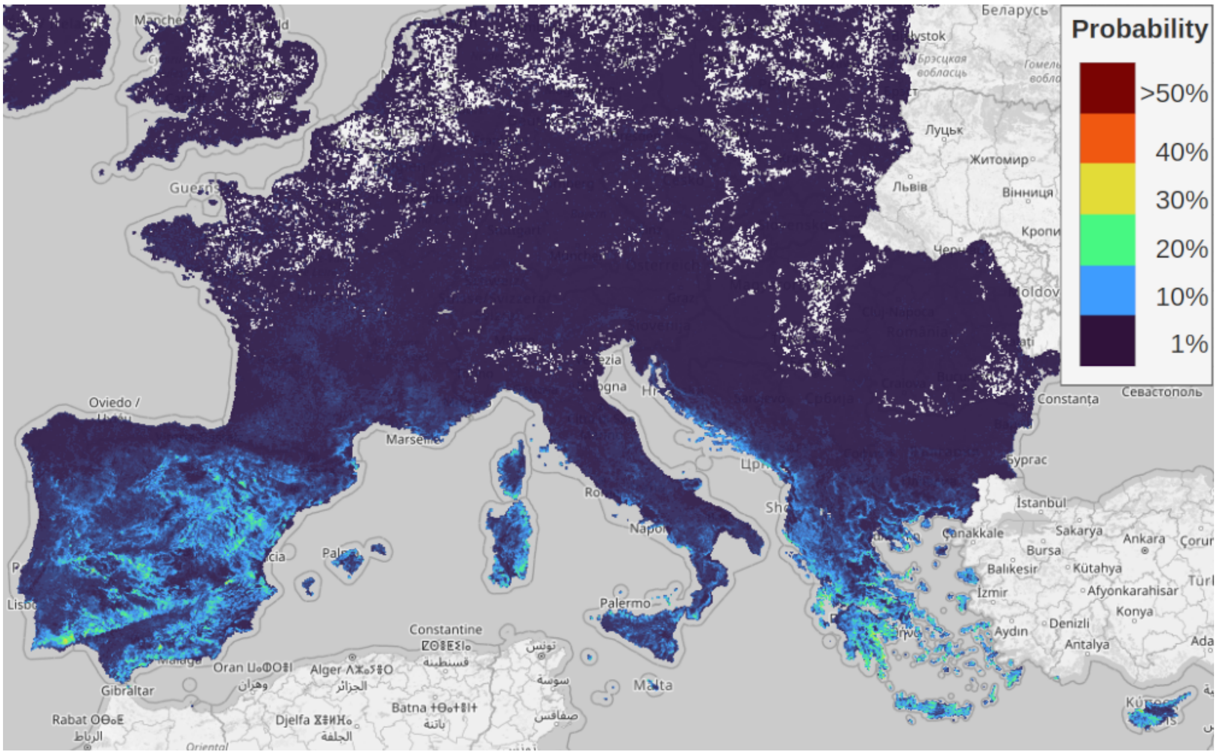


Figure 5.1: Predicted probabilities for Sclerophyllous vegetation. The geographical spread of the values matches that of CORINE, but it was rarely predicted confidently enough to avoid being overshadowed by common classes such as *Coniferous woodland.*

The proportional mapping approach demonstrated in Chapter 4 may provide a solution to a discrepancy between given class proportions and mapped hard-class proportions, as long as some level of meaningful probability is predicted in the right locations. In that chapter, we mapped the 8 level 1 LUCAS land cover classes across five years and five countries, and compared the performance of models trained only on the country they mapped with the performance of a model that was trained on data from multiple European countries. We found that proportional maps based on predictions by the general model were of similar accuracy to maps based on probabilities predicted by the less biased local models. This suggests that it is possible to train one model that can recognize many classes. If such a model is well-calibrated within each class and the top 5% of predicted probabilities, for e.g., *Sclerophyllous vegetation*, are indeed the most likely to actually be that class—even if they are overshadowed by more common classes—local area estimates can be used to force them to the forefront in areas where such classes are known to be more numerous.

In conclusion, leveraging training data from across a wide temporal and geographical range can strongly improve the accuracy and generalization of land cover classification. Models trained on a diverse temporal dataset consistently outperformed those limited to a single year's data, even when restricted to being trained on similarly sized datasets. Integrating training data in such ways can cause model bias that can strongly reduce accuracy, but these can be countered by incorporating area statistics during the hard-class classification process.

5.1.3 How do the number and type of classes in a legend affect the accuracy of land cover classification?

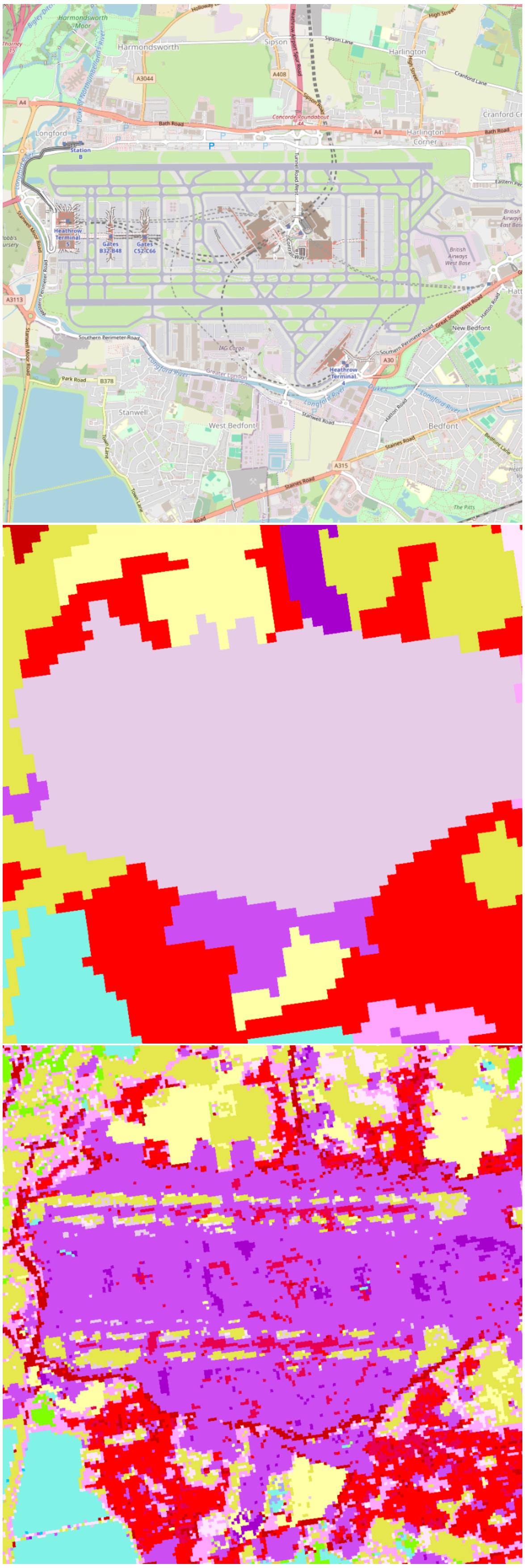
In Chapter 3, we found large differences in hard-class accuracy between the three different levels of the CORINE legend: At level 3, only 10 out of 43 classes were mapped with an F1-score above 0.5 (*Discontinuous urban fabric, Industrial or commercial units, Non-irrigated arable land, Rice fields, Broad-leaved forest, Coniferous forest, Bare rocks, Glaciers and perpetual snow, Peat bogs, and Water bodies*). At level 2, this was 9 out of 14 classes, and at level 1, all 5 classes. At level 3, we can see a positive relationship between the number of samples ('support' in table 3.10) and the F1-score, although there are classes with many samples that still scored poorly. There are a number of error sources:

Firstly, some of these classes are 'mixed types' such as *Complex cultivation patterns*, *Agriculture with significant natural vegetation* and *Mixed forest*. Some other classes occur mostly at a sub-pixel level, in effect being mixed classes with whatever class they border. For instance, *Roads and rail networks and associated land* in Chapter 3. Stretches of road or rail infrastructure that are wider than 30 m are relatively rare, and this type of land cover is often very close to other classes such as buildings. If there are classes in the same (level of the) legend, these classes 'compete', even if a prediction for any of them would be true. Second, some LULC classes have the same land cover, but differ in land **use**, such as *Pasture*, *Natural grasslands*, and *Airports* (which also have significant grasslands, see Fig. 5.2). Properly distinguishing those classes may require higher temporal and spectral resolution or feature engineering to detect differences in mowing policy or grass species.

The accuracy/detail trade-off

These mixed class and land use errors largely disappeared when aggregating classes to a level in the hierarchy where mixed nature or land use distinctions ceased to matter, such as *313: Mixed Forests* to *CLC 31: Forests and semi-natural areas*. This only works for classes that are placed in a logical order in the legend, for instance, *Pastures* and *Natural grasslands* are considered completely different LULC types even at the highest level of the legend, where they are aggregated to *Agricultural areas* and *Forest and semi-natural areas*, respectively. In the S2GLC legend, however, we both aggregated these grass classes to *Herbaceous vegetation*, which was much more effective. This applied to the S2GLC legend in general: When we summed the probabilities of all classes according to the S2GLC scheme, we achieved similar or higher accuracy than the S2GLC land cover maps. This means that training a model on many classes, even more classes than 'needed' for a specific use case, does not intrinsically harm its accuracy in a simpler legend, even when the land cover predictions are inaccurate at the highest level of detail.

In conclusion, this thesis proves that the number and type of classes within a land cover classification legend can adversely influence classification accuracy. However, aggregating predictions into a smaller set of less ambiguous classes can negate this impact. Therefore, pursuing the classification of diverse and detailed land cover maps—without the objective of achieving minimum accuracy for each class at the highest level of thematic detail—holds merit.



🟥	Continuous Urban Fabric
🟥	Discontinuous Urban Fabric
🟪	Industrial or Commercial Units
🟥	Road and rail networks
⬜	Port areas
⬜	Airports
🟪	Mineral extraction sites
🟫	Dump sites
🟪	Construction sites
🟪	Green urban areas
⬜	Sport and leisure facilities
🟨	Non-irrigated arable land
🟨	Permanently irrigated land
🟨	Rice fields
🟧	Vineyards
🟧	Fruit trees and berry plantations
🟧	Olive groves
🟨	Pastures
🟨	Annual crops associated with permanent crops
🟨	Complex cultivation patterns
🟨	Land principally occupied by agriculture
🟧	Agro-forestry areas
🟩	Broad-leaved forest
🟩	Coniferous forest
🟩	Mixed forest
🟩	Natural grasslands
🟩	Moors and heathland
🟩	Sclerophyllous vegetation
🟩	Transitional woodland-shrub
⬜	Beaches, dunes, sands
⬜	Bare rocks
🟩	Sparsely vegetated areas
⬛	Burnt areas
🟩	Glaciers and perpetual snow
🟦	Inland marshes
🟦	Peat bogs
🟦	Salt marshes
⬜	Salines
🟦	Intertidal flats
🟦	Water courses
🟩	Water bodies
🟩	Coastal lagoons
🟩	Estuaries
⬜	Sea and ocean

Figure 5.2: Left: Heathrow Airport, London, UK, as represented by OpenStreetMap (top), CORINE Land Cover (center), and the predictions of Chapter 3. Right: The 43-class CORINE Land Cover legend. Note that while the predictions replicate the patterns of the actual airport infrastructure much more closely, they are classified as *Industrial and commercial units*, *Pastures*, and *Roads*.

5.1.4 What is the effect of enforcing design-based class proportions on map accuracy?

In Chapter 4, we created annual maps of 5 European countries. For each country, we predicted probabilities with a local model that was trained only on data from that country. We also predicted probabilities for each country with a model that was trained on a larger, pan-European dataset. We then made hard-class maps of each country and year in two ways: by maximum probability assignment and with a novel algorithm called Iterative Mapping of Probabilities (IMP).

The IMP algorithm was developed as a tool to produce maps that match existing area estimates without sacrificing accuracy. Our experiments in Chapter 4 show that classifications by IMP tend to be *more accurate* than those by maximum probability assignment, especially in the case of models that are biased to over- and underpredict certain classes.

Results showed that maximum probability maps using the probability of the local models tended to be more accurate than those by the general model, but this was not the case when using IMP instead of maximum probability assignment. The maps created by IMP were however of equal or better accuracy than those made with maximum probability assignment, while having class proportions as estimated by Eurostat. The effect of IMP on accuracy can be summarized as sacrificing precision (user accuracy) of some classes, to raise recall (producer's accuracy), with a net benefit to higher weighted F1-scores.

Future research

Essentially, IMP quantifies the bias of a model when compared to a given class proportion estimate and uses the iteration probability thresholds to impose its own bias on the model predictions, minimizing the difference between precision and recall. While class-wise precision and recall values of proportional maps were closer to each other than those of maximum probability maps, they were not identical. While Eurostat area estimates are considered to be quite reliable, a direct class count from a dataset is by definition 100% accurate. It is likely that using these counts could further harmonize precision and recall on a class-by-class basis. If this is true, IMP can potentially be used to quantify either how representative a validation dataset is for its study area, or the accuracy of a class proportion (area) estimate. Additional experiments that did not make it into the thesis provide strong suggestions that this is true, and we will keep researching this in the near future.

Chapter 4 showed that IMP provided a greater accuracy improvement to models that were trained on data with a different class distribution than the area they mapped. This means that IMP can be used to optimize the predictions of large land cover models to a regional context without requiring new training or predictions. It also means that IMP can learn the bias of a model if the model is validated in an area for which class proportions can be estimated; the quantification of this bias can then be stored, and used to make proportional maps of areas for which class proportions are not available. This too is a promising line of research that will be continued after this thesis.

If reliable and trusted validation data exists for an area, IMP can be used to compare the accuracy of different area estimates by making hard-class maps for each area estimate. The area estimate that matches the reality on the ground most closely should allow IMP to produce the most accurate map. When used this way, IMP could be used to settle disputes about quantities of land cover and land cover change.

Wider applicability

The ability of IMP to quantify and correct model bias extends its usage beyond land cover classification. It can be used for any type of machine learning task where class proportions can be either estimated, targeted, or dictated, and where bias must be quantified. Any field that utilizes machine learning models and faces challenges with class imbalance, representation bias, or requires models to perform accurately across diverse and potentially underrepresented groups could use the algorithm to enhance model fairness, accuracy, and generalization.

For example, in predictive policing and recidivism prediction, biases in the training data can perpetuate and amplify societal biases. Reducing the bias in such models could contribute to fairer decision-making in the criminal justice system [14, 56]. Financial institutions often use machine learning models for credit scoring and risk assessment. The training data can be biased due to historical decisions and social demographics, potentially leading to unfair assessments. A post-hoc correction algorithm could mitigate these biases, leading to fairer credit scoring and risk assessment models [37, 132].

5.2 Reflection and outlook

This thesis presents a transparent, reproducible framework for combining data from various sources to create detailed, consistent maps. Its chapters show examples of how to combine Earth observation and land cover data from multiple sources, pre-process and harmonize them, and make annual maps with many classes. It shows how to deal with the problems of model bias toward some classes, which may arise when some land cover classes have more data available compared to others.

Combining several datasets from different times and regions to create large and rich training datasets is a challenging task. It requires knowledge of available data sources, technical skills, and extensive spatial, temporal, and thematic harmonization work. This thesis shows that such a dataset can be used to train a model that classifies many classes. While the accuracy at full thematic depth is lacking compared to those of other recent products, we showed that they are of comparable accuracy to similar products when we reduce the number of classes, like CORINE level 1, or include a legend more optimized for remote sensing like S2GLC. We also show that there are benefits to training a model on data from different times, even without performing any kind of time-series analysis. Finally, we show that it is possible to incorporate class proportions such as area estimates into the workflow to create land cover maps that not only match these proportions, but are more accurate than maximum probability maps.

We have published the code and data that we used and produced in this thesis in the hope that it is useful to other modelers, especially those who are not experts in remote sensing. The analysis-ready feature space and harmonized training datasets can be used by anyone to improve our method and maps. The predicted probabilities of Chapter 3 can be used as input for ensembles to map other types of land cover, potentially much more accurate than the probabilities themselves, just like has been suggested by the creators of DynamicWorld [24].

This thesis not only contributes to the field of remote sensing by enhancing methods for land cover mapping but also sheds light on the role of data integration and methodological innovation in addressing current limitations. Nonetheless, the journey toward comprehensive, dynamic, detailed representation of the Earth's surface is far from complete. As we stand on the threshold of new discoveries and technological frontiers, we are propelled to ask a fundamental question: What is necessary to map *everything, everywhere, all the time*?

5.2.1 Mapping everything: On detailed hierarchical legends

It is, of course, unfeasible at best to map *everything*. However, there is a clear use case for having detailed and accurate maps that scientists and policy-makers can use for their specific use cases, for example, to distinguish wetland types for bird conservation [66] or peat bogs [246] for insect biodiversity. But it extends beyond nature; different categories and qualities of the urban environment are frequently investigated and compared due to their effects on human well-being [141], socio-economic inequality [263], and large-scale analyses on the sustainability of social-environmental systems [38]. Especially different vegetation types are becoming increasingly important. Unfortunately, these are also the most variable in space and time, while simultaneously being hard to distinguish, especially on a high thematic resolution. Any class that is mapped needs to be properly represented, both conceptually and in the data. In general, to map LULC at high thematic resolution, we need:

1. A hierarchical legend with meaningful definitions that work for both machines and people;

2. Training data and area estimates that are compatible with said legend;

3. A feature space where all classes can be distinguished.

The Stuff of Legends

In this thesis, I attempted (in Chapter 3) reproduce the CORINE land cover legend, and experienced the limitations of mixed land use / land cover classes, and the contextual overlap between some classes like pastures, industrial buildings, and airports (see Fig 5.2). While land use can cause much confusion to some machine learning models, it is essential, perhaps even more so than land cover, to locate and quantify. These limitations can likely be overcome by optimizing the legend and disaggregating land use and land cover in a fine-grained way.

Fortunately, much work has already been done on this. The discussion around the ambiguities of CLC and its incompatibilities to other nomenclatures, like LUCAS, inspired the formation of the Eionet Action Group on Land monitoring in Europe (EAGLE) group, which developed a method to quantify explicit aspects of land cover components (e.g., points, pixels or polygons on a map) that can be combined to assign a meaningful land use / land cover label. The EAGLE system presents a shift from single-class classification of mapped units, and instead characterize them by their land **cover**, land **use**, and land **characteristics** (See Fig. 5.3).

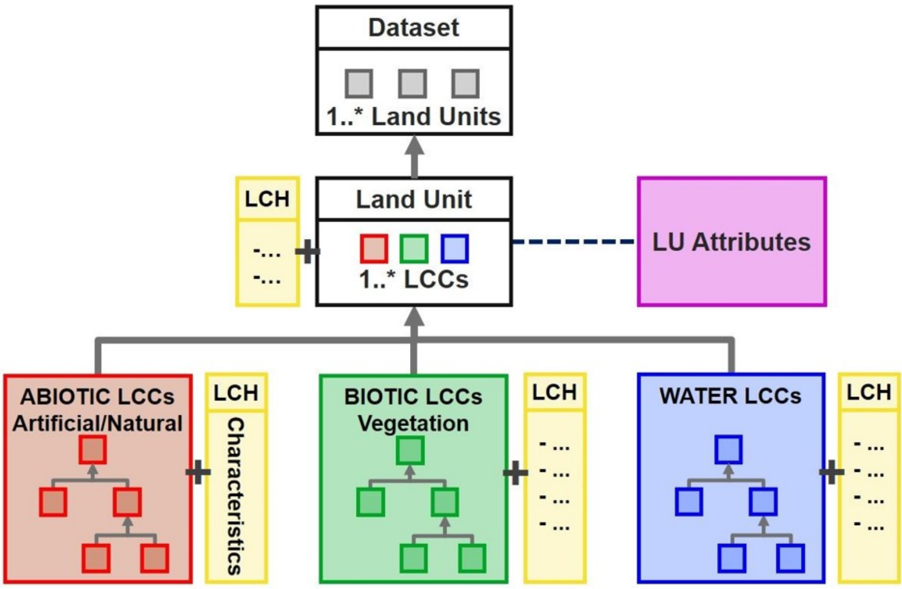
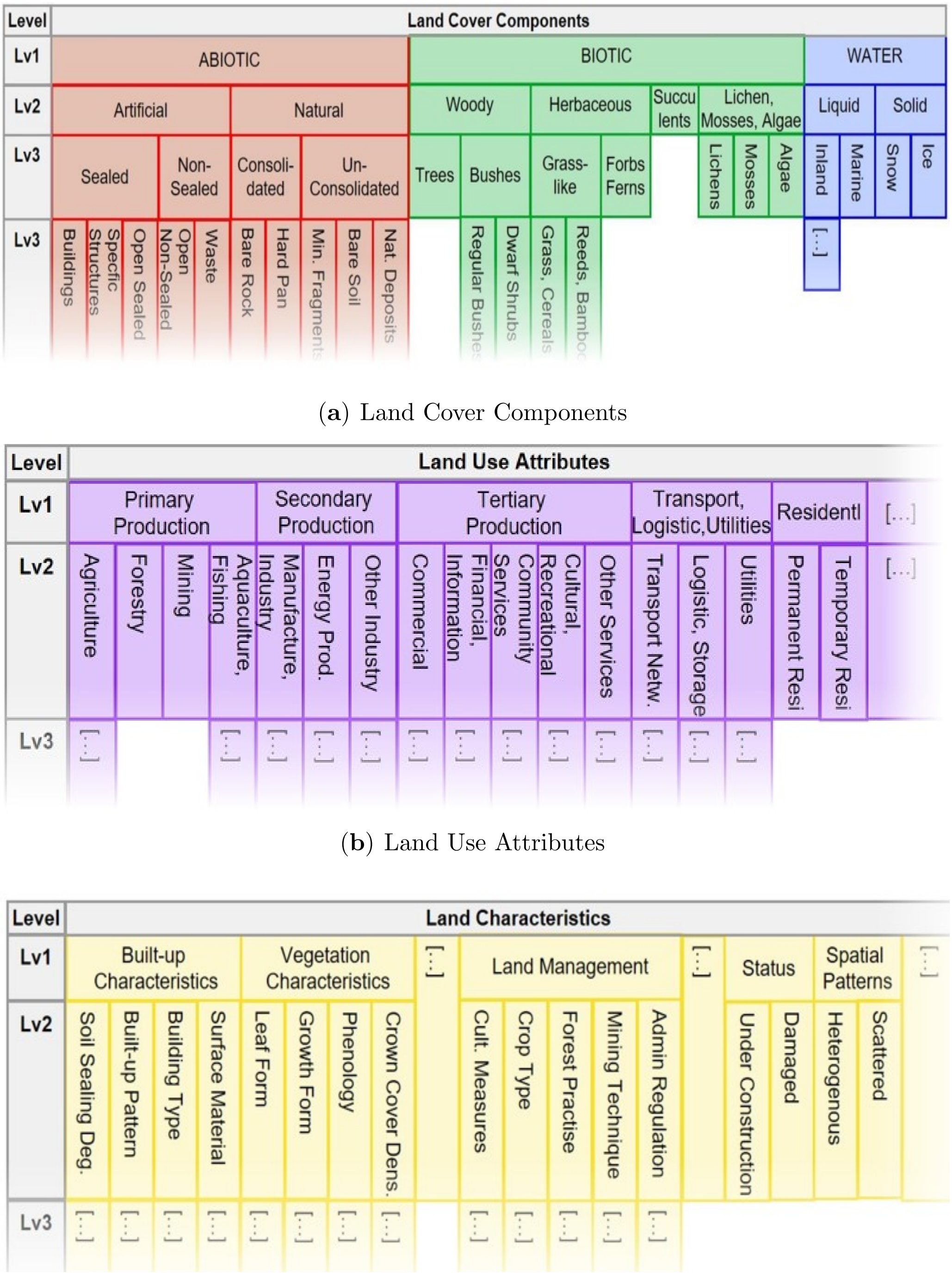


Figure 5.3: Classification structure of the EAGLE concept: Land Cover (LC), Land Use (LU), and Land Characteristics (LCH) are quantified separately, and can be combined in multiple ways to provide a LULC label to a mapped unit (e.g., a pixel or polygon). Source: `https://land.copernicus.eu/en/eagle`

These components are (or consist of) hierarchical legends, and each mapped unit is represented by a combination of these aspects instead of being classified by a single label. For example, *Land Cover* consists of *Abiotic*, *Biotic*, and *Water* land cover categories that each have subtypes (see Figs. 5.3and 5.4a) such as *Woody* vegetation being split up in *Trees* and *Bushes*. Examples of land *characteristics* are the *frequency of surface water presence through time* [198] and *Canopy height* [208] (See Fig. 5.4c). Land *use* is defined at the highest level as the sector of the human economy a piece of land is used for, such as *Primary production*, and then further defined as e.g., *Forestry* and *Agriculture*, in turn, *Agriculture* then contains many different agricultural practices and purposes (See Fig. 5.4b).



(a) Land Cover Components

(b) Land Use Attributes

(c) Land Characteristics

Figure 5.4: A simplified representation of the EAGLE concept and how it can be used to divide land units into categories based on hierarchical combinations of land cover, land use, and water presence. Source: `https://land.copernicus.eu/en/eagle`

I have shown that aggregating predictions to simpler and/or more optimal legends can drastically improve their accuracy. It would not be surprising if the greater confusion at higher thematic detail has motivated other attempts to create CORINE land use / land cover predictions to pre-emptively reduce their level of detail. But what if you could selectively reduce the detail of predictions that are probably mistakes?

Using prediction uncertainty to allow a model to predict *nothing* is called *Selective classification*, but this would cause gaps in a map. While no one wants mistakes on their map, why would you throw out the baby with the bathwater to get rid of them?

A well-designed hierarchy can be used to *save the baby* in the case of uncertain predictions at high thematic detail. For example, it may be hard to correctly distinguish the class of a mixed *Wheat* and *Grass* pixel if it's on the edge of a field, or at the start of the growing season. Prediction uncertainty metrics could be used to aggregate such pixels to a sensible common class, such as *Herbaceous vegetation*. The prediction would then 1) be correct instead of incorrect, 2) be useful instead of empty and 3) indicate prediction uncertainty in a way that humans can easily comprehend.

Doing this requires a reliable way to quantify prediction uncertainty, in the explicit sense of *"the probability that this specific pixel is a misclassification"*. Our first attempt, the *"model deviance"* from Chapter 3, is calculated by taking the standard deviation of predicted probabilities by each learner in an ensemble. While we found no significant relationship to the chance that a given pixel prediction is correct, it turned out to be a useful proxy for indicating the distance from the nearest training point of a given class in the feature space. However, there are several metrics that have a stronger theoretical foundation to quantify prediction uncertainty. The most straight-forward and often-used is the highest probability among predicted probabilities. This metric is only reliable when the model is well-calibrated [184], because models can also be "confidently incorrect". Calderón-Loor et al. [32] and Bonannella et al. [20] used the *margin of victory*, which is a useful metric for how confident the model is between classes. Conformal prediction promises statistical guarantees of correct error estimation [3] and is recently finding traction in the land cover community [242, 279]. Finally, in Chapter 4, we found indications that there is a robust relationship between the iteration at which a pixel is classified, and how likely it is to be correct.

Training Data Montage

No single training dataset currently contains all classes that can be mapped. For example, not even detailed legends like LUCAS have unambiguous exclusive labels for increasingly important objects such as solar panels. To make internally consistent maps with a unified approach, it is therefore essential that existing and future open LULC-annotated data can be integrated in a modular and flexible way, especially without 1) adverse effects from class imbalance in the resulting dataset and 2) sacrificing detail from aggressive legend harmonization.

The effect of class imbalance from dataset concatenation can be countered by applying IMP or similar processes that adjust for model bias. The ability of IMP to correct the bias of a model post-hoc means that models don't need to be trained on datasets that respect class balance, like other mapping [138, 294], as long as accurate area estimates are available. This means it can be combined with automated training data generation techniques that don't respect class distributions, such as extracting points from polygons *en masse*, as was done in this thesis. It also means that training datasets with different distributions and classes can be easily combined.

Using a disaggregated set of land cover and land use legends, in line with the EAGLE concept, will allow us to more easily combine existing training datasets. Because different biotic, abiotic, and water land cover classes, as well as land use classes, can simultaneously be present on a pixel, a multi-label classification system would be a logical choice [255]. Any actual LULC class in the legend can then be calculated by summing the probabilities of all labels that contribute to it. This can even work for different levels of a hierarchical legend [297]. Multiple predicted labels from different levels in a hierarchical legend can be combined to represent a single meaningful class. For example, two land cover classes *Herbaceous vegetation*, *Grass* and two land use classes *Primary production* and *Grazing* can be combined to represent 'pasture'.

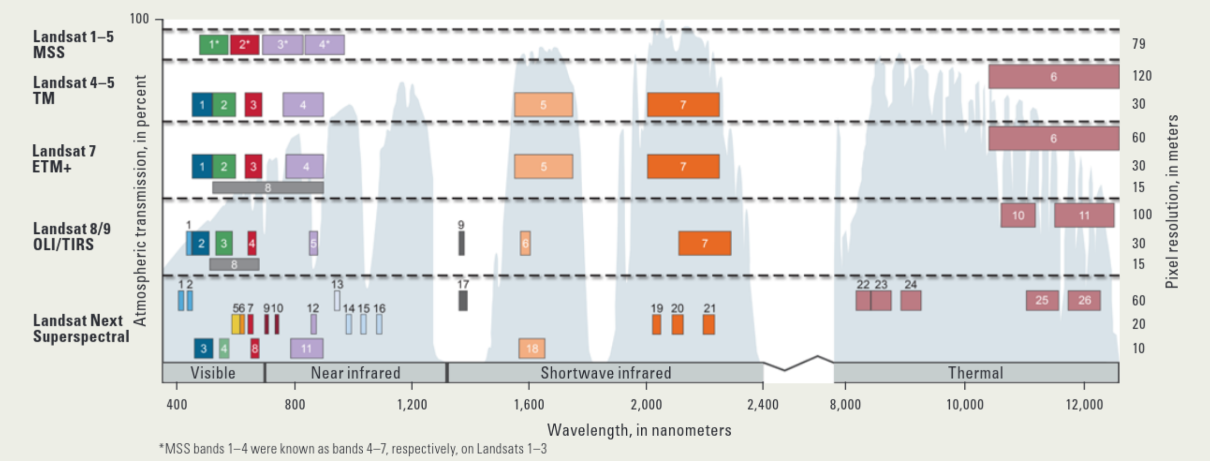
Featuring: Space

However, regardless of how meaningful, effective, and detailed its legend, a LULC map can only be as good as the data used to detect them.

In this thesis, we primarily used a temporally aggregated version of the long-running Landsat archives due to its objective to make consistent time-series of maps and to enable the use of training data from legacy datasets such as CORINE. While the Landsat archives provide a consistent dataset spanning decades, its relatively low spatial, spectral, and temporal resolution limited the classification accuracy of several classes.

For example, different vegetation types are often best distinguished based on their profile in the electromagnetic spectrum [105, 182, 306]. Furthermore, crop types, in particular, benefit from high temporal resolution [61, 306]. Their growth cycles, during which their appearance and spectral profile changes drastically, combined with temporal dynamics such as crop rotation, fallow periods, and other practices, require a high frequency of EO data. Making use of Sentinel-2 imagery, with its revisit time of 10 days and higher spatial resolution (10 m), would have limited the time range of the maps to 2015 but might improve performance, as suggested in Chapter 2. If the emphasis is on mapping longer time series, the framework could be significantly improved by incorporating the NASA Harmonized Landsat and Sentinel-2 (HLS), which combines the longevity of the Landsat program with the high temporal resolution of the Sentinel program [41].

In the longer term, Landsat Next, scheduled to start providing data in 2030, promises substantial improvements while ensuring compatibility with both its own predecessors and other systems (See Fig. 5.5). The 26 spectral bands of this new iteration will match the 11 'heritage' bands of previous Landsat programs but will also contain 5 bands with similar spatial and spectral characteristics as Sentinel-2, improving revisit time from 16 to 6 days [277]. Similarly, ESA's Copernicus Hyperspectral Imaging Mission for the Environment (CHIME) will have over 200 spectral bands and a revisit time of 12.5 days, revolutionizing hyperspectral Earth observation [185]. Meanwhile, the German Aerospace Center's recently launched Environmental Mapping and Analysis Program (EnMAP) system, with 228 hyperspectral bands and a revisit time of only 4 days, is particularly suitable for distinguishing vegetation types such as crops and tree species. It is already available and being used to map LULC, showcasing the utility of high spectral and temporal resolution in current remote sensing applications [143, 251].



*MSS bands 1–4 were known as bands 4–7, respectively, on Landsats 1–3

Figure 5.5: Comparison of pixel resolution and spectral bands between Landsat Next and previous Landsat satellites. From: [277]

5.2.2 Mapping everywhere: Moving beyond Europe

The framework presented in this thesis relies on openly available land cover samples for training and validation, and benefits greatly from accurate area estimates. Thanks to the efforts of European institutions, such as Eurostat, the EEA and JRC, such resources exist for Europe. Its high levels of collaboration, organization and prosperity also facilitate other projects that can be good sources of training data, such as the impressive EUBUCCO [172] building dataset.

Most other continents, however, are not similarly fortunate. For example, the LUCAS land cover observations [49] have allowed many global maps to be validated and compared in Europe [83, 286], but this is more difficult to do in a standardized way in countries and continents that have a less detailed, reliable, or otherwise representative validation dataset.

Furthermore, land cover and land use represent the diverse ways humans interact with the environment, influenced by geographical, cultural, and economic factors that vary widely across regions. To extend the proposed framework to other continents and make global maps at high thematic detail, such differences need to be acknowledged and accounted for. To do this, we need global samples and statistics, and supplement them with data that is meaningful on a local scale.

Global differences

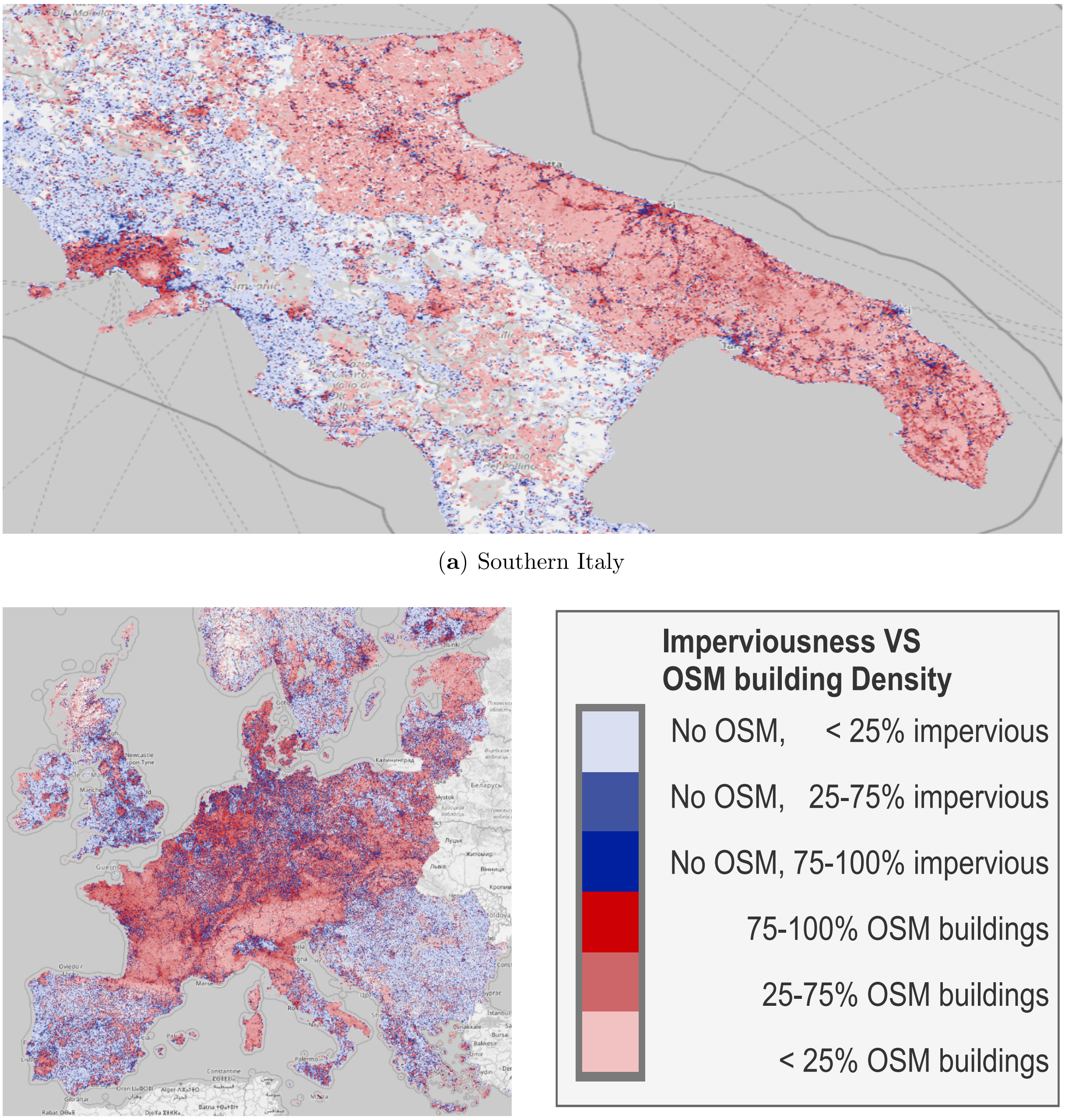
A global map needs to reflect the significant differences between land cover and land use types across various regions. These differences stem from diverse administrative norms, capacities, environmental contexts, and land use needs.

For instance, the classification systems and the types of crops that are prevalent in different biomes vary widely. In Europe, the distinction between crops is typically made in July to capture peak growing seasons [61, 306], whereas the monsoon cycle determines India's Kharif (June–October) and Rabi (November–April) seasons. Projects like WorldCereal address these challenges by creating a collection of regional maps, each tailored to the unique agricultural seasons of the area it represents [284].

Global data

To map relevant LULC classes globally (or at least in other continents than Europe), training and validation data from those areas is needed. For this purpose, we can combine generic global training datasets like the Dynamic World training data [260] and GLANCE [247] with global datasets that represent a few detailed classes such as WorldCereal [21] and NASA CropHarvest [274].

Volunteered geographical data is another important source of detailed class labels. For example, OpenStreetmap was used to extract training data and map 14 CORINE classes more accurately than CORINE Land Cover in most cases [232]. In 2016 OSM's road network was 83% complete, and more than 40% of countries, including several in the developing world, had a fully mapped street network [9]. Still, its coverage for other objects is far from perfect, as even in Europe, the extent to which buildings and other objects are annotated is heavily biased to areas that are more prosperous or more popular with prosperous tourists (See Fig. 5.6a).



(**a**) Southern Italy

(**b**) Europe

Figure 5.6: Extracted OpenStreetMap buildings (in red), supplemented by Copernicus High Resolution Imperviousness data (in blue), used in Chapter 3. The red areas indicate where buildings are annotated in OSM, while blue indicates impervious areas that are most likely buildings, but were missed by OSM. It shows clear differences between country borders (e.g., between France and Spain, Austria and Slovenia), regions (e.g., Puglia and Campagna in Southern Italy), and the urban/rural divide. For a more detailed explanation of this dataset, see Chapter 3.

The detail of such a global dataset can then be enriched with localized datasets that represent the landscape of different regions in locally relevant classes, such as CATLC [84]. In Section 5.2.1, we already suggest ways that a combination of multi-label classification and IMP can be used to combine diverse datasets with incompatible legends and class balances. In this context, it would mean that a model analyzing a dense sclerophyllous forest in the Mediterranean can predict *Forest* based on DynamicWorld training data [260], *Dense Forest of sclerophylls* based on data from CatLC [84], and *Sclerophyllic vegetation* from Corine Land Cover and S2GLC [129].

To effectively combine polygon and point data sources, a point sampling technique such as the one used in this thesis can be applied, creating a harmonious point dataset. Although this limits the number of modeling techniques, especially more modern ones such as convolutional neural networks, interesting work has been done on *weakly supervised image classification* [118]. This technique allows powerful deep learning image segmentation models to be trained without fully annotated pixel maps, instead only requiring labels about the classes that are present *somewhere in the image* (See Fig. 5.7). Applying this technique to land cover classification would ameliorate the need for costly wall-to-wall annotations that are currently used in many benchmark datasets such as CatLC [84] and BigEarthNet [254], and would improve the value of legacy and novel point-wise observations such as LUCAS [46] and GeoWiki [78].

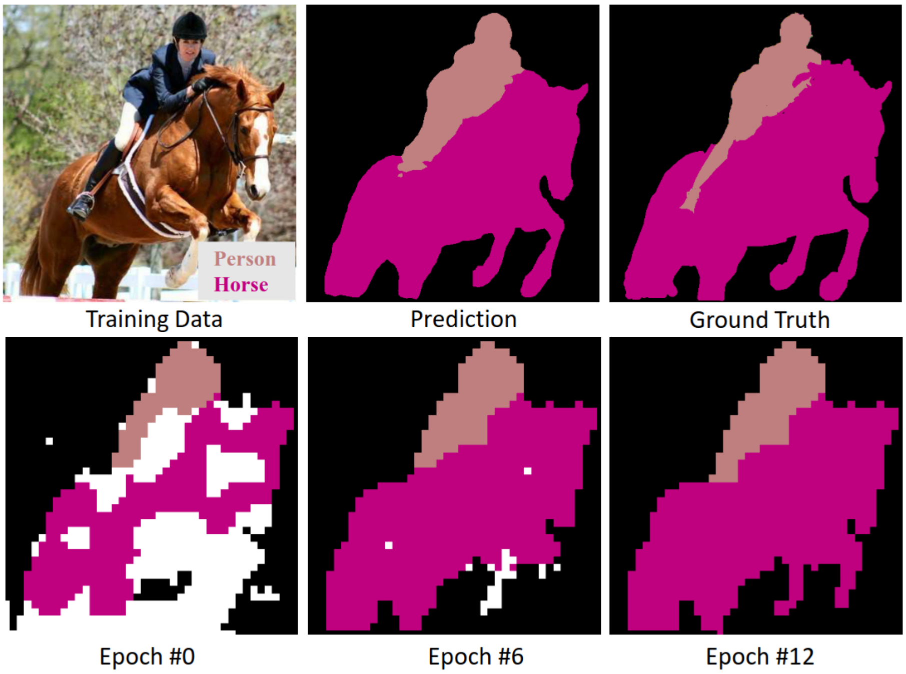


Figure 5.7: Example of the training process of a weakly supervised semantic segmentation model, from Huang et al. [118]. The top row shows a training image that is only annotated with the classes that are present in the image, as well as the predicted and true class membership of each pixel. The bottom row shows the performance of the model at various steps (epochs) in the training process.

Global statistics

It is imperative to counter the effects of class imbalance that can be caused by combining different datasets, or by using data from a large region (See Chapter 3 and Section 5.1.2). In Chapter 4, we have shown that this is possible by using area statistics from official sources. This is not a novel concept: You et al. [311] used multiple datasets, including area statistics from governmental organizations, to map the area of twenty different crop types. What is novel, however, is that IMP might be generalizable through space and time, when the predictions are made in a consistent feature space.

IMP offers a promising way to address class imbalance and model bias in land cover classification, even when detailed area statistics are not universally available. The algorithm tackles model bias by analyzing cut-off probability values for each class across iterations. The key idea lies in generalizability: if the confusion patterns between classes are similar across regions, IMP might be applicable even without area estimates. You could store the probability threshold for each class at each iteration and apply them to probabilities predicted by the same model for a different area. This can be explored in a European context by mapping neighboring NUTS2 areas of the ones we have already mapped, using the cutoff values of their originally mapped neighbor. We can then 'pixel count', and compare the counts to the area estimates. This could make the method applicable for areas that are less meticulously quantified than Europe. This will still require area estimations to be done in some representative locations, but does not require continent- or country-wide surveys.

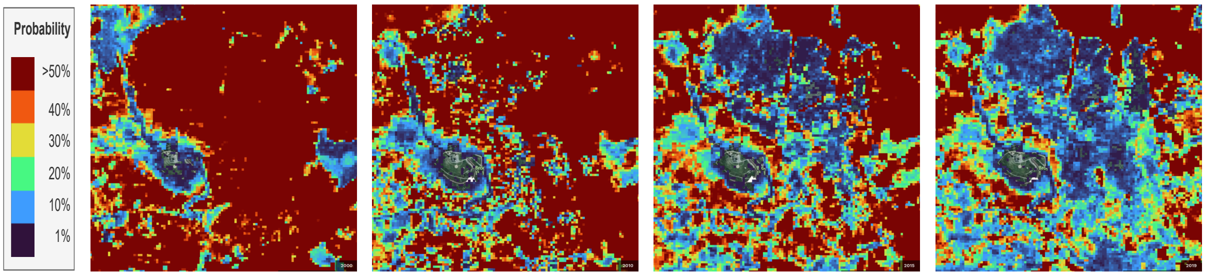
5.2.3 Mapping all the time: Temporal resolution and range

Approaches that map many things will need a rich feature space and will likely not be useful for rapid detection of, for example, illegal deforestation or natural disasters.

Long-term analysis

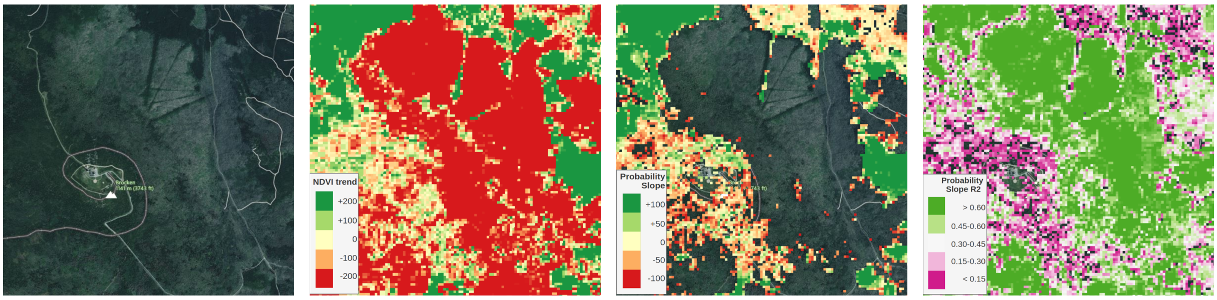
They will be useful for assessing long-term trends though. In Chapter 3, we used regression on the predicted probability time-series per pixel to analyze the long-term trends predicted by our model (See Fig. 5.9). While the method we used had problems with extremely negative slopes being quantified as ’no data’ (see Fig. 5.9c), the slope of the harmonized NDVI trend (Fig. 5.9b) and annual predicted probabilities for *Coniferous forests* accurately shows the mass dieback of Norway spruce trees following a bark beetle infestation in the area [171].

More sophisticated and robust pixel-wise trend analysis techniques can be used to quantify gradual processes, like competition and succession of different plant species, which is becoming more relevant in the light of global climate change [20]. Having long-term, well-calibrated predicted probabilities for many classes, based on a rich, harmonized feature space, would allow many different users to conduct their own analysis on the classes that are relevant to their use case, such as drivers of deforestation [165] or mapping and assessing wetlands ecosystem services [74]. In this light, there is a unique value to primarily relying on long-running datasets with a high likelihood of future continuity, such as the Landsat program.



(**a**) 2000 (**b**) 2010 (**c**) 2015 (**d**) 2019

Figure 5.8: Predicted probabilities for coniferous forest (Norway spruce) near Mt. Brocken, Germany in four years. A bark beetle infestation caused large-scale dieback, which is reflected in a decrease of probability over time.



(**a**) High-resolution imagery of the east side of Mt. Brocken. Source: Microsoft Bing

(**b**) 20-year trend slope of Landsat NDVI values.

(**c**) Probability slope for *Coniferous forest.*

(**d**) R2 of probability slope for *Coniferous Forest.*

Figure 5.9: Dieback of coniferous forest (Norway spruce) near Mt. Brocken, Germany, as represented by data generated in Chapters 2 and 3. Note that the *"Probability Slope"* has nodata exactly where it should be highly negative and has a high confidence, but that the *"NDVI slope"* correctly displays the area where the dieback occurred.

Temporal resolution

Creating long-term land cover maps that capture temporal dynamics is a complex task. While annual maps seem intuitive, they struggle to represent the changing nature of land cover classes like crops, which undergo various phenological stages [227], fallow periods [264], and other transitions [221] throughout the year. The *"correct"* long-term map period depends on the specific application and the land cover classes of interest.

However, to help work with time periods of six months or more, one possible approach involves using multi-temporal classes in the legend system. These classes could represent sequential land cover types within a single year. For example, *Cotton-Wheat* could indicate an area planted with cotton followed by wheat in the same year. By incorporating this strategy, annual or seasonal maps can still be informative. These maps would depict the dominant land cover for a specific period while acknowledging the sequential changes that occur throughout the year. This approach provides a balance between detail and usability, without overwhelming users with excessive maps.

Increasing the temporal resolution to capture frequent changes comes with trade-offs. Highly detailed maps (e.g., daily or weekly) can become computationally expensive. For time-series methods that require observing the entire land cover cycle, a time period with a time resolution that captures the full cycle is crucial. Many crops have cycles lasting 3-4 months, suggesting that bi-weekly or monthly resolutions might be more suitable for these cases. Evergreen broadleaf forests face much less changes along the entire year. In this case, monthly or bi-monthly resolution might be more appropriate. This allows the model to *"see"* the complete transformation of the land cover throughout the season or the year and improve classification accuracy.

5.2.4 Mapping it all at once? A final word on models

In each of the previous sections about classes and legends (5.2.1), training data and feature space (5.2.2), and the role temporal resolution and coverage (5.2.3), I have only briefly mentioned modeling techniques that would be appropriate to deal with their specific issues. While each separate question has separate answers, the modeling part of this synthesis must be answered holistically. This final section presents my perspectives on a modeling approach to make detailed, global maps that have useful and meaningful LULC classes.

I have previously mentioned that different models and feature spaces are optimal for different types of land cover, that global mapping would benefit from an adaptable weakly supervised system that can combine the diverse offer of open datasets, and that the type of classes is intrinsically linked to the temporal resolution of the mapping approach, especially when going beyond mapping land *cover* only.

A modeling approach is needed that:

1. Classifies many different LULC classes by incorporating:
   (a) Spatial context (e.g., for airports and urban green);
   (b) Temporal dynamics (e.g., for agricultural practices and flood plains)
2. Can be understood and trusted by scientists and policy-makers;
3. Can be trained on mixed point data with diverse overlapping legends;
4. Can deal with potentially imbalanced training dataset.

That is why I propose, and hope to contribute to, the following approach:

Hierarchical Multi-label Classification for Legendary Learning

The LULC legend should have many different classes in a single hierarchical legend with many levels. The model should be trained with a loss function that varies its penalty based on where errors happen in the hierarchy: For example, errors between tree species should be penalized less harshly than errors between buildings and trees. Previous work on constraining predictions to correct hierarchy paths exists and is called *Hierarchical multi-label classification* [297].

Pixel-based uncertainty estimates

The final output of the model must have pixel-based uncertainty estimates so that subsequent modeling and decision-making can properly weigh the risks and rewards of using the data.

An additional benefit of having a robust uncertainty metric is that likely errors in the LULC classification can be aggregated to higher levels in the hierarchical legend. This requires the legend to be designed in such a way that confusion between similar classes is minimized by ascending to higher levels. This can be done in two ways. The first method is a top-down, human-assumed way like EAGLE, where splits happen on a conceptual basis between biotic and abiotic at the top, and between different species of the same vegetation types at the bottom of the legend. The second method is by analyzing the confusion matrix of initial predictions, and forming the aggregation decision tree purely based on which classes are more likely confused by the model. The first method has the advantage of being more interpretable by humans, while the second method has the advantage of likely being more effective at minimizing error by reducing detail. In either case, pixel-based prediction uncertainty must be derived using a robust metric (See section 5.2.1).

An Interpretable Land Cover and Characteristic Ensemble

Several models are trained to classify land cover types, such as water and herbaceous vegetation, and land characteristics such as height and photosynthetic activity. These models are trained on point samples on a diverse, modular set of training data to make monthly predictions on a pixel-by-pixel basis. For some classes, especially crops, it is essential to incorporate their temporal dimension, while for others, this is not necessary. The predictions by these models will together form the feature space for a meta-learner that learns from their spatial and temporal context. An added benefit of this approach is it mirrors hybrid modeling techniques, which integrate domain-specific knowledge and physical laws into the ML framework, enhancing model transparency and interpretability. Thus optimizing the balance between effectiveness and trustworthiness makes them ideal input for further scientific analysis and decision-making [71].

Spatiotemporal LULC Metalearner

The spatiotemporal datacube of monthly LC/LCH predictions is then used as the feature space for a weakly supervised (see Section 5.2.2) meta-learner performing hierarchical multi-label classification. The training data of this meta-learner will be multidimensional windows in the input datacube that overlap with annotated points: these will function similarly to labels that are given to pictures in weakly supervised approach in computer vision.

The meta-learner will incorporate the spatial and temporal context of the land cover / land characteristic data cube. This will allow it to simultaneously detect the spatial context and monthly dynamics that are necessary to correctly characterize, for example, different types of urban environments (suburban, metropolitan, urban green) and agricultural practices, respectively. There is overwhelming evidence [8, 24, 228, 306, 307] that Deep Learning (DL) architectures are at the forefront of these innovations, and is therefore the most likely branch of machine learning models to be used at this step. This is a currently rapidly developing field, especially in the context of crop classification [8]. For example, using 3D convolutions [307], sequential recurrent encoders [228] have recently shown excellent results. Deep learning methods are powerful but complex to explain and understand by humans, and therefore not well-trusted by scientists and policy-makers. However, by purely relying on conceptually straight-forward land cover and land characteristic estimations, the decisions of such a model remain easy to interpret and evaluate [71].

Global Foundations, local knowledge

This model is then trained on a large diverse set of collected open data points from around the world, and used to create annual LULC maps at unprecedented thematic detail. This will be a computationally intensive process, and it would be prohibitively costly for many users to reproduce. While the detailed baseline map produced by the base model may be enough for many use cases, investigations into niche classes might need even more detail.

However, thanks to a technique called *transfer learning*, large, expensive, well-trained models can be fine-tuned for a new task, using a small set of samples [94, 276] from a study area, whether it be a survey site, province, country, or continent. This will allow scientists to use small, fast, cheap data collection efforts in their own legend to get accurate and meaningful data. Retraining is much cheaper and faster, so it saves time, money, and effort. The current state of the art in this field is Foundation Models: massive models that are pre-trained in a non- or self-supervised way on large amounts of data, recognizing patterns and establishing connections through many training cycles. They are then re-trained for a fraction of the time and cost, and have been found to outperform single-purpose models in many fields, from medicine [176] and cell biology [45] to meteorology [179]. While foundation models are an extremely novel concept outside of computer vision, the Prithvi foundation model shows promising results on geographical tasks: after extensive unsupervised pre-training on Harmonized Landsat-Sentinel 2 (HLS) data [41], it needed relatively little training data and time to be fine-tuned to perform well on various tasks, including flood mapping and crop segmentation [127] and query-based remote sensing image retrieval [17]. Whether a foundational model is truly the answer to geographical science's questions remains to be seen, as relying on a single model for many tasks makes all these tasks vulnerable to flaws and biases in the foundation [18]. The advantage, however, is that such a foundation model can be published as open source, or integrated in cloud-based computing efforts, such as openEO [196], to ensure equal opportunity access to users across the globe, not just organizations with access to powerful computation.

In this way, anyone with a question could supply the system with annotated points from their use case and specify the classes they are interested in. This would allow local communities to leverage the computational power of institutional research.

References

[1] E. Alcaras, D. Costantino, F. Guastaferro, C. Parente, and M. Pepe. “Normalized Burn Ratio Plus (NBR+): A New Index for Sentinel-2 Imagery”. *Remote Sensing* 14.7 (2022). DOI: 10.3390/rs14071727.

[2] A. Alvera Azcarate, A. Barth, M. Rixen, and J.-M. Beckers. “Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature”. English. *Ocean Modelling* 9.4 (2005). DOI: 10.1016/j.ocemod.2004.08.001.

[3] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. *Prediction-Powered Inference*. 2023. arXiv: 2301.09633 [cs, q-bio, stat]. (Visited on 2023).

[4] O. Arino, J. J. Ramos Perez, V. Kalogirou, S. Bontemps, P. Defourny, and E. Van Bogaert. *Global land cover map for 2009 (GlobCover 2009)*. European Space Agency (ESA) & Université catholique de Louvain (UCL), 2012.

[5] L. Aune-Lundberg and G.-H. Strand. “The content and accuracy of the CORINE Land Cover dataset for Norway”. *International Journal of Applied Earth Observation and Geoinformation* 96 (2021), 102266.

[6] V. Avitabile, M. Schultz, N. Herold, S. de Bruin, A. K. Pratihast, C. P. Manh, H. V. Quang, and M. Herold. “Carbon Emissions from Land Cover Change in Central Vietnam”. *Carbon Management* 7.5-6 (2016), 333–346. DOI: 10.1080/17583004.2016.1254009. (Visited on 2023).

[7] A. Banskota, N. Kayastha, M. J. Falkowski, M. A. Wulder, R. E. Froese, and J. C. White. “Forest monitoring using Landsat time series data: A review”. *Canadian Journal of Remote Sensing* 40.5 (2014), 362–384.

[8] V. Barriere, M. Claverie, M. Schneider, G. Lemoine, and R. d’Andrimont. *Boosting Crop Classification by Hierarchically Fusing Satellite, Rotational, and Contextual Data*. 2023. arXiv: 2305.12011 [cs.CV].

[9] C. Barrington-Leigh and A. Millard-Ball. “The world’s user-generated road map is more than 80% complete”. *PloS one* 12.8 (2017), e0180698.

[10] F. Batista e Silva, C. Lavalle, and E. Koomen. "A procedure to obtain a refined European land use/cover map". *Journal of Land Use Science* 8.3 (2013), 255–283.

[11] P. Benevides, N. Silva, H. Costa, H. Costa, F. D. Moreira, F. Moreira, D. Moraes, M. Castelli, Ü. Halik, and M. Caetano. "Land cover mapping at national scale with Sentinel-2 and LUCAS: a case study in Portugal". *Remote Sensing* (2021). DOI: 10.1117/12.2598789.

[12] J. Bengtsson, J. Bullock, B. Egoh, C. Everson, T. Everson, T. O'connor, P. O'farrell, H. Smith, and R. Lindborg. "Grasslands—more important for ecosystem services than you might think". *Ecosphere* 10.2 (2019), e02582.

[13] K. Berger, M. Machwitz, M. Kycko, S. C. Kefauver, S. Van Wittenberghe, M. Gerhards, J. Verrelst, C. Atzberger, C. van der Tol, A. Damm, U. Rascher, I. Herrmann, V. S. Paz, S. Fahrner, R. Pieruschka, E. Prikaziuk, M. L. Buchaillot, A. Halabuk, M. Celesti, G. Koren, E. T. Gormus, M. Rossini, M. Foerster, B. Siegmann, A. Abdelbaki, G. Tagliabue, T. Hank, R. Darvishzadeh, H. Aasen, M. Garcia, I. Pôças, S. Bandopadhyay, M. Sulis, E. Tomelleri, O. Rozenstein, L. Filchev, G. Stancile, and M. Schlerf. "Multi-Sensor Spectral Synergies for Crop Stress Detection and Monitoring in the Optical Domain: A Review". *Remote Sensing of Environment* 280 (2022), 113198. DOI: 10.1016/j.rse.2022.113198.

[14] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. "Fairness in criminal justice risk assessments: The state of the art". *Sociological Methods & Research* 50.1 (2021), 3–44.

[15] P. Bhugra, B. Bischke, C. Werner, R. Syrnicki, C. Packbier, P. Helber, C. Senaras, A. S. Rana, T. Davis, W. De Keersmaecker, D. Zanaga, A. Wania, R. Van de Kerchove, and G. Marchisio. "Rapidai4Eo: Mono-and Multi-Temporal Deep Learning Models for Updating the Corine land Cover Product". In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium.* 2022, 2247–2250. DOI: 10.1109/IGARSS46834.2022.9883198.

[16] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones. "mlr: Machine Learning in R". *The Journal of Machine Learning Research* 17.1 (2016), 5938–5942.

[17] B. Blumenstiel, V. Moor, R. Kienzler, and T. Brunschwiler. "Multi-Spectral Remote Sensing Image Retrieval Using Geospatial Foundation Models". *arXiv preprint arXiv:2403.02059* (2024).

[18] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. "On the opportunities and risks of foundation models". *arXiv preprint arXiv:2108.07258* (2021).

[19] C. Bonannella, T. Hengl, J. Heisig, L. Parente, M. N. Wright, M. Herold, and S. de Bruin. "Forest tree species distribution for Europe 2000-2020: mapping potential

and realized distributions using spatiotemporal Machine Learning". *PeerJ* 10 (2022), e13728. DOI: 10.7717/peerj.13728.

[20] C. Bonannella, T. Hengl, L. Parente, and S. de Bruin. "Biomes of the world under climate change scenarios: increasing aridity and higher temperatures lead to significant shifts in natural vegetation". *PeerJ* 11 (2023), e15593. DOI: https://doi.org/10.7717/peerj.15593.

[21] H. Boogaard, A. Pratihast, J. C. Laso Bayas, S. Karanam, S. Fritz, K. V. Tricht, J. Degerickx, and S. Gilliams. "WorldCereal open global harmonized reference data repository (CC-BY licensed data sets)" (2023).

[22] M. Bossard, J. Feranec, J. Otahel, et al. *CORINE land cover technical guide: Addendum 2000*. Vol. 40. Copenhagen: European Environment Agency, 2000.

[23] L. Breiman. "Random forests". *Machine learning* 45.1 (2001), 5–32.

[24] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. B. Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwer, M. Weisse, F. Stolle, C. Hanson, O. Guinan, R. Moore, and A. M. Tait. "Dynamic World, Near real-time global 10 m land use land cover mapping". *Scientific Data* 9.1 (2022), 1–17. DOI: 10.1038/s41597-022-01307-4.

[25] D. J. Brus, G. M. Hengeveld, D. J. J. Walvoort, P. W. Goedhart, A. H. Heidema, G. J. Nabuurs, and K. Gunia. "Statistical mapping of tree species over Europe". en. *European Journal of Forest Research* 131.1 (2012), 145–157. DOI: 10.1007/s10342-011-0513-5. (Visited on 2023).

[26] A. Bryn, G.-H. Strand, M. Angeloff, and Y. Rekdal. "Land Cover in Norway Based on an Area Frame Survey of Vegetation Types". *Norsk Geografisk Tidsskrift - Norwegian Journal of Geography* 72.3 (2018), 131–145. DOI: 10.1080/00291951.2018.1468356. (Visited on 2023).

[27] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets. "Copernicus global land cover layers – Collection 2". *Remote Sensing* 12.6 (2020), 1044.

[28] O. Buck, C. Haub, S. Woditsch, M. Lindemann, L. Kleinwillinghöfer, G. Hazeu, B. Kosztra, S. Kleeschulte, S. Arnold, and M. Hölzl. *Analysis of the LUCAS nomenclature and proposal for adaptation of the nomenclature in view of its use by the Copernicus land monitoring services*. Technical Report, European Environment Agency and the European Environment Information and Observation Network. Copenhagen: EEA - European Environment Agency, 2015.

[29] G. Büttner. "CORINE land cover and land cover change products". In: *Land use and land cover mapping in Europe*. Springer, 2014, 55–74.

[30] J. Buus-Hinkler, B. U. Hansen, M. P. Tamstorf, and S. B. Pedersen. "Snow-vegetation relations in a High Arctic ecosystem: Inter-annual variability inferred

from new monitoring and modeling concepts". *Remote Sensing of Environment* 105.3 (2006), 237–247.

[31] A. Buyantuyev and J. Wu. "Effects of thematic resolution on landscape pattern analysis". *Landscape Ecology* 22.1 (2007), 7–13.

[32] M. Calderón-Loor, M. Hadjikakou, and B. A. Bryan. "High-resolution wall-to-wall land-cover mapping and land change assessment for Australia from 1985 to 2015". *Remote Sensing of Environment* 252 (2021), 112148.

[33] G. Castilla, K. Larkin, J. Linke, and G. J. Hay. "The impact of thematic resolution on the patch-mosaic model of natural landscapes". *Landscape Ecology* 24.1 (2009), 15–23.

[34] R. Cazzolla Gatti, P. B. Reich, J. G. Gamarra, T. Crowther, C. Hui, A. Morera, J.-F. Bastin, S. De-Miguel, G.-J. Nabuurs, J.-C. Svenning, et al. "The number of tree species on Earth". *Proceedings of the National Academy of Sciences* 119.6 (2022), e2115329119.

[35] G. Ceccherini, G. Duveiller, G. Grassi, G. Lemoine, V. Avitabile, R. Pilli, and A. Cescatti. "Abrupt increase in harvested forest area over Europe after 2015". *Nature* 583.7814 (2020), 72–77.

[36] B. Chatenoux, J.-P. Richard, D. Small, C. Roeoesli, V. Wingate, C. Poussin, D. Rodila, P. Peduzzi, C. Steinmeier, C. Ginzler, A. Psomas, M. E. Schaepman, and G. Giuliani. "The Swiss data cube, analysis ready data archive using earth observations of Switzerland". *Scientific data* 8.1 (2021), 1–11. DOI: 10.1038/s41597-021-01076-6.

[37] I. Chen, F. D. Johansson, and D. Sontag. "Why is my classifier discriminatory?" *Advances in neural information processing systems* 31 (2018).

[38] J. Chen, R. John, J. Yuan, E. A. Mack, P. Groisman, G. Allington, J. Wu, P. Fan, K. M. De Beurs, A. Karnieli, et al. "Sustainability challenges for the social-environmental systems across the Asian Drylands Belt". *Environmental Research Letters* 17.2 (2022), 023001.

[39] J. Chen, J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, et al. "Global land cover mapping at 30 m resolution: A POK-based operational approach". *ISPRS Journal of Photogrammetry and Remote Sensing* 103 (2015), 7–27.

[40] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 2016, 785–794.

[41] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, and C. Justice. "The Harmonized Landsat and Sentinel-2 surface reflectance data set". *Remote sensing of environment* 219 (2018), 145–161.

[42] T. Conway. “The impact of class resolution in land use change models”. *Computers, Environment and Urban Systems* 33.4 (2009), 269–277.

[43] R. Costanza et al. “The value of the world’s ecosystem services and natural capital”. *World Environ.* 25 (1997), 3–15.

[44] R. Costanza et al. “Changes in the global value of ecosystem services”. *Global Environ. Chang.* 26 (2014), 152–158.

[45] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. “scGPT: toward building a foundation model for single-cell multi-omics using generative AI”. *Nature Methods* (2024), 1–11.

[46] R. d’Andrimont, A. Verhegghen, M. Meroni, G. Lemoine, P. Strobl, B. Eiselt, M. Yordanov, L. Martinez-Sanchez, and M. van der Velde. “LUCAS Copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the European Union”. *Earth System Science Data* 13.3 (2021), 1119–1133.

[47] R. d’Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eiselt, A. Palmieri, P. Dominici, J. Gallego, H. I. Reuter, C. Joebges, G. Lemoine, and M. van der Velde. “Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union”. *Scientific Data* 7.1 (2020), 1–15. DOI: 10.1038/s41597-020-00675-z.

[48] R. d’Andrimont, A. Verhegghen, G. Lemoine, P. Kempeneers, M. Meroni, and M. Van Der Velde. “From parcel to continental scale–A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations”. *Remote sensing of environment* 266 (2021), 112708.

[49] R. d’Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eiselt, A. Palmieri, P. Dominici, J. Gallego, H. I. Reuter, C. Joebges, G. Lemoine, et al. “Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union”. *Scientific data* 7.1 (2020), 352.

[50] N. Davidson. “How much wetland has the world lost? Long-term and recent trends in global wetland area”. *Mar. Freshw. Res.* 65 (2014), 936–941.

[51] A. Defazio, F. Bach, and S. Lacoste-Julien. “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives”. *Advances in neural information processing systems* (2014), 1646–1654.

[52] P. Defourny, G. Kirches, C. Brockmann, M. Boettcher, M. Peters, S. Bontemps, C. Lamarche, M. Schlerf, and M. Santoro. *Land cover CCI: Product User Guide Version 2.4.* Vol. 2. 325. European Space Agency (ESA), UCL-Geomatics, 2012, 10–1016.

[53] A. Demirkaya, J. Chen, and S. Oymak. “Exploring the Role of Loss Functions in Multiclass Classification”. In: *2020 54th Annual Conference on Information Sciences and Systems (CISS).* 2020, 1–5. DOI: 10.1109/CISS48834.2020.1570627167.

[54] U. N. S. Division. *Global Indicator Framework for the Sustainable Development Goals and Targets of the 2030 Agenda for Sustainable Development.* 2023.

[55] J. A. Dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falao. “Multiscale classification of remote sensing images”. *IEEE Transactions on Geoscience and Remote Sensing* 50.10 (2012), 3764–3775.

[56] J. Dressel and H. Farid. “The accuracy, fairness, and limits of predicting recidivism”. *Science advances* 4.1 (2018), eaao5580.

[57] D. Duarte, C. Fonte, H. Costa, and M. Caetano. “Thematic Comparison between ESA WorldCover 2020 Land Cover Product and a National Land Use Land Cover Map”. *Land* 12.2 (2023), 490. DOI: `10.3390/land12020490`.

[58] P. Dugan. *Wetlands in Danger—A World Conservation Atlas.* New York, NY, USA: Oxford University Press, 1993.

[59] G. Duveiller, L. Caporaso, R. Abad-Viñas, L. Perugini, G. Grassi, A. Arneth, and A. Cescatti. “Local biophysical effects of land use and land cover change: towards an assessment tool for policy makers”. *Land Use Policy* 91 (2020), 104–382. DOI: `10.1016/j.landusepol.2019.104382`.

[60] ESA. *Copernicus DEM - Global and European Digital Elevation Model (COP-DEM).* DLR e.V. 2010–2014 and ©Airbus Defense and Space GmbH 20140–2018. Provided under COPERNICUS by the European Union and ESA; all rights reserved. DLR, 2018. DOI: `10.5270/ESA-c5d3d65`.

[61] T. Esch, A. Metz, M. Marconcini, and M. Keil. “Differentiation of crop types and grassland by multi-scale analysis of seasonal satellite data”. *Land Use and Land Cover Mapping in Europe: Practices & Trends* (2014), 329–339.

[62] *Estimating area, area change and their uncertainties.* 2022. (Visited on 2022).

[63] European Commission, Eurostat. *Land Use/Cover Area frame Survey (LUCAS) – Eurostat Database.* `https://ec.europa.eu/eurostat/web/lucas/database`. Accessed: 2024-01-05. 2024.

[64] Eurostat. *Land cover statistics.* `https://ec.europa.eu/eurostat/statistics-explained/index.php/Land_cover_statistics`. Accessed on 22 Feb 2024. 2021.

[65] Eurostat. *Agricultural production - crops.* `https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops`. Accessed on 22 Feb 2024. 2023.

[66] J. Fan, X. Wang, W. Wu, W. Chen, Q. Ma, and Z. Ma. “Function of restored wetlands for waterbird conservation in the Yellow Sea coast”. *Science of the Total Environment* 756 (2021), 144061.

[67] FAO. *The State of the World’s Forests 2022. Forest pathways for green recovery and building inclusive, resilient and sustainable economies.* Rome, FAO, 2022.

[68] M. Feng and Y. Bai. “A global land cover map produced through integrating multi-source datasets”. *Big Earth Data* 3.3 (2019), 191–219.

[69] J. Feranec, G. Jaffrain, T. Soukup, and G. Hazeu. “Determining changes and flows in European landscapes 1990–2000 using CORINE land cover data”. *Applied geography* 30.1 (2010), 19–35.

[70] J. Feranec, T. Soukup, G. Hazeu, and G. Jaffrain. *European landscape dynamics: CORINE land cover data.* CRC Press, 2016.

[71] A. Ferchichi, A. B. Abbes, V. Barra, and I. R. Farah. “Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review”. *Ecological Informatics* 68 (2022), 101552.

[72] Y. Finegold Antonia and A. Ortmann. *Map Accuracy Assessment and Area Estimation: A Practical Guide.* en. National Forest Monitoring and Assessment Working Paper 46. Rome, Italy: FAO, 2016. (Visited on 2023).

[73] P. Fisher, A. J. Comber, and R. Wadsworth. “Land use and land cover: contradiction or complement”. *Re-presenting GIS* 85 (2005), 98.

[74] E. Fitoka, M. Tompoulidou, L. Hatziiordanou, A. Apostolakis, R. Höfer, K. Weise, and C. Ververis. “Water-related ecosystems’ mapping and assessment based on remote sensing techniques and geospatial analysis: The SWOS national service case of the Greek Ramsar sites and their catchments”. *Remote Sensing of Environment* 245 (2020), 111795.

[75] J. A. Foley, R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, et al. “Global consequences of land use”. *Science* 309.5734 (2005), 570–574.

[76] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd. “Training Set Size Requirements for the Classification of a Specific Class”. *Remote Sensing of Environment* 104.1 (2006), 1–14. DOI: 10.1016/j.rse.2006.03.004.

[77] D. Frantz. “FORCE—Landsat+ Sentinel-2 analysis ready data and beyond”. *Remote Sensing* 11.9 (2019), 1124. DOI: 10.3390/rs11091124.

[78] S. Fritz, I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. Van der Velde, F. Kraxner, and M. Obersteiner. “Geo-Wiki: An online platform for improving global land cover”. *Environmental Modelling & Software* 31 (2012), 110–123.

[79] F. J. Gallego. “Remote sensing and land cover area estimation”. en. *International Journal of Remote Sensing* 25.15 (2004), 3019–3047. DOI: 10.1080/01431160310001619607. (Visited on 2023).

[80] J. Gallego. *Copernicus land services to improve EU statistics.* eng. LU: Publications Office of the European Union, 2017. (Visited on 2023).

[81] J. Gallego and C. Bamps. "Using CORINE land cover and the point survey LUCAS for area estimation". *International Journal of Applied Earth Observation and Geoinformation* 10.4 (2008), 467–475.

[82] B.-C. Gao. "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space". *Remote sensing of environment* 58.3 (1996), 257–266.

[83] Y. Gao, L. Liu, X. Zhang, X. Chen, J. Mi, and S. Xie. "Consistency analysis and accuracy assessment of three global 30-m land-cover products over the European Union using the LUCAS dataset". *Remote Sensing* 12.21 (2020), 3479.

[84] C. García, O. Mora, F. Pérez-Aragüés, and J. Vitrià. "CatLC: Catalonia Multiresolution Land Cover Dataset". *Scientific Data* 9.1 (2022), 554. DOI: 10.1038/s41597-022-01674-y. (Visited on 2024).

[85] M. L. García and V. Caselles. "Mapping burns and natural reforestation using Thematic Mapper data". *Geocarto International* 6.1 (1991), 31–37.

[86] D. García-Álvarez, J. Lara Hinojosa, F. J. Jurado Pérez, and J. Quintero Villaraso. "Global General Land Use Cover Datasets with a Time Series of Maps". In: *Land Use Cover Datasets and Validation Tools: Validation Practices with QGIS*. Springer International Publishing Cham, 2022, 287–311. DOI: 10.1007/978-3-030-90998-7_15.

[87] R. Gardner and M. Finlayson. *Global Wetland Outlook: State of the World's Wetlands and Their Services to People*. Ramsar Convention. Gland, Switzerland, 2018.

[88] R. Gardner et al. *State of the World's Wetlands and Their Services to People: A Compilation of Recent Analyses*. Social Science Electronic Publishing. Gland, Switzerland, 2015.

[89] A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[90] R. Ghorbani and R. Ghousi. "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques". *IEEE Access* 8 (2020). Conference Name: IEEE Access, 67899–67911. DOI: 10.1109/ACCESS.2020.2986809.

[91] R. Giblett. *Postmodern Wetlands: Culture, History, Ecology*. Edinburgh, UK: Edinburgh University Press, 1996.

[92] G. Giuliani, B. Chatenoux, A. De Bono, D. Rodila, J.-P. Richard, K. Allenbach, H. Dao, and P. Peduzzi. "Building an earth observations data cube: lessons learned from the Swiss data cube (SDC) on generating analysis ready data (ARD)". *Big Earth Data* 1.1-2 (2017), 100–117. DOI: 10.1080/20964471.2017.1398903.

[93] G. Giuliani, B. Chatenoux, T. Piller, F. Moser, and P. Lacroix. "Data Cube on Demand (DCoD): Generating an earth observation Data Cube anywhere in the world". *International Journal of Applied Earth Observation and Geoinformation* 87 (2020), 102035. DOI: 10.1016/j.jag.2019.102035.

[94] Y. Hamrouni, E. Paillassa, V. Chéret, C. Monteil, and D. Sheeren. "From local to global: A transfer learning-based approach for mapping poplar plantations at national scale using Sentinel-2". *ISPRS Journal of Photogrammetry and Remote Sensing* 171 (2021), 76–100.

[95] J. T. Hancock and T. M. Khoshgoftaar. "CatBoost for big data: an interdisciplinary review". *Journal of big data* 7.1 (2020), 1–45.

[96] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. Stehman, S. J. Goetz, T. R. Loveland, et al. "High-resolution global maps of 21st-century forest cover change". *Science* 342.6160 (2013), 850–853.

[97] M. C. Hansen, P. V. Potapov, A. H. Pickens, A. Tyukavina, A. Hernandez-Serna, V. Zalles, S. Turubanova, I. Kommareddy, S. V. Stehman, X.-P. Song, and A. Kommareddy. "Global land use extent and dispersion within natural land cover using Landsat data". *Environmental Research Letters* 17.3 (2022), 034050.

[98] K. L. Harper, C. Lamarche, A. Hartley, P. Peylin, C. Ottlé, V. Bastrikov, R. San Martín, S. I. Bohnenstengel, G. Kirches, M. Boettcher, et al. "A 29-year time series of annual 300-metre resolution plant functional type maps for climate models". *Earth System Science Data Discussions* 2022 (2022), 1–37.

[99] L. Hawker, P. Uhe, L. Paulo, J. Sosa, J. Savage, C. Sampson, and J. Neal. "A 30 m global map of elevation with forests and buildings removed". *Environmental Research Letters* 17.2 (2022), 024016. DOI: 10.1088/1748-9326/ac4d4f.

[100] H. He and E. A. Garcia. "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), 1263–1284. DOI: 10.1109/TKDE.2008.239.

[101] T. Hengl, L. Leal Parente, J. Krizan, and C. Bonannella. *Continental Europe Digital Terrain Model at 30 m resolution based on GEDI, ICESat-2, AW3D, GLO-30, EUDEM, MERIT DEM and background layers.* Zenodo, 2020. DOI: 10.5281/zenodo.4724549.

[102] T. Hengl, L. Leal Parente, J. Krizan, and C. Bonannella. *Continental Europe Digital Terrain Model at 30 m Resolution Based on GEDI, ICESat-2, AW3D, GLO-30, EUDEM, MERIT DEM and Background Layers.* 2021. DOI: 10.5281/zenodo.4724549.

[103] T. Hengl, L. Leal Parente, J. Križan, and C. Bonannella. *Continental Europe digital terrain model at 30 m resolution based on GEDI, ICESat-2, AW3D, GLO-30,*

EUDEM, MERIT DEM and background layers. Zenodo, 2021. DOI: 10.5281/zenodo.4724549.

[104] T. Hengl, J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, et al. "SoilGrids250m: Global gridded soil information based on machine learning". *PLoS one* 12.2 (2017), e0169748.

[105] A. Hennessy, K. Clarke, and M. Lewis. "Hyperspectral Classification of Plants: A Review of Waveband Selection Generalisability". *Remote Sensing* 12.1 (2020). DOI: 10.3390/rs12010113.

[106] M. Herold, P. Mayaux, C. Woodcock, A. Baccini, and C. Schmullius. "Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets". *Remote Sensing of Environment* 112.5 (2008), 2538–2556.

[107] M. Herold, P. Mayaux, C. E. Woodcock, A. Baccini, and C. Schmullius. "Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets". *Remote Sensing of Environment*. Earth Observations for Terrestrial Biodiversity and Ecosystems Special Issue 112.5 (2008), 2538–2556. DOI: 10.1016/j.rse.2007.11.013. (Visited on 2023).

[108] M. Herold, L. See, N.-E. Tsendbazar, and S. Fritz. "Towards an Integrated Global Land Cover Monitoring and Mapping System". en. *Remote Sensing* 8.12 (2016), 1036. DOI: 10.3390/rs8121036. (Visited on 2023).

[109] D. Hillger, T. Kopp, T. Lee, D. Lindsey, C. Seaman, S. Miller, J. Solbrig, S. Kidder, S. Bachmeier, T. Jasmin, et al. "First-light imagery from Suomi NPP VIIRS". *Bulletin of the American Meteorological Society* 94.7 (2013), 1019–1029.

[110] C. Homer, J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, J. Wickham, et al. "Completion of the 2001 national land cover database for the counterminous United States". *Photogrammetric engineering and remote sensing* 73.4 (2007), 337.

[111] C. Homer, J. Dewitz, S. Jin, G. Xian, C. Costello, P. Danielson, L. Gass, M. Funk, J. Wickham, S. Stehman, et al. "Conterminous United States land cover change patterns 2001–2016 from the 2016 national land cover database". *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020), 184–199.

[112] C. Hong, J. A. Burney, J. Pongratz, J. E. Nabel, N. D. Mueller, R. B. Jackson, and S. J. Davis. "Global and regional drivers of land-use emissions in 1961–2017". *Nature* 589.7843 (2021), 554–561.

[113] P. Horvath, R. Halvorsen, T. Simensen, and A. Bryn. "A comparison of three ways to assemble wall-to-wall maps from distribution models of vegetation types". *GIScience & Remote Sensing* 58.8 (2021). Publisher: Taylor & Francis

_eprint: https://doi.org/10.1080/15481603.2021.1996313, 1458–1476. DOI: 10.1080/15481603.2021.1996313. (Visited on 2023).

[114] P. Horvath, R. Halvorsen, F. Stordal, L. M. Tallaksen, H. Tang, and A. Bryn. "Distribution Modelling of Vegetation Types Based on Area Frame Survey Data". *Applied Vegetation Science* 22.4 (2019), 547–560. DOI: 10.1111/avsc.12451. (Visited on 2023).

[115] B. Hosseiny, A. M. Abdi, and S. Jamali. "Urban land use and land cover classification with interpretable machine learning–A case study using Sentinel-2 and auxiliary data". *Remote Sensing Applications: Society and Environment* 28 (2022), 100843.

[116] R. A. Houghton, J. I. House, J. Pongratz, G. R. Van Der Werf, R. S. DeFries, M. C. Hansen, C. L. Quéré, and N. Ramankutty. "Carbon emissions from land use and land-cover change". *Biogeosciences* 9.12 (2012), 5125–5142.

[117] S. Hu, Z. Niu, Y. Chen, L. Li, and H. Zhang. "Global wetlands: Potential distribution, wetland loss, and status". *Sci. Total Environ.* 586 (2017), 319–327.

[118] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang. "Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, 7014–7023. DOI: 10.1109/CVPR.2018.00733. (Visited on 2024).

[119] A. Huete. "A soil-adjusted vegetation index (SAVI)". *Remote Sensing of Environment* 25.3 (1988), 295–309. DOI: https://doi.org/10.1016/0034-4257(88)90106-X.

[120] A. R. Huete. "A soil-adjusted vegetation index (SAVI)". *Remote sensing of environment* 25.3 (1988), 295–309.

[121] P. Hurskainen, H. Adhikari, M. Siljander, P. Pellikka, and A. Hemp. "Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes". *Remote sensing of environment* 233 (2019), 111354.

[122] S. Ibrahim, M. Landa, O. Pešek, L. Brodskỳ, and L. Halounová. "Machine Learning-Based Approach Using Open Data to Estimate PM2.5 over Europe". *Remote Sensing* 14.14 (2022), 3392. DOI: 10.3390/rs14143392.

[123] S. Ibrahim, M. Landa, O. Pešek, K. Pavelka, and L. Halounova. "Space-time machine learning models to analyze COVID-19 pandemic lockdown effects on aerosol optical depth over Europe". *Remote Sensing* 13.15 (2021), 3027. DOI: 10.3390/rs13153027.

[124] M. Immitzer, M. Neuwirth, S. Böck, H. Brenner, F. Vuolo, and C. Atzberger. "Optimal Input Features for Tree Species Classification in Central Europe Based on Multi-Temporal Sentinel-2 Data". *Remote Sensing* 11.22 (2019). DOI: 10.3390/rs11222599.

[125] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. "Operational high resolution land cover map production at the country scale using satellite image time series". *Remote Sensing* 9.1 (2017), 95.

[126] IPCC. "Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen". *Cambridge University Press* In Press (2021).

[127] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, et al. "Foundation models for generalist geospatial artificial intelligence". *arXiv preprint arXiv:2310.18660* (2023).

[128] L. L. Janssen and H. Middelkoop. "Knowledge-based crop classification of a Landsat Thematic Mapper image". *International Journal of Remote Sensing* 13.15 (1992), 2827–2837.

[129] M. Jenerowicz, E. Krätzschmar, P. Schauer, E. Gromny, R. Malinowski, M. Krupiński, S. Lewiński, M. Rybicki, and C. Wojtkowski. *Validation dataset for Land Cover Map of Europe 2017.* data set. 2021. DOI: 10.1594/PANGAEA.934197.

[130] S. Jin and S. A. Sader. "Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances". *Remote sensing of Environment* 94.3 (2005), 364–372.

[131] E. Józsa, S. Á. Fábián, and M. Kovács. "An evaluation of EU-DEM in comparison with ASTER GDEM, SRTM and contour-based DEMs over the Eastern Mecsek Mountains". *Hungarian Geographical Bulletin* 63.4 (2014), 401–423. DOI: 10.15201/hungeobull.63.4.3.

[132] F. Kamiran and T. Calders. "Data preprocessing techniques for classification without discrimination". *Knowledge and information systems* 33.1 (2012), 1–33.

[133] J. O. Kaplan, K. M. Krumhardt, E. C. Ellis, W. F. Ruddiman, C. Lemmen, and K. K. Goldewijk. "Holocene carbon emissions as a result of anthropogenic land cover change". *The Holocene* 21.5 (2011), 775–791.

[134] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby. "Global Land Use / Land Cover with Sentinel 2 and Deep Learning". In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS.* 2021, 4704–4707. DOI: 10.1109/IGARSS47720.2021.9553499. (Visited on 2024).

[135] C. H. Key and N. C. Benson. "Measuring and remote sensing of burn severity". In: *Proceedings joint fire science conference and workshop.* Vol. 2. University of Idaho and International Association of Wildland Fire. Moscow, ID, 1999, 284.

[136] C. H. Key and N. C. Benson. "Landscape assessment (LA)". In: *FIREMON: Fire effects monitoring and inventory system. Gen. Tech. Rep. RMRS-GTR-164-CD.*

Vol. 164. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station, 2006.

[137] M. Kilibarda, T. Hengl, G. B. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić, and B. Bajat. "Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution". *Journal of Geophysical Research: Atmospheres* 119.5 (2014), 2294–2313.

[138] L. Kleinewillinghöfer, P. Olofsson, E. Pebesma, H. Meyer, O. Buck, C. Haub, and B. Eiselt. "Unbiased Area Estimation Using Copernicus High Resolution Layers and Reference Data". en. *Remote Sensing* 14.19 (2022). Number: 19 Publisher: Multidisciplinary Digital Publishing Institute, 4903. DOI: 10.3390/rs14194903. (Visited on 2023).

[139] K. T. Korhonen. *A new article in the journal Nature overestimates the increase of forest harvesting in Europe.* Accessed: 2021-07-04. Helsinki, Finland: Natural Resources Institute Finland, 2020.

[140] P. Koshute, J. Zook, and I. McCulloh. *Recommending Training Set Sizes for Classification.* 2021. DOI: 10.48550/arXiv.2102.09382. arXiv: 2102.09382 [cs]. (Visited on 2023).

[141] C. Krekel, J. Kolbe, and H. Wüstemann. "The greener, the happier? The effect of urban land use on residential well-being". *Ecological economics* 121 (2016), 117–127.

[142] M. Landa, L. Brodsky, L. Parente, **M. Witjes**, and T. Hengl. *Multi-year harmonized land cover samples based on LUCAS and CORINE datasets.* Version v0.1. Zenodo, 2021. DOI: 10.5281/zenodo.4740691.

[143] C. Lekka, G. P. Petropoulos, and S. E. Detsikas. "Appraisal of EnMAP hyperspectral imagery use in LULC mapping when combined with machine learning pixel-based classifiers". *Environmental Modelling & Software* 173 (2024), 105956.

[144] M. Lesiv, M. Dürauer, I. Georgieva, A. Bilous, J. Laso Bayas, and S. Fritz. "Global reference data set for validating ESA WorldCereal temporary cropland extent" (2023).

[145] A. Levering. "Landscape quality assessments using deep learning". PhD thesis. Wageningen University, 2024.

[146] C. Li, Z. Ma, L. Wang, W. Yu, D. Tan, B. Gao, Q. Feng, H. Guo, and Y. Zhao. "Improving the accuracy of land cover mapping by distributing training samples". *Remote Sensing* 13.22 (2021), 4594.

[147] L. Liangyun, G. Yuan, Z. Xiao, C. Xidong, and X. Shuai. *A Dataset of Global Land Cover Validation Samples.* Version v1. Zenodo, 2019. DOI: 10.5281/zenodo.3551995.

[148] H. Liu, P. Gong, J. Wang, N. Clinton, Y. Bai, and S. Liang. “Annual dynamics of global land cover and its long-term changes from 1982 to 2015”. *Earth System Science Data* 12.2 (2020), 1217–1243.

[149] H. Liu, P. Gong, J. Wang, X. Wang, G. Ning, and B. Xu. “Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020-iMap World 1.0”. *Remote Sensing of Environment* 258 (2021), 112364. DOI: `10.1016/j.rse.2021.112364`.

[150] H. Q. Liu and A. Huete. “A feedback based modification of the NDVI to minimize canopy background and atmospheric noise”. *IEEE transactions on geoscience and remote sensing* 33.2 (1995), 457–465.

[151] L. Liu, X. Zhang, Y. Gao, X. Chen, X. Shuai, and J. Mi. “Finer-Resolution Mapping of Global Land Cover: Recent Developments, Consistency Analysis, and Prospects”. *Journal of Remote Sensing* 2021 (2021).

[152] X. Liu, S. Trogisch, J.-S. He, P. A. Niklaus, H. Bruelheide, Z. Tang, A. Erfmeier, M. Scherer-Lorenzen, K. A. Pietsch, B. Yang, et al. “Tree species richness increases ecosystem carbon storage in subtropical forests”. *Proceedings of the Royal Society B* 285.1885 (2018), 12–40.

[153] Y. Liu, X. Hou, X. Li, B. Song, and C. Wang. “Assessing and predicting changes in ecosystem service values based on land use/cover change in the Bohai Rim coastal zone”. *Ecological Indicators* 111 (2020), 106004.

[154] R. Lovelace, J. Nowosad, and J. Muenchow. *Geocomputation with R*. Chapman & Hall/CRC The R Series. CRC Press, 2019.

[155] F. Löw, U. Michel, S. Dech, and C. Conrad. “Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines”. *ISPRS journal of photogrammetry and remote sensing* 85 (2013), 102–119.

[156] M. Lu, M. Appel, and E. Pebesma. “Multidimensional arrays for analysing geoscientific data”. *ISPRS International Journal of Geo-Information* 7.8 (2018), 313. DOI: `10.3390/ijgi7080313`.

[157] R. Lucas, N. Mueller, A. Siggins, C. Owers, D. Clewley, P. Bunting, C. Kooymans, B. Tissott, B. Lewis, L. Lymburner, and G. Metternicht. “Land cover mapping using digital earth Australia”. *Data* 4.4 (2019), 143. DOI: `10.3390/data4040143`.

[158] Y. Luo, Z. Zhang, L. Zhang, J. Han, J. Cao, and J. Zhang. “Developing high-resolution crop maps for major crops in the european union based on transductive transfer learning and limited ground data”. *Remote Sensing* 14.8 (2022), 1809.

[159] Y. Luo, K. Guan, J. Peng, S. Wang, and Y. Huang. “STAIR 2.0: A Generic and Automatic Algorithm to Fuse Modis, Landsat, and Sentinel-2 to Generate 10 m,

Daily, and Cloud-/Gap-Free Surface Reflectance Product". *Remote Sensing* 12.19 (2020). DOI: 10.3390/rs12193209.

[160] M. D. Mahecha, F. Gans, G. Brandt, R. Christiansen, S. E. Cornell, N. Fomferra, G. Kraemer, J. Peters, P. Bodesheim, G. Camps-Valls, J. F. Donges, W. Dorigo, L. M. Estupinan-Suarez, V. H. Gutierrez-Velez, M. Gutwin, M. Jung, M. C. Londoño, D. G. Miralles, P. Papastefanou, and M. Reichstein. "Earth system data cubes unravel global multivariate dynamics". *Earth System Dynamics* 11.1 (2020), 201–234. DOI: 10.5194/esd-11-201-2020.

[161] R. Malinowski, S. Lewiński, M. Rybicki, E. Gromny, M. Jenerowicz, M. Krupiński, A. Nowakowski, C. Wojtkowski, M. Krupiński, E. Krätzschmar, et al. "Automated production of a land cover/use map of Europe based on Sentinel-2 imagery". *Remote Sensing* 12.21 (2020), 3523.

[162] R. Malinowski, S. Lewiński, M. Rybicki, E. Gromny, M. Jenerowicz, M. Krupiński, A. Nowakowski, C. Wojtkowski, M. Krupiński, E. Krätzschmar, et al. "Automated production of a land cover/use map of Europe based on Sentinel-2 imagery". *Remote Sensing* 12.21 (2020), 3523.

[163] R. K. Mark. *Multidirectional, oblique-weighted, shaded-relief image of the Island of Hawaii.* 92-422. US Geological Survey, 1992. DOI: 10.3133/ofr92422.

[164] "Markets can save forests". *Nature* 452.7184 (2008). DOI: https://doi.org/10.1038/452127a.

[165] R. N. Masolele, D. Marcos, V. De Sy, I.-O. Abu, J. Verbesselt, J. Reiche, and M. Herold. "Mapping the diversity of land uses following deforestation across Africa". *Scientific Reports* 14.1 (2024), 1681.

[166] T. Mastelic, J. Lorincz, I. Ivandic, and M. Boban. "Aerial imagery based on commercial flights as remote sensing platform". *Sensors* 20.6 (2020), 1658.

[167] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". *The bulletin of mathematical biophysics* 5.4 (1943), 115–133.

[168] S. K. McFeeters. "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features". *International journal of remote sensing* 17.7 (1996), 1425–1432.

[169] A. Mellor, S. Boukir, A. Haywood, and S. Jones. "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin". *ISPRS Journal of Photogrammetry and Remote Sensing* 105 (2015), 155–168. DOI: 10.1016/j.isprsjprs.2015.03.014. (Visited on 2023).

[170] H. Meyer and E. Pebesma. "Predicting into unknown space? Estimating the area of applicability of spatial prediction models". *Methods in Ecology and Evolution* 12.9 (2021), 1620–1633.

[171] P. Meyer, P. Janda, M. Mikoláš, V. Trotsiuk, F. Krumm, H. Mrhalová, M. Synek, J. Lábusová, D. Kraus, J. Brandes, et al. "A matter of time: self-regulated tree regeneration in a natural Norway spruce (Picea abies) forest at Mt. Brocken, Germany". *European journal of forest research* 136 (2017), 907–921.

[172] N. Milojevic-Dupont, F. Wagner, F. Nachtigall, J. Hu, G. B. Brüser, M. Zumwald, F. Biljecki, N. Heeren, L. H. Kaack, P.-P. Pichler, et al. "EUBUCCO v0. 1: European building stock characteristics in a common and open database for 200+ million individual buildings". *Scientific Data* 10.1 (2023), 147.

[173] Z. Mingguo, C. Qianguo, and Q. Mingzhou. "The effect of prior probabilities in the maximum likelihood classification on individual classes". *Photogrammetric Engineering & Remote Sensing* 75.9 (2009), 1109–1117.

[174] S. M. Mirmazloumi, M. Kakooei, F. Mohseni, A. Ghorbanian, M. Amani, M. Crosetto, and O. Monserrat. "ELULC-10, a 10 m European Land Use and Land Cover Map Using Sentinel and Landsat Data in Google Earth Engine". *Remote Sensing* 14.13 (2022), 3041. DOI: 10.3390/rs14133041.

[175] M. S. Monmonier. *How to Lie with Maps.* eng. Third. Chicago, IL: The University of Chicago Press, 2018.

[176] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. "Foundation models for generalist medical artificial intelligence". *Nature* 616.7956 (2023), 259–265.

[177] A. Mouratidis and D. Ampatzidis. "European Digital Elevation Model Validation against Extensive Global Navigation Satellite Systems Data and Comparison with SRTM DEM and ASTER GDEM in Central Macedonia (Greece)". *ISPRS International Journal of Geo-Information* 8.3 (2019). DOI: 10.3390/ijgi8030108.

[178] C. Mucher, K. Steinnocher, F. Kressler, and C. Heunks. "Land cover characterization and change detection for environmental monitoring of pan-Europe". *International Journal of Remote Sensing* 21.6-7 (2000), 1159–1181.

[179] S. K. Mukkavilli, D. S. Civitarese, J. Schmude, J. Jakubik, A. Jones, N. Nguyen, C. Phillips, S. Roy, S. Singh, C. Watson, et al. "Ai foundation models for weather and climate: Applications, design, and implementation". *arXiv preprint arXiv:2309.10808* (2023).

[180] R. Müller, S. Kornblith, and G. Hinton. *When Does Label Smoothing Help?* 2020. arXiv: 1906.02629 [cs.LG].

[181] G. Myburgh and A. van Niekerk. "Impact of Training Set Size on Object-Based Land Cover Classification: A Comparison of Three Classifiers". *International Journal of Applied Geospatial Research* 5.3 (2014). MAG ID: 2010669231 S2ID: b9ecc37554e527125e76627138df63d75c488264, 49–67. DOI: 10.4018/ijagr.2014070104.

[182] E. Neinavaz, M. Schlerf, R. Darvishzadeh, M. Gerhards, and A. K. Skidmore. “Thermal infrared remote sensing of vegetation: Current status and perspectives”. *International Journal of Applied Earth Observation and Geoinformation* 102 (2021), 102415.

[183] P. Neis and D. Zielstra. “Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap”. *Future Internet* 6.1 (2014), 76–106. DOI: `10.3390/fi6010076`. (Visited on 2024).

[184] A. Niculescu-Mizil and R. Caruana. “Predicting good probabilities with supervised learning”. In: *Proceedings of the 22nd international conference on Machine learning*. ICML ’05. New York, NY, USA: Association for Computing Machinery, 2005, 625–632. DOI: `10.1145/1102351.1102430`. (Visited on 2023).

[185] J. Nieke, L. Despoisse, A. Gabriele, H. Weber, H. Strese, N. Ghasemi, F. Gascon, K. Alonso, V. Boccia, B. Tsonevska, et al. “The copernicus hyperspectral imaging mission for the environment (CHIME): an overview of its mission, system and planning status”. *Sensors, Systems, and Next-Generation Satellites XXVII* 12729 (2023), 21–40.

[186] OECD Development Assistance Committee. *Guidelines for Aid Agencies for Improved Conservation and Sustainable Use of Tropical and Subtropical Wetlands*. Paris, France, 1996.

[187] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder. “Good practices for estimating area and assessing accuracy of land change”. en. *Remote Sensing of Environment* 148 (2014), 42–57. DOI: `10.1016/j.rse.2014.02.015`. (Visited on 2022).

[188] P. Olofsson, G. M. Foody, S. V. Stehman, and C. E. Woodcock. “Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation”. *Remote Sensing of Environment* 129 (2013), 122–131. DOI: `10.1016/j.rse.2012.10.031`. (Visited on 2023).

[189] K. Owens. *Global issues—National Politics: Comparing Wetland Protection Policies and Perceptions in the Netherlands and the United States*. `https://www.researchgate.net/profile/Katharine_Owens/publication/242150504_Global_issues_-_National_Politics_Comparing_wetland_protection_policies_and_perceptions_in_the_Netherlands_and_the_United_States/links/0f31752fa4e0b79df5000000.pdf`. 2001.

[190] M. Palahi, R. Valbuena, C. Senf, N. Acil, T. A. Pugh, J. Sadler, R. Seidl, P. Potapov, B. Gardiner, L. Hetemäki, et al. “Concerns about reported harvests in European forests”. *Nature* 592.7856 (2021), E15–E17.

[191] L. Parente, T. Hengl, J. Krizan, M. Landa, L. Brodsky, and **M. Witjes**. *Input Dataset for Gap Filling and Land-Cover Mapping Using Eumap Library - 2000 to 2020*. 2020. DOI: `10.5281/zenodo.4311598`.

[192] L. Parente, **M. Witjes**, T. Hengl, M. Landa, and L. Brodsky. *Continental Europe Land Cover Mapping at 30m Resolution Based CORINE and LUCAS on Samples.* 2021. DOI: `10.5281/zenodo.4725429`.

[193] J. Paulsson, S. Claesson, J. Fridman, and H. Olsson. "Incorrect figures on harvested forests in Nature article". *SLU news* (2020). Accessed: 2021-07-04.

[194] N. K. Pavlis, S. A. Holmes, S. C. Kenyon, and J. K. Factor. "The development and evaluation of the Earth Gravitational Model 2008 (EGM2008)". *Journal of geophysical research: solid earth* 117.B4 (2012). DOI: `10.1029/2011JB008916`.

[195] T. Payn, J.-M. Carnus, P. Freer-Smith, M. Kimberley, W. Kollert, S. Liu, C. Orazio, L. Rodriguez, L. N. Silva, and M. J. Wingfield. "Changes in planted forests and future global implications". *Forest Ecology and Management* 352 (2015), 57–67.

[196] E. Pebesma, W. Wagner, P. Soille, M. Kadunc, N. Gorelick, M. Schramm, J. Verbesselt, J. Reiche, M. Appel, J. Dries, et al. "openEO: an open API for cloud-based big Earth Observation processing platforms". In: *EGU General Assembly Conference Abstracts*. 2018, 4957.

[197] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. "Scikit-learn: Machine learning in Python". *Journal of machine learning research* 12 (2011), 2825–2830.

[198] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward. "High-resolution mapping of global surface water and its long-term changes". *Nature* 540 (2016), 418–. DOI: `10.1038/nature20584`.

[199] D. Pflugmacher, A. Rabe, M. Peters, and P. Hostert. "Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey". *Remote Sensing of Environment* (2019). DOI: `10.1016/j.rse.2018.12.001`.

[200] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage. "Sentinel-2 data for land cover/use mapping: A review". *Remote Sensing* 12.14 (2020), 2291.

[201] N. Picard, J.-M. Leban, J.-M. Guehl, E. Dreyer, O. Bouriaud, J.-D. Bontemps, G. Landmann, A. Colin, J.-L. Peyron, and P. Marty. "Recent increase in European forest harvests as based on area estimates (Ceccherini et al. 2020a) not confirmed in the French case". *Annals of Forest Science* 78.1 (2021), 1–5.

[202] R. A. Pielke Sr, G. Marland, R. A. Betts, T. N. Chase, J. L. Eastman, J. O. Niles, D. D. S. Niyogi, and S. W. Running. "The influence of land-use change and landscape dynamics on the climate system: relevance to climate-change policy beyond the radiative effect of greenhouse gases". *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360.1797 (2002), 1705–1719.

[203] C. Pigaiani and F. Batista E Silva. "The LUISA Base Map 2018". KJ-NA-30663-EN-N (online) (2021). DOI: 10.2760/503006(online).

[204] R. G. Pontius and C. D. Lippitt. "Can Error Explain Map Differences Over Time?" en. *Cartography and Geographic Information Science* 33.2 (2006), 159–171. DOI: 10.1559/152304006777681706. (Visited on 2023).

[205] R. G. Pontius and M. Millones. "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment". *Journal of remote sensing* (2011). DOI: 10.1080/01431161.2011.552923.

[206] P. Potapov, M. C. Hansen, I. Kommareddy, A. Kommareddy, S. Turubanova, A. Pickens, B. Adusei, A. Tyukavina, and Q. Ying. "Landsat analysis ready data for global land cover and land cover change mapping". *Remote Sensing* 12.3 (2020), 426.

[207] P. Potapov, M. C. Hansen, A. Pickens, A. Hernandez-Serna, A. Tyukavina, S. Turubanova, V. Zalles, X. Li, A. Khan, F. Stolle, et al. "The global 2000-2020 land cover and land use change dataset derived from the Landsat archive: first results". *Frontiers in Remote Sensing* 3 (2022), 856903.

[208] P. Potapov, X. Li, A. Hernandez-Serna, A. Tyukavina, M. C. Hansen, A. Kommareddy, A. Pickens, S. Turubanova, H. Tang, C. E. Silva, J. Armston, R. Dubayah, J. B. Blair, and M. Hofton. "Mapping global forest canopy height through integration of GEDI and Landsat data". *Remote Sensing of Environment* 253 (2021), 112165. DOI: 10.1016/j.rse.2020.112165.

[209] S. M. Powers and S. E. Hampton. "Open science, reproducibility, and transparency in ecology". *Ecological Applications* 29.1 (2018), e01822. DOI: 10.1002/eap.1822.

[210] C. Programme. *Copernicus Land Monitoring Service: Pan-European High Resolution Layers.* 2023.

[211] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. "CatBoost: unbiased boosting with categorical features". *Advances in neural information processing systems* 31 (2018).

[212] J. Qi, A. Chehbouni, A. R. Huete, Y. H. Kerr, and S. Sorooshian. "A modified soil adjusted vegetation index". *Remote sensing of environment* 48.2 (1994), 119–126.

[213] V. C. Radeloff, D. P. Roy, M. A. Wulder, M. Anderson, B. Cook, C. J. Crawford, M. Friedl, F. Gao, N. Gorelick, M. Hansen, et al. "Need and vision for global medium-resolution Landsat and Sentinel-2 data products". *Remote Sensing of Environment* 300 (2024), 113918.

[214] C. A. Ramezan, T. A. Warner, A. E. Maxwell, Bradley S. Price, and B. S. Price. "Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data". *Remote Sensing* 13.3 (2021), 368. DOI: 10.3390/rs13030368.

[215] Ramsar Convention Bureau. *Wetlands Values and Functions.* Gland, Switzerland: Ramsar Convention Bureau, 2001.

[216] Ramsar Convention Bureau. *Wetlands for Our Future: Act Now to Prevent, Stop, and Reserve Wetland Loss.* https://www.ramsar.org/news/press-release-wetlands-for-our-future-act-now-to-prevent-stop-and-reverse-wetland-loss. Accessed on 22 May 2019. 2015.

[217] R. Raši. *State of Europe's Forests 2020.* Accessed: 2021-10-05. Ministerial conference on the protection of forests in Europe, 2020.

[218] S. J. Riley, S. D. DeGloria, and R. Elliot. "Index that quantifies topographic heterogeneity". *intermountain Journal of sciences* 5.1-4 (1999), 23–27.

[219] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure". *Ecography* 40.8 (2017), 913–929.

[220] C. Robinson, L. Hou, K. Malkin, R. Soobitsky, J. Czawlytko, B. Dilkina, and N. Jojic. "Large Scale High-Resolution Land Cover Mapping With Multi-Resolution Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019.

[221] S. L. Rodríguez, L. G. van Bussel, and R. Alkemade. "Classification of agricultural land management systems for global modeling of biodiversity and ecosystem services". *Agriculture, Ecosystems & Environment* 360 (2024), 108795.

[222] R. Rodríguez-Pérez, M. Vogt, and J. Bajorath. "Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds". *Journal of Chemical Information and Modeling* 57.4 (2017), 710–716. DOI: 10.1021/acs.jcim.7b00088.

[223] M. O. Román, Z. Wang, Q. Sun, V. Kalb, S. D. Miller, A. Molthan, L. Schultz, J. Bell, E. C. Stokes, B. Pandey, et al. "NASA's Black Marble nighttime lights product suite". *Remote Sensing of Environment* 210 (2018), 113–143.

[224] E. Romijn, M. Herold, B. Mora, S. Briggs, F. M. Seifert, and M. Paganini. "Monitoring Progress towards Sustainable Development Goals The Role of Land Monitoring" (2016).

[225] F. Rossi, J. Breidenbach, S. Puliti, R. Astrup, and B. Talbot. "Assessing Harvested Sites in a Forested Boreal Mountain Catchment through Global Forest Watch". *Remote Sensing* 11.5 (2019). DOI: 10.3390/rs11050543.

[226] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering, et al. "Monitoring vegetation systems in the Great Plains with ERTS". *NASA Spec. Publ* 351.1 (1974), 309.

[227] M. Rußwurm, N. Courty, R. Emonet, S. Lefèvre, D. Tuia, and R. Tavenard. "End-to-end learned early classification of time series for in-season crop type mapping". *ISPRS Journal of Photogrammetry and Remote Sensing* 196 (2023), 445–456.

[228] M. Rußwurm and M. Körner. "Multi-temporal land cover classification with sequential recurrent encoders". *ISPRS International Journal of Geo-Information* 7.4 (2018), 129.

[229] O. E. Sala, F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig, et al. "Global biodiversity scenarios for the year 2100". *Science* 287.5459 (2000), 1770–1774.

[230] M. H. R. Sales, S. de Bruin, C. Souza, and M. Herold. "Land Use and Land Cover Area Estimates From Class Membership Probability of a Random Forest Classification". *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). Conference Name: IEEE Transactions on Geoscience and Remote Sensing, 1–11. DOI: 10.1109/TGRS.2021.3080083.

[231] M. Schneider, T. Schelte, F. Schmitz, and M. Körner. "Eurocrops: The largest harmonized open crop dataset across the european union". *Scientific Data* 10.1 (2023), 612.

[232] M. Schultz, J. Voss, M. Auer, S. Carter, and A. Zipf. "Open land cover from OpenStreetMap and remote sensing". *International journal of applied earth observation and geoinformation* 63 (2017), 206–213.

[233] S. Seabold and J. Perktold. "statsmodels: Econometric and statistical modeling with Python". In: *9th Python in Science Conference*. 2010.

[234] Secretariat of the Convention on Biological Diversity. *Indicators for the Strategic Plan for Biodiversity 2011-2020 and the Aichi Biodiversity Targets*. Tech. rep. 2016.

[235] C. Senf, D. Pflugmacher, Y. Zhiqiang, J. Sebald, J. Knorn, M. Neumann, P. Hostert, and R. Seidl. "Canopy mortality has doubled in Europe's temperate forests over the last three decades". *Nature Communications* 9.1 (2018), 1–8.

[236] C. Senf and R. Seidl. "Mapping the forest disturbance regimes of Europe". *Nature Sustainability* 4.1 (2021), 63–70.

[237] G. Seni and J. Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Synthesis lectures on data mining and knowledge discovery. Morgan & Claypool Publishers, 2010.

[238] K. Shahi, H. Z. Shafri, E. Taherzadeh, S. Mansor, and R. Muniandy. "A novel spectral index to automatically extract road networks from WorldView-2 satellite imagery". *The Egyptian Journal of Remote Sensing and Space Science* 18.1 (2015), 27–33. DOI: 10.1016/j.ejrs.2014.12.003.

[239] B. Shivakumar and S. Rajashekararadhya. "Spectral similarity for evaluating classification performance of traditional classifiers". In: *2017 International Conference on*

Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE. 2017, 1999–2004.

[240] T. Shumba, A. De Vos, R. Biggs, K. J. Esler, J. M. Ament, and H. S. Clements. "Effectiveness of private land conservation areas in maintaining natural land cover and biodiversity intactness". *Global Ecology and Conservation* 22 (2020), e00935.

[241] A. Sikora. "European Green Deal–legal and financial challenges of the climate change". In: *Era Forum*. Vol. 21. Springer. 2021, 681–697.

[242] G. Singh, G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson. *Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction*. 2024. arXiv: `2401.06421 [cs.LG]`.

[243] R. Smardon. *Sustaining the Worlds Wetlands: Setting Policy and Resolving Conflicts*. New York, NY, USA: Springer, 2009.

[244] X.-P. Song, M. C. Hansen, S. V. Stehman, P. V. Potapov, A. Tyukavina, E. F. Vermote, and J. R. Townshend. "Global land change from 1982 to 2016". *Nature* 560.7720 (2018), 639–643.

[245] A. M. Sparks, I. Bouhamed, L. Boschetti, I. Z. Gitas, and C. Kalaitzidis. "Mapping Arable Land and Permanent Agriculture Extent and Change in Southern Greece Using the European Union LUCAS Survey and a 35-Year Landsat Time Series Analysis". *Remote Sensing* (2022). DOI: `10.3390/rs14143369`.

[246] K. Spitzer and H. V. Danks. "Insect biodiversity of boreal peat bogs". *Annu. Rev. Entomol.* 51 (2006), 137–161.

[247] R. Stanimirova, K. Tarrio, K. Turlej, K. McAvoy, S. Stonebrook, K.-T. Hu, P. Arévalo, E. L. Bullock, Y. Zhang, C. E. Woodcock, et al. "A global land cover training dataset from 1984 to 2020". *Scientific Data* 10.1 (2023), 879.

[248] S. V. Stehman. "Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change". *Remote Sensing Letters* 3.2 (2012). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431161.2010.541950, 111–120. DOI: `10.1080/01431161.2010.541950`. (Visited on 2023).

[249] S. V. Stehman. "Estimating area from an accuracy assessment error matrix". *Remote Sensing of Environment* 132 (2013), 202–211. DOI: `10.1016/j.rse.2013.01.016`. (Visited on 2023).

[250] S. V. Stehman. "Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes". *International Journal of Remote Sensing* 35.13 (2014), 4923–4939. DOI: `10.1080/01431161.2014.930207`. (Visited on 2023).

[251] T. Storch, H.-P. Honold, S. Chabrillat, M. Habermeyer, P. Tucker, M. Brell, A. Ohndorf, K. Wirth, M. Betz, M. Kuchler, et al. "The EnMAP imaging spectroscopy mission towards operations". *Remote Sensing of Environment* 294 (2023), 113632.

[252] A. H. Strahler, L. Boschetti, G. M. Foody, M. A. Friedl, M. C. Hansen, M. Herold, P. Mayaux, J. T. Morisette, S. V. Stehman, and C. E. Woodcock. "Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps". *European Communities, Luxembourg* 51.4 (2006), 1–60.

[253] A. H. Strahler. "The Use of Prior Probabilities in Maximum Likelihood Classification of Remotely Sensed Data". *Remote Sensing of Environment* 10.2 (1980), 135–163. DOI: 10.1016/0034-4257(80)90011-5. (Visited on 2023).

[254] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl. "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]". *IEEE Geoscience and Remote Sensing Magazine* 9.3 (2021), 174–180.

[255] G. Sumbul and B. Demİr. "A Deep Multi-Attention Driven Approach for Multi-Label Remote Sensing Image Classification". *IEEE Access* 8 (2020), 95934–95946. DOI: 10.1109/ACCESS.2020.2995805.

[256] S. Sy and B. Quesada. "Anthropogenic land cover change impact on climate extremes during the 21st century". *Environmental Research Letters* 15.3 (2020), 034002.

[257] V. D. Sy, M. Herold, F. Achard, V. Avitabile, A. Baccini, S. Carter, J. G. P. W. Clevers, E. Lindquist, M. Pereira, and L. Verchot. "Tropical Deforestation Drivers and Associated Carbon Emission Factors Derived from Remote Sensing Data". *Environmental Research Letters* 14.9 (2019), 094022. DOI: 10.1088/1748-9326/ab3dc6. (Visited on 2023).

[258] Z. Szantoi, G. N. Geller, N.-E. Tsendbazar, L. See, P. Griffiths, S. Fritz, P. Gong, M. Herold, B. Mora, and A. Obregón. "Addressing the need for improved land cover map products for policy support". *Environmental Science & Policy* 112 (2020), 28–35. DOI: 10.1016/j.envsci.2020.04.005. (Visited on 2023).

[259] T. Tadono, H. Ishida, F. Oda, S. Naito, K. Minakawa, and H. Iwamoto. "Precise global DEM generation by ALOS PRISM". *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2.4 (2014), 71. DOI: 10.5194/isprsannals-II-4-71-2014.

[260] A. M. Tait, S. P. Brumby, S. B. Hyde, J. Mazzariello, and M. Corcoran. *Dynamic World Training Dataset for Global Land Use and Land Cover Categorization of Satellite Imagery*. Data Set. 2021. DOI: 10.1594/PANGAEA.933475.

[261] J. Takaku, T. Tadono, K. Tsutsui, and M. Ichikawa. "Quality improvements of AW3D global DSM derived from ALOS prism". In: *IGARSS 2018-2018 IEEE*

International Geoscience and Remote Sensing Symposium. IEEE. 2018, 1612–1615. DOI: `10.1109/IGARSS.2018.8518360`.

[262] L. Tang and G. Shao. "Drone remote sensing for forestry research and practices". *Journal of Forestry Research* 26 (2015), 791–797.

[263] Y. Tian, E. van Leeuwen, N.-e. Tsendbazar, C. Jing, and M. Herold. "Urban green inequality and its mismatches with human demand across neighborhoods in New York, Amsterdam, and Beijing". *Landscape Ecology* 39.3 (2024), 1–20.

[264] X. Tong, M. Brandt, P. Hiernaux, S. Herrmann, L. V. Rasmussen, K. Rasmussen, F. Tian, T. Tagesson, W. Zhang, and R. Fensholt. "The forgotten land use class: Mapping of fallow fields across the Sahel using Sentinel-2". *Remote Sensing of Environment* 239 (2020), 111598.

[265] J. R. Townshend, J. G. Masek, C. Huang, E. F. Vermote, F. Gao, S. Channan, J. O. Sexton, M. Feng, R. Narasimhan, D. Kim, et al. "Global characterization and monitoring of forest cover using Landsat data: opportunities and challenges". *International Journal of Digital Earth* 5.5 (2012), 373–397.

[266] K. E. Trenberth. "What are the seasons?" *Bulletin of the American Meteorological Society* 64.11 (1983), 1276–1282.

[267] A. Treves. "Best available science and the reproducibility crisis". *Frontiers in Ecology and the Environment* 20.9 (2022), 495–495. DOI: `10.1002/fee.2568`.

[268] Y. Trisurat, H. Shirakawa, and J. M. Johnston. "Land-use/land-cover change from socio-economic drivers and their impact on biodiversity in Nan Province, Thailand". *Sustainability* 11.3 (2019), 649.

[269] K. Tröltzsch, J. Van Brusselen, and A. Schuck. "Spatial occurrence of major tree species groups in Europe derived from multiple data sources". en. *Forest Ecology and Management* 257.1 (2009), 294–302. DOI: `10.1016/j.foreco.2008.09.012`. (Visited on 2023).

[270] N. Tsendbazar, M. Herold, L. Li, A. Tarko, S. De Bruin, D. Masiliunas, M. Lesiv, S. Fritz, M. Buchhorn, B. Smets, R. Van De Kerchove, and M. Duerauer. "Towards operational validation of annual global land cover maps". en. *Remote Sensing of Environment* 266 (2021), 112686. DOI: `10.1016/j.rse.2021.112686`. (Visited on 2023).

[271] N.-E. Tsendbazar, S. de Bruin, and M. Herold. "Integrating global land cover datasets for deriving user-specific maps". *International Journal of Digital Earth* 10.3 (2017), 219–237. DOI: `10.1080/17538947.2016.1217942`. (Visited on 2023).

[272] N. Tsendbazar, S. De Bruin, and M. Herold. "Assessing global land cover reference datasets for different user communities". *ISPRS Journal of Photogrammetry and Remote Sensing* 103 (2015), 93–114.

[273] N. Tsendbazar, M. Herold, S. De Bruin, M. Lesiv, S. Fritz, R. Van De Kerchove, M. Buchhorn, M. Duerauer, Z. Szantoi, and J.-F. Pekel. “Developing and applying a multi-purpose land cover validation dataset for Africa”. *Remote Sensing of Environment* 219 (2018), 298–309.

[274] G. Tseng, I. Zvonkov, C. L. Nakalembe, and H. Kerner. “Cropharvest: A global dataset for crop-type classification”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* 2021.

[275] C. J. Tucker. “Red and photographic infrared linear combinations for monitoring vegetation”. *Remote sensing of Environment* 8.2 (1979), 127–150.

[276] D. Tuia, C. Persello, and L. Bruzzone. “Domain adaptation for the classification of remote sensing data: An overview of recent advances”. *IEEE geoscience and remote sensing magazine* 4.2 (2016), 41–57.

[277] U.S. Geological Survey. *Landsat Next.* English. Tech. rep. 2024-3005. Reston, VA, 2024, 2. DOI: `10.3133/fs20243005`.

[278] V. UNFCCC. “Adoption of the Paris agreement”. *Proposal by the President* 282 (2015), 2.

[279] D. Valle, R. Izbicki, and R. V. Leite. “Quantifying Uncertainty in Land-Use Land-Cover Classification Using Conformal Statistics”. *Remote Sensing of Environment* 295 (2023), 113682. DOI: `10.1016/j.rse.2023.113682`. (Visited on 2023).

[280] R. Van De Kerchove, D. Zanaga, W. Keersmaecker, N. Souverijns, J. Wevers, C. Brockmann, A. Grosu, A. Paccini, O. Cartus, M. Santoro, M. Lesiv, I. Georgieva, S. Fritz, S. Carter, N.-E. Tsendbazar, L. Li, M. Herold, and O. Arino. “ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data”. In: *AGU Fall Meeting Abstracts.* Vol. 2021. 2021, GC45I–0915.

[281] C. Van Rijsbergen. *Information Retrieval.* Butterworth Heinemann, 1980.

[282] C. Van Rijsbergen. *Information Retrieval.* Butterworth Heinemann, 1980.

[283] T. Van Thinh, P. Cao Duong, K. Nishida Nasahara, et al. “How Does Land Use/Land Cover Map’s Accuracy Depend on Number of Classification Classes?” *SOLA* 15 (2019), 28–31.

[284] K. Van Tricht, J. Degerickx, S. Gilliams, D. Zanaga, M. Battude, A. Grosu, J. Brombacher, M. Lesiv, J. C. L. Bayas, S. Karanam, S. Fritz, I. Becker-Reshef, B. Franch, B. Mollà-Bononad, H. Boogaard, A. K. Pratihast, and Z. Szantoi. “WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping”. *Earth System Science Data Discussions* 2023 (2023), 1–36. DOI: `10.5194/essd-2023-184`.

[285] A. Veldkamp and E. F. Lambin. “Predicting land-use change”. *Agriculture Ecosystems and Environment* (2001).

[286] Z. S. Venter, D. N. Barton, T. Chakraborty, T. Simensen, and G. Singh. "Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover". *Remote Sensing* 14.16 (2022), 4101. DOI: 10.3390/rs14164101.

[287] Z. S. Venter and M. A. Sydenham. "Continental-scale land cover mapping at 10 m resolution over Europe (ELC10)". *Remote Sensing* 13.12 (2021), 2301.

[288] P. H. Verburg, K. Neumann, and L. Nol. "Challenges in using land use and land cover data for global change studies". en. *Global Change Biology* 17.2 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2010.02307.x, 974–989. DOI: 10.1111/j.1365-2486.2010.02307.x. (Visited on 2023).

[289] A. Verhegghen, R. d'Andrimont, F. Waldner, and M. van der Velde. "Accuracy Assessment of the First Eu-Wide Crop Type Map with Lucas Data" (2021). DOI: 10.1109/igarss47720.2021.9553758.

[290] L. Vilar, J. Garrido, P. Echavarría, J. Martínez-Vega, and M. Martín. "Comparative analysis of CORINE and climate change initiative land cover maps in Europe: Implications for wildfire occurrence estimation at regional and local scales". *International Journal of Applied Earth Observation and Geoinformation* 78 (2019), 102–117. DOI: 10.1016/j.jag.2019.01.019.

[291] F. Vuolo, M. Neuwirth, M. Immitzer, C. Atzberger, and W.-T. Ng. "How much does multi-temporal Sentinel-2 data improve crop type classification?" *International Journal of Applied Earth Observation and Geoinformation* 72 (2018), 122–130. DOI: 10.1016/j.jag.2018.06.007.

[292] A. M. J.-C. Wadoux, D. J. Brus, and G. B. M. Heuvelink. "Sampling Design Optimization for Soil Mapping with Random Forest". *Geoderma* 355 (2019), 113913. DOI: 10.1016/j.geoderma.2019.113913. (Visited on 2023).

[293] J. Wagemann, S. Siemen, B. Seeger, and J. Bendix. "Users of open Big Earth data — An analysis of the current state". *Computers & Geosciences* 157 (2021), 104916. DOI: 10.1016/j.cageo.2021.104916.

[294] F. Waldner, D. De Abelleyra, S. R. Verón, M. Zhang, B. Wu, D. Plotnikov, S. Bartalev, M. Lavreniuk, S. Skakun, N. Kussul, G. Le Maire, S. Dupuy, I. Jarvis, and P. Defourny. "Towards a set of agrosystem-specific cropland mapping methods to address the global cropland diversity". *International Journal of Remote Sensing* 37.14 (2016), 3196–3231. DOI: 10.1080/01431161.2016.1194545. (Visited on 2023).

[295] F. Waldner and P. Defourny. "Where can pixel counting area estimates meet user-defined accuracy requirements?" en. *International Journal of Applied Earth Observation and Geoinformation* 60 (2017), 1–10. DOI: 10.1016/j.jag.2017.03.014. (Visited on 2023).

[296] X. Wang, C. Wu, D. Peng, A. Gonsamo, and Z. Liu. “Snow cover phenology affects alpine vegetation growth dynamics on the Tibetan Plateau: Satellite observed evidence, impacts of different biomes, and climate drivers”. *Agricultural and Forest Meteorology* 256 (2018), 61–74.

[297] J. Wehrmann, R. Cerri, and R. Barros. “Hierarchical multi-label classification networks”. In: *International conference on machine learning.* PMLR. 2018, 5075–5084.

[298] K. Winkler, R. Fuchs, M. Rounsevell, and M. Herold. “Global land use changes are four times greater than previously estimated”. en. *Nature Communications* 12.1 (2021), 2501. DOI: `10.1038/s41467-021-22702-2`. (Visited on 2023).

[299] **M. Witjes**, M. Herold, and S. de Bruin. *Iterative Mapping of Probabilities: A data fusion framework for generating accurate land cover maps that match area statistics.* 2024.

[300] **M. Witjes**, M. Herold, and S. de Bruin. *Iterative Mapping of Probabilities.* 2024. DOI: `10.5281/zenodo.10641340`.

[301] **M. Witjes**, L. Parente, C. J. van Diemen, T. Hengl, M. Landa, L. Brodský, L. Halounova, J. Križan, L. Antonić, C. M. Ilie, V. Craciunescu, M. Kilibarda, O. Antonijević, and L. Glušica. “A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat”. *PeerJ* 10 (2022), e13573. DOI: `10.7717/peerj.13573`.

[302] **M. Witjes**, L. Parente, J. Križan, T. Hengl, and L. Antonić. “Ecodatacube. eu: Analysis-ready open environmental data cube for Europe”. *PeerJ* 11 (2023), e15478. DOI: `https://doi.org/10.7717/peerj.15478`.

[303] WorldCereal Project Team. *WorldCereal Reference Data Module.* `https://worldcereal-rdm.geo-wiki.org/`. Accessed: 2024-03-18. 2024.

[304] M. A. Wulder, D. P. Roy, V. C. Radeloff, T. R. Loveland, M. C. Anderson, D. M. Johnson, S. Healey, Z. Zhu, T. A. Scambos, N. Pahlevan, M. Hansen, N. Gorelick, C. J. Crawford, J. G. Masek, T. Hermosilla, J. C. White, A. S. Belward, C. Schaaf, C. E. Woodcock, J. L. Huntington, L. Lymburner, P. Hostert, F. Gao, A. Lyapustin, J.-F. Pekel, P. Strobl, and B. D. Cook. “Fifty years of Landsat science and impacts”. *Remote Sensing of Environment* 280 (2022), 113195. DOI: `10.1016/j.rse.2022.113195`.

[305] X. Xiong and J. J. Butler. “MODIS and VIIRS calibration history and future outlook”. *Remote Sensing* 12.16 (2020), 2523.

[306] J. Xu, J. Yang, X. Xiong, H. Li, J. Huang, K. Ting, Y. Ying, and T. Lin. “Towards interpreting multi-temporal deep learning models in crop mapping”. *Remote Sensing of Environment* 264 (2021), 112599.

[307] Z. Xu, K. Guan, N. Casler, B. Peng, and S. Wang. “A 3D convolutional neural network method for land cover classification using LiDAR and multi-temporal Landsat imagery”. *ISPRS journal of photogrammetry and remote sensing* 144 (2018), 423–434.

[308] D. Yamazaki, D. Ikeshima, J. Sosa, P. D. Bates, G. H. Allen, and T. M. Pavelsky. “MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset”. *Water Resources Research* 55.6 (2019), 5053–5073. DOI: 10.1029/2019WR024873.

[309] D. Yamazaki, D. Ikeshima, R. Tawatari, T. Yamaguchi, F. O’Loughlin, J. C. Neal, C. C. Sampson, S. Kanae, and P. D. Bates. “A high-accuracy map of global terrain elevations”. *Geophysical Research Letters* 44.11 (2017), 5844–5853. DOI: 10.1002/2017GL072874.

[310] R. Yokoyama, M. Shirasawa, and R. J. Pike. “Visualizing topography by openness: A new application of image processing to digital elevation models”. *Photogrammetric Engineering and Remote Sensing* 68.3 (2002), 257–265.

[311] L. You, S. Wood, U. Wood-Sichra, and W. Wu. “Generating global crop distribution maps: From census to grid”. *Agricultural Systems* 127 (2014), 53–60.

[312] F. Yuan, A. Lewis, A. Leith, T. Dhar, and D. Gavin. “Analysis Ready Data for Africa”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. 2021, 1789–1791. DOI: 10.1109/IGARSS47720.2021.9554019.

[313] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, et al. “ESA WorldCover 10 m 2021 v200” (2022).

[314] C. Zhang and Y. Ma. *Ensemble Machine Learning: Methods and Applications*. Springer New York, 2012.

[315] X. Zhang, L. Liu, X. Chen, Y. Gao, S. Xie, and J. Mi. “GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery”. *Earth System Science Data Discussions* (2020), 1–31.

[316] D. Zhao, Y. Hou, Z. Zhang, Y. Wu, X. Zhang, L. Wu, X. Zhu, and Y. Zhang. “Temporal resolution of vegetation indices and solar-induced chlorophyll fluorescence data affects the accuracy of vegetation phenology estimation: A study using in-situ measurements”. *Ecological Indicators* 136 (2022), 108673. DOI: 10.1016/j.ecolind.2022.108673.

[317] H. Zhao, Z. Yang, L. Di, and Z. Pei. “Evaluation of temporal resolution effect in remote sensing based crop phenology detection studies”. In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2011, 135–150. DOI: 10.1007/978-3-642-27278-3_16.

[318] W. Zhou, Y. Qian, X. Li, W. Li, and L. Han. "Relationships between land cover and the surface urban heat island: seasonal variability and effects of spatial and thematic resolution of land cover data on predicting land surface temperatures". *Landscape ecology* 29.1 (2014), 153–167.

[319] X. Zhu, D. Liu, and J. Chen. "A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images". *Remote Sensing of Environment* 124 (2012), 49–60. DOI: `https://doi.org/10.1016/j.rse.2012.04.019`.

[320] Z. Zhu, A. L. Gallant, C. E. Woodcock, B. Pengra, P. Olofsson, T. R. Loveland, S. Jin, D. Dahal, L. Yang, and R. F. Auch. "Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative". *ISPRS Journal of Photogrammetry and Remote Sensing* 122 (2016), 206–221. DOI: `10.1016/j.isprsjprs.2016.11.004`. (Visited on 2023).

Summary

This thesis aims to contribute towards efforts made in large-scale land cover mapping, with an emphasis on the benefits of combining several datasets from different sources and of different types. It presents different steps of a methodology to extract training data from multiple rich human-annotated datasets and overlay them on Earth observation data from diverse sources. It furthermore details the challenges and benefits of creating land cover maps that navigate the trade-off between spatial, temporal, and thematic resolution, as well as quantity and allocation accuracy.

Chapter 2 focuses on the benefits and challenges of harmonizing and combining large-scale spatiotemporal datasets for land cover mapping, most of which were used in the following chapters. The chapter details the work that went into creating, harmonizing, and imputing multiple Earth observation datasets (Landsat, Sentinel-2, and a new 30m resolution DTM) covering Europe. It introduces and describes the imputation algorithm TMWM that was used to impute the Landsat data, and validates its accuracy in a spatiotemporally explicit way. It then explores how combining the different datasets improves the accuracy of land cover classification models. Lastly, it shows that models trained on samples from a longer time range can generalize better to years that they have not been trained on.

Chapter 3 focuses on the production of annual land use / land cover maps of Europe for 2000-2020. It details the steps taken to harmonize and clean the training data from multiple openly available sources (CORINE, LUCAS) into a legend with 43 classes. A thorough accuracy assessment using cross-validation and an independent set of S2GLC validation points describes how well the model generalizes across space and time, and quantifies the trade-off imposed by having a legend with high thematic resolution. Results show that the maps have similar accuracy as other current continental-scale maps at low thematic resolution, and that a more detailed legend introduces more errors.

Chapter 4 introduces IMP, an algorithm that uses land cover area estimates to iteratively classify land cover from existing probabilities, producing maps whose class proportions match the input estimates. It details the algorithm and showcases its use by mapping five European countries in five years. The accuracy of the maps is compared with maps created using highest likelihood classification. Results show that the proportional maps do not only have more accurate class proportions, but equal or better accuracy than

highest likelihood maps. We also compare the accuracy and proportions of maps based on probabilities predicted by models trained on data representative of the area of interest, and probabilities predicted by a general model trained on large parts of Europe. Results show that maps based on general model predictions reach more accurate class proportions, while maps based on local model predictions are slightly more accurate. Finally, it presents an unintentional finding that the iterations at which the algorithm classifies certain pixels is related to the accuracy of those pixels, suggesting that it can be used to generate pixel-level accuracy estimates.

Acknowledgements

The four years that I worked on this thesis were without a doubt the most challenging and transformative of my adult life. I like to think that I have come out stronger and wiser, but this is hardly to my own credit. As such, I have a large list of people to thank:

Firstly, my promotor **Martin Herold**, Shrewd Strategic Sage of Spatial Science, for providing exactly the right words and ideas at exactly the right moments. I hope to one day reach a level of wisdom and insight that lets me do the same, and pay it forward.

The relentless **Tomislav Hengl**, Meritous Mad Mapper of Mayhem, whose bottomless drive and energy are the the primary reason I even did all this. I hope to keep butting heads, ideas, and cups with you for years to come.

The sagacious **Sytze de Bruin**, Merciless Masticator of my Many Mistakes. I vividly remember the first time I came to him for advice on my MSc thesis: His capacity for good advice and tough, but constructive criticism has been a continuous boon to everything I've written since. May the scorching glare of his unsalted criticism keep perfecting all he beholds.

The impressive **Leandro Parente**, Meticulous Master of Music, Meat, and Many Methods, for sharing his technical expertise, level head, and professional perspectives, and of course for blasting his beautiful beats.

Honorary supervisor, **Ichsani Wheeler**, Energetic Eater of Elephants. I would not have finished this thesis without her guidance, mental support, babysitting, gyoza, tough love, deadlines, reminders of meaning, and unlimited cheerful banter.

My Precious Paranymphs, Dainius Masiliūnas and Xuemeng Tian, Both Bright Brainiacs Facing Fantastic Futures. Dainius, the absolute smartest and most lovable person at GRS, was there when I started my GIS master's many years ago, and I am proud to have him at my side during my defense and all its preparations. I have not known Xuemeng equally long, but the two years that we have shared at OpenGeoHub have been filled with excellent dinners and conversations, and her perspective and mindset never cease to inspire me. I am supremely thankful to them both for chasing and motivating me to sort out the defense, despite their own daunting deadlines.

Rolf Simoes, Zestful Zen Zealot of Zenodo, whose help, advice, and indestructible optimism made the last stressful months and hours bearable, and with whom I hope to collaborate for years to come.

Maya Kroodsma, Expedient Editor Extraordinaire and Erstwhile Eega. I am infinitely grateful to her for finding truth in all aspects of my life, and, more tangibly, for the immense amount of mental, logistical and editorial support during the writing of this thesis. These last four years have been an intense journey for us. However, if journeys are really about the friends you make along the way, I can count myself singularly fortunate to have made such a formidable friend.

Carsten Lygteskov Hansen, Benevolent Blackbeard and Beer Buddy, whose advice, encouragement and mentorship got me both into and through this mess.

Luka Antonic, Coolest Co-author and Courageous Co-host, for being there at 2AM, whether writing about Sentinel-2 resampling, giving me gratuitous goulash, twerking to turbofolk, or smashing spaceships to smithereens.

All the wonderful people at **OpenGeoHub**, past and present, for the savage (after)parties, sword fights, glass-shattering barbecues, dangerous delicacies, epically cutthroat game nights (and mornings), road trips, cookie rounds, and endless laughter.

Most notably, **Carmelo Bonannella**, Charming Buccaneer of Freezing Forests and Majestic Mountains, a 'PhD Twin' whose erudition dwarfs my own, and whose will and fortitude I can only dream of approximating. I always enjoy realizing how much we have in common despite being so different. Thank you being the best Blue Oni I could have wished for.

Last but not least, **Ron van Lammeren**, for making my first encounters with spatial analysis such a positive experience by means of his seemingly infinite time, memory and affection for all GIS students.

About the author

As a young child, Martijn used to draw giant maps with innumerable amounts of things on them, even transforming map-drawing into a multi-table boardgame with his friends. Beleaguered by a broad array of interests, he tried out several educational programmes, from entrepeneurship and political science to forestry. He finally came full circle with a return to mapmaking, and from there, remote sensing.

Peer-reviewed Journal Publications

[5] **M. Witjes**, L. Parente, C. J. van Diemen, T. Hengl, M. Landa, L. Brodskỳ, L. Halounova, J. Križan, L. Antonić, C. M. Ilie, V. Craciunescu, M. Kilibarda, O. Antonijević, and L. Glušica. “A spatiotemporal ensemble machine learning framework for generating land use/land cover time-series maps for Europe (2000–2019) based on LUCAS, CORINE and GLAD Landsat”. *PeerJ* 10 (2022), e13573. DOI: `10.7717/peerj.13573`.

[6] **M. Witjes**, L. Parente, J. Križan, T. Hengl, and L. Antonić. “Ecodatacube. eu: Analysis-ready open environmental data cube for Europe”. *PeerJ* 11 (2023), e15478. DOI: `https://doi.org/10.7717/peerj.15478`.

Other Scientific Publications

[1] M. Landa, L. Brodsky, L. Parente, **M. Witjes**, and T. Hengl. *Multi-year harmonized land cover samples based on LUCAS and CORINE datasets.* Version v0.1. Zenodo, 2021. DOI: `10.5281/zenodo.4740691`.

[2] L. Parente, T. Hengl, J. Krizan, M. Landa, L. Brodsky, and **M. Witjes**. *Input Dataset for Gap Filling and Land-Cover Mapping Using Eumap Library - 2000 to 2020.* 2020. DOI: `10.5281/zenodo.4311598`.

[3] L. Parente, **M. Witjes**, T. Hengl, M. Landa, and L. Brodsky. *Continental Europe Land Cover Mapping at 30m Resolution Based CORINE and LUCAS on Samples.* 2021. DOI: `10.5281/zenodo.4725429`.

[4] **M. Witjes**, M. Herold, and S. de Bruin. “Iterative Mapping of Probabilities: A data fusion framework for generating accurate land cover maps that match area statistics”. Accepted for publication. 2024.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review / project proposal (4.5 ECTS)

- Automated machine learning for spatial and spatio-temporal data cubes: predicting land use and land cover and landscape dynamics at regional and global scales.

Post-graduate courses (7.5 ECTS)

- Summer school; OpenGeoHub (2020, 2021, 2022)
- Uncertainty propagation in spatial environmental modelling; PE&RC (2022)
- Quarto; VLAG (2023)

Deficiency, refresh, brush-up courses (6.6 ECTS)

- Geo for Good; Google (2020)
- Workshop on big data and artificial intelligence in earth observation; European Commission (2020)
- Introduction to computer science; Harvard University (2021)

Laboratory training and working visits (2.1 ECTS)

- Species distribution modelling with deep learning; BIOMAC, Amsterdam, the Netherlands (2021)
- Combining land cover mapping with official statistics; Eurostat, Brussels, Belgium (2023)

- OEMC Co-development sessions; WUR & OpenGeoHub (2023)

Competence, skills and career-oriented activities (2.6 ECTS)

- Project and time management; WGS (2021)
- PhD Workshop carousel; PE&RC (2023)
- Mindful productivity; WGS (2023)
- Reviewing a scientific manuscript; WGS (2023)

Scientific integrity/ethics in science activities (0.3 ECTS)

- Scientific integrity; WGS (2021)

PE&RC Annual meetings, seminars and the PE&RC weekend (2.1 ECTS)

- PE&RC Weekend for first years (2021)
- Midterm retreat (2022)
- PE&RC day (2022)

Discussion groups / local seminars / other scientific meetings (4.5 ECTS)

- GeoHarmonizer scientific meetings (2020–2021)
- SoilMacroFauna discussion groups & meetings (2021–2024)
- OpenEarthMonitor seminars (2022–2024)

International symposia, workshops and conferences (3.8 ECTS)

- Open Data Science Europe workshop; oral presentation; Wageningen, the Netherlands (2021)
- SoilMacroFauna Workshop; oral presentation; Leipzig, Germany (2022)
- Living Planet Symposium; oral and poster presentation; Bonn, Germany (2022)

Societally relevant exposure (0.6 ECTS)

1. Guest lecture at primary school: satellites, land cover, climate change (2023)
2. Guest lecture at CineScience, Heerenstraat Theater Wageningen: Land cover & climate change (2024)

Lecturing / supervision of practicals / tutorials (4.8 ECTS)

- Geoscripting: Using Python for Geospatial data (2021)
- Supervision of Msc student project on gap filling methods; Statistical University of Muenchen, Germany (2021)
- Land cover classification with deep learning (2022)
- Using STAC and scikit-map for land cover classification (2023)

This research received funding from Grant Agreement Connecting Europe Facility (CEF) Telecom project 2018-EU-IA-0095 by the European Union (”GeoHarmonizer”) and from the European Commission Horizon Europe research project ”Open-Earth-Monitor Cyberinfrastructure” (grant agreement No. 101059548). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Cover design by Martijn Witjes