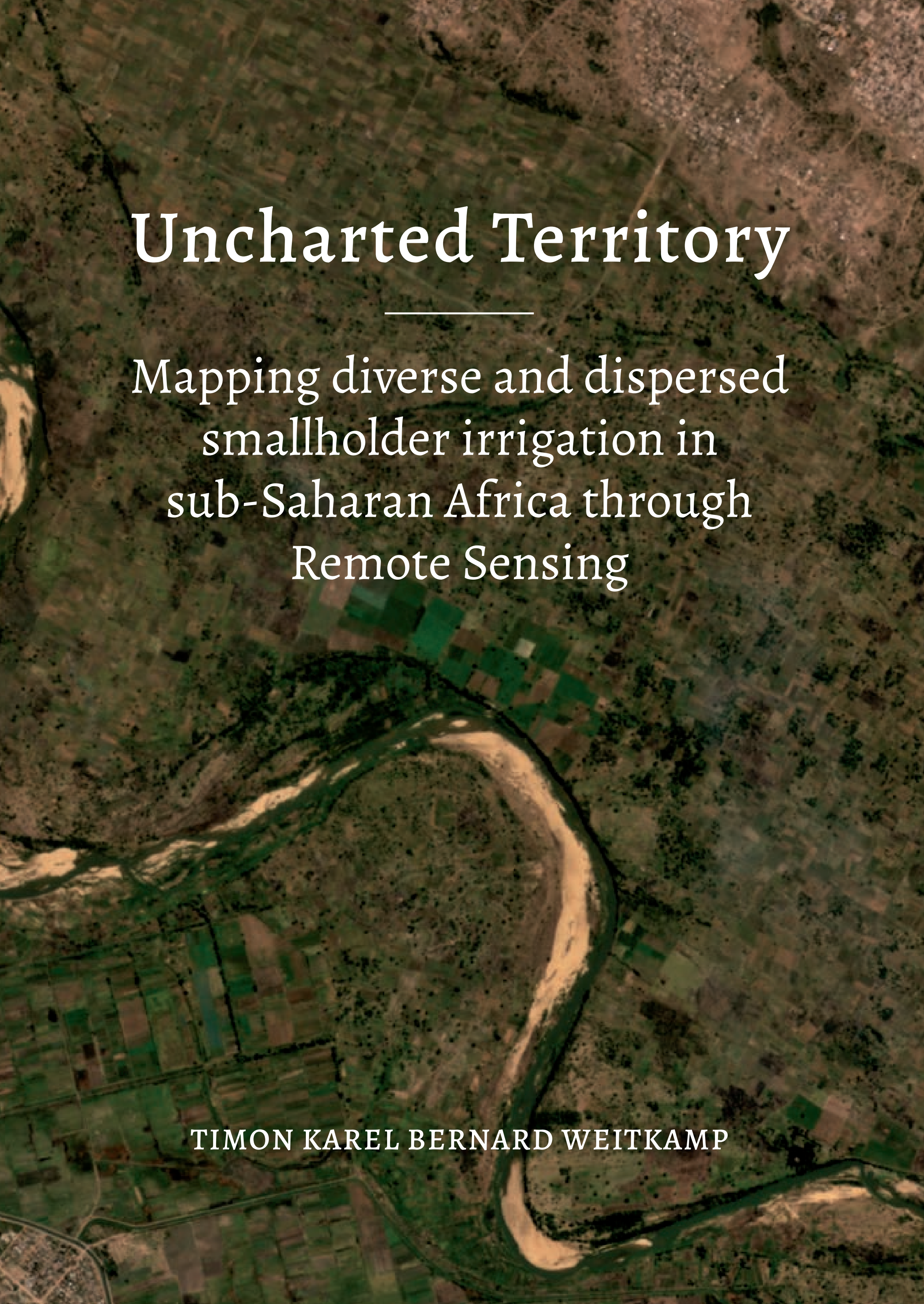


Uncharted Territory

Mapping diverse and dispersed smallholder irrigation in sub-Saharan Africa through Remote Sensing

TIMON KAREL BERNARD WEITKAMP

Propositions

1. The emphasis on map accuracy often overshadows its contextual relevance.
(this thesis)

2. Failed experiments contribute more to learning than successful ones.
(this thesis)

3. Funding is too often geared towards finding problems to a solution.

4. The quest for unique findings results in increasingly niche research.

5. The return on investment from integrating sports into the workday exceeds the associated financial costs

6. The benefits of a 40-hour workweek are vastly overrated.

Propositions belonging to the thesis, entitled

Uncharted Territory: mapping diverse and dispersed smallholder irrigation in sub-Saharan Africa through Remote Sensing

Timon Karel Bernard Weitkamp
Wageningen, 2 February 2024

Uncharted Territory: mapping diverse and dispersed smallholder irrigation in sub-Saharan Africa through Remote Sensing

Timon Karel Bernard Weitkamp

Thesis committee

Promotors

Prof. Dr Charlotte de Fraiture
Professor of Hydraulic Engineering for Land and Water Development
IHE Delft Institute for Water Education and Wageningen University & Research

Co-promotors

Dr Gert Jan Veldwisch
Associate professor, Water Resources Management group
Wageningen University & Research

Dr Poolad Karimi
Associate professor, Water and Agriculture group
IHE Delft Institute for Water Education

Other members

Prof. Dr Kirsten de Beurs, Wageningen University & Research
Dr Jonathan Denison, Independent consultant, Port Elizabeth, South Africa
Dr Marloes Mul, IHE Delft Institute for Water Education
Prof. Dr Hans Komakech, The Nelson Mandela African Institute of Science and Technology, Arusha, Tanzania

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE)

Uncharted Territory: mapping diverse and dispersed smallholder irrigation in sub-Saharan Africa through Remote Sensing

Timon Karel Bernard Weitkamp

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 2 February 2024
at 11 a.m. in the Omnia Auditorium

Timon Karel Bernard Weitkamp
Uncharted Territory: mapping diverse and dispersed smallholder irrigation in sub-Saharan Africa through Remote Sensing
212 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)
With references, with summary in English and Dutch

ISBN 978-94-6447-969-0
DOI https://doi.org/10.18174/641428

I wisely started with a map

J.R.R. Tolkien

Table of contents

Chapter 1

General Introduction

1. From top-down agricultural development…

After gaining independence, governments in sub-Saharan Africa (SSA) have historically continued to build large-scale public irrigation schemes, where a central authority supervised tenant farmers (Harrison, 2018; Veldwisch et al., 2009). However, such scheme designs were primarily focused on meeting the needs of European colonial powers and increasing cash crop production rather than addressing the food security needs of local populations (Bjornlund et al., 2020). After independence, African countries inherited these irrigation systems, with their development accelerating in the 1960s because of investments from multilateral donors such as the World Bank and African Development Bank (Harmon et al., 2023; Higginbottom et al., 2021). However, the countries often could not maintain these systems due to a lack of resources and expertise, resulting in disrepair and abandonment. These schemes were primarily designed based on engineering considerations and often failed to achieve their objectives. Smaller-scale public irrigation schemes suffered a similar fate as they were often built with little consultation with local communities and no consideration for their traditional farming practices.

Nevertheless, there has been a renewed interest in irrigated agriculture in SSA since the turn of the century, driven by the need to address agricultural development and food security challenges (Wiggins & Lankford, 2019). With the projected population of the continent expected to exceed 2 billion within the next 25 years (Statista, 2022), it is necessary to expand and intensify irrigation substantially to meet the region's food requirements without relying heavily on increased imports by 2050.

This challenge is intensified by the vulnerability of agriculture to climate change, as rainfed agriculture provides little mitigation against erratic rains and climate shocks. Governments in SSA are responding to these threats and prospects by setting ambitious targets for expanding irrigation, increasing farm productivity, and alleviating poverty (African Union, 2020). However, there is a risk in the prevailing narrative that suggests that irrigation is most effectively managed in schemes that require external expertise, financing, and engineering (Harrison, 2018), which can be seen in plans for new and rehabilitated schemes in several countries in SSA. Despite the optimistic prospects of expanding irrigated areas and promoting economic development and food security, large-scale irrigation projects in SSA following independence have generally fallen short of achieving the expected benefits. The underperformance of these projects can be attributed to a range of factors, including governance deficiencies, recurrent cycles of construction and refurbishment, high costs of development, inadequate management practices, limited access to rural finance, and the high expense of fertilisers (Higginbottom et al., 2021; Redicker et al., 2022). Furthermore, systems focused solely on cultivating low-value cereal crops to ensure food security seldom generate sufficient funds to sustain themselves (Redicker et al., 2022; van Rooyen et al., 2017).

Nevertheless, the ambitious targets for irrigated agricultural production set by governments may be well on their way to being achieved, though not primarily through public schemes. Smallholder farmers' own irrigation initiatives are contributing much more to food security than official statistics show (Woodhouse et al., 2017a).

2. … to bottom-up irrigation development

Farmers have been driving irrigated agriculture in SSA for a long time by establishing new areas, expanding existing ones, or improving them (Veldwisch et al., 2019a) without (initial) external support (Nkoka et al., 2014a). Irrigation furrow systems existed in Kenya, Tanzania, Zimbabwe, and Mozambique well before colonial times (e.g., Adams & Carter, 1987; Bolding et al., 1996). These systems were not solely shaped by infrastructure but also influenced by social and customary networks (Nkoka et al., 2014a). As early as the 1970s and the following decades, various donor, financial and public agencies were aware of the importance of small-scale irrigation for food production and promoted it; yet many of these initiatives failed as they were often not farmer-centred nor context-specific (Harmon et al., 2023).

While government irrigation development policies were often ineffective, farmers kept taking matters into their own hands. They continued investing in irrigation independently, often relatively unnoticed by official institutions. Only when researchers highlighted the widespread dynamics of farmer-led irrigation development was the attention of governments and development agencies drawn to these grassroots efforts (Harmon et al., 2023). This development process was named Farmer-led Irrigation Development (FLID), characterised by farmers all over SSA being the leading actors in initiating, operating, maintaining and usually constructing irrigation infrastructure, using local materials and ideas to improve their crop yields and income (Beekman et al., 2014a; de Fraiture & Giordano, 2014; Nkoka et al., 2014a; Veldwisch et al., 2019a; Woodhouse et al., 2017a). Farmers often adopt a commercially oriented approach, investing in and managing irrigation infrastructure to enhance their productivity and income. However, farmers do not operate in isolation. Throughout this process, they rely on and are influenced by diverse actors, including neighbouring farmers, agro-dealers, traders, agricultural extension agents, irrigation engineers, administrative authorities, and local and national policymakers (Woodhouse et al., 2017a). Consequently, farmers prioritise cultivating high-value (cash) crops like tomato or cabbage, which offer greater profitability than staple foods. FLID is not a homogenous practice, though, nor is it an irrigation category. Its extent and nature vary widely depending on the region, crops, topography, and socioeconomic conditions (Figure 1). Farmers use a variety of methods to irrigate their fields, such as weirs to divert water, flood and drainage management, bucket irrigation or irrigation with small motor pumps (Woodhouse et al., 2017a), primarily to generate income by selling the produce (De Bont, Liebrand, et al., 2019; Veldwisch et al., 2019a).

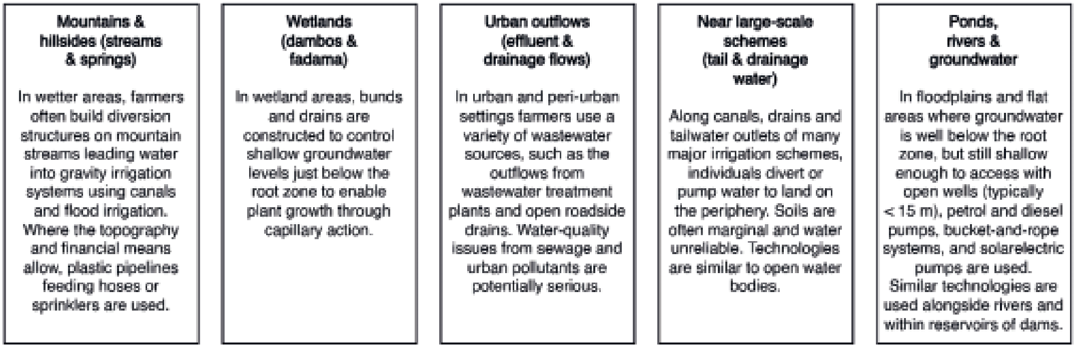


Figure 1: Locations and situations where FLID is prominent (adopted from Izzi et al., 2021)

3. The invisibility of FLID

While individual smallholder farmers typically cultivate small plots of land, the combined area under irrigation in SSA is substantial, encompassing hundreds of thousands of hectares (Beekman et al., 2014a; Venot et al., 2021; Woodhouse et al., 2017a). Irrigation in SSA is undergoing a noteworthy expansion as a result of FLID-processes. However, this is not always acknowledged by state agencies, development organisations, or researchers (Beekman et al., 2014a; De Bont, Liebrand, et al., 2019; Venot et al., 2021), both as a result of its heterogeneity and due to the common (technical) narrative of what irrigation is. I will briefly explain these two aspects in the following paragraphs.

The fragmented and small-scale nature of irrigation developed through FLID makes it challenging for governments to count and detect, leading to under-reporting in official statistics. An agricultural census (where officials visit almost every field in a country) might capture this, but this is rarely done (Wiggins & Lankford, 2019). These sporadic measuring moments also mean that when underperforming schemes are abandoned and disappear, the non-updated official record still includes it as an irrigated area (Wiggins & Lankford, 2019). Often, the organisational capacity, budget and/or staff are lacking to monitor large areas and to fill the databases with up-to-date information on irrigated areas developed by farmers (De Bont, Liebrand, et al., 2019).

Furthermore, as smallholder farmers often do not conform to developmental ideals and do not use "modern irrigation technologies", their practices might be seen as illegal, inferior, or irrelevant despite their significant contributions to higher-level government goals such as food security (De Bont et al., 2019). The lack of technical and organisational capacities is intertwined with the working cultures, narratives, and politics of irrigation development in SSA, leading to the invisibility of farmer's initiatives (Venot et al., 2021). Even when public institutes recognise the existence of smallholder irrigation, it is often perceived as backward and needing modernisation by external expertise (de Bont & Veldwisch, 2020; Hounkonnou et al., 2012).

However, the recent publication of the "Farmer-led irrigation development guide" by the World Bank (Izzi et al., 2021) signifies a shift in perspective, with a growing emphasis on promoting farmer-led irrigation development within investment portfolios, next to the continued presence of large-scale irrigation projects (Harmon et al., 2023).

4. Estimating the extent of smallholder irrigation

Gaining insight into farmer-led irrigation development (FLID) dynamics begins with understanding where, when, and how much area farmers irrigate. However, obtaining this information is challenging due to the fragmented and small-scale nature of irrigated agriculture and the varied definitions of irrigation.

Over the past decade, multiple studies have investigated the dynamics of FLID and tried to estimate or extrapolate numbers to regional or national scales. Studies usually involved interviews (De Bont, Komakech, et al., 2019; Duker et al., 2023) combined with desk studies (Hornun & Bolwig, 2020), comparing governmental statistics with import numbers on pumps, for example (de Fraiture & Giordano, 2014; Namara et al., 2014; Woodhouse et al., 2017a), or participatory mapping with farmers and extension workers (Beekman et al., 2014a). Although anecdotal and sometimes on relatively small scales (e.g., tens of interviews or hectares), these studies show that the extent and activities of farmer-led irrigation are expanding and that the total irrigated area is often more significant than that of the public irrigation schemes.

However, as it is difficult to do extensive interview studies or collect data over large areas, these (smaller) anecdotal studies may seem just that, anecdotal.

One alternative and promising approach to map irrigated agriculture over large areas is by combining satellite imagery and machine learning with field observations. Mapping irrigated areas through remote sensing (RS) derived images involves grouping pixels in the image into classes based on their spectral similarity and dissimilarity. Multispectral RS uses the principle that different materials (or land cover types) reflect and absorb different wavelengths of electromagnetic radiation (i.e., sunlight) with varying intensities (Figure 2). This variation is called a spectral signature and can be used to identify and classify different land cover types. Satellite sensors measure the reflected or emitted electromagnetic radiation.

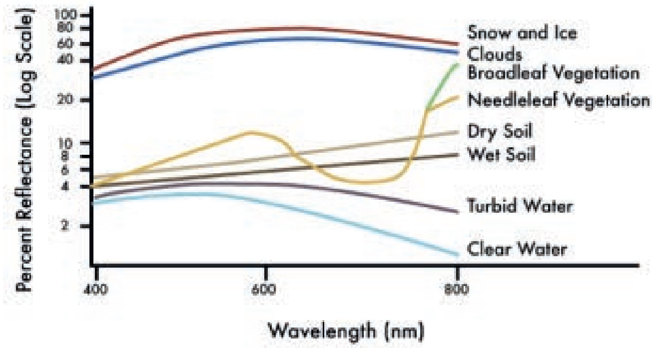


Figure 2 Example spectral signatures for different land covers (Source: ARSET - Fundamentals of Remote Sensing | NASA Applied Sciences, n.d.).

The leaves of vegetation strongly absorb visible light, in particular red light (~660 nm) and blue (~490 nm) and reflect green (~560 nm) and near-infrared light (~830 nm). Our eyes cannot see the infrared spectrum, so we see healthy vegetation as green. The mapping of irrigated agriculture in SSA often relies on the distinction between green irrigated crops, which receive sufficient water, and surrounding vegetation that turns brown or dies off due to water scarcity in the dry season.

To classify the pixels, machine learning algorithms are often used. These algorithms learn from a set of training data, which are typically ground-truth data collected and labelled through field observations. The training data contain examples of different land cover types. They are used to extract the pixel reflectance values per band for those areas. The machine learning algorithm then uses these spectral properties per class to recognise patterns in the data, which it can then apply to the rest of the satellite image to classify all pixels into the various land cover classes. The resulting thematic map shows the spatial distribution of the land classes present in the training data.

Several RS-based studies have estimated cropland and irrigated areas in Africa over the past years at local scale (for example, Fujihara et al., 2020; Magidi et al., 2021; Meier & Mauser, 2023; Traoré et al., 2019; Venot et al., 2021; Wellington & Renzullo, 2021) up to the regional and continental scale (Salmon et al., 2015; Vogels et al., 2019; Xiong et al., 2017). Table 1 shows some of these studies for the continent and the Horn of Africa to illustrate the wide variety in total cropland and irrigated area estimated. The table's primary purpose is not to present the exact numerical values but to demonstrate that remote sensing-based classification of irrigated areas is complex. The various producers of these studies employed different methodologies, including variations in the spatial resolution of the maps, satellite sensor selection, and algorithmic approaches. Additionally, the studies differ in their interpretation of what falls under the definition of irrigation. Thus, the table underscores the challenges and discrepancies in classifying irrigated areas using remote sensing techniques.

Table 1 Examples of continent and regional studies that use remote sensing to classify croplands and irrigated areas.

Study	**Product name**	**Resolution (m)**	**Cropland area (Mha)**	**Irrigated area (Mha)**	**Irrigated cropland as percentage of total cropland**	**Year**	**Area**
Salmon et al. 2015	GRIPC	500	202	13	6%	2015	Africa
Xiong et al. 2017	GFSAD250	250	296	24	8%	2014	Africa
Vogels et al. 2019 – the study compares the different products and their irrigated area extent.	RS study by Vogels et al. 2019	10	41,67	28	67%	2017	Horn of Africa
	IAAA	250	50,94	22,39	44%	2010	
	Globcover 2009	300	37,97	0,0004	0,001%	2009	
	GRIPC	500	22,08	1,15	5%	2015	
	GFSAD 1000 (GIAM)	1000	1000	9,93	1%	2009	
	AQUASTAT	variable	24,41	2,33	10%		

5. Problem statement and research questions

Mapping smallholder irrigation in SSA by means of remote sensing imageries has several challenges due to its heterogeneous and dynamic nature.

Although land cover classes can be distinguished based on spectral signatures, in practice, there is considerable variation within and overlap between classes when it comes to measuring irrigated agriculture:

- Different land cover but a similar spectral signature: instead of directly measuring irrigation, the measurement relies on the crops' spectral response to soil moisture. Typically, irrigated crops appear green, while surrounding natural vegetation turns brown during the dry season as soil moisture depletes. However, misclassification can occur in areas with sufficient soil moisture, such as near rivers or wetlands, where both natural vegetation and irrigated crops may remain green.

- Same land cover but a different spectral signature: spectral signatures of a land cover can also become mixed when fields contain weeds or different types of crops or when the spatial resolution of satellite imagery exceeds the size of individual fields, covering multiple crops or non-cropland vegetation. Moreover, remote sensing may not detect crops with a low leaf area that are, however, being irrigated.

- Complex shapes and arrangement of fields: the complexity of the landscape and arrangement of irrigated fields further influence mapping accuracy. Fields are often small and irregularly shaped, and intercropping practices, variations in agronomic activities, and differences in planting, harvesting, and irrigation timings occur.

- Meaning of irrigation: the subjective definition of irrigation can introduce data collection and classification biases, affecting field and remote sensing-based statistics.

Despite these challenges, remote sensing presents several advantages for mapping irrigated agriculture. It offers wide spatial coverage, facilitating the monitoring of trends across different temporal and spatial scales, particularly in regions where ground-based data is scarce. Remote sensing assists in prioritising field visits, enables consistent analysis of historical and near-real-time data, and is easily accessible. Furthermore, distinguishing between different classes can be achieved by considering factors such as image acquisition timing, vegetation colour variations, the maximum level of "greenness," and notable changes such as harvesting.

This thesis examines the production of RS maps and their ability to recognise and depict irrigated agriculture. While RS cannot directly measure farmer-led irrigation, it can capture the diverse and dispersed nature of small-scale irrigated agriculture, which requires subsequent interpretation through fieldwork and local and expert knowledge. Although RS offers a promising approach for mapping irrigated agriculture, it is crucial to be aware of potential pitfalls that might be overlooked. Therefore, **this research investigates the mapping of the spatial-temporal extent of irrigated agriculture in SSA using RS data and how modelling choices influence these maps**. Essentially, this thesis focuses on identifying and avoiding these pitfalls. To achieve this, the research addresses four key research questions (RQs) related to the process of irrigation classification:

- **RQ 1:** How have recent RS-based irrigation mapping projects in SSA consciously and unconsciously defined and classified irrigated agriculture, and how do these choices impact irrigation mapping?

- **RQ 2**: How does the selection of algorithms and composite lengths influence the accuracy of predicting irrigated agriculture in various landscapes and cropping systems?

- **RQ 3**: How does the size and composition of training data impact the accuracy of predicting irrigated agriculture in diverse landscapes and cropping systems?

- **RQ 4**: What approaches can enable the successful application of models trained on one area to other areas, minimising the need for extensive field data collection?

To reach the aim of this research, I use four case studies in Mozambique, Chokwe and Xai-Xai in Gaza province and Manica and Catandica in Manica province. These areas were chosen for their diverse agroecological characteristics and the presence of irrigated agriculture, including small-scale and large-scale systems. These characteristics make them a suitable playground for investigating irrigated agriculture mapping.

The next section further elaborates on the challenges, and Section 7 outlines the remaining chapters of this thesis.

6. Conceptualisation of research objective: uncertainties in RS-based estimates

Although Figure 2 indicates the basic differences in reflection based on which different land cover classes can be distinguished, there is a lot of variation and overlap between and within classes. Additionally, satellite sensors do not measure irrigated agriculture itself; instead, satellites measure the crops' spectral response to sufficient soil moisture.

This is one of the primary sources of confusion between classes indicating irrigated agriculture and natural vegetation – it is assumed that irrigated crops remain green in periods when the surrounding natural vegetation dies off and turns brown. However, in areas with sufficient soil moisture, such as near rivers or wetlands, the natural vegetation will also remain green and potentially be classified as irrigated agriculture.

To overcome this, the timing of the used images and statistics over the growing periods becomes relevant, allowing for the distinction between irrigated croplands and the surrounding natural vegetation. In other words, aspects such as how fast the greening and browning of corps versus surrounding vegetation happens, the maximum 'greenness', or abrupt changes (such as harvesting) can be used to distinguish classes with similar spectral signatures further.

Although this in itself is challenging, mapping smallholder adds multiple dimensions due to the heterogeneous and dynamic nature of FLID processes, including the small, irregularly shaped fields with in-class variance as a result of inter- and mix-cropping systems and variability in the timing of agronomic activities such as planting, harvesting, and irrigation (Bey et al., 2020; Nabil et al., 2020; Rufin et al., 2022). I will show this with a few examples in the following four subsections (based on Weitkamp & Beekman (2022)).

6.1. Spectral signatures

In a world optimal for RS, all irrigated fields would be kept free of weeds, and clear boundaries between crop types and surrounding vegetation would be visible, as the four images below show (Figure 3). Each (section of the) field has its own crop, free of weeds. The beans in Figure 3A have a different green hue (i.e., spectral signature) than the lettuce in Figure 3B or the cabbage in Figure 3C. This is also visible in Figure 3D, where three crops are visible and can be easily distinguished by the naked eye. The spectral signatures of each crop type are more clearly defined this way and more unique (i.e., less overlap).

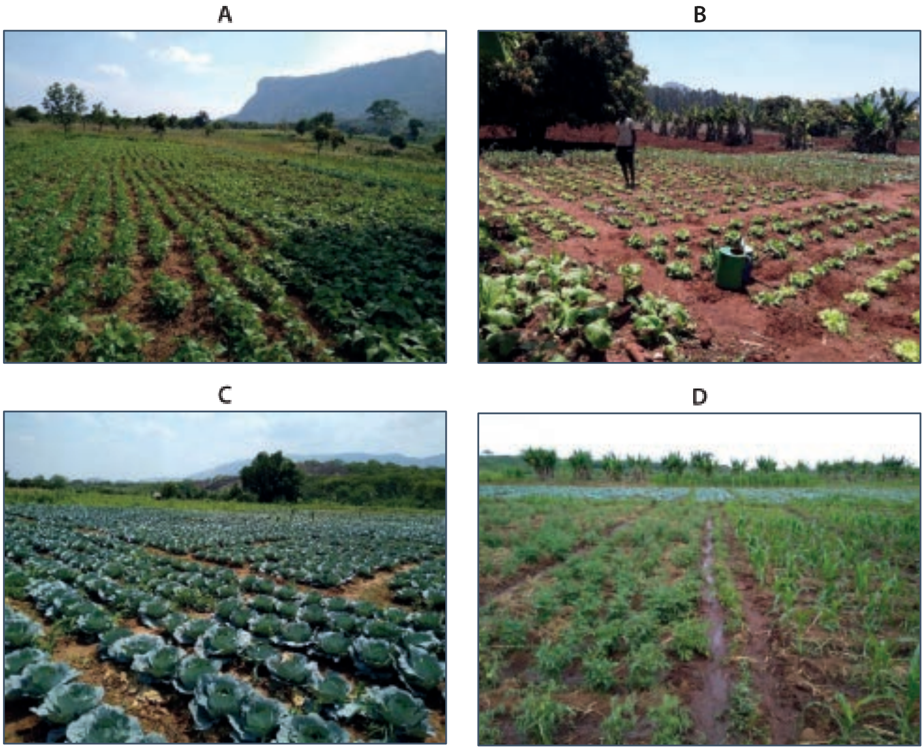


Figure 3 Fields clear of weeds have purer spectral signatures and are easier to identify with RS. A) beans, B) lettuce, C) cabbage, D) tomato (foreground left), maise (foreground right) and cabbage (background). Photos by Timon Weitkamp.

However, this ideal scenario is not always the case. In some instances, fields contain more weeds than crops, as illustrated in the following four photos featuring croplands of beans, cabbage, maise, and pumpkin with a considerable weed presence (Figure 4). It is not difficult to imagine the confusion between the four fields. Additionally, some farmers have agroforestry systems in which multiple crops and trees are grown on the same field, leading to even more confusion. In contrast to the "pure" spectral signature of the fields shown earlier in Figure 3, the spectral signatures of the weedy or multi-crop fields appear more "mixed," with greater overlap between them. Essentially, multiple "crop-weed" or "crop-crop" combinations lead to the same spectral signature.

A further complication in class confusion is natural vegetation areas, which contain the same weeds present in agricultural fields besides larger shrubs and trees (Figure 5). If the agricultural fields contain enough weeds or the crop's green hue is similar to that of a weed, the spectral signature will be similar to that of natural vegetation.

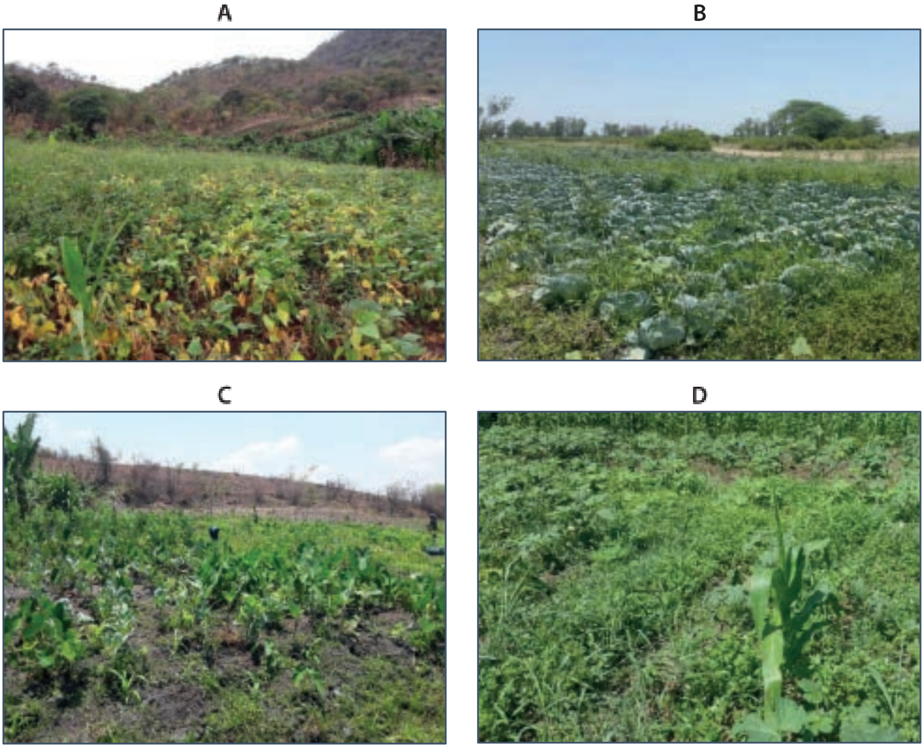


Figure 4 These fields contain many weeds and mix the spectral signature; consequently, these fields are harder to recognise as croplands by machine learning algorithms. Beside weeds, these fields contain beans (A), cabbage (B), maise (C), and pumpkin (D). Photos by Timon Weitkamp.

Figure 5 A) Natural vegetation can have the identical spectral signatures as croplands with mixed signatures (i.e., fields with many weeds). B) Natural vegetation grows on fallow fields (old maise stalks visible on the ground). Photos by Timon Weitkamp.

6.2. Spatial resolution

The above examples demonstrate that different classes can possess overlapping characteristics but also that similar classes can possess different characteristics. In addition to the type of vegetation being classified, pixel coverage is also crucial. The following two fields (Figure 6) are smaller than the smallest open-source pixel size (Sentinel-2, 10x10 meter), which implies that the pixel includes non-cropland vegetation (with mixed spectral signatures). Another consideration is that the field may not be covered by a single pixel in the first place but by part of two or more pixels, further "diluting" the spectral signature of the crop with non-crop characteristics. In Figure 6A, this includes the stream and banana plant in the background within the pixel. In Figure 6B, a single pixel may cover parts of two different fields with different crops, further mixing the spectral signature as one crop may be planted or harvested at a different time or have a different growing length.

Figure 6 Pixels covering small fields contain more non-crop spectral signatures. A) The stream and banana plant and B) multiple crops will also be covered by the pixel classified as irrigated agriculture. Photos by Timon Weitkamp.

Additionally, the shape of irrigated fields and their spatial arrangement (clustered or with non-cropland classes in between), or the complexity of the landscape, influence how easy it is to map irrigated agriculture (Meier & Mauser, 2023).

Figure 7 demonstrates how spatial resolution affects the visibility of four fields with different sizes, shapes, and crops in a 30 × 30-meter area. With a 30 × 30-meter resolution image, only one dominant field (Field 2) can be seen. When the resolution is increased to 10 × 10-meters, three additional fields (Fields 1, 3, and 4) become visible, but some fields still dominate the pixel. Using a 2.5 × 2.5-meter resolution image reveals almost all field contours. In all three scenarios, the areas of each field are calculated, and as resolution decreases from 30 to 2.5 meters, the area of the dominant field (Field 2) decreases while the other fields' areas increase.

The significance of the small and irregular nature of smallholder fields is easily overlooked, as demonstrated by Figure 7

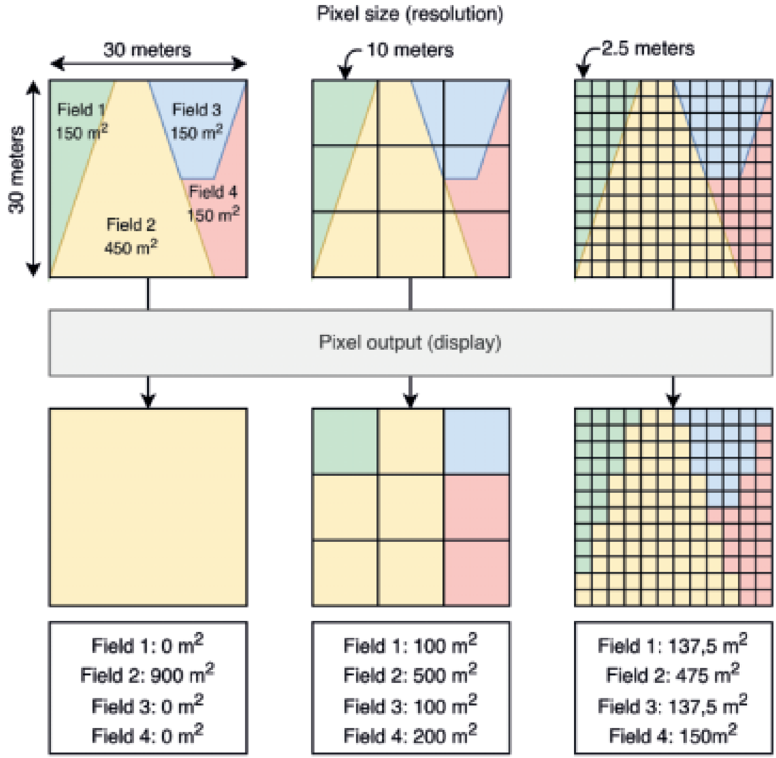


Figure 7 How spatial resolution determines if fields can be 'seen' or not. The area of fields 1, 3 and 4 are 150 m^2 whilst field 2 is 450 m^2. Source: own compilation.

6.3. Crop type and soil cover

RS may not always be able to detect all crops; crops in stages without sufficient leaf area may not be detectable. For instance, the onions shown in Figure 8A do not cover a substantial portion of the soil, even when maturing. Consequently, a pixel over this area would probably display as bare soil, as the amount of bare soil in the pixel's spectral reflectance is greater than the amount of vegetation. The second image (Figure 8B) depicts cabbage in its early growth stages in the foreground and background. The satellite image can only pick up this crop when it has grown for several weeks and covers enough bare soil to affect the spectral reflectance of that pixel. At the same time, there is a small patch of paprika, which will mix the signal of the surrounding cabbage pixels.

Figure 8 Crops with small leaves show as bare soil on satellite images. A) onions have small leaves and do not cover much of the soil. B) young cabbage does not cover much soil either but will eventually. Photos by Timon Weitkamp.

6.4. Areas with high water tables

Another factor that makes RS-based classification of irrigated agricultures challenging is areas with a high water table. As explained in previous sections, farmers often use buckets, cans, or furrows to apply water to crops, in addition to the more widely known sprinkler and drip irrigation. However, there are also areas where excess water needs to be drained before non-rice crops can be grown. The first two photos in Figure 9 (A and B) depict irrigated regions near Xai-Xai, Mozambique, that employ canals and sluices for active water management. During the dry season, these fields can be prepared and sowed with horticulture crops that thrive in areas with a high water table. However, spectrally speaking, these fields resemble fallow fields after rice cultivation (Figure 9C), where natural vegetation starts growing during the dry season, also benefiting from the high water table. Reeds (Figure 9D) also flourish in water-rich areas and remain green for a long time during the dry season, making them prone to misclassification as rice.

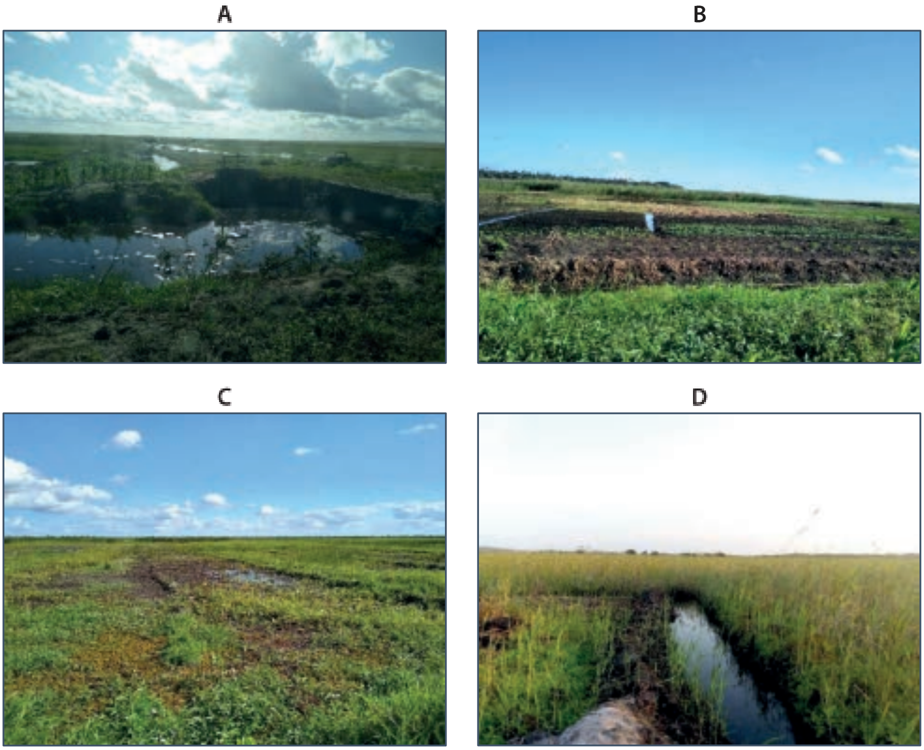


Figure 9 Areas with high water tables also contain much natural vegetation that remains green during the dry season, like crops. A) Area with high water table. B) Similar area but fields are being prepared and cropped when sufficiently drained. C) Grassland with ponding water. C) Reeds. Photos by Timon Weitkamp.

6.5. Definition of irrigation

RS analyses rely on using defined categories of land use. Still, FLID often does not align with these conventional boundaries (Figure 1). The reality of farming practices rarely fits neatly into these categories (Woodhouse et al., 2017a). Therefore, efforts to generate irrigation data are heavily influenced by the map makers' perception of what constitutes irrigation (often modernistic) rather than truly reflecting actual farming practices (De Bont et al., 2019; Venot et al., 2021). This applies to both RS-based and field-based data collection methods. Both field-based and RS-based statistics can be (willingly or not) biased due to human choices. For instance, official field-based statistics may not account for informal or unauthorised irrigation systems, while RS-based statistics can be affected by the classification of irrigation types and methodological choices. Throughout the thesis, I use the word 'irrigation' to refer to any form of (active) water management that involves applying or draining water from fields.

6.6. Summarising the uncertainties

Mapping irrigated agriculture can be challenging for many reasons, from the (sufficient) availability of satellite data and processing capacity to the varying landscapes agriculture takes place into the many interpretations of when irrigation is irrigation. The examples from the above sub-sections illustrate the heterogeneity of the phenomenon being mapped and the many considerations it takes to map it. In Figure 10, I have tried to summarise many of these aspects in a conceptual representation to show the various elements that come together when using RS for mapping irrigated areas. It also shows that the users have their own set of elements through which they interpret the map, further complicating the whole interplay.

The main body of this thesis focuses on the individuals responsible for the production of maps (*Production side*) and the intended users (*Application side*) and, to a lesser extent, the subject being mapped, which in this case is irrigated agriculture (*Feature side*). The map maker and user are influenced by overlapping elements, such as their knowledge of the area, their understanding of what constitutes irrigation and their broader stakes and interests. However, as these do not coincide, their interpretation of the map will differ.

Furthermore, the map maker's decision-making process is shaped by various factors, including available financial resources, time constraints, model selections, data availability, and institutional guidelines. Conversely, the map user's perspective is influenced by their comprehension of the methodologies employed in remote sensing-based map generation and their familiarity with such maps. This dynamic interaction between the map maker and the reader can lead to mutual influence. The map maker may align with the user's objectives, and the user may gain deeper insights into the intricacies of the map creation process.

Considering this framework, the interrelationship and interplay among the four research questions become more evident. RQ1 examines the relationship between the mapmaker, the mapped feature, and their interaction. In contrast, RQ2 focuses on the choices made regarding models and data in the mapmaking process. RQ3 delves into the practical aspects of collecting field data and its associated challenges. Lastly, RQ4 investigates the scalability of models and the transfer of knowledge from one area to another, encompassing both personal expertise and the insights gained by the model.

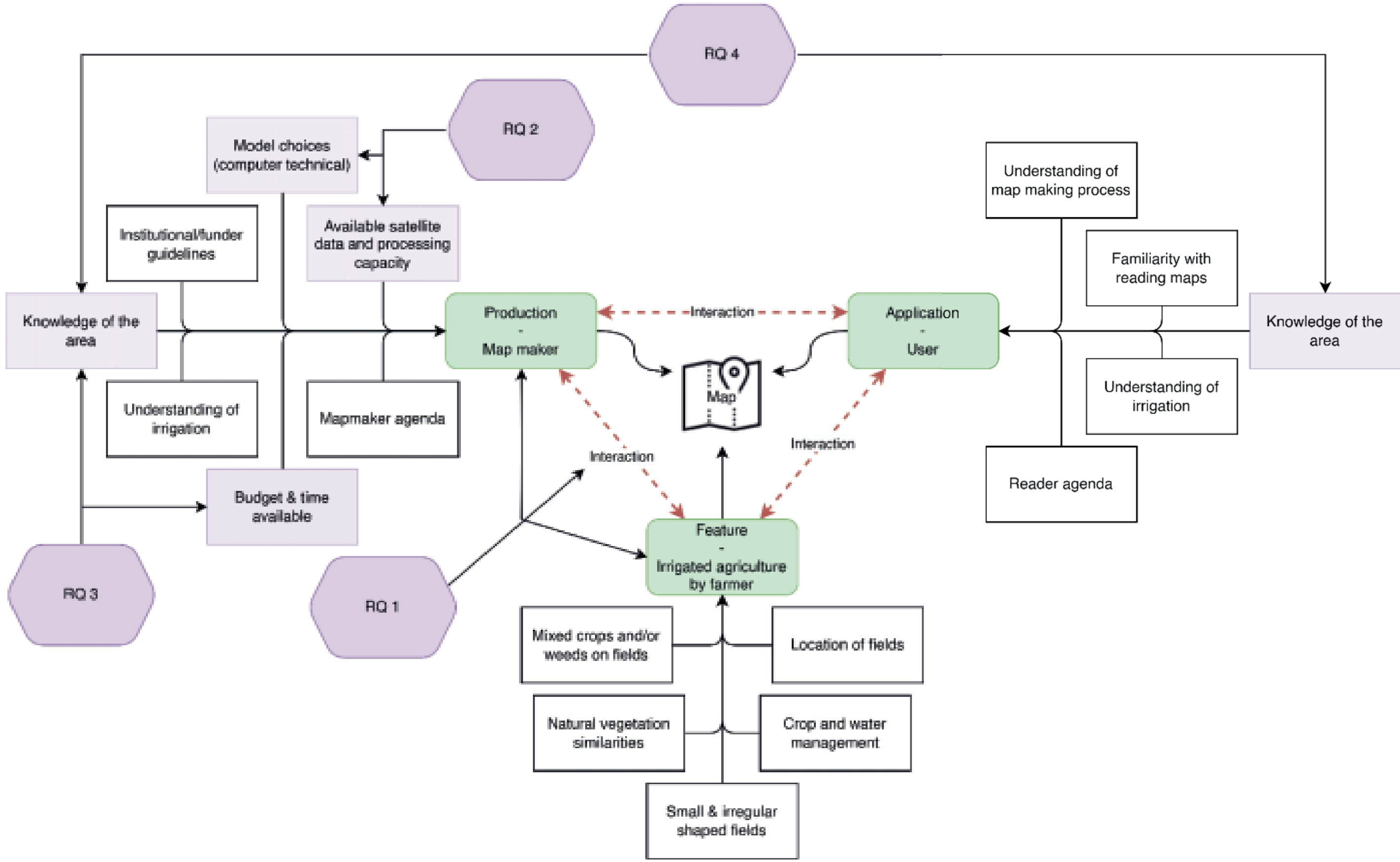


Figure 10 Schematic overview of the many aspects to consider when making and reading a map of irrigated agriculture.

7. Contributions and thesis outline

This thesis consists of six chapters, including this introductory chapter. Chapters 2-5 respectively address the four research questions formulated in the previous section, and the final chapter (Chapter 6) reflects on the findings of this study.

In Chapter 2, I look at common RS classification steps that all mapping studies go through and determine the authors' choices, consciously or unconsciously, based on what was reported. Specifically, I reviewed literature on studies that mapped irrigated agriculture in SSA from 2015 to 2022. The overabundance of options and possibilities, such as study extent, the sensor used, data collection strategies or classes used, made it difficult to compare the studies directly. To compare the studies, I developed a framework that shows what classification steps a study goes through and how certain choices might influence/bias the final results. This allowed me to compare the studies on the classification process and outcome rather than the exact choices or scripts, as those are context-dependent. Parameter values used in one area might not be directly applicable to another. However, the logic behind using those parameter values can be reasoned and transferred. This chapter answers **RQ 1**: *How have recent RS-based irrigation mapping projects in SSA consciously and unconsciously defined and classified irrigated agriculture, and how do these choices impact irrigation mapping?*

Where in Chapter 2, I examine the information reported in published articles and their potential biases, in the subsequent two chapters I delve deeper into specific classification steps to demonstrate how particular modelling choices can directly influence the resulting outcomes. Specifically, Chapter 3 focuses on the algorithm and satellite data employed, while Chapter 4 examines the training data aspects in greater detail.

Chapter 3 investigates how four different algorithms and four different satellite data composite lengths classify irrigated agriculture in four study areas with different climates, landscapes, and farming practices, which cover various farmer-led agriculture practices and contexts found in SSA. **RQ 2** is answered here: *How does the selection of algorithms and composite lengths influence the accuracy of predicting irrigated agriculture in various landscapes and cropping systems?* To create composite images, multiple satellite observations of a specific area are merged to form a single, representative image. This merging of measurements from different observations results in comprehensive datasets, which facilitate detailed analysis of large areas of interest. Different composite lengths can be created by merging more or fewer images (i.e., longer or shorter time). In this chapter, I show how different algorithms and input data result in different maps. However, there are areas of overlap, or 'hotspots', where all models agreed irrigated agriculture could be found. I continue by exploring how these agreement maps – maps showing how many models classified the pixels as irrigated agriculture – can be used to visualise both large-scale and small-scale agriculture, but also that the combination of models allows for determining which areas are irrigated with greater certainty.

Despite the comprehensive reporting of the algorithm and input data, the training data used to train the classification models often has a greater impact on the results obtained. However, this is hardly reported on. In Chapter 1, I describe how smallholder irrigation can be underrepresented in data collection due to its often dispersed and dynamic nature and where it can be found, but also due to the perception of what irrigation is by the data collector. In Chapter 4, I explore these biases through different scenarios. I examine if fewer data would yield acceptable results, how the training data composition matters, and what would happen if the data collection focuses too much on irrigated agriculture. I also describe a method to determine if the amount of training data is large enough and the composition, which others can use to determine if further data collection is needed. Here, I answer **RQ 3**: *How does the size and composition of training data impact the accuracy of predicting irrigated agriculture in diverse landscapes and cropping systems?*

Chapters 2-4 each cover specific steps in the classification process within a specific geographical area, but they do not look at the useability of the developed models in a different area. Chapter 5, therefore, examines the transferability of models trained in one area and applied to another. Field data collection is expensive and time-consuming; hence, using pre-trained models for new areas can be cost-effective. However, unsurprisingly by now, the heterogeneous landscapes and irrigation practices mean the models likely overlook irrigated areas when transferring models. In this chapter, I explore if models can be transferred to areas with different climates and landscapes, but also if combining different features into one model improves the generalisability of the model – the model has seen more options of where irrigated agriculture can be found. Although model transfer saves time and effort, it comes at the cost of accuracy, as there are likely classes in the *new* area that are not found in the area on which the model was trained. Hence, I also explore how (dis)similarity in areas can be expressed and how additional data collection in the most dissimilar areas improves the final results. This chapter addresses **RQ 4**: *What approaches can enable the successful application of models trained on one area to other areas, minimising the need for extensive field data collection?*

Finally, Chapter 6 gives a summary of the main findings of this thesis. It also delves into the implications and shortcomings and discusses future research directions.

Chapter 2

Towards transparent reporting in Remote Sensing-based mapping of irrigated agriculture in sub-Saharan Africa: development of a reporting framework

Timon Weitkamp, Gert Jan Veldwisch, Poolad Karimi, Charlotte de Fraiture

1. Abstract

Irrigation is critical for intensifying and expanding agriculture in sub-Saharan Africa (SSA). Policymakers increasingly use remote sensing-based techniques to identify previously unknown irrigated areas. As smallholder irrigation practices in SSA vary widely depending on the type of crops, plot sizes, irrigation methods and landscapes, how maps are made that depict their extent becomes more important. We have identified methodological choices in at least eight essential domains for classification or irrigated agriculture sampling design, labelling protocol sets of classes, field data collection, predictor variables, algorithm adequacy, input variables, accuracy assessment, map seasonality, and code and data sharing. This study demonstrates and systematises how these choices affect classification in a reporting framework. We found that none of the reviewed articles sufficiently documented all classification steps when applying the framework. Although the reasons for not reporting are unknown, the lack of explicitly made choices hampers a proper evaluation of irrigated agriculture's extent, particularly smallholder irrigation. Ultimately, this may reinforce the impression that smallholder irrigation is irrelevant because it does not appear on maps. Finally, we conclude that sharing extensively documented irrigation mapping methodologies promotes the adoption of best practices across different regions or countries. Policymakers and practitioners can learn from successful experiences and avoid repeating mistakes made in other contexts. This approach advances irrigation practices worldwide by fostering collaboration and knowledge exchange.

2. Introduction

Irrigation is critical for both the expansion and intensification of agriculture in sub-Saharan Africa (SSA) to mitigate erratic climate conditions and reduce dependence on erratic seasonal rainfall. With a mere 4% of the cultivated area reportedly irrigated (FAO 2021), there is considerable potential for expanding the irrigated area. On the other hand, statistics on the extent of smallholder irrigation under-estimate how much land is irrigated (Beekman et al., 2014b; Veldwisch et al., 2019b; Venot et al., 2021; Woodhouse et al., 2017b).

These smallholder irrigation practices, i.e. irrigation on relatively small farms, are often initiated, operated, maintained and constructed by local people using local materials and ideas, referred to as 'farmer-led irrigation development' (FLID) (Nkoka et al., 2014b, p. 2), as opposed to irrigated areas developed or initiated by the state or agro-industries. Irrigation development by farmers themselves takes place in diverse contexts and is dynamic in nature (see Box 1), which makes it very difficult for authorities to keep track and stay informed of its ever-changing extent, often scattered widely and challenging to reach. Invariably, the responsible institutes do not have the organisational capacity, budget or staff to monitor large areas (de Bont et al., 2019).

Another reason behind the phenomenon of under-reporting is that these same institutes often see smallholder farmers' irrigation practices as illegal, inferior or irrelevant, as they do not conform to developmental ideals or do not employ 'modern irrigation technologies', even though smallholder farmers vastly contribute to higher-level government goals such as food security (de Bont et al., 2019). It is often a combination of land lease rights, size of farms, and water abstraction that allow farmers to irrigate on the boundary margins between legal and illegal. The state cannot monitor, develop, or support all these farmers whilst the farmers support local food demands. Thus, the lack of technical and organisational capacities is intertwined with the working cultures, narratives and politics of irrigation development in SSA, leading to the invisibility of farmers' practices (Venot et al., 2021). Even if public institutes recognise smallholder irrigation, it is often seen as backward and needing conversion by external expertise (de Bont & Veldwisch, 2020; Hounkonnou et al., 2012). However, irrigated areas developed by smallholder farmers develop faster and more cost-efficiently than those developed by external expertise (Beekman et al., 2014b). In fact, the attitude of investment agencies and donors is changing. For instance, the World Bank is in the process of recognising the role played by FLID and is developing plans to support it (Izzi et al., 2021). There is a growing emphasis on promoting farmer-led irrigation development within investment portfolios, despite the continued presence of large-scale irrigation projects (Harmon et al., 2023)

The use of remote sensing (RS) for monitoring smallholder irrigated areas is a progressive alternative to conventional periodic census surveys, often based on known (permanent) public and large-scale irrigation systems (Venot et al., 2021). RS can greatly help monitor irrigation, as it potentially reduces the need for extensive field visits and their associated costs, hardware and staffing needs. Over the last twenty years, developments in algorithms, the spatial and temporal resolution of satellite imagery, and the use of time series have improved, boosting methodological developments (Massari et al., 2021). With RS techniques, modellers can interpret earth surface reflections to identify agricultural fields, land cover change over time, specific regions of irrigated agriculture in large landscapes, and information on irrigation timing (Massari et al., 2021; Ozdogan et al., 2010). However, mapping irrigated lands using RS is a complex technical process in which the output maps' accuracy and reliability greatly depend on how the task is executed (Ozdogan et al., 2010). If irrigation maps are to be accurate and reliable, it is essential that makers of the map address the following challenges: 1) the interpretation of 'irrigation', 2) classification of distinct categories of land use and land cover, and 3) reproducibility and transparency:

1. How modellers and field staff understand and interpret smallholder irrigation plays a significant role in RS-based classification, just as it does in classifications based on conventional census surveys. Mapping is a process of interpreting a reality through a model rather than 'mirroring' nature. Accordingly, multiple interpretations can exist simultaneously (Comber et al., 2005), even for a single area, solely because modellers and field staff have diverse backgrounds and experiences. To effectively map irrigated areas, modellers must know when and where irrigation is applied, which is often site-specific, meaning that the classification choices are also site-specific (Ozdogan et al., 2010). Both through field data collection and ground truthing/validation, data-generation processes reflect modellers' understanding of what irrigation is (Venot et al., 2021), laying the basis for irrigation classification. In practice, it is often the interpretation by the public irrigation institutes that is reflected in maps, as their staff collects the field observations required for the training of the model.

2. The strongly heterogeneous and dynamic landscapes in which smallholder irrigation often occurs complicate the distinction between different classes. The small fields, inter- and mixed-cropping systems, and variability of irrigation timing, method and quantity can often not be captured through satellite images' sometimes insufficient spatial and temporal resolution (Bégué et al., 2018; Veldwisch et al., 2019b). The classification maps depend on categories that are discrete and mutually exclusive, yet it is challenging to develop categories that capture a continuous mosaic landscape (Foody, 2021), such as where smallholder

irrigation and other spectrally similar categories like natural vegetation can be found.

3. Even when modellers address the above two challenges, the lack of transparency of choices made in classification studies can significantly limit the usefulness and uptake of the resulting maps beyond the original study. Modellers need to make choices during the classification process, which are often subjective and different for each research project. Therefore, modellers must describe and acknowledge uncertainty in models and analyse their conceptualisations, values and assumptions regarding the model's parameters and construction (Melsen et al., 2018). This allows modellers to analyse and discuss irrigation data and data-generation methods in relation to the narratives and politics of irrigation development in SSA (Venot et al., 2021). In doing so, models will become more transparent, and the results can be reproduced and validated.

All three processes described above contain mechanisms that may lead to missing out on irrigated agriculture, particularly irrigation by smallholders. However, there is no easy way to evaluate if and how irrigated agriculture may have been missed.

This study aims to set a first step in systematically reporting on all the steps that classification studies go through, starting with smallholder irrigation in SSA. Additionally, we aimed to take the first step towards promoting systematic reporting on all the classification steps involved in such studies. We developed a framework that allows modellers, reviewers, editors, and funders to evaluate if all relevant aspects are reported quickly. We also highlight in what way underreporting on choices could affect the results.

We first describe how we selected publications that mapped irrigated agriculture in SSA in recent years (Section 3) and the framework we developed to make modelling choices explicit (Section 4). We then analyse to what extent recent RS studies on irrigation in SSA report on these choices (Section 5). In the final sections, we discuss the implications of not reporting on these steps on irrigation extent and policy (Section 6) and conclude the study in Section 7.

Box 1: Illustration of common smallholder irrigation practices in SSA and how they can be missed by officials

Furrow irrigation in mountainous areas: Furrow irrigation often uses water diverted from (semi-)permanent mountain streams and is usually found in mountainous SSA regions. One stream typically provides water to several furrow irrigation systems, which are often interlinked directly or indirectly (seepage or return flows to the stream), resulting in complex socio-hydrological networks. These systems grow or shrink depending on periods of more or less than average rainfall, and population dynamics can lead to a reconfiguration of the furrows changing in time and space.

Pumps from open water and groundwater: Farmers pump water from open water bodies and shallow groundwater for horticulture production. Over the past 20 years, pumps have become increasingly available and affordable. Sand river aquifers (groundwater systems of sandy deposits in river beds) have been used for crop production for many centuries, although this water source is under-utilised. Depending on the pump's capacity to transport water, the fields do not have to be close to the water source. The pump's flow also determines the area that the farmer can irrigate, which can be relatively small for solar-powered pumps (0.5 ha) to multiple hectares for petrol or electric pumps. Water and energy costs can limit pump usage.

Shallow groundwater in valley bottoms and well-drained depressions: Farmers also grow crops (often vegetables) on relatively wet valley bottoms (*baixas* or *dambos*) in regions that are usually dry for a large part of the year. Farmers grow crops for their own consumption, and for the market, either with residual moisture in the soil or from water in a natural drain or in shallow wells up to 5 metres deep. Farmers access this water with watering cans/buckets or pumps (treadle or petrol). There may be too much water during the wet season, in which case drainage canals are dug. During the dry season, the residual water content and shallow wells can become low, in which case the cropped area also decreases.

How states and officials in SSA view smallholder irrigation can explain why smallholder irrigation is often not taken into account (de Bont et al., 2019). FLID is sometimes ignored by government officials, while it contributes to development goals, such as regional and national food security. Smallholder farmers make active investments in inputs: they use pumps, fertilisers, improved seeds and pesticides to increase production. However, government officials may use a narrow interpretation of 'good' irrigation. Good irrigation, based on the prevailing view, happens when the system is planned, designed, or managed by a trained engineer, in line with political priorities determined by the government – often without understanding the relationship between the designed scheme and the farmers 'existing use of land and water (Venot et al., 2021, p. 13). This leads to labelling irrigation systems developed by smallholders as 'sub-optimal' and inefficient. It excludes these areas from the strategy to increase agricultural production and leaves them out of policy. This contributes to irrigation officials overlooking a rapid and widespread irrigation development process initiated by small-scale farmers. There are also practical reasons why policy does not feature smallholders, namely the weak technical capacity and limited budgets within government agencies, which reduce the potential to respond adaptively and instead depend increasingly on standard technological approaches to irrigation development. Consequently, engineers do not identify the irrigation schemes built by farmers as adaptive, cost-effective measures, even though they are aware of them (de Bont et al., 2019). Not recognising the phenomenon, together with a lack of capacity to visit these areas, excludes smallholder irrigation from the mapping methodology.

Sources: de Bont et al. (2019); Duker et al. (2020); Venot et al. (2021); Woodhouse et al. (2017)

3. Methodology

3.1. Development of a framework for analysis

Although all modellers of (irrigation) classification studies go through roughly the same steps, they do not always document these steps and the corresponding choices in the final result. This lack of documentation on the exact methodological choices makes comparing studies difficult.

We have developed a framework that allows us to compare classification studies that are methodologically different in multiple ways. The framework focuses on transparency, reproducibility, and the implications of not reporting on those steps. In this study, we look for the identification of irrigation as a discrete category. In reality, agricultural practices combine rainfall and other water sources in various ways, and it is sheer impossible to distinguish between rainfed and irrigated agriculture strictly. We neither move beyond the general categorisation of irrigation to recognise different types of irrigation systems or field application methods such as distinguished in the AQUASTAT database of the Food and Agriculture Organisation of the United Nations (FAO 2021; see also Venot et al. 2021). Our understanding of irrigation refers to any form of (active) water management that involves applying or draining water from fields, independent of the infrastructure used, how it is applied, or who applies it.

We summarised commonly used classification processes into eight essential steps based on literature and personal experiences. We then assessed how common preferences could influence the representation of smallholder irrigation. The assessment of these eight steps can be found in Annex 1. Each classification step has a key question that can be answered with 'fully reported', 'partially reported' and 'not reported' to give insight into the choices reflected in the in different studies.

By going through the eight classification steps (and their questions), a researcher, modeller or policymaker can quickly evaluate if and how modelling choices were made explicitly. In the results section (Section 4), we use this tool to assess existing literature on smallholder agriculture mapping in SSA to demonstrate its potential for making model choices explicit. Annex 2 shows the applied framework results.

3.2. Selecting articles

We selected literature on irrigated agriculture classification through the database of Scopus. Our search focused on the title, abstract and keywords of English-language articles from 2015 to May 2022. We chose 2015 as this was just before Sentinel 2 was launched and made it possible to map at a higher spatial and temporal resolution. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) method and applied the following criteria:

- LULC mapping (search terms: '*classification*' OR '*mapping*').
- Irrigation or irrigated agriculture had to be present as a class (search terms: '*irrigated agriculture*', '*irrigation*', OR '*cropland(s)*').
- Mapping the extent of irrigated agriculture using satellite-derived imagery (search terms: AND NOT 'UAV').

We did not include the names of satellite constellations (Sentinel, MODIS, Landsat, etc.) or the type of RS data (optical, radar) in the search criteria.

We designed an automated general query on Scopus to extract those results that were most likely to satisfy the selection criteria, which resulted in 646 results. The titles and abstracts were then screened based on the exclusion criteria stated above, as well as the geographical location of the study, after which we assessed the full texts, resulting in 22 records for the final review. ; The geographical location (i.e., study areas not in SSA) or the lack of an irrigation class were the main reasons for exclusion in this step. After this final step, we included three articles to the list that were not captured by the query but fit the requirements. These three articles were known to us through other queries.

We grouped the literature according to their different objectives to highlight the importance of the 'irrigation' class to a study:

- Agriculture (n=8): mapping of agriculture (croplands), of which irrigated and rainfed agriculture are two of the classes.
- Irrigation (n=8): mapping of irrigated agriculture is the main focus.
- LULC (n=3): other land uses/land cover are mapped besides rainfed and irrigated

agriculture. Generally, there are many classes.
- Other (n=7): land use is mapped, but the goal is more on the methodology than the maps.

3.3. Analysis of the papers using the framework

Using the framework and table described in Annexes 1 and 2, we analysed what the 25 articles reported for each classification step with regard to potentially identifying smallholder irrigation (Table 1 in the Annex). Note that this framework is not meant to compare the selected articles with the 'perfect method' of mapping irrigated agriculture. Instead, the framework is used to compare *if* and *how* the authors have documented the specific steps, regardless of whether it is the "right" choice for that situation.

4. A framework for making modelling choices explicit

Image classification is the process by which a modeller assigns areas with similar spectral signatures to land use classes, commonly by using classification algorithms. Modellers typically draw on their individual experience and expertise when making decisions on processing paths, algorithms and sensors (Khatami et al., 2016). Even though these decisions have marked influences on the model's output, they are seldom made explicit. Morales-Barquero et al. (2019) reviewed 304 papers on how verifiable (i.e., reproducible, transparent, well documented) the accuracy assessment was. They found that two-thirds insufficiently reported this aspect. Although the authors note that surveys and interviews are needed to explore why decisions are not always reported, they also point out, in line with Castilla (2016), that behaviour will change if editors and reviewers demand better reporting of the choices. One practical way to do this is with a framework that makes modelling choices explicit. This section describes the main decisions needed for the different steps of the classification process and the potential consequences for the visibility of smallholder irrigation, though we believe the steps can also be applied to broader remote sensing topics.

Most commentators (Foody et al., 2016; Olofsson et al., 2014; S. Stehman & Foody, 2019; Stehman & Wickham, 2020) argue that RS studies should include three main elements for rigorous accuracy assessment: sample design, response design and analysis. Elaborating on these three main elements, we define eight common classification steps grouped into three overarching elements for reporting (Figure 1): i) training and validation data, ii) model, and iii) presentation.

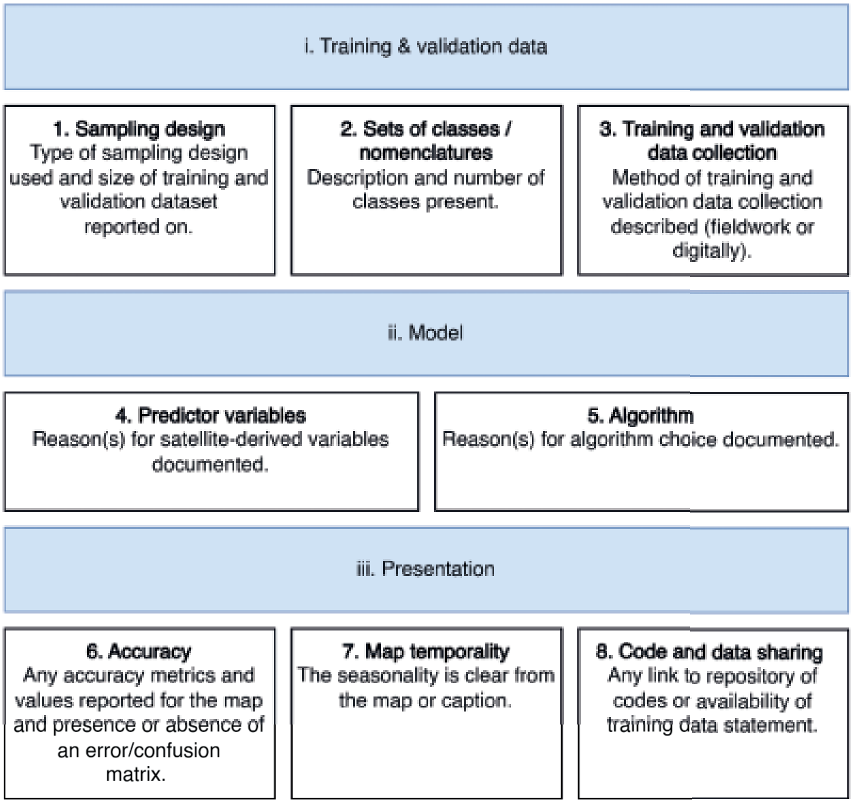


Figure 1 Framework overview containing the eight steps divided over three elements.

We will first describe the individual steps that make up this framework, after which we will apply them to the selected articles in Section 5.

4.1. Element 1: Training & validation data

This first element contains all the preparation steps for the eventual classification related to training and validation data. It includes information about the classes and field data collection.

4.1.1. Step 1 – Sampling design

The sampling design is the protocol for selecting the sample pixels or polygons that will form the basis of the accuracy assessment (Olofsson et al., 2014). In other words, the sampling design defines when, where, how many and what type of samples are collected (Elmes et al., 2020). Stehman and Czaplewski (1998) were among the first to state that modellers should properly document this to enable the reproducibility of study results. Morales-Barquero et al. (2019) found that the topic is still relevant. Unfortunately, this is not a priority in most classification studies. For example, Ye, Pontius, and Rakshit (2018) found that only one-third of the 209 articles they reviewed mentioned how samples were selected.

Two main methods of sampling are commonly employed in RS studies: purposive or opportunistic sampling and probability sampling. With purposive or opportunistic sampling, samples are collected opportunistically, for example, near a road or in known areas. The other method is probability sampling (using a simple random, stratified random, systematic, or clustered design), in which all pixels have an equal chance of being selected. Both sampling designs are acceptable for training the (machine learning) classification model; however, a probability sampling design should be used for the accuracy assessment (validation) (Stehman & Foody, 2019).

The sampling of field data is a trade-off between practicality, such as available time, ease of travel, access and budget, and representativeness of the classes in the sampling. Accounting for the spatial and temporal variability in all classes (e.g. how grassland greens and browns during the year) is essential (Johannsen & Daughtry, 2009), which requires field visits to understand (step 3). Homogeneous landscapes, such as large agricultural fields, may require fewer samples and can be more spread out than the complex landscapes where smallholder agriculture occurs. Although the amount of training data is small compared to the eventual maps, numerous studies have found that the sample size and quality influence the classification accuracy more than the algorithm used (Elmes et al., 2020). Large and accurate training datasets are generally preferable, although they may not always be feasible because of limited time or access or interpretation constraints (Maxwell et al., 2018).

4.1.2. Step 2 – Sets of classes

Sets of classes - or nomenclatures – are not standardised, and variations exist because of political or technical choices (Comber et al., 2004). Consequently, each research on irrigated agriculture has a slightly different understanding of the concept. Each new context or research defines its own classes and what they mean, making it hard to compare different results, and each combination of data and processing methods constrains how those sets of classes are classified (Baudoux et al., 2021). The set of classes and the spatial resolution of the map are mutually dependent (Homer et al., 2020). Thus, the class nomenclatures must be documented, allowing others to understand whether and how smallholder irrigation can be identified.

Nevertheless, a class definition alone is insufficiently precise to ensure replicability (Yu et al., 2014). To address this, modellers should document the data creation methods and share likely sources of error and potential uncertainties (Elmes et al., 2020). Nonetheless, the uncertainty in training data (if labels are correct) is rarely assessed or reported, and the accuracy of datasets is seldom questioned (Foody et al., 2016). In addition, the person labelling the data can also (un)consciously mislabel it. Equally, they may not recognise smallholder irrigated areas as a legal form of irrigation, labelling them instead as 'homestead gardens' or 'uncultivated areas'.

Choosing and defining the sets of classes requires defining sharp boundaries, while the change in landscape is often gradual. As a result, a pixel often contains a mixture of different classes. Although each location may have a 'best' category, others may also be suitable (Woodcock & Gopal, 2000). Sharp class boundaries can cause problems when using discrete class categories (Elmes et al., 2020). Even though agricultural fields have relatively sharp boundaries between fields and other land covers, pixels covering these boundaries remain mixed. Boundary pixels may cover part of a field and bushland, both with different spectral signatures, resulting in this pixel containing mixed spectral responses. Note that this effect largely depends on the resolution of the satellite imagery. A pixel size of 0.3x0.3 metres (the size of individual crops) will be less mixed than a pixel of 250x250 metres (the size of many crops, natural vegetation and non-vegetation classes). Consequently, the spectral response within the smaller pixel will be from that crop alone (more 'pure'), whereas the large pixel will have a mixed signal from multiple types of vegetation and non-vegetation classes. This influences the class label/sets of classes that can be used – the larger pixels can only cover general classes, such as *cropland*, whereas the smaller pixels can contain sub-class labels, such as *cropland – tomato (irrigated)*.

4.1.3. Step 3 – Field data collection

Understanding and defining the relationship between the biophysical variables (such as soil cover or chlorophyll content) and objects' spectral responses is essential for selecting variables to observe and measure in the field and for selecting representative samples (Campbell & Wynne, 2011). Understanding these relationships best requires data from the field.

The study's objective will largely determine how often the modeller or data collector will visit a field. A study that maps the extent of agriculture during a 3-month window may have enough data during one visit, however making the same map for an 8-month window will require multiple visits, as the crops are in different phenological stages with different spectral responses; fields may only be cropped once instead of twice (resulting in fallow fields); and not all fields will be irrigated throughout the season. A study mapping the annual extent will require visiting the field multiple times per year for multiple years. Through interviews with farmers and other (personal) observations, the modeller gets a more in-depth understanding of when, where and how irrigation occurs. Understanding this irrigation context will considerably improve the ability to interpret RS data and can positively contribute to the classification process.

Besides improving the understanding of spatial and temporal variability and the social contexts that drive these variations, going into the field also allows for physical field data collection for training and validation. Improved (open source) technology, such as ODK (Open Data Kit) Collect (ODK, 2023), makes it possible to gather more data cheaply and easily with more collectors, which requires a sound sampling design and labelling protocol so that all collectors have the same understanding. Alternatively, training data is collected digitally by drawing polygons and using high-resolution images. However, this method is sensitive to misclassification, as determining the class of mixed pixels is often difficult.

4.2. Element 2: Model

This element contains the steps related to the modelling/classification of maps, such as the algorithm and what satellite data was included.

4.2.1. Step 4 – Predictor variables

The algorithm (step 5) learns how to separate classes based on different (satellite-derived) predictor variables (this step) and training data (step 3). The main idea behind providing more variables to the algorithm is to separate the classes better. However, adding too much information might even decrease the accuracy if insufficient training data characterises the increased complexity associated with the feature space's larger dimensionality (Maxwell et al., 2018). Maxwell et al. (2018) add that even if the accuracy is not decreased, it may be desirable to use fewer variables to simplify the model, perhaps for reproducibility, simplicity, or speed. Understanding how the various input variables can describe irrigated agriculture through field data collection can help the modeller determine which variables are suitable for the classification.

It is good practice for the modeller to understand how the variables may represent certain classes. There are hundreds of vegetation indices alone; sometimes, their differences seem minor. Documenting why a particular variable is included in the model asks the modeller to consciously consider if that variable is relevant in the first place. A final consideration is that when the mapped object is smaller than the pixel size, the sensor may not be adequate, such as when irrigated fields are smaller than the pixel size.

4.2.2. Step 5 – Algorithm

An algorithm seeks to separate classes into the feature space provided by RS images (Foody, 2021). This model learns how to see different classes based on the input training data, which can later be applied to areas that need to be classified. Choosing an algorithm for classification is difficult, not only because there are so many but also because the literature seems contradictory. A possible explanation is that the different study procedures may not be comparable (Maxwell et al., 2018). However, even with similar procedures, the 'best' algorithm is not easy to determine (Lawrence & Moran, 2015), and the authors suggest evaluating multiple algorithms.

Comparing results from different studies informs us about the applicability of an algorithm. However, the most suitable algorithm is case-specific and depends on the classes mapped, the nature of training data, and the predictor variables (Maxwell et al., 2018) – i.e., all the previous steps. The algorithm choice may also be based on personal preferences, project requirements, or software limitations, which should be disclosed as the reason for using the algorithm. Nevertheless, experimenting with multiple algorithms is needed to determine the most suitable classifier (Maxwell et al., 2018), and reporting this supports the final choice.

Classification accuracy may also be affected by user-defined parameters of the algorithm. Although the default values of such parameters can be sufficient, experimenting with different values is needed to determine that the best classification has been chosen (Maxwell et al., 2018). Some algorithms require many user-defined parameters to be set, whereas others only require a few. Parameter settings should be documented even if the default values are used, as they are often case-specific. As there is no 'best' algorithm for mapping smallholder irrigation, experimenting with different algorithms and parameters and evaluating the model accuracies and maps is recommended to find which algorithm can best distinguish irrigated agriculture from other spectrally similar classes. Documenting these steps will inform others about this; however, note that the reasons behind the higher classifications remain a mystery in the black box with most algorithms.

4.3. Element 3: Presentation

The final element is about presenting these results, specifically, the map (in combination with the accuracy assessment), which shows the spatial distribution of all the classes defined in the first steps.

4.3.1. Step 6 – Accuracy

Different training data sets, classification algorithms and input variables will produce different results for the same region. Which result to use in the end will depend on assessing the map's accuracy and uncertainty (i.e. with a confidence interval), which, besides indicating the quality of the map, also provides a means to enhance its usefulness (Elmes et al., 2020; Foody, 2009). Acknowledging potential limitations of the assessment provides map users with an informed understanding of the results' accuracy. In contrast, a lack of transparency would give a false impression of reliability (Stehman & Foody, 2019). These potential limitations are the considerations and reasoning for choices of the previous steps and give context to the accuracy assessment.

A map's accuracy depends on the reference dataset used in its training (G. Foody et al., 2016). Map accuracy is assessed by evaluating the agreement between the estimated map and those of the validation data, which is often summarised in the confusion matrix (Elmes et al., 2020). It shows more than just the accuracy, as a high accuracy can be achieved by merely allocating all training data to the most abundant class and not considering rare classes, although such a map would not be of much use (G. M. Foody, 2020; He & Garcia, 2009). The confusion matrix also shows the accuracy per class, increasing the interpretation of the map's results (Foody, 2020). Any research involving classification techniques or maps evaluated exclusively regarding overall accuracy may be unreliable. Instead, this metric should be used with other metrics, such as user or producer accuracy (Shao et al., 2019).

4.3.2. Step 7 – Map temporality

This step is relevant for dynamic land classes throughout the year, such as (smallholder) irrigated agriculture. Irrigated areas increase and decrease in size as the water availability increases or decreases, which changes during the irrigation season. A map showing the extent of irrigated agriculture at the start of the irrigation season likely shows more irrigation than one from the end of the season.

4.3.3. Step 8 – Code and data sharing

A lack of transparency in reporting poses credibility issues, which in turn hinder the comparison and usefulness of maps. This concern becomes more significant with the increasing complexity and automation of remote sensing analysis across various disciplines and applications. To address this concern, it is crucial to publish the elements necessary for accurate assessment (steps 1-6), enabling remote sensing scientists to evaluate the reliability of new methods and modelling techniques (Morales-Barquero et al., 2019). The availability of model code and data for scientific practice would increase transparency, facilitate building on existing theories, and allow testing under different conditions and areas (Melsen et al., 2017). Together with the use of open-source software and data, a study can genuinely be reproduced (Elmes et al., 2020; Stehman & Foody, 2019).

In other words, we would be able to know *why* specific models work better for identifying irrigated agriculture, independent of the specific case study, rather than only knowing *that* a model works, unable to build on that knowledge.

5. Analysis of reported modelling choices in recent RS studies on irrigation in SSA

Following the selection of 25 recent RS studies on irrigation in SSA (Section 3), we analysed them using our framework for making modelling choices explicit (Section 4). We aimed to conduct a content analysis of the modelling choices and their implications. However, we noticed that there is minimal reporting on these choices. We, therefore, start this section with an analysis of which modelling choices these papers report on, categorised as *fully reported, partially reported* (one or more sub-steps missing), and *not reported* (no information present). The guidelines for this categorisation can be found in Annex 1. Figure 2 presents these categories per article (A) and per modelling step (B); details can be found in Annex 2.

Figure 2A reveals that several articles provide scant information on their modelling choices, rendering them of limited value as we cannot build upon their knowledge effectively. The only discernible detail is that, under specific circumstances, irrigated agriculture could be classified. However, we lack insights into the model and data, preventing us from understanding why the model worked or assessing its accuracy conclusively. Figure 2B

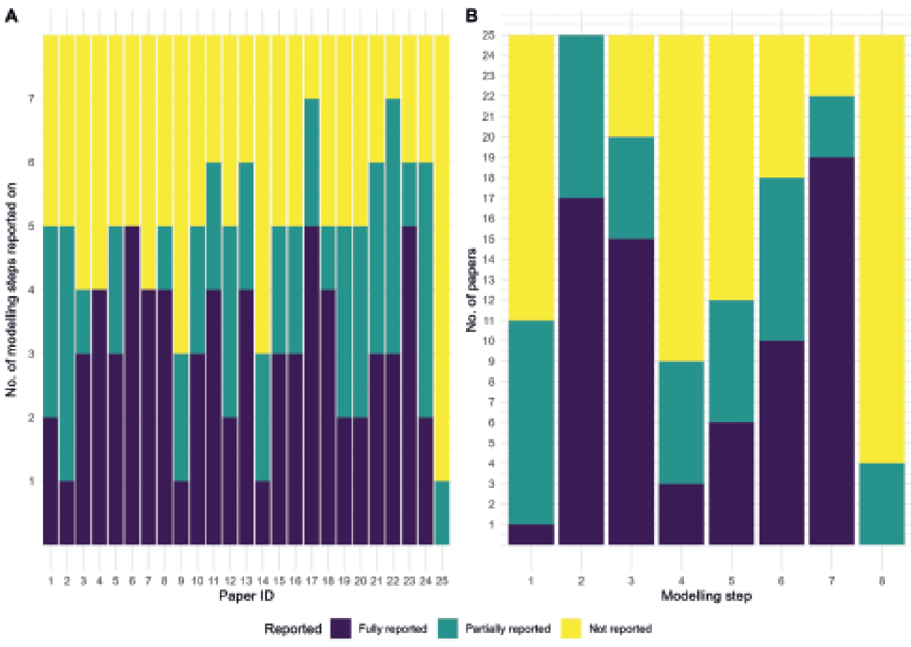


Figure 2A: Assessment of the modelling steps that the papers reported on. Figure 1B: number of papers that reported on modelling steps. Modelling steps: 1) Sampling design, 2) Sets of classes, 3) Field data collection, 4) Input variables, 5) Algorithm, 6) Accuracy, 7) Map temporality, 8) Code and data sharing.

shows that the map temporality, sets of classes and field data collection are most commonly reported on. In contrast, the code and data sharing, sampling design and input variables are least reported on.

In the remainder of this section, we describe what information the papers documented.

1. Sampling design

Type of sampling design used and size of training and validation dataset reported on.

Of the 25 papers, only one fully reported on the sampling design: the authors used a randomised strategy for unbiased samples, with 100 samples per class, and additional samples were collected for less frequent classes (such as irrigated agriculture), although the sampling design for the additional samples was not explained. Ten papers partially reported their sampling design, either opportunistic or random. Opportunistic sampling was used based on the accessibility of the fields, field size, representativeness of the fields and/or authors' knowledge of the area, but there was no information on the number or distribution of samples. In other cases, the authors provided a map of where samples were taken, showing sampling near roads (indicating opportunistic methods), or a confusion matrix was provided later on, but with no reporting on the sampling design. In total, 14 papers did not clearly report on the number of samples collected and the sampling strategy.

2. Sets of classes

Description and number of classes present.

Seventeen of the articles fully report on the classes and their descriptions; the other eight articles only present the class names without a description (partial report). All articles had the "irrigated agriculture" class, but the descriptions varied. For example, "areas equipped for irrigation" and "areas being green in the dry season" are used, meaning two completely different things. Areas equipped for irrigation are not necessarily used for irrigation; think of large irrigation schemes of which only the first fields receive enough water to irrigate, whereas the other fields do not (but are still equipped for irrigation). The second definition assumes anything green has to be irrigated agriculture, but areas with high water tables or near streams are also green during the dry season, creating a source of confusion.

3. Field data collection

Training and validation data collected in-situ or digitally.

Fifteen papers fully reported on data collection (either fieldwork or digital) and their rationale, namely for understanding the local context, deciding on the classes present, collecting reference data, or because the objective was to compare maps over decades. Most data were collected during the peak cropping season, with either one or multiple visits spread over the season. Three papers partially reported on this element: two mentioned knowing the area even though no field data was collected, and one did collect field data but gave no specifics on this. No articles did not report on any aspect.

4. Predictor variables

Reason(s) for satellite-derived variables documented.

The fourth step is deciding which variables adequately represent the different classes and, more importantly, allow the algorithm to differentiate these classes. Six papers fully reported why they used certain variables: from the spatio-temporal resolution of some satellites suitable for capturing smallholder irrigation (Sentinels) to monitoring land cover changes requiring a long time series (Landsat). Usually, the NDVI was used because of its suitability for monitoring vegetation (based on other studies). However, often no alternative indices were mentioned as tested or mentioned as potentially useful. One paper used the normalised difference wetness index (NDWI) to separate classes because of the irrigation method (spate irrigation). Six papers partially reported on the reasoning, whereas 13 papers only mentioned the use of variables but without reasoning.

5. Algorithm

Reason(s) for algorithm choice documented.

Although all papers document the specific algorithm they used (and sometimes in which software), they do not always share their reason(s) for selecting a particular algorithm – only three papers mention that it was based on how suitable the algorithm was deemed for detecting irrigation. The first paper used a decision tree because of the area's non-complexity, based on experimental values for temperature and mean normalised difference vegetation index (NDVI). The second paper's author (or authors) based their choice on experiments with multiple algorithms. The third paper used a knowledge-based method which required thresholding and local expert knowledge. The authors of six papers partially reported their choice of algorithm, either based on other studies that use it to classify agriculture (often random forest) or because it is a common method (maximum likelihood classification). The remaining 16 papers either gave no reason for using the algorithm or gave general arguments, such as its common use in RS, ease of understanding the structure, or robustness, without explicitly mentioning if experiments were done to confirm these reasons. Five articles discussed their parameter settings; consequently, we can only assume the default settings were used, which might not be the appropriate settings in those contexts.

6. Accuracy

Any accuracy metrics and values reported for the map and the presence or absence of an error/confusion matrix.

Ten papers fully reported the error matrix, either in pixels misclassified or as a total area percentage. Eight papers reported class-specific accuracies but not confusions between classes (i.e., no error matrix), and seven reported only the overall accuracy or the kappa (labelled 'not reported'). Most papers excluded details such as class-specific accuracy. Consequently, the quality of a map, particularly in terms of the extent of irrigated agriculture and its confusion with other classes, can only be fully understood for seven papers.

7. Map temporality

The seasonality is clear from the map or caption.

Nineteen papers fully report the seasonality of the map (either a specific date or a month), although often, this has to be deduced from when the satellite data or field reference data is collected. In contrast, three papers partially report on this by only showing a year, making it unclear if the map shows the maximum extent of irrigated agriculture or perhaps the minimum. The remaining three papers do not report on this step.

8. Code and data sharing

Any link to the repository of codes or availability of training data statement

No articles shared links to a repository where code and data were available. Four articles linked to the code used, or could share it on request; however, the vast majority, 21 articles, did not link to any code or data.

6. Discussion

In this section, we use the three main elements in the framework to interpret our results.

6.1. Element 1: Training and validation data

Given that more than half of the reviewed articles lacked adequate reporting on the sampling design, we are left with mere speculation regarding the sample collection locations and potential biases. Consequently, the resulting maps have the potential to both overestimate and underestimate the extent of irrigated agriculture. The outcome largely hinges on the study's specific objective (whether it emphasises smallholders or not) and the approach taken by the data collectors (having a narrow or broad understanding of irrigation).

Irrigated agriculture might be underestimated when a narrow technical definition of irrigation is used, dismissing small scattered fields. Policymakers may see untapped potential based on the map and plan interventions in seemingly unused land. However, on-site visits might reveal inventive irrigation methods. This mismatch between the policymaker's expectations and the actual situation can erode their trust in the map as a reliable tool for decision-making. As a result, the policymaker might begin to use the map less frequently or even dismiss its value altogether. The sets of classes are reported in 17 of the 25 reviewed articles, indicating that most authors see the importance of reporting.

In total, 15 of the reviewed articles reported on field data collection. Authors who inadequately report on field data collection methods fail to reveal the representativeness of their data concerning local irrigation practices. Relying solely on pre-existing knowledge of known irrigated areas may result in an inadequate representation of smaller, intricate agricultural landscapes. Relying on past experiences, assuming that the relationships learned from other areas apply universally, may lead to overestimating or underestimating irrigated areas due to the variability of smallholder irrigation across locations and time.

6.2. Element 2: Modelling choices

Our findings reveal that most papers rely on variables based on existing literature, predominantly NDVI, without reporting on exploring other indices that may be more suitable for their study area. This prevailing assumption may hinder accurately capturing complex landscape characteristics, particularly in smallholder farming areas where natural vegetation may resemble croplands. While feature selection methods exist to filter variables during the modelling phase, certain variables may prove more sensitive to on-ground features than others. For example, the normalised difference vegetation index (NDVI) is commonly used in classification studies. However, it may lack the sensitivity to distinguish irrigated fields from other vegetation types (Ozdogan et al., 2010). An NDVI value of 0.8 can be observed in dense forests, grasslands, or well-managed crops. Consequently, the model faces difficulty differentiating between various classes using NDVI alone. To address this limitation, additional or other variables may improve the classification.

The algorithm has not been documented in 13 of the 25 articles. Not reporting on the algorithm mainly limits the reproducibility of the results, i.e., the same map cannot be remade, as the algorithm and it is parameter values will most likely be different from the one in the original study. The implication is that the location and extent of irrigated agriculture will differ every time the model is rerun. Even where data collection and predictor variables are documented, there will be variation, albeit more limited. The map used as the basis for decision-making is chosen randomly, as any other variation of the map could have been selected due to the unknown algorithm settings. This uncertainty diminishes trust in the eventual product.

6.3. Element 3: Presentation

Ten of the 25 articles reported on multiple accuracy metrics. Smallholder irrigated agriculture is often considered a rare category due to its relatively small size compared to more common land-use classes. Studies that overlook class-specific accuracies present a misleading sense of reliability. However, merely citing (high) accuracy figures does not guarantee quality; this aspect is frequently misunderstood (Braun, 2021). Braun further asserts that if accuracy values are accepted uncritically, and the criteria for selecting them in relation to the relevant image classes are disregarded, it could lead to inflated political significance, particularly for less common land-use systems. Conversely, low accuracy values for challenging-to-distinguish classes, such as smallholder irrigation, should not be perceived as failures but as honest assessments. Attempting to tune accuracy values by arbitrarily aggregating classes or modifying training sites to meet published thresholds in the literature is ultimately futile. In such cases, it might be more purposeful to acknowledge that remote sensing may not be the most suitable approach to address the scientific question at hand, and instead, knowledge could be better generated by increasing fieldwork if feasible.

Nineteen articles reported on the map temporality, suggesting a somewhat more common practice.

No article shared both data and code, although four articles shared code. This highlights that making methods publicly available is not a common practice within the field. It is crucial to provide the exact code and settings used in the original creation to ensure transparency and reproducibility. The absence of this information conceals the assumptions made by the modeller, which may not be evident to users. Additionally, sharing code allows the article to focus on the implications and interpretation of the map while explicitly referencing the methods in the code. As a result, the article becomes less of a technical document and revolves more around discussing the results.

6.4. Implications to irrigation policy and practice of undocumented steps

The inadequate reporting on classification choices, as demonstrated in this study, not only raises credibility concerns but also hinders the comparability of maps and limits the overall usefulness of the maps (Morales-Barquero et al., 2019). Consequently, this may perpetuate the notion that smallholder irrigation is inconsequential, as maps may not adequately represent it. The modeller's perspective of irrigation influences the classification process, and our framework can bring these biases to light.

The representation of an area depends on the choices made during the classification steps. These choices are not trivial but can have some real-life consequences. RS can be used to classify 'under-utilised' areas (marginal lands), and the maps that are produced can be used as a tool for negotiation. For example, Nalepa and Bauer (2012) compared four studies that used RS to map available land (marginal lands) for biofuel crops (based on biophysical parameters) for potential investment, but note that these marginal lands are often not as 'empty' as the maps show; the RS analyses do not consider the socio-economic dynamics of those 'empty' marginal lands. Consequently, there is a possibility that smallholder farmers on those marginal lands are pushed to even lower-quality land. There are more examples in which land has been classified as suitable for certain activities because they were 'empty' but were actually in use (e.g. Exner et al. (2015); Nalepa, Short Gianotti, and Bauer (2017)), illustrating that RS-derived maps can influence perceptions and actions, with (unintended) consequences for the people living off that land. These papers focused on how RS determines what land is available (i.e. empty) for large-scale (often foreign) investment, with dire consequences for the farmers on those lands.

RS has many advantages. Accurate information on the location and extent of smallholders could play a vital role in providing support when and where needed the most (Izzi et al., 2021). Investments and interventions in these often-neglected areas can go a long way in improving the livelihoods of those engaged. This requires knowing their whereabouts through transparent and reproducible approaches that can be relied upon for progress monitoring purposes. When classification methods are trained and utilised in one field, their transferability to other areas becomes smoother when all relevant steps are openly reported. By sharing the challenges faced and unsuccessful methods, potential obstacles can be anticipated and avoided, enabling others to learn from the experiences and pursue alternative approaches.

On the contrary, withholding specific classification steps requires individual mapmakers to independently experiment and validate methods, which may lead to redundant efforts. Moreover, it creates a false sense of reliability as crucial information remains undisclosed (Stehman & Foody, 2019).

6.5. Limitations of the review

The analysis presented in this study has certain limitations that influence the scope of the results. While we are confident that our analysis was based on a representative sample, it is important to acknowledge that search strategies involve a degree of subjectivity. Due to the vast extent of land cover mapping literature, we cannot dismiss the possibility that employing a different set of keywords and combinations could have resulted in a varied literature sample, potentially leading to a higher or lower percentage of papers considered reproducible. Including other variations on the word irrigated agriculture might have yielded other results.

We specifically focused our analysis on the reviewed literature, as this is often where the standard is established. From an external perspective, it might appear that if papers of high-ranking journals do not report on certain elements, others may feel less inclined to do so. However, this practice does not necessarily guarantee the reproducibility and transparency required for robust research.

While we considered all reporting steps equally in our analysis, it is essential to recognise that not all steps have an equal impact on the extent of irrigated agriculture depicted in the maps. The training data, for instance, has a more substantial influence on the results than the algorithm used (Maxwell et al., 2018). Consequently, our developed framework is not designed to rank and directly compare studies against each other based on a numerical score. For instance, a study scoring 7 out of 8 is not necessarily deemed more reliable than a study scoring 2 out of 8.

Instead, the framework's primary purpose is to highlight the extent of reporting on each step. A study scoring 7 out of 8 indicates that it can be better interpreted within its specific settings than a study scoring 2 out of 8. This way, the framework allows for a qualitative assessment of the reporting comprehensiveness. It provides valuable context for understanding the results within each study's unique context.

While we have addressed all the primary classification and presentation steps with our eight-step framework, this may not be an exhaustive list. Specific steps contain sub-steps that could rightfully be considered individual main steps. For instance, Training and validation data could benefit from further expansion, particularly regarding the three steps mentioned.

As more authors adopt and report on all the steps using this framework or a similar one, the importance of specific steps may increase, while others may become less relevant. Consequently, the framework is subject to evolution and refinement over time, with the possibility of adding or removing steps based on the collective understanding and experience of researchers. Continuous improvement and adaptation of such frameworks are vital to ensure the comprehensive assessment of classification and presentation procedures in research.

As there is no fixed structure in reporting, it is often difficult to find the elements in the texts for a conclusive answer in the framework. Consequently, we could answer the questions by interpreting ambiguous formulations – for example, map temporality. An option to mitigate this would have been to interview the authors of the 25 articles on their choices.

7. Conclusion

Producing quality land use maps using remote sensing requires careful, accurate, and transparent choices. We define eight classification steps that need careful reporting, including sampling design, sets of classes, field data collection, predictor variables, algorithm, accuracy, map seasonality, and code and data sharing. Reviewing 25 articles, we found that none sufficiently documented all classification steps. The limited number of papers we found during the literature search indicates that mapping smallholder irrigation is an emerging field of study, making the recommendations in this study even more important. Although the reasons for not reporting are unknown, the lack of explicitly made choices hampers a proper evaluation of irrigated agriculture's extent, particularly smallholder irrigation. There are numerous ways in which the classification choices can influence the accuracy of irrigated area mapping. Making choices explicit in the classification process will allow others to use relevant parts for their own study and assess the likelihood that the extent of irrigated agriculture is reasonable, over-estimated or under-estimated. Because maps are always social constructs and subjective abstractions of reality, being transparent about a study's implicit and explicit choices, reasoning, and interpretations is good practice.

Furthermore, making the elements used in the classification process public and accessible is crucial. It enables remote sensing scientists to assess the dependability of new methods and modelling techniques by providing essential information (Morales-Barquero et al. 2019). All authors working on remote sensing-derived maps go through the first seven steps of the developed framework, whether consciously or not. However, it is apparent that steps remain undocumented, at least in the final publication. A full explanation for this behaviour will require further study involving survey techniques and interviews with remote sensing scientists.

Creating the framework is just the initial step; its effectiveness relies on diverse actors adopting it in various ways. Drawing inspiration from other scientific fields, like hydrology (Stagge et al., 2019), authors can utilise the framework as a self-assessment checklist, ensuring the inclusion of data, models, and code in their work before submitting it for publication. As echoed by other scholars (Castilla, 2016; Morales-Barquero et al., 2019), if journals impose requirements for reporting on all elements, it is likely to drive a positive change in behaviour. This, in turn, can influence practitioners who do not actively publish, as they will reference articles that adhere to complete reporting and follow the same structured approach.

Moreover, journals, funders, and institutions can employ the framework to assess the presence of data, models, and code in new submissions, offering feedback to authors and making it a prerequisite for submission. To encourage compliance, journals may consider introducing special recognition or awards for papers that exemplify best practices in documenting irrigation maps. Journals already encourage authors to publish data, code and other additional information in the annexes. Such an approach would motivate authors and research teams to adopt these practices as it enhances their reputation and increases the visibility of their work.

The extensive documentation of all classification steps in irrigation maps carries several positive implications for irrigation policy and practice. Carefully made maps may help inform governments of areas with and without agricultural activity, allowing them to support farmers who need it and minimise wasted effort on areas where farmers are not actively practising irrigation. At the same time, it avoids allocating presumably 'empty' areas to other uses despite smallholders being active there.

Sharing extensively documented irrigation mapping methodologies promotes the adoption of best practices across different regions or countries. Policymakers and practitioners can learn from successful experiences and avoid repeating mistakes made in other contexts. This approach advances irrigation practices worldwide by fostering collaboration and knowledge exchange.

8. Annex

Framework

Steps	Criteria for assessment	Possible answers	Reasoning and impact on smallholder irrigation visibility
1. Sampling design	Type of sampling design used and size of training and validation dataset reported on.	Fully reported	Sampling design and dataset size fully shared. *The documentation provides enough information to assess if the field data is biased towards certain irrigation practices or areas or not. This allows the reader to determine if the results are helpful.*
		Partially reported	Either 1) the sampling design is reported on or has to be derived form available maps, or 2) size is reported, but not both elements. *A different number of samples collected on other areas will capture the heterogeneity of smallholder farming differently, with smaller sample sizes and more opportunistic sampling potentially missing most of the smallholder fields.*
		Not reported	No direct information reported on sampling strategy or dataset size. *Without basic information on the sampling design, it is not clear if the results are biased. Small sample sizes and collection through opportunistic sampling will likely not capture all different irrigation practices.*
2. Sets of classes / nomenclatures	Description and number of classes present.	Fully reported	All classes fully described. *Although the labels might not explicitly describe (smallholder) irrigation, the classes, their description and the labelling protocol allow people with different backgrounds to understand the results through the same lens.*
		Partially reported	Either 1) classes are described, 2) number of classes provided, but not both elements. *Although some information is provided, it is incomplete. Readers and decision-makers will use their own definitions to interpret the results and decide whether the map is reliable, based on what they know of the area.*
		Not reported	No information reported on the classes or size. *With no documentation, the reader must use their own interpretation or definition of a class, which is probably different from what the author intended. A reader with a technical background will view smallholder irrigation through a different lens than a reader with a social background. Consequently, if a reader uses the method of the article for a different study, different class definitions will be used when going into the field, either missing irrigated areas that were included in the initial result or including more irrigated areas.*

Framework continued

Steps	Criteria for assessment	Possible answers	Reasoning and impact on smallholder irrigation visibility
3. Training and validation data collection	Method of training and validation data collection described (fieldwork or digitally).	Fully reported	Description present on how data is collected and that knowledge of the area is present. *The field data can be related to recent smallholder activities in the field. However, there is always a chance that some activities are not captured in the data and hence can be missed in the final results.*
		Partially reported	Indirect information is presented, however not enough to know (as a reader) how data was collected and if prior knowledge was present. *The author shows that (some of) the complexity of the landscape is understood, but it is based mainly on past experiences and assumptions. Consequently, newer, sometimes hybrid forms of smallholder activities may be missed in the final results.*
		Not reported	No information is presented on data collection and understanding the area. *We may assume that choices are based on earlier experiences, or the researcher does not have any regional experience suitable for this context. Smallholder irrigation is context-dependent, and some regions can have unique (hybrid) forms of irrigation activities. By not documenting if the author knows the local context, we cannot determine if the final results cover all irrigation areas and forms.*
4. Predictor variables	Reason(s) for satellite-derived variables documented.	Fully reported	The reasoning behind input variables is described. *The choices are backed by similar studies or experiments and are likely optimised for irrigation mapping. The bias (over- or under-estimation) of this extent is minimised.*
		Partially reported	Information is presented on 1) why a sensor is used, or 2) why variables are used, but not both. *In other circumstances, the data might be suitable for detecting irrigated agriculture, but it may not be the case for this study. Consequently, the extent of irrigated agriculture is likely biased (over- or under-estimated).*
		Not reported	No information is presented on why the variables are used. *It remains uncertain if the data is suitable for identifying irrigated agriculture, especially if more classes are spectrally similar in time. Consequently, the extent of irrigated agriculture is likely biased (over- or under-estimated).*

Framework continued

Steps	Criteria for assessment	Possible answers	Reasoning and impact on smallholder irrigation visibility
5. Algorithm	Reason(s) for algorithm choice documented..	Fully reported	The reason(s) for using the algorithm(s) is described and relates to the study objective. *The reason is either based on similar studies or from experimenting, and the model is optimised for irrigation mapping. The bias (over- or under-estimation) of this extent is minimised.*
		Partially reported	The reason(s) are documented, but are not related to the objective (e.g. often used). *Although the algorithm may not have been tested on irrigation mapping in this or other studies, the reasoning behind using it is still documented. This allows a reader to estimate the bias.*
		Not reported	No reason(s) are documented, only choice of algorithm is shared. *Besides knowing that a particular classifier was used, it remains unclear why and if a classifier is suitable for the task. The final result may, however, over- or under-estimate the extent of irrigated agriculture.*
6. Accuracy	Any accuracy metrics and values reported for the map and presence or absence of an error/ confusion matrix	Fully reported	The error matrix, user/producer accuracy, and the values are all presented. *This ultimately shows how the overall method and model can map irrigated agriculture without seeing the map itself.*
		Partially reported	Only general accuracies are presented, no class specific information is available. No error matrix. *No class-specific conclusions can be made, but overall, the accuracy gives some indication of the results. However, overall accuracy can be biased with imbalanced training data. Where irrigated agriculture is a smaller class, a high overall accuracy likely under-estimates the accuracy of irrigated agriculture.*
		Not reported	Neither error matrix or class specific accuracies reported on. *With no class-specific information, the results cannot reliably be used for decision-making. It also remains unclear how much bias there is in the extent of irrigated agriculture.*
7. Map seasonality	The seasonality is clear from the map or caption.	Fully reported	The caption or supporting text makes clear what season(s) the map covers. *The map shows the extent of irrigated agriculture (with some error margin) and season. This allows for better decision-making, especially if it concerns smallholder farmers.*
		Partially reported	Map is presented, but seasonality has to be derived from acquisition dates. *Although somewhat clear, the map representation can be misleading, as the extent of irrigated agriculture likely changes throughout the year. This over- or under-representation*

Framework continued

Steps	Criteria for assessment	Possible answers	Reasoning and impact on smallholder irrigation visibility
			can result in deciding to focus, or not, on specific areas, consequently excluding irrigated areas from external support.
		Not reported	Only a map is presented of the study area. *As irrigated agriculture is dynamic in time and space, it remains unclear if the map shows the maximum or minimum extent of irrigated agriculture.*
8. Code and data sharing	Any link to repository of codes or availability of training data statement	Fully reported	All data can easily be accessed. *Links and/or data available without additional login requirements. All steps can be reproduced.*
		Partially reported	Either 1) code or 2) data is shared, or 3) requires login details (i.e. not open access). *Code or data is shared, but it is not possible to reproduce the complete study.*
		Not reported	No sources presented. *No code or data is shared.*

Framework results

	1 **Sampling design**	2 **Sets of classes / nomen-clatures**	3 **Field data collection**	4 **Input variables**	5 **Algorithm**	6 **Accuracy assessment**	7 **Map temporality**	8 **Availability training data and codes**
Author	*Type of sampling design used and size of training and validation dataset*	*Description of classes and number of classes.*	*Training and validation data collected in-situ or digitally*	*Reason(s) for satellite-derived variables explained*	*Parameter settings. Reason for algorithm choice and documented.*	*Any accuracy metrics and values reported for the map and presence or absence of an error/confusion matrix*	*Map shown*	*Any link to repository of codes or availability of training data statement*
(Eckert, 2016)	No	Yes	Only timing shared	No reason	RF, No agri reason. Parameter settings discussed	Only PA and UA given per class	Yes	No
(Lebourgeois et al., 2017)	Opportunistic: accessibility, size field, representa-tiveness; No, No (bc opportunistic)	Only classes mentioned	Yes, peak cropping season	No reason	RF, No agri reason	Only PA and UA given per class	Partially, from image acquisition dates	No
(Eckert et al., 2017)	No	Yes	Yes, only Feb	Landsat to go back in time, No NDVI reason	RF, No agri reason	No	Yes	No
(Xiong et al., 2017)	No	Yes	Yes, peak cropping season	No, only MODIS NDVI mentioned	Yes	No	Yes	No

Framework results continued

	1 Sampling design	2 Sets of classes / nomen-clatures	3 Field data collection	4 Input variables	5 Algorithm	6 Accuracy assessment	7 Map temporality	8 Availability training data and codes
(Cai et al., 2017)	Opportunistic (No reason), mainly irrigation (No specifics)	Yes	Yes	Modis & Landsat were used	MLC, easy to incorporate prior knowledge	No	Yes	No
(Knauer et al., 2017)	Random (unbiased), 100 per class (literature) + extra for irrigation	Yes	Yes	No reason	RF, No agri reason	Yes	Yes	No
(Aredehey et al., 2018)	No, only class distribution	Yes	No	Modis, No reason	No	Yes	Yes	No
(Ghebreamlak et al., 2018)	Partially (equal distribution, No strategy)	Yes	No	Landsat, spatio-temporal res; mndwi wet area; ndvi veg sensitive	DT, Yes	No	Yes	No
(Vogels, de Jong, Sterk, Douma, et al., 2019)	No	Yes	No, but references to similar area	No reason	RF, No agri reason	No, only OA per class	No	No
(Msigwa et al., 2019)	No (map shows opportunistic, focus irrigation)	Yes	Yes, august	Landsat, No reason; NDVI No reason	MLC, used a lot	Only PA and UA given per class	Yes, date mentioned	No

Framework results continued

	1 Sampling design	2 Sets of classes / nomen-clatures	3 Field data collection	4 Input variables	5 Algorithm	6 Accuracy assessment	7 Map temporality	8 Availability training data and codes
(Nhamo et al., 2020)	Opportunistic (knowledge of area) + random (validation), focus on irrigation	Yes	Yes, dry season	Yes, Landsat dry season for 3 years , NDVI	MLC	No	Yes	On request
(Tena et al., 2019)	No	Yes	Done, but no further details	Landsat to go back in time	MLC	Only PA and UA given per class	No, only year	No
(Vogels, de Jong, Sterk, & Addink, 2019)	No, but differences in class	Yes	No	Spot6 ndvi, No reason, obia Yes reason	RF	Yes	Yes	Results as TIF available
(Landmann et al., 2019)	No, but equal distribution	Only classes mentioned	No	Landsat ndvi, No reason	RF, suitable for agro. No parameters.	Yes	No, seems permanent map	No
(Abera et al., 2019)	No, but unequal distribution	Yes	Yes, during dry season	Landsat to go back in time	MLC, common classifier but Nothing about agri	Only PA and UA given per class	No, only year	No
(Traoré et al., 2019)	No, but unequal distribution	Only classes mentioned	No, fieldwork from different study	Landsat, Dates correspond with irrigation activity; timeserie	Authors have tried in different study, perhaps it is mentioned there	Yes in % Not absolute values	Yes, date mentioned	No
(Sedano et al., 2019)	Size shared, semi-random	Yes	Yes	No reason	Yes	Yes	Yes	No

Framework results continued

	1 Sampling design	2 Sets of classes / nomen-clatures	3 Field data collection	4 Input variables	5 Algorithm	6 Accuracy assessment	7 Map temporality	8 Availability training data and codes
(Mbaabu et al., 2019)	No description (seems opportunistic)	Yes	Yes	No reason	RF, No agri reason. Parameter settings discussed	Yes	Yes	No
(Fujihara et al., 2020)	No (map: opportunis-tic), No (Not same per class)	Only classes mentioned	Yes, to get to know the area & decide classes	Some: Landsat, spatio-temporal res; NDVI common used, EVI for irrigation, SAVI for dry areas, other to test water related; thermal different per class	No, but DT is easy to understand so effects might be better understood? (result of study, Not requisit of model)	Yes, but validation data Not gather randomly (Not mentioned at least). No details on how well irrigation is classified other than kappa/OA	Yes, same period different years	No
(Bey et al., 2020a)	Yes (training: opportunistic, validation: random but limited..), No (but random)	Only classes mentioned	Yes, 1x wet and 1x dry. Opportunisti-cially for heterogeneity and under-stand classes	Landsat (time); NDVI (?); composite: testing which is best	RF, No explanation why	No, PA and UA given per class with minimlal explanation present confusions	Yes, map per date shown (4 dates)	No
(Ouattara et al., 2020)	Yes, purposive, distribution Not explained.	Only classes mentioned	Yes	No reason, testing many variables	RF, No agri reason	Yes	Yes	No

Framework results continued

	1 Sampling design	2 Sets of classes / nomen-clatures	3 Field data collection	4 Input variables	5 Algorithm	6 Accuracy assessment	7 Map temporality	8 Availability training data and codes
(Wellington & Renzullo, 2021)	Yes, purposive, distribution Not explained.	Only classes mentioned	Yes	No reason	RF, No agri reason. Parameter settings discussed	Yes	Yes	Code available, No data
(Venot et al., 2021)	Yes, random, distribution Not explained	Yes	Yes	Yes	RF, No agri reason	Yes	Yes	No, dead link
(Magidi et al., 2021)	Yes, random, distribution Not explained	Yes	Somewhat	NDVI method explained, No reason for why NDVI	RF, No agri reason.	No	Yes	Code available, No data
(Muluneh et al., 2022)	No	Only classes mentioned	No	No	No	No	No	No

Chapter 3

Mapping irrigated agriculture in fragmented landscapes of sub-Saharan Africa: an examination of algorithm and composite length effectiveness

This chapter is published as:
Weitkamp, T., Veldwisch, G. J., Karimi, P., & de Fraiture, C. (2023).

Mapping irrigated agriculture in fragmented landscapes of sub-Saharan Africa: An examination of algorithm and composite length effectiveness.

International Journal of Applied Earth Observation and Geoinformation, 122, 103418. https://doi.org/10.1016/j.jag.2023.103418

1. Abstract

Accurately identifying irrigated areas is crucial for sustainable development, food security, and effective land and water resource management. However, incomplete or outdated national estimates of irrigated areas underestimate the extent of it, particularly among smallholders. This study aimed to address this issue by investigating the impact of different algorithms and composite lengths on predicting irrigated agriculture in four study areas in Mozambique. The study found that the choice of algorithm and composite length notably impacted the accuracy of identifying irrigation. Shorter composite lengths, such as 2-monthly or 3-monthly composites, were more effective in identifying irrigation in fragmented and dynamic landscapes, while longer composite lengths were better suited to stable classes and homogeneous landscapes. Artificial neural networks, support vector machines, and random forests were all effective algorithms for classifying irrigation. However, the study emphasised the importance of considering hotspots and agreement maps when identifying irrigation. Agreement maps combine the classification results of multiple models, providing better insights into the core areas of irrigated agriculture and allowing for a better understanding of irrigation dynamics and policy decision-making, particularly among smallholder systems. This research provides valuable insights for those working on remote sensing-based irrigation mapping and monitoring in sub-Saharan Africa, focusing on identifying smallholder irrigation with greater certainty.

2. Introduction

Obtaining accurate information about irrigation is vital for making informed decisions about land and water resource management for food security and sustainable development (Bofana et al., 2020; Wellington & Renzullo, 2021). Unfortunately, national estimates of irrigated areas are often based on limited on-ground surveys or low-resolution remote sensing data for large-scale applications (Wellington & Renzullo, 2021). The available information is often outdated or incomplete (Beekman et al., 2014b; Espey, 2019; Venot et al., 2021). Furthermore, limited budgets prevent officials from conducting regular in-person agriculture monitoring (Ajaz et al., 2019; de Bont et al., 2019; C. Ramezan et al., 2019).

African smallholder agriculture is a complex system that often takes place on small, irregular-shaped fields with in-class variance such as inter- and mix-cropping systems and variability in the timing of agronomic activities such as planting, harvesting and irrigation (Bégué et al., 2018; Izzi et al., 2021; Veldwisch et al., 2019). It is often found in mosaic landscapes where agriculture and natural vegetation alternate over short distances, resulting in frequent changes in land cover/use over short distances.

Distinguishing irrigated from rainfed agriculture or natural vegetation can be challenging, particularly in areas where soil moisture does not quickly deplete, such as near streams or in wetlands, which may have similar soil moisture patterns as irrigated croplands.

Despite the challenges of accurate mapping, quantifying and monitoring irrigation practices, remote sensing (RS) imagery has become popular for land use classification. Evaluating how different machine learning algorithms perform in classifications is one of the most studied aspects of land use classifications (Marín Del Valle & Jiang, 2022), of which the random forest (RF), support vector machine (SVM), artificial neural networks (ANN) and k-nearest neighbours (k-NN) are among the most mature and widely used (Maxwell et al., 2018; Sheykhmousa et al., 2020; Thanh Noi & Kappas, 2017). RF is popular for its ease of use and high accuracy (Belgiu & Drăguţ, 2016), while SVM is often chosen due to its ability to perform well with few training samples (Mountrakis et al., 2011). ANN is frequently used when detecting trends or patterns is difficult, and with the increase in computation power, it is being utilised more frequently (Abdolrasol et al., 2021). The k-NN classifier, although simple, has been found to compete with more complex classifiers in terms of performance (Abu Alfeilat et al., 2019). However, few studies compare two or more algorithms in the field of (smallholder) irrigation mapping.

Simple methods to use satellite data for classification are through single images or composites (Gella et al., 2021). Composites are widely used to generate cloud-free spatially consistent images from satellite time series, and can be created based on summary measures extracted from the time series (Khatami et al., 2020), such as mean, minimum, or maximum pixel values. Vegetation phenology can be characterised by creating shorter composites such as monthly or seasonal composites (Bey et al., 2020b; Khatami et al., 2020; Kumar et al., 2022). However, using them could reduce the classification accuracies because they contain less information than, for example, time series data (Marín Del Valle & Jiang, 2022), although contrasting findings suggest that the opposite effect is also possible (Hasenbein et al., 2022). Alternatively, the temporal variation can be captured by calculating the geometric median, which preserves high-dimensional relationships between spectral bands, and three median absolute deviation statistics of temporal variation (Roberts et al., 2017, 2018). These composites and statistics have successfully been used in classifying irrigated croplands in Zimbabwe (Wellington & Renzullo, 2021) and seem promising for our study.

Enough studies have already investigated the effect of different machine learning algorithms or composites in land use classification. Bey et al. (2020) found high accuracy using the median composite with RF for mapping smallholder croplands in Mozambique, Abubakar et al. (2020) achieved high accuracy in mapping maise fields in Nigeria with RF and SVM but used single images instead of composites. Furthermore, Bofana et al. (2020) compared four algorithms using combined seasonal input data but did not explore other composite lengths. However, to our knowledge, no study exists in which different algorithms and composite lengths are compared over the same study area. This study examines how different algorithms and composite lengths affect the accuracy of predicting irrigated agriculture in Mozambique. The research evaluates four classifiers (RF, SVM, ANN, and k-NN) and four composite lengths (12-monthly, 6-monthly, 3-monthly, and 2-monthly) and introduces "agreement maps" to show core areas of irrigated agriculture surrounded by an uncertainty zone. These maps can combine the strengths of multiple models and reduce the possibility of false positives. This unique method focuses on specific class distribution and classification certainty.

3. Materials and methods

We analyse the impact of different algorithms and composite lengths on the accuracy of irrigated agriculture in two stages (Figure 1). Firstly, we test four algorithms and select the one with the highest accuracy. Secondly, we test this algorithm with different composite lengths, limiting each phase to one study area per province. We present maps of classifications and measures of accuracy for each combination of algorithm, composite length, and study area. A new method for consolidating the results by identifying hotspots is introduced. Table 1 summarises the different classifications, with each combination of algorithm and composite referred to as a distinct model (16 models in total).

Table 1 Overview of the different models/classifications.

	Algorithm	Composite	Study areas
Phase 1	RF	2x6	Catandica & Xai-Xai
	SVM	2x6	Catandica & Xai-Xai
	k-NN	2x6	Catandica & Xai-Xai
	ANN	2x6	Catandica & Xai-Xai
Phase 2	RF	1x12	Manica & Chokwe
	RF	2x6	Manica & Chokwe
	RF	4x3	Manica & Chokwe
	RF	6x2	Manica & Chokwe

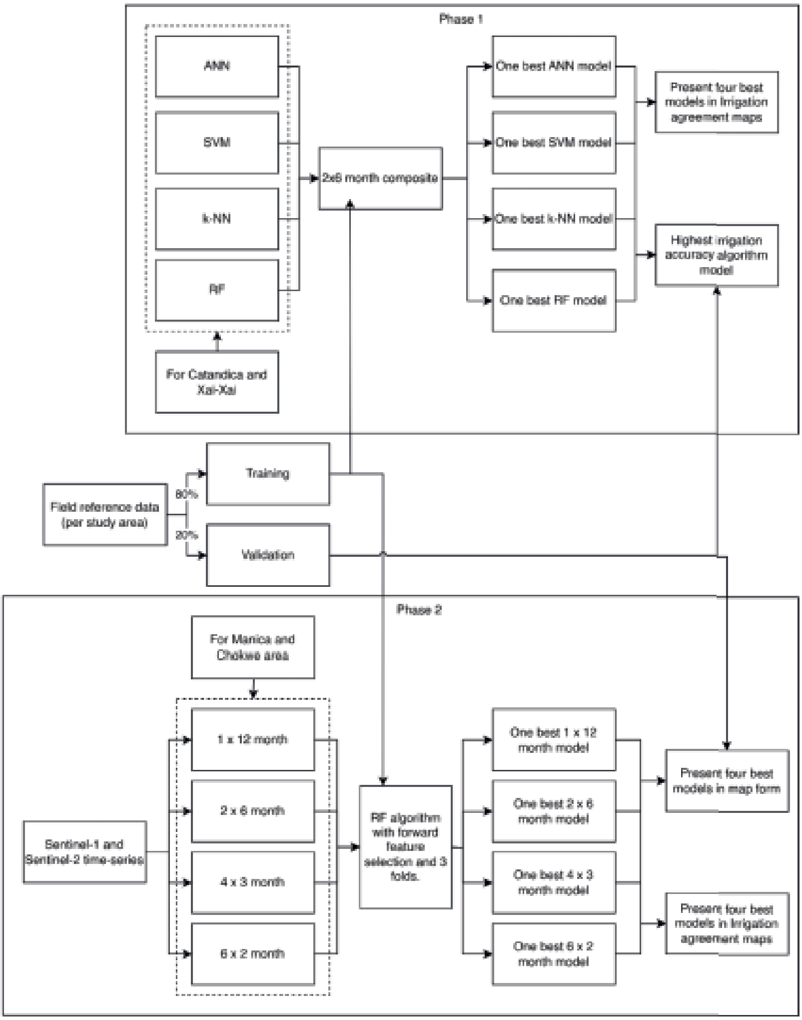


Figure 1 Flow chart illustrating the two phases and methods used per phase.

3.1. Study area

This study was conducted in four areas in Mozambique: Chokwe and Xai-Xai in Gaza province and Manica and Catandica in Manica province (Figure 2). These areas were chosen for their diverse agroecological characteristics and the presence of irrigated agriculture, including small-scale and large-scale systems. The case studies covered approximately 40x40 km in size.

Mozambique's rainy season occurs from November to April, with peak rainfall between December and February (Figure 3). Chokwe receives 650 mm/year (Kajisa & Payongayong, 2011), Xai-Xai receives 950 mm/year (Brandt et al., 2009), and both Manica and Catandica receive 1100 mm/year (Gumbo et al., 2021; Weemstra et al., 2014). Irrigation occurs during the dry season, with two cycles occurring roughly from April to July and August to November.

In Manica province, the landscape is mountainous, with small streams serving as irrigation sources. Farmers redirect the water into earthen canals called "furrows" and use sprinkler irrigation, small pumps, and bucket irrigation. These systems are smaller than those in Gaza province and vary based on water availability. Horticultural crops are irrigated during the dry season, while maise is grown during the rainy season.

In Manica province, the landscape is mountainous, with small streams serving as irrigation sources. Farmers redirect the water into earthen canals called "furrows" and use sprinkler irrigation, small pumps, and bucket irrigation. These systems are smaller than those in Gaza province and vary based on water availability. Horticultural crops are irrigated during the dry season, while maise is grown during the rainy season.

In Gaza province, there are both large- and small-scale irrigation systems along the banks of the Limpopo River. Flooding is a common practice, and pumps are used to access higher areas. Near Xai-Xai, there are irrigated areas with shallow groundwater tables that require drainage after the rainy season. Horticulture and maise are common crops in the irrigation season, while rice and maise dominate the rainy season.

Figure 2 The four study areas in Mozambique, from top to bottom: Catandica, Manica (Manica province), Chokwe, and Xai-Xai (Gaza province). See Annex 2 for detailed classifications per study area.

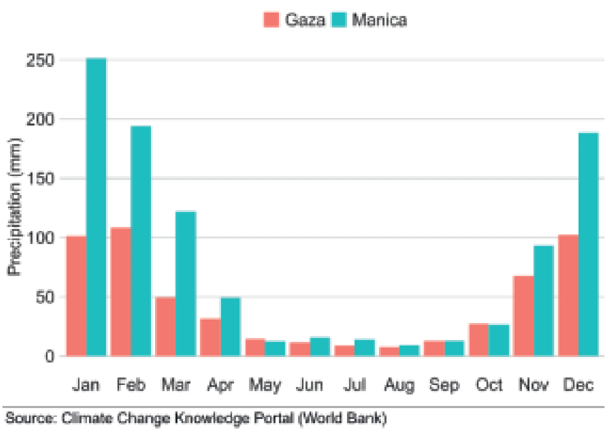


Figure 3 Mean monthly precipitation (1991-2020) per province. Irrigation occurs during the dry season, with a first cycle roughly from April to July and a second from roughly August to November

3.2. Sampling design, labelling protocol, and field data sampling

Field reference data were collected from August to November 2020 using three different sampling strategies: random clustered, opportunistic, and regular clustered designs. A random clustered sampling design was initially used to minimise travel time, but resulted in overlapping polygons and limited samples of irrigated agriculture. An opportunistic sampling design was used to gather more irrigated agriculture samples specifically, while a regular clustered design was used to prevent overlap and ensure sufficient polygon size for Sentinel pixels. The collected data was cleaned and analysed, resulting in 823 unevenly distributed polygons among different classes and areas. Hard-to-reach areas were mapped manually. Table 2 describes the classes following the ESA WorldCover definition (Zanaga et al., 2022), while Table 3 provides the number of polygons and total hectares per class and area. The classes cropland, grassland, shrubland, and tree cover were labeled in the field using Open Data Kit (ODK) Collect, a smartphone application that allows rapid and scalable field data collection (ODK collect, 2022).

The cropland classes (irrigated and rainfed) were distinguished by the period In which crops were actively grown, specifically during the rainy season (water is primarily supplied through rainfall) or the dry season (water is actively managed on the fields, either by applying or by draining water).

Table 2 Class descriptions

Cropland irrigated	Croplands under management mainly during the dry season. Any active form of water management is considered, from drainage to application through buckets.
Cropland rainfed	Croplands under management mainly during the wet season
Tree cover	Natural vegetation comprises mainly trees and dense undergrowth.
Shrubland	Natural vegetation comprising of mainly low shrubs, grasses, and some scattered trees.
Grassland	Natural vegetation of primarily grass.
Wetland	Natural vegetation that is submerged part of the year (mainly during the rainy season and first part of the dry season).
Water	Water bodies and rivers.
Built-up area	Man-made surfaces and built-up areas, including bare areas such as sand (no vegetation).

Table 3 Polygon distribution and size (hectares) per area and class of the collected field data.

	Catandica		Manica		Xai-Xai		Chokwe	
	# polygons	**hectares**	**# polygons**	**hectares**	**# polygons**	**hectares**	**# polygons**	**hectares**
Cropland irrigated	45	16,4	58	10,2	157	38,3	68	166
Cropland rainfed	34	10,9	32	7	19	5,8	48	40,4
Tree cover	9	148	19	104	9	37,2	15	12,5
Shrubland	25	89,5	20	11,3	28	26	104	187
Grassland	0	0	0	0	52	111	0	0
Wetland	0	0	0	0	6	27	12	144
Water	0	0	9	113	9	42,6	5	17,2
Built-up area	10	3,4	10	5,6	10	18,1	10	11,5
Total	123	268,2	148	251,1	290	306	262	578,6

3.3. Input variables: Data collection and preprocessing – Digital Earth Africa

Satellite data for the four areas were collected within the Digital Earth Africa (DEA) 'sandbox', which provides access to Open Data Cube products in a Jupyter Notebook environment[1]. Sentinel-1 and 2 geomedian products (a robust high-dimensional statistic like the normal median that maintains relationships between spectral bands, DEA, 2021; Roberts et al., 2018) were generated at 10-meter resolution for four different composite lengths (one 12-monthly, two 6- monthly, four 3-monthly, and six 2-monthly), covering October 2019 – September 2020, corresponding to the hydrological year (wet and dry season). Images with more than 30% cloud cover (Sentinel 2) were filtered out. A 6-month composite means that all acceptable satellite images are mosaiced into a single geomedian composite, over which specific statistics and indices are calculated.

From Sentinel-2 we calculated the Normalised Difference Vegetation Index (NDVI), Bare Soil Index (BSI), and Normalised Difference Water Index (NDWI), using the DEA indices package for the Sentinel-2 composites (Wellington & Renzullo, 2021), while the Chlorophyll Index Red-Edge (CIRE) (Gitelson et al., 2005; Segarra et al., 2020) was calculated in R. Three second-order statistics (Median Absolute Deviations (MADs)) were also calculated, which are change statistics based on the geomedian: the Euclidean (EMAD, based on Euclidean distance), Spectral (SMAD, based on cosine distance), and Bray-Curtis (BCMAD, based on Bray-Curtis dissimilarity) MADs (Roberts et al., 2018). Wellington & Renzullo (2021) used

1 Sandbox link and explanation can be found on https://docs.digitalearthafrica.org/en/latest/sandbox/index.html

these change statistics, as well as a few of the indices in their classification of irrigated areas, with success. We used these indices and statistics to cover the different phases of croplands, from browning (BSI) to greening (NDVI, CIRE), the NDWI for water detection, while the MADs are suitable for change detection, particularly for irrigation (Wellington & Renzullo, 2021).

We also used Sentinel-1, specifically the VV and VH bands, and calculated the Radar Vegetation Index (RVI). These have also been used in recent agriculture mapping studies (Abubakar et al., 2020; Gella et al., 2021; Venot et al., 2021). The VV polarisation data is sensitive to soil moisture, whereas the VH polarisation data is more sensitive to volume scattering, which depends strongly on the geometrical alignment and characteristics of the vegetation. Therefore, VH data has a limited potential for estimating soil moisture compared to VV data but higher sensitivity to vegetation (Gao et al., 2018). The RVI can be used to separate soil from vegetation (Jennewein et al., 2022; Mandal et al., 2020). Additionally, the study area experiences frequent cloud cover for parts of the year, and the synthetic-aperture radar (SAR)

Table 4 Overview of variables per composite time-length

Group	Variable	Equation
Sentinel-2	Blue	
	Green	
	Red	
	Near Infrared (NIR)	
	Red-edge 1 (RE1)	
	Red-edge 2 (RE2)	
	Shortwave Infrared 1 (SWIR1)	
	Shortwave Infrared 2 (SWIR2)	
Indices S2	Normalised Difference Vegetation Index (NDVI)	(NIR - Red)/(NIR + Red)
	Normalised Difference Water Index (NDWI)	(NIR – SWIR1)/(NIR + SWIR1)
	Bare Soil Index (BSI)	((Red + SWIR1) - (NIR + Blue))/((Red + SWIR1) + (NIR + Blue))
	Chlorophyll index (CI)	(NIR / Red Edge 1) - 1
Temporal variation	3 MADS S2	See Roberts et al. (2018) and Wellington and Renzullo (2021) for more details on equations
Sentinel-1	VV	
	VH	
Indices S1	RVI	x VH / (VV + VH)

data is less affected by cloud cover. As a result, the SAR composites of the cloudy seasons contain fewer missing observations and improve classification results, as radiofrequency radiation from SAR can penetrate through clouds.

All bands and indices were merged into one dataset, forming an 18-variable dataset (Table 4). This was done per composite length (4 lengths) and per area (4 areas).

3.4. Classification

3.4.1. Conceptual description of the machine-learning algorithms

We used four different algorithms, namely a radial support vector machine (SVM), random forest (RF), artificial neural networks (ANN), and k-nearest neighbours (k-NN). We used the *caret* package (Kuhn, 2008), which uses the free statistical software tool R and allows for systematically comparing different algorithms and composites in a standardised method. The scripts can be found on GitHub.

Since our focus is on the application of the algorithms rather than the theoretical aspects of their design, we provide only a short description of each algorithm.

- Support vector machines (SVMs) split the classes by fitting an optimal separating hyperplane (OSH) between classes using the training samples within feature space (i.e., all the pixel band values within the training sample) and to maximise the margins between OSH and the closest training samples (the support vectors) (Mountrakis et al., 2011).
- Random forest (RF) is an ensemble learning technique that generates many random decision trees that are then aggregated to compute a classification (Belgiu & Drăguţ, 2016).
- Artificial neural network (ANN) design is based on the biological nervous systems, which is where their name comes from. An ANN is made up of neurons, which are organised in layers. The key characteristic of an ANN is that all neurons in one layer are connected to all neurons in all adjacent layers, and these connections have weights (Abdolrasol et al., 2021).
- The k-NN classifier is different from the other classifiers. Instead of producing a model, each unknown sample is directly compared against the original training data and is assigned to the most common class of the k training samples that are nearest in the feature space to the unknown sample (Maxwell et al., 2018).

3.4.2. Spatial folds and parameter settings

The polygon shapefiles and images were read into R, after which all pixel values for all variables were extracted. After extracting the pixel values, the field data was split into 80% training and 20% validation data using a fixed seed number (i.e., the same data used in each model's training and validation), stratified per landcover class. The *CreateSpacetimeFolds* from the *caret* package (Kuhn, 2019) was used to create three spatial folds, meaning all pixels within a polygon remain together in either the training or testing phase, instead of some pixels within the same polygon being used for training, and their neighboring pixels being used for testing. This reduces spatial overfitting, i.e., it avoids over-optimistic models (Meyer et al., 2018a). Five cross-validation folds were used during the training phase (*caret::ffs()*). These scripts can be found on GitHub.

The *caret::ffs()* function, or forward feature selection, first trains a model with two predictors using all possible pairs of predictor variables, after which the best initial model is kept. Iteratively, a new predictor is added to the model, and again the best combination is kept. This process stops when there is no increase in model performance. This function reduces the complexity of the model; however, combining all predictors takes time. Doubling the number of variables results in roughly four times as many sub-models to process.

All hyperparameters were tuned through the *tuneLength* (in *caret::ffs()*) option, which generated five random tuning parameter combinations. Manual hyperparameter setting was considered but not used. The classification model with the highest overall accuracy was used to predict the entire extent of each site.

3.5. Accuracy/error assessment

We evaluated the performance of the models using a range of metrics, including overall map accuracy, user accuracy, and producer accuracy. These metrics were calculated using the unbiased accuracy assessment method described by Olofsson et al. (2014) and the *mapac* package in R (Pflugmacher, 2022).

To assess the models, we used a cross-validation approach, in which the training data was split into folds, and the model with the highest result was compared to 20% of the validation data (the same 20% in each run). The results for each model were then reported in a confusion matrix. It is important to note that the overall accuracy can be biased towards the most abundant class in the training data. Therefore, it is useful also to consider the user's and producer's accuracies, which provide more detailed information about the model's performance for a specific class.

3.6. Presentation of results

Using multiple models to assess irrigated agriculture is crucial, but defining boundaries can vary. To address uncertainty, "irrigation hotspots" can be identified as areas where irrigation is known to exist but cannot be accurately measured. "Agreement maps" combine model classifications, showing consensus on irrigation locations. A value of 4 out of 4 models signifies unanimous identification, while 1 out of 4 models means only one identified irrigation.

4. Results

In the first section, we explore the influence of the different algorithms (using only the 2x6-month composite). In the second section, we explore the influence of composite length on the visibility of irrigated agriculture (using only the rf algorithm).

4.1. Comparison: algorithms

We use the 2x6-monthly composites to compare how well irrigated agriculture is classified using different algorithms for Catandica and Xai-Xai regions. This composite length is used because of the balance between a low number of parameters (i.e., computation time) and acceptable accuracies.

4.1.1. Accuracies and classifications for different algorithms

The results in Table 5 show the accuracies of various models that use different algorithms for classifying irrigated agriculture in two study areas. The knn algorithm had low user and producer accuracy (7-26%) in both areas but had higher overall accuracy due to its good performance in identifying tree cover in Catandica and grassland in Xai-Xai. The nnet and svmRadial algorithms had very high accuracy (95-99%) in Catandica, while the rf algorithm had reasonable accuracy (80-85%) in both areas, and the svmRadial algorithm had reasonable accuracy (75-85%) in Xai-Xai. The overall accuracies were higher than the class-specific accuracies, indicating that certain classes, such as dense and shrubland in Catandica and grassland in Xai-Xai, were more prevalent. The confusion matrices in Annex 1 show that in Xai-Xai, irrigated agriculture was mainly confused with grassland and shrubland, while in Catandica, it was mostly confused with both light and tree cover (for the rf classification only).

Figure 4 demonstrates that while the nnet and svm algorithms have similar levels of accuracy, they produce different maps of irrigated agriculture. The svm algorithm shows more irrigation on the western side of the map, while the nnet algorithm shows more clusters of irrigation following streams. While the rf and nnet algorithms have different levels of

Table 5 User, producer and overall accuracy for irrigated agriculture for the algorithm models.

		Accuracy		
		Producer's	**User's**	**Overall**
Catandica	knn	7.4 ± 0.6	25.8 ± 2.0	61.1 ± 0.5
	nnet	99.6 ± 0.3	97.7 ± 0.7	98.4 ± 0.2
	rf	80.3 ± 1.7	79.3 ± 1.9	93.7 ± 0.3
	svmRadial	93.5 ± 1.1	94.9 ± 1.0	98.1 ± 0.2
Xai-Xai	knn	10.6 ± 0.6	18.5 ± 1.1	41.9 ± 0.4
	nnet	85.8 ± 0.9	91.0 ± 0.8	91.6 ± 0.3
	rf	85.9 ± 0.9	86.2 ± 1.0	91.8 ± 0.3
	svmRadial	74.3 ± 1.0	84.8 ± 1.0	85.3 ± 0.3

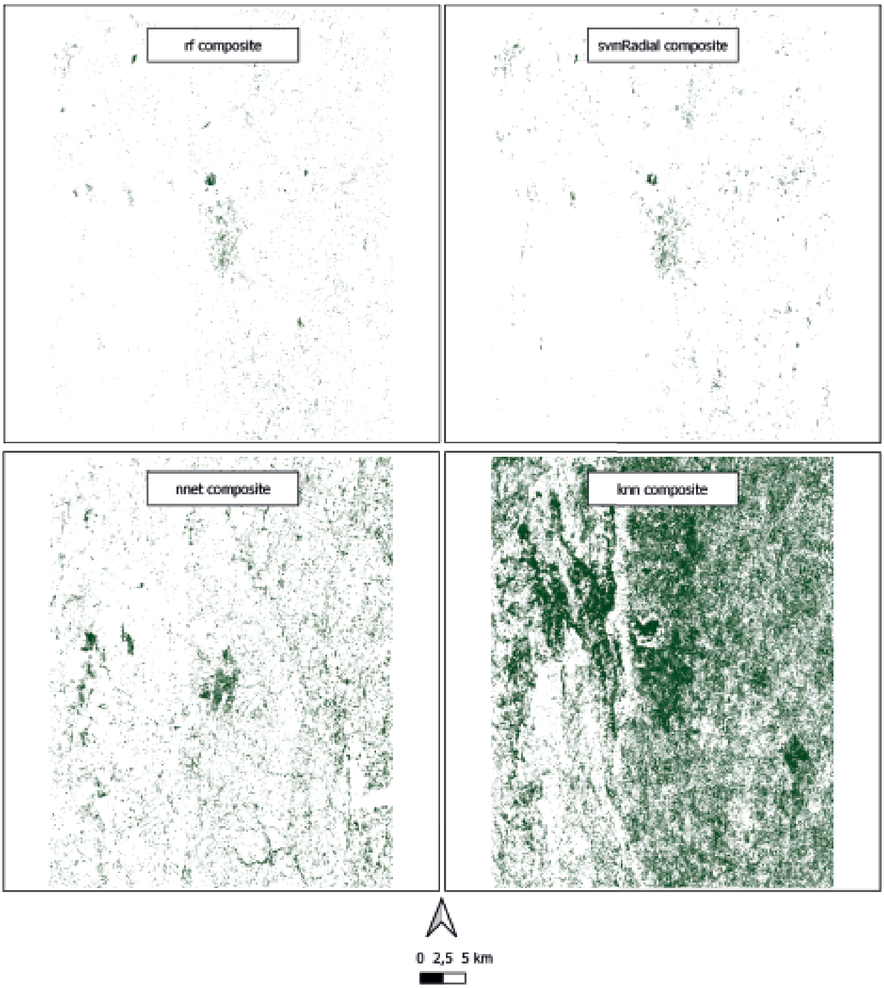


Figure 4 Extent of irrigated agriculture per algorithm (tlbr: rf, svm, ann, knn) for Catandica.

accuracy, the maps they produce are similar. The knn algorithm greatly overestimates the extent of irrigated agriculture, with almost the entire map showing this class except for areas of tree cover in the bottom left corner. All four algorithms also incorrectly classify trees in Catandica town (located in the centre of the map, see Annex 2 for more details) and some rock outcroppings (not present in the training data) as irrigated agriculture.

In Xai-Xai, the rf and nnet algorithms have similar levels of accuracy, and their classified maps are also similar (Figure 5). However, both of these algorithms, as well as the svm algorithm, incorrectly classify many individual trees in towns (located in the bottom right quadrant of the map, see Annex 2 for more details) and groups of trees in predominantly

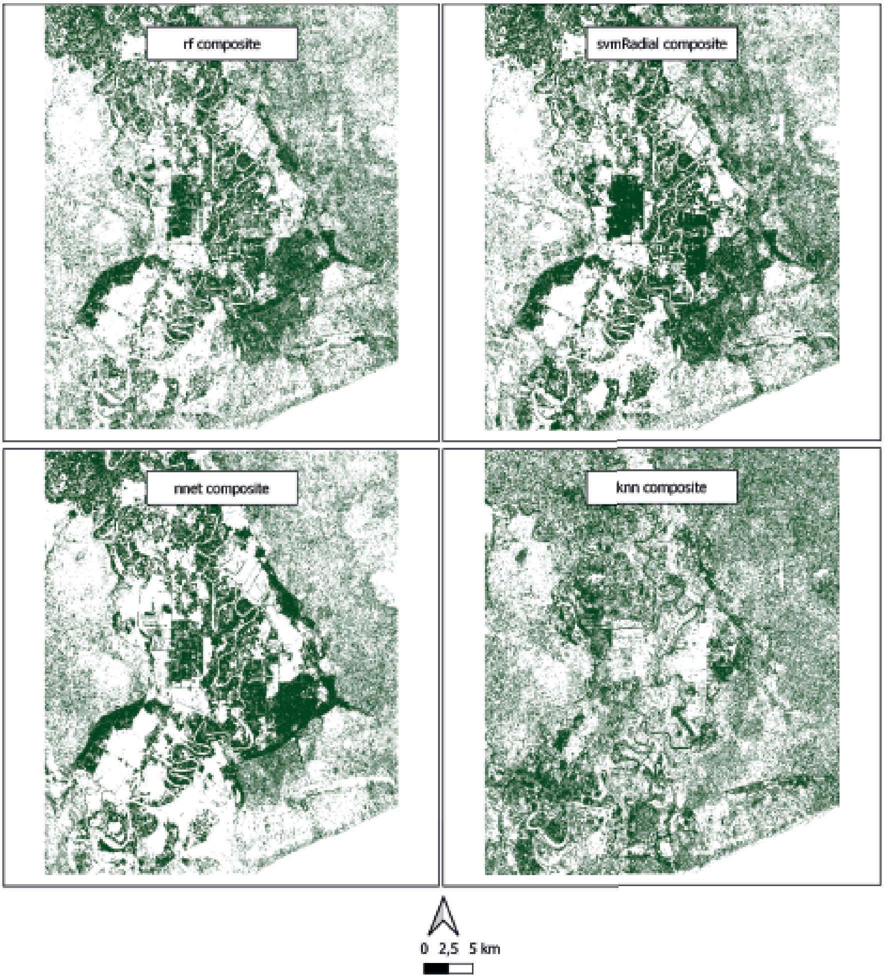


Figure 5 Extent of irrigated agriculture per algorithm (tlbr: rf, svm, nnet, knn) for Xai-Xai.

rainfed agriculture areas (on the east and west sides of the map, outside of the Limpopo river valley) as irrigated areas. The maps produced by these three algorithms show many areas of irrigated agriculture along the edges of the valley and the river. In contrast, the map produced by the knn algorithm shows no clear structures that follow the landscape.

4.1.2. Irrigation agreement maps

The knn algorithm tends to overestimate the area of irrigated agriculture, making it unsuitable to use in agreement maps for visualising hotspots accurately. Consequently, we will exclude its results and only consider the outcomes from the remaining three algorithms. By overlaying the estimated maps from these algorithms, which identify the irrigation class, in an 'agreement map,' we can identify hotspots (Figure 6).

In the top inset map (A), smallholder irrigation is near Catandica's urban region, correctly classified as irrigation by all three algorithms. However, the algorithms wrongly classify most of the urban trees as irrigation, and their boundaries differ slightly, leading to some areas with uncertainty. We call the pixels where all models agree (3/3 in this case) the *core areas*,

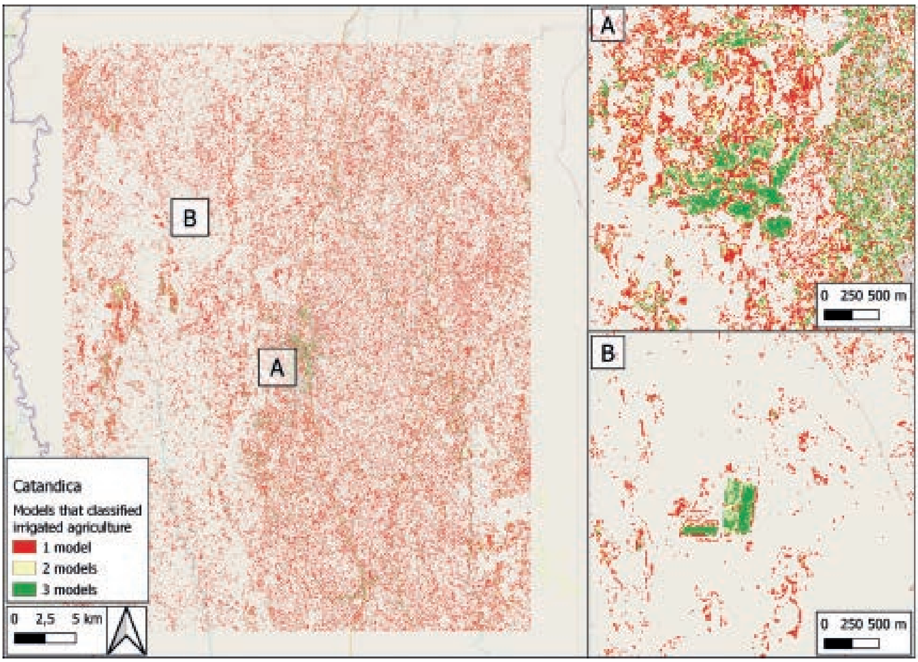


Figure 6 Map of Catandica showing how many of the models classified a pixel as irrigated agriculture, and two zoom-ins of a smallholder irrigation scheme (A) and part of the larger tea plantation (B). The values in the legend show how many models classified a pixel as irrigated agriculture: 3 means agreement in 3 models.

and the pixels surrounding these core areas the *uncertainty zone*. In the bottom inset map (B), the three algorithms accurately identify most of a tea plantation as irrigated agriculture, but some minor patches are classified differently by one or two algorithms. The knn algorithm, not included in this figure, classified all the surrounding grasslands as irrigated areas (Figure 4), overestimating the extent and location of irrigated agriculture.

In Xai-Xai (Figure 7), we excluded knn results from the agreement map. In area A (top inset map), smallholder irrigated fields have clusters of '3 models', indicating agreement between the results, but with less certain areas in between. The bottom right part of area A, an urban area (Xai-Xai), has most of the trees misclassified as irrigated area. Area B (bottom inset map) shows a large, irrigated rice scheme (Hubei-Gaza Rice project). There is a major cluster of irrigated agriculture recognised by all models in the centre of this map, but the remaining fields are only identified by one or two of the algorithms.

The main overview map also shows that there are a lot of irrigated areas in the northeast quadrant, which are mostly misclassified pixels (1 model); this area has a higher elevation

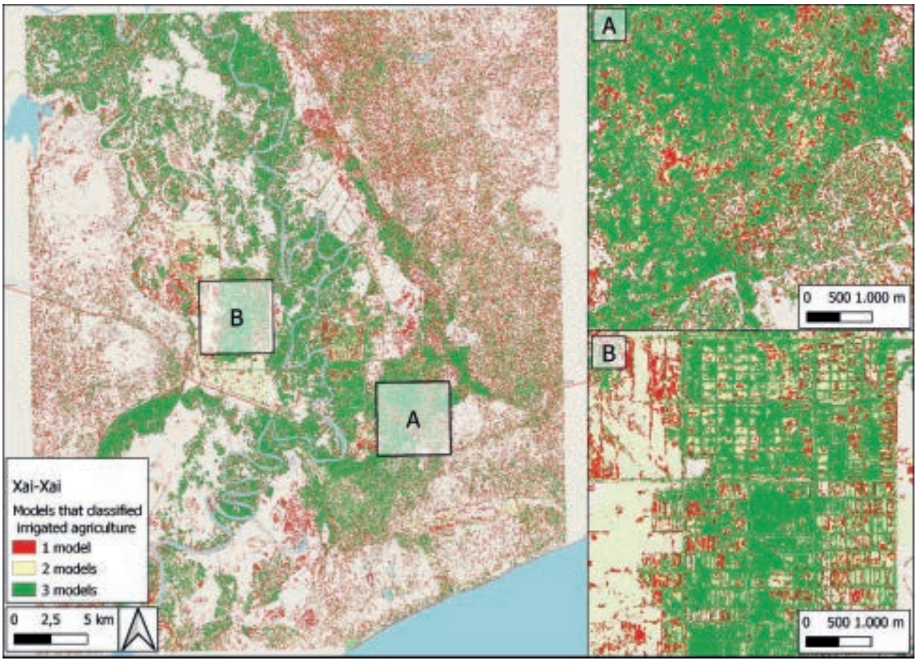


Figure 7 Map of Xai-Xai showing how many of the models classified a pixel as irrigated agriculture, and two zoom-ins of a smallholder irrigation scheme (A), and part of the large rice irrigation scheme (B). The values in the legend show how many models classified a pixel as irrigated agriculture: 3 means agreement in 3 models.

(+20 meters) than most of the irrigated fields (which are in the Limpopo valley), where we primarily find rainfed agriculture, small patches of tree cover, shrubland, and urban areas.

Table 6 Accuracy of hotspot values and total number of hectares classified.

			Training data			
		Agreement	# pixels correctly classified	Total # pixels classified	% correctly classified	Total hectares
Algorithms	Catandica	1 model	26	242	11%	23124
		2 models	189	200	95%	6660
		3 models	2145	2145	100,0%	1205
	Xai-Xai	1 model	355	2255	16%	20872
		2 models	839	1315	64%	17744
		3 models	4656	4704	99%	26537

Table 6 shows hotspot accuracy and classified hectares for Catandica and Xai-Xai, with three categories based on the agreement between models: 3 models refers to three models classifying the same pixel as irrigated agriculture. The table shows that 3 models pixels are almost 100% correctly classified as irrigated agriculture, indicating high confidence in the core hotspots. However, there is an uncertainty zone surrounding the core areas. In Catandica, the 2 models ring is still accurate, while in Xai-Xai, only two-thirds of the pixels were accurately classified. Pixels identified by only one model are usually incorrect and can be excluded from final assessments.

4.2. Comparison: composite lengths

Here we present the results of the different composite lengths using the RF. We used this algorithm because of its high computation speed, ease of use, and widespread use within the community.

4.2.1. Accuracies and classifications for different algorithms

Table 7 displays the accuracies of various models that used different composite lengths to classify irrigated agriculture in two study areas. All models had high overall accuracies (above 95%). A single 12-month composite may not be sufficient to capture the differences between irrigated agriculture, rainfed agriculture, and shrubland in a complex landscape, such as the one found in Manica. This composite performs better in the slightly less complex landscape of Chokwe. Based solely on overall accuracy, Chokwe should be classified using the 2x6-month composites, while Manica should be classified using the 6x2-month composites. However, doubling the number of variables results in roughly four times as many sub-models to process, with only a limited increase in accuracy. Additionally, accuracy alone is insufficient to base conclusions on, as discussed in Section 3.1.

Confusion of irrigated agriculture in Chokwe was mostly with shrubland in all models (Annex 1). In Manica, irrigated agriculture was confused with several classes, primarily rainfed agriculture, followed by shrubland.

Table 7 User, producer, and overall accuracy for irrigated agriculture for the algorithm models.

		Accuracy		
		Producer's	User's	Overall
Chokwe	12m	98.9 ± 0.2	98.0 ± 0.2	96.1 ± 0.2
	3m	97.5 ± 0.3	95.7 ± 0.3	94.9 ± 0.2
	6m	99.7 ± 0.1	99.3 ± 0.1	98.0 ± 0.1
	2m	98.0 ± 0.2	96.2 ± 0.3	95.9 ± 0.2
Manica	12m	73.9 ± 2.0	74.8 ± 2.3	94.5 ± 0.3
	3m	92.8 ± 1.3	90.2 ± 1.6	98.2 ± 0.2
	6m	90.4 ± 1.5	91.2 ± 1.6	98.3 ± 0.2
	2m	94.8 ± 1.2	91.8 ± 1.5	98.8 ± 0.1

Figure 8 shows the extents of irrigated agriculture for the four composites for Manica. At first glance, the four results seem similar, with irrigated agriculture following the rivers and slopes of the mountains. However, the urban area of Messica (located at the bottom centre of the map, see Annex 2 for more details) contains trees that have been misclassified as irrigated agriculture. Compared to Chokwe, Manica shows more small-scale irrigation spread out over the landscape.

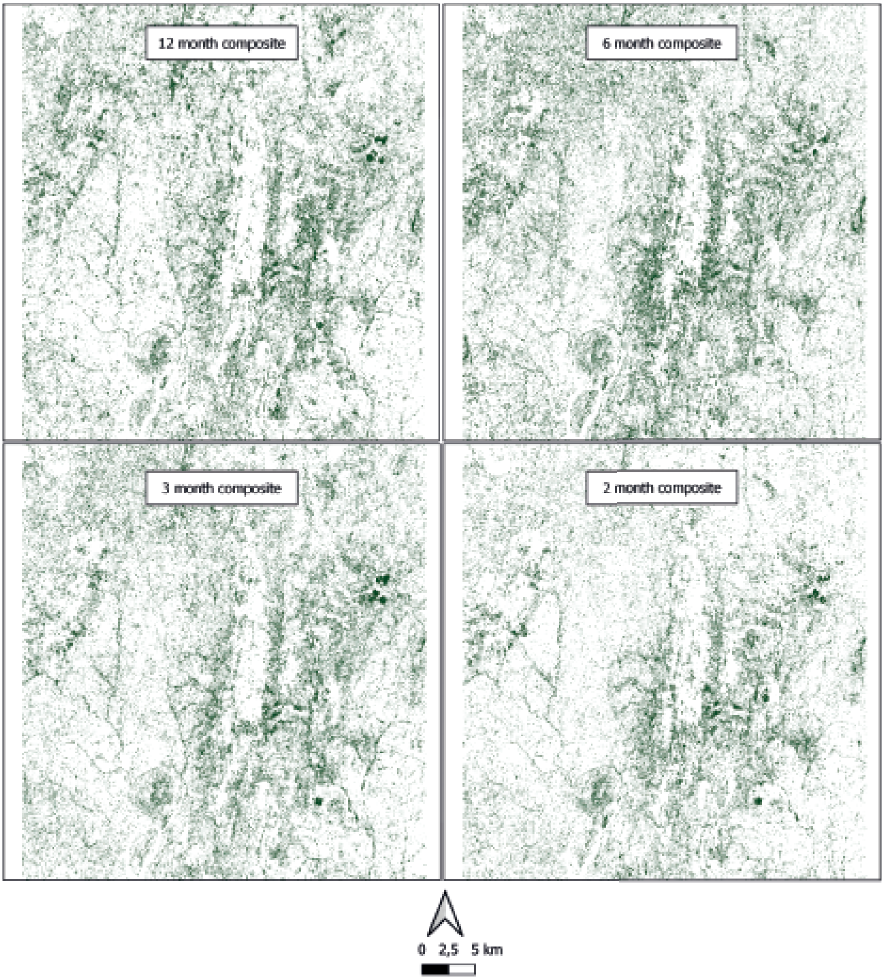


Figure 8 Extent of irrigated agriculture per composite length (tlbr: 12, 6, 3 &2 month) for Manica.

Figure 9 shows the similarities in the extent of irrigated agriculture in Chokwe and reveals that most of the fields at the head end of the Chokwe irrigation scheme are classified as irrigated agriculture and are actively cultivated, while the tail end shows less irrigated agriculture – this reflects the actual situation well. The 3-month and 2-month composites follow the same trends but show a smaller overall area of irrigated agriculture. The 6-month composite stands out from the other three in its lower misclassification of shrubland in the map's top right and bottom left parts. The other composites show small clusters of irrigated agriculture in these areas, which are not present in the 6-month composite. The urban area of Chokwe (located at the centre of the map, see Annex 2 for more details) hardly shows any irrigated agriculture, similar to the other three study areas.

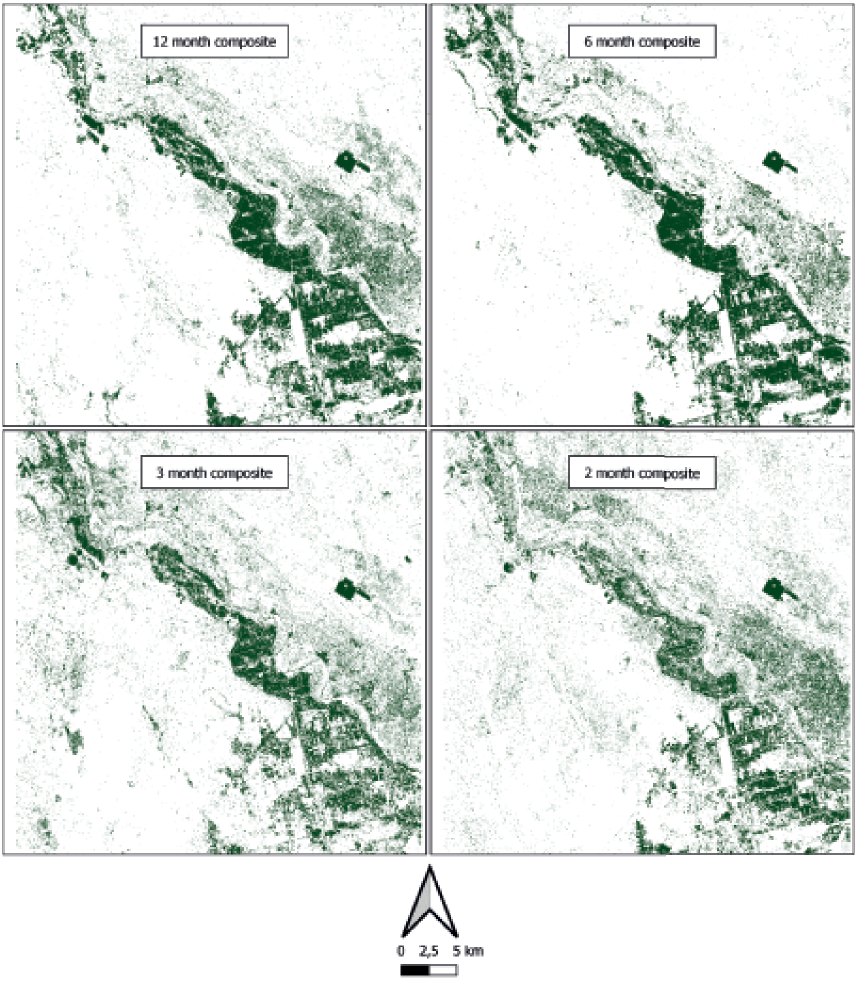


Figure 9 Extent of irrigated agriculture in green per composite length (tlbr: 12, 6, 3, 2 month) for Chokwe.

4.2.2. Irrigation agreement maps

In the main map of Manica (Figure 10), we can see irrigation occurring in riverbeds and near mountains, with some large clusters of fields as well as many small patches. Area A (top inset map) shows an area with two known, clearly delineated smallholder irrigation schemes. Some core areas (4 models) are surrounded by areas that gradually change from 3 models to 1 model in a short distance, the uncertainty zone. Area B (bottom inset map) focuses on a few centre pivots (circular shapes). It shows that only parts of these fields are labelled with 4 models – an agreement by all four models – but as all pixels of the field are irrigated by the centre pivot, we would expect all pixels of those fields to be labelled irrigation by all four models. If we had used only one classification, these areas would not have been very recognisable as centre pivots.

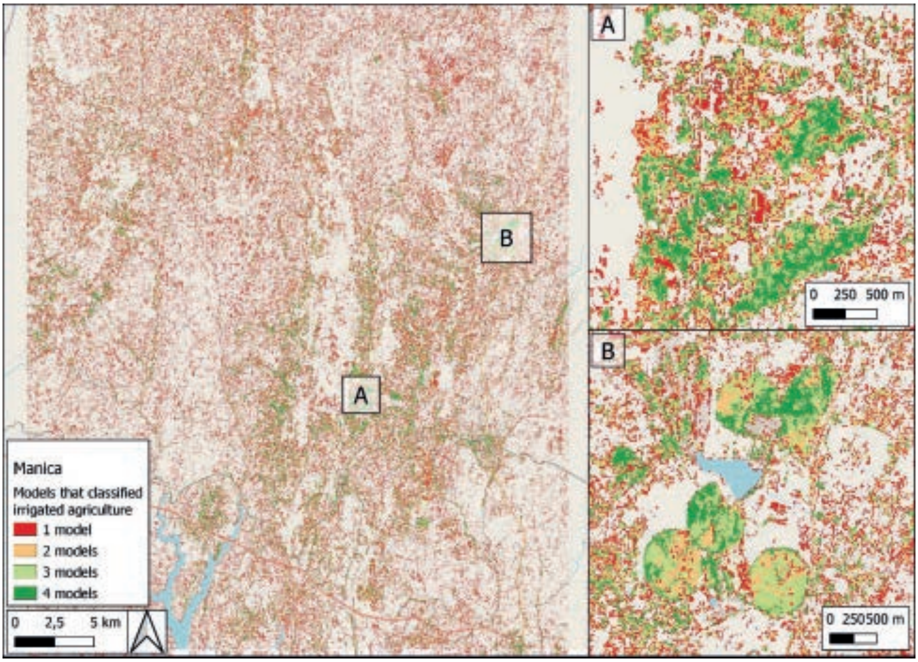


Figure 10 Map of Manica showing how many of the models classified a pixel as irrigated agriculture, and two zoom-ins of a smallholder irrigation scheme (A), and part of an irrigated estate by means of center pivots (circular shapes within area B). The values in the legend show how many models classified a pixel as irrigated agriculture: 4 means agreement in 4 models.

In Chokwe (Figure 11), we see a similar pattern of core area and uncertainty zone. The map clearly shows the large-scale Chokwe irrigation scheme along the Limpopo River's south bank and some smaller schemes on the north bank, such as area A (top inset map). It also shows that some of the models have identified irrigated agriculture on islands in the river (1 model), which is certainly possible but may be natural vegetation that has been misclassified. This area also contains clusters of trees in predominantly rainfed areas that have been misclassified as irrigated agriculture (1 model). Area B (bottom inset map) box highlights part of the Chokwe irrigation scheme, of which we know only part is still actively used.

Table 8 summarises the accuracy of the classification of irrigated agriculture in Manica and Chokwe using different composite models. In Manica, the 3 and 4 models agreement achieved 100% accuracy, while the 1 model and 2 models (uncertainty zone) had lower accuracy rates of 1.40% and 64.20%, respectively. In Chokwe, the 4 models achieved 100% accuracy, while the 1 model and 2 models had accuracy rates of 1.40% and 84.60%, respectively.

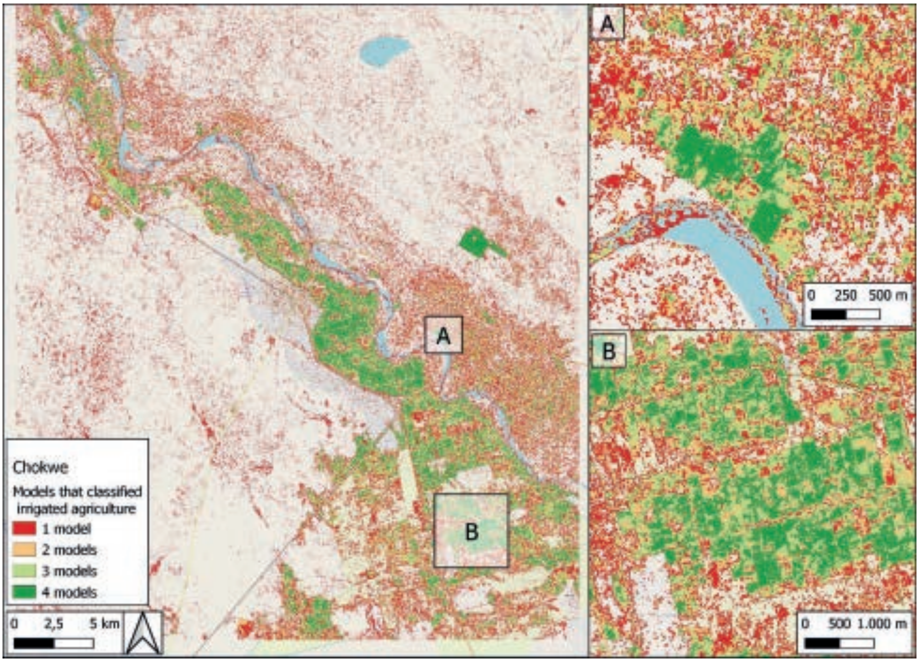


Figure 11 Map of Chokwe showing how many of the models classified a pixel as irrigated agriculture, and two zoom-ins of a smallholder irrigation scheme (blue area) and part of the large-scale Chokwe irrigation scheme (red area). The values in the legend show how many models classified a pixel as irrigated agriculture: 4 means agreement in 4 models.

Table 8 Accuracy of hotspot values and total number of hectares classified.

			Training data			
		Agreement between models	Irrigated agriculture	Total pixels classified	Irrigation correctly classified	Total hectares
Composites	Manica	1 model	6	440	1,40% 20444	
		2 models	61	95	64,20%	9565
		3 models	396	396	100,00%	5795
		4 models	1259	1259	100,00%	3289
	Chokwe	1 model	3	212	1,40% 16866	
		2 models	22	26	84,60%	8212
		3 models	370	370	100,00%	6736
		4 models	18199	18199	100,00%	5389

5. Discussion and recommendations

We examined how different composite lengths and algorithms affect the accuracy of remote sensing-based models in identifying irrigated agriculture in four distinct study areas. Our analysis of 16 models revealed that the composite length and algorithm choice can significantly impact the results. Therefore, it is necessary to integrate the results of various models to account for model-specific biases. The following sections discuss our key findings.

5.1. Algorithm

Our study found that the choice of algorithm can greatly impact the accuracy of remote sensing-based models in identifying irrigated agriculture. Our experiments showed that ANNs, SVMs, and RFs effectively classified irrigated areas. However, there was no straightforward "best" algorithm, as all achieved user, producer, and overall accuracies ranging from 80% to 95%.

Based on the agreements and differences observed between the different algorithm maps, we recommend using at least three algorithms and focusing on hotspots to consider both the heterogeneous and homogeneous parts of the landscape in the model. Additional research could assess the algorithmic sensitivity to the diverse methods employed in farmer-led irrigation. This could be accomplished by analysing the performance of the models in scenarios where training data from these farmers are either excluded or included, allowing for a comparison between the two.

5.2. Composite length

The study found that composite length is crucial in accurately identifying irrigated agriculture in diverse landscapes. Shorter composites are better for complex landscapes, while longer composites are sufficient for homogeneous ones. It is important to consider composite length when creating remote sensing-based models and to focus on hotspots. The 6-month and 3-month composites are promising options due to their lower computation time and data size. Using agreement maps incorporating multiple composites enhances the visibility of features like centre pivots.

Further investigation could centre on determining the optimal selection of months to include or exclude in the composite. In the current research, a 12-month dataset was used, distributed across various composite lengths. However, it is worth exploring the possibility of achieving comparable results by solely utilising the dry season months. This approach may offer the advantage of requiring less data and reducing model complexity.

5.3. Model agreement method: hotspot maps

Our analysis revealed that combining models with different composite lengths and algorithms can improve the accuracy of identifying irrigated agriculture. Hotspot maps provide valuable information for decision-making and prioritising targeted field surveys or management decisions. For complex landscapes with dynamic and heterogeneous classes, combining models can provide better insights into the core areas of hotspots. We recommend including at least three models to improve the accuracy of the core areas.

5.4. Reflection: other aspects that likely influenced our results

Our findings suggest using multiple composite lengths to capture the dynamic nature of irrigated agriculture. Shorter composites (quarterly or bi-monthly) are necessary to identify highly dynamic classes like irrigated agriculture accurately, while longer composites (annual or seasonal) may be more effective for stable classes like tree cover and urban areas. Focusing on specific periods, such as the end of the rainy season and the start of the dry season, can also help capture changes in irrigation and vegetation patterns.

We chose variables and composite statistics based on previous studies on mapping irrigated agriculture (Elwan et al., 2022; Lebourgeois et al., 2017; Wellington & Renzullo, 2021; Xie et al., 2019). Our aim was not to determine the "best" variables or statistics, as this is context-dependent. Different combinations of variables were important for different runs, and the geomad statistic was sufficient to show the influence of composite length and algorithm choice. Although these methods have the potential to improve accuracies further, our results were already high, which raises the question of whether more effort should be focused on field data collection or improving models at optimal performance.

The training data was collected during the dry season, and the labels for rainfed agriculture were based on leftover maise stalks and shrubland. The training data may have been imbalanced, with fewer samples for less prevalent classes. The regular clustered sampling design was used due to the tradeoff between complete random data collection and travel time. The data collected through abandoned strategies were still used, but the overall size was small. Some areas, such as bare rocks and sand, were included into the built-up class, which resulted in inaccurate classification by the algorithms.

The study was conducted over four areas chosen because of their differences in weather, topography, and agricultural uses. We hoped to capture various irrigation circumstances but undoubtedly missed some practices. Hence the findings on composite and algorithm use may be helpful for some areas of Mozambique but less so further away. For example, Wellington & Renzullo (2021) found that the annual composite was optimal for classifying irrigated agriculture in Zimbabwe.

6. Conclusion

We investigated the impact of different composite lengths and algorithms on the accuracy of remote sensing-based models for identifying irrigated agriculture in four sub-Saharan African study areas. Our findings showed that the choice of algorithm and composite length can considerably affect model outcomes. We found that SVMs, RFs, and ANNs were effective in classifying irrigated areas, while the k-nearest neighbour algorithm was ineffective in this task. Shorter composite lengths, such as 2-monthly or 3-monthly composites, were more effective for identifying irrigated agriculture in complex and dynamic landscapes, while longer composite lengths were more appropriate for stable classes.

Our study also highlighted the importance of considering hotspots and agreement maps when identifying irrigated agriculture. Combining the outputs of various models into agreement maps can provide better insights into the core areas and uncertainty zones of hotspots. These findings can help decision-makers remotely situated to understand irrigation dynamics better.

7. Annex

7.1. Annex 1: Confusion matrices

7.1.1. *Catandica*

Comparison algorithm - Catandica 6m knn composite

	Reference class (sample counts)					Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Map	Ref.	Producer's	User's	F-score
Bu	5	3	88	9	6	111	23	21.74 ± 8.44	4.50 ± 1.97	7.46 ± 2.75
Dv	18	2977	209	35	2	3241	3371	88.31 ± 0.48	91.85 ± 0.48	90.05 ± 0.34
Ia	-	115	121	232	1	469	1633	7.41 ± 0.55	25.80 ± 2.02	11.51 ± 0.70
Lv	0	276	1042	773	26	2117	1279	60.44 ± 1.02	36.51 ± 1.05	45.52 ± 0.86
Ra	0	0	173	230	3	406	38	7.89 ± 4.36	0.74 ± 0.43	1.35 ± 0.71

Overall accuracy = 61.14 ± 0.45

Comparison algorithm - Catandica 6m nnet composite

	Reference class (sample counts)					Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Map	Ref.	Producer's	User's	F-score
Bu	110	-	-	0	1	111	111	99.10 ± 0.89	99.10 ± 0.90	99.10 ± 0.63
Dv	-	3207	-	34	0	3241	3237	99.07 ± 0.17	98.95 ± 0.18	99.01 ± 0.12
Ia	1	5	458	2	3	469	460	99.57 ± 0.31	97.65 ± 0.70	98.60 ± 0.39
Lv	0	25	1	2074	17	2117	2123	97.69 ± 0.32	97.97 ± 0.31	97.83 ± 0.22
Ra	0	0	1	13	392	406	413	94.92 ± 1.05	96.55 ± 0.91	95.73 ± 0.70

Overall accuracy = 98.38 ± 0.16

Comparison algorithm - Catandica 6m rf composite

	Reference class (sample counts)					Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Map	Ref.	Producer's	User's	F-score
Bu	84	-	1	1	25	111	92	91.30 ± 2.81	75.68 ± 4.07	82.76 ± 2.70
Dv	-	3165	38	38	0	3241	3293	96.11 ± 0.32	97.66 ± 0.27	96.88 ± 0.21
Ia	-	55	372	38	4	469	463	80.35 ± 1.68	79.32 ± 1.87	79.83 ± 1.26
Lv	0	73	39	1970	35	2117	2081	94.67 ± 0.47	93.06 ± 0.55	93.85 ± 0.36
Ra	8	0	13	34	351	406	415	84.58 ± 1.57	86.45 ± 1.70	85.51 ± 1.15

Overall accuracy = 93.66 ± 0.3

Comparison algorithm - Catandica 6m svmRadial composite

	Reference class (sample counts)					Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Map	Ref.	Producer's	User's	F-score
Bu	109	-	-	0	2	111	109	100.00 ± 0.00	98.20 ± 1.26	99.09 ± 0.64
Dv	-	3220	2	19	0	3241	3251	99.05 ± 0.17	99.35 ± 0.14	99.20 ± 0.11
Ia	-	8	445	13	3	469	476	93.49 ± 1.08	94.88 ± 1.02	94.18 ± 0.74
Lv	0	23	13	2064	17	2117	2101	98.24 ± 0.28	97.50 ± 0.34	97.87 ± 0.22
Ra	0	0	16	5	385	406	407	94.59 ± 1.09	94.83 ± 1.10	94.71 ± 0.77

Overall accuracy = 98.09 ± 0.17

7.1.2. Xai-Xai

Comparison algorithm - Xai-Xai 6m knn composite

	Reference class (sample counts)								Σ		Accuracy		
Map class	Bu	Dv	Gr	Ia	Lv	Ra	Wa	We	Map	Ref	Producer's	User's	F-score
Bu	440	-	-	3	-	-	0	0	443	1477	29.79 ± 0.51	99.32 ± 0.39	45.83 ± 0.60
Dv	-	827	18	58	-	-	0	135	1038	1138	72.67 ± 1.10	79.67 ± 1.25	76.01 ± 0.83
Gr	158	70	1290	1025	2	-	0	31	2576	2143	60.20 ± 0.80	50.08 ± 0.99	54.67 ± 0.67
Ia	48	34	516	223	-	-	0	384	1205	2109	10.57 ± 0.60	18.51 ± 1.12	13.46 ± 0.57
Lv	200	10	250	168	-	-	0	7	635	9	0.00 ± 0.00	0.00 ± 0.00	-
Ra	101	-	17	19	-	-	0	0	137	41	0.00 ± 0.00	0.00 ± 0.00	-
Wa	524	0	0	419	6	41	0	0	990	0	-	0.00 ± 0.00	-
We	6	197	52	194	1	0	0	609	1059	1166	52.23 ± 1.13	57.51 ± 1.52	54.74 ± 0.93

Overall accuracy = 41.93 ± 0.44

Comparison algorithm - Xai-Xai 6m nnet composite

	Reference class (sample counts)								Σ		Accuracy		
Map class	Bu	Dv	Gr	Ia	Lv	Ra	Wa	We	Map	Ref	Producer's	User's	F-score
Bu	437	-	-	3	1	2	0	0	443	439	99.54 ± 0.32	98.65 ± 0.55	99.09 ± 0.32
Dv	-	963	22	29	19	-	1	4	1038	1010	95.35 ± 0.64	92.77 ± 0.80	94.04 ± 0.52
Gr	-	6	2439	77	32	-	0	22	2576	2662	91.62 ± 0.49	94.68 ± 0.44	93.13 ± 0.33
Ia	1	8	59	1097	26	4	0	10	1205	1278	85.84 ± 0.89	91.04 ± 0.82	88.36 ± 0.61
Lv	1	28	115	46	397	47	0	1	635	529	75.05 ± 1.60	62.52 ± 1.92	68.21 ± 1.32
Ra	-	-	6	18	53	60	0	0	137	113	53.10 ± 4.09	43.80 ± 4.24	48.00 ± 3.05
Wa	0	0	0	0	0	0	990	0	990	991	99.90 ± 0.10	100.00	99.95
We	0	5	21	8	1	0	0	1024	1059	1061	96.51 ± 0.55	96.69 ± 0.55	96.60 ± 0.39

Overall accuracy = 91.64 ± 0.28

Comparison algorithm - Xai-Xai 6m rf composite

	Reference class (sample counts)								Σ		Accuracy		
Map class	Bu	Dv	Gr	Ia	Lv	Ra	Wa	We	Map	Ref	Producer's	User's	F-score
Bu	429	-	1	7	1	5	0	0	443	436	98.39 ± 0.59	96.84 ± 0.83	97.61 ± 0.51
Dv	-	945	17	25	19	1	0	31	1038	1009	93.66 ± 0.74	91.04 ± 0.89	92.33 ± 0.58
Gr	-	22	2464	48	28	-	0	14	2576	2656	92.77 ± 0.46	95.65 ± 0.40	94.19 ± 0.31
Ia	1	9	91	1039	44	8	0	13	1205	1209	85.94 ± 0.91	86.22 ± 0.99	86.08 ± 0.67
Lv	2	6	76	65	448	31	0	7	635	568	78.87 ± 1.54	70.55 ± 1.81	74.48 ± 1.22
Ra	4	-	3	12	25	93	0	0	137	138	67.39 ± 3.47	67.88 ± 3.99	67.64 ± 2.64
Wa	0	0	0	0	0	0	990	0	990	990	100.00 ± 0.00	100.00	100.00
We	0	27	4	13	3	0	0	1012	1059	1077	93.96 ± 0.70	95.56 ± 0.63	94.76 ± 0.47

Overall accuracy = 91.8 ± 0.29

Comparison algorithm - Xai-Xai 6m svmRadial composite

	Reference class (sample counts)								Σ		Accuracy		
Map class	Bu	Dv	Gr	Ia	Lv	Ra	Wa	We	Map	Ref	Producer's	User's	F-score
Bu	410	-	2	10	9	12	0	0	443	432	94.91 ± 1.01	92.55 ± 1.25	93.71 ± 0.81
Dv	-	873	34	44	-	-	0	87	1038	957	91.22 ± 0.87	84.10 ± 1.13	87.52 ± 0.73
Gr	-	28	2361	150	14	-	0	23	2576	2835	83.28 ± 0.55	91.65 ± 0.54	87.27 ± 0.39
Ia	-	23	141	1022	5	3	0	11	1205	1376	74.27 ± 1.00	84.81 ± 1.03	79.19 ± 0.73
Lv	13	1	266	97	202	44	0	12	635	257	78.60 ± 2.36	31.81 ± 1.85	45.29 ± 1.91
Ra	9	-	24	9	27	68	0	0	137	127	53.54 ± 3.80	49.64 ± 4.27	51.52 ± 2.90
Wa	0	0	0	0	0	0	990	0	990	1001	98.90 ± 0.33	100.00	99.45
We	0	32	7	44	0	0	11	965	1059	1098	87.89 ± 0.90	91.12 ± 0.87	89.48 ± 0.63

Overall accuracy = 85.25 ± 0.35

7.1.3. *Chokwe*

Comparison composite length - Chokwe 12m rf composite

	Reference class (sample counts)							Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	We	Map	Ref.	Producer's	User's	F-score
Bu	262	-	1	2	-	0	0	265	265	98.87 ± 0.65	98.87 ± 0.65	98.87 ± 0.46
Dv	-	319	6	123	-	0	2	450	344	92.73 ± 1.36	70.89 ± 2.14	80.35 ± 1.47
Ia	-	7	3634	45	21	0	1	3708	3675	98.88 ± 0.17	98.00 ± 0.23	98.44 ± 0.14
Lv	3	17	13	4324	61	0	26	4444	4643	93.13 ± 0.33	97.30 ± 0.24	95.17 ± 0.21
Ra	-	1	17	87	935	0	12	1052	1029	90.86 ± 0.86	88.88 ± 0.97	89.86 ± 0.65
Wa	0	0	0	0	0	381	0	381	381	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
We	0	0	4	62	12	0	3050	3128	3091	98.67 ± 0.20	97.51 ± 0.28	98.09 ± 0.17

Overall accuracy = 96.11 ± 0.16

Comparison composite length - Chokwe 6m rf composite

	Reference class (sample counts)							Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	We	Map	Ref.	Producer's	User's	F-score
Bu	268	-	-	1	1	0	0	270	268	100.00 ± 0.00	99.26 ± 0.52	99.63 ± 0.26
Dv	-	337	1	108	-	0	0	446	347	97.12 ± 0.89	75.56 ± 2.03	84.99 ± 1.33
Ia	-	3	3680	18	4	0	0	3705	3692	99.67 ± 0.09	99.33 ± 0.13	99.50 ± 0.08
Lv	-	7	6	4375	57	0	6	4451	4550	96.15 ± 0.26	98.29 ± 0.19	97.21 ± 0.16
Ra	-	-	5	43	988	0	1	1037	1050	94.10 ± 0.70	95.27 ± 0.66	94.68 ± 0.48
Wa	0	0	0	0	0	381	0	381	381	100.00 ± 0.00	100.00	100.00
We	0	0	0	5	0	0	3122	3127	3129	99.78 ± 0.08	99.84 ± 0.07	99.81 ± 0.06

Overall accuracy = 98.02 ± 0.11

Comparison composite length - Chokwe 3m rf composite

	Reference class (sample counts)							Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	We	Map	Ref.	Producer's	User's	F-score
Bu	272	-	-	7	1	0	0	280	274	99.27 ± 0.51	97.14 ± 1.00	98.19 ± 0.57
Dv	-	316	12	120	1	0	4	453	348	90.80 ± 1.50	69.76 ± 2.16	78.90 ± 1.49
Ia	-	7	3533	84	56	0	10	3690	3624	97.49 ± 0.25	95.75 ± 0.33	96.61 ± 0.21
Lv	2	22	52	4261	85	0	20	4442	4635	91.93 ± 0.36	95.93 ± 0.30	93.89 ± 0.23
Ra	-	1	21	123	942	0	2	1089	1088	86.58 ± 0.96	86.50 ± 1.04	86.54 ± 0.71
Wa	0	0	0	0	0	381	0	381	381	100.00 ± 0.00	100.00	100.00
We	0	2	6	40	3	0	3018	3069	3054	98.82 ± 0.19	98.34 ± 0.23	98.58 ± 0.15

Overall accuracy = 94.92 ± 0.18

Comparison composite length - Chokwe 2m rf composite

	Reference class (sample counts)							Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	We	Map	Ref.	Producer's	User's	F-score
Bu	266	-	-	3	4	0	0	273	267	99.63 ± 0.37	97.44 ± 0.96	98.52 ± 0.52
Dv	-	337	8	129	-	0	0	474	362	93.09 ± 1.30	71.10 ± 2.08	80.62 ± 1.42
Ia	-	1	3546	124	13	0	1	3685	3619	97.98 ± 0.23	96.23 ± 0.31	97.10 ± 0.20
Lv	1	24	42	4281	61	0	34	4443	4618	92.70 ± 0.35	96.35 ± 0.28	94.49 ± 0.22
Ra	-	-	23	47	977	0	1	1048	1055	92.61 ± 0.77	93.23 ± 0.78	92.91 ± 0.55
Wa	0	0	0	0	0	382	0	382	382	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
We	0	0	0	34	0	0	3092	3126	3128	98.85 ± 0.19	98.91 ± 0.19	98.88 ± 0.13

Overall accuracy = 95.9 ± 0.17

7.1.4. Manica

Comparison composite length - Manica 12m rf composite

	Reference class (sample counts)						Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	Map	Ref.	Producer's	User's	F-score
Bu	150	-	1	2	1	0	154	153	98.04 ± 1.10	97.40 ± 1.28	97.72 ± 0.85
Dv	-	2286	13	11	1	0	2311	2333	97.99 ± 0.28	98.92 ± 0.22	98.45 ± 0.18
Ia	-	20	264	20	49	0	353	357	73.95 ± 1.96	74.79 ± 2.31	74.37 ± 1.51
Lv	-	27	40	204	57	0	328	272	75.00 ± 2.31	62.20 ± 2.68	68.00 ± 1.86
Ra	3	0	39	35	155	0	232	263	58.94 ± 2.42	66.81 ± 3.09	62.63 ± 1.93
Wa	0	0	0	0	0	2412	2412	2412	100.00 ± 0.00	100.00	100.00

Overall accuracy = 94.49 ± 0.26

Comparison composite length - Manica 6m rf composite

	Reference class (sample counts)						Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	Map	Ref.	Producer's	User's	F-score
Bu	153	-	-	2	2	0	157	159	96.23 ± 1.47	97.45 ± 1.26	96.84 ± 0.97
Dv	-	2300	8	2	1	0	2311	2322	99.05 ± 0.20	99.52 ± 0.14	99.29 ± 0.12
Ia	1	9	291	9	9	0	319	322	90.37 ± 1.54	91.22 ± 1.58	90.80 ± 1.11
Lv	-	13	11	290	5	0	319	310	93.55 ± 1.34	90.91 ± 1.61	92.21 ± 1.05
Ra	5	0	12	7	211	0	235	228	92.54 ± 1.66	89.79 ± 1.98	91.14 ± 1.30
Wa	0	0	0	0	0	2454	2454	2454	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Overall accuracy = 98.34 ± 0.16

Comparison composite length - Manica 3m rf composite

	Reference class (sample counts)						Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	Map	Ref.	Producer's	User's	F-score
Bu	155	-	-	1	5	0	161	159	97.48 ± 1.22	96.27 ± 1.49	96.87 ± 0.97
Dv	-	2307	3	6	1	0	2317	2321	99.40 ± 0.16	99.57 ± 0.14	99.48 ± 0.10
Ia	1	8	323	9	17	0	358	348	92.82 ± 1.31	90.22 ± 1.57	91.50 ± 1.03
Lv	-	5	9	266	7	0	287	297	89.56 ± 1.65	92.68 ± 1.54	91.10 ± 1.13
Ra	3	1	13	15	206	0	238	236	87.29 ± 2.01	86.55 ± 2.21	86.92 ± 1.49
Wa	0	0	0	0	0	2421	2421	2421	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Overall accuracy = 98.2 ± 0.17

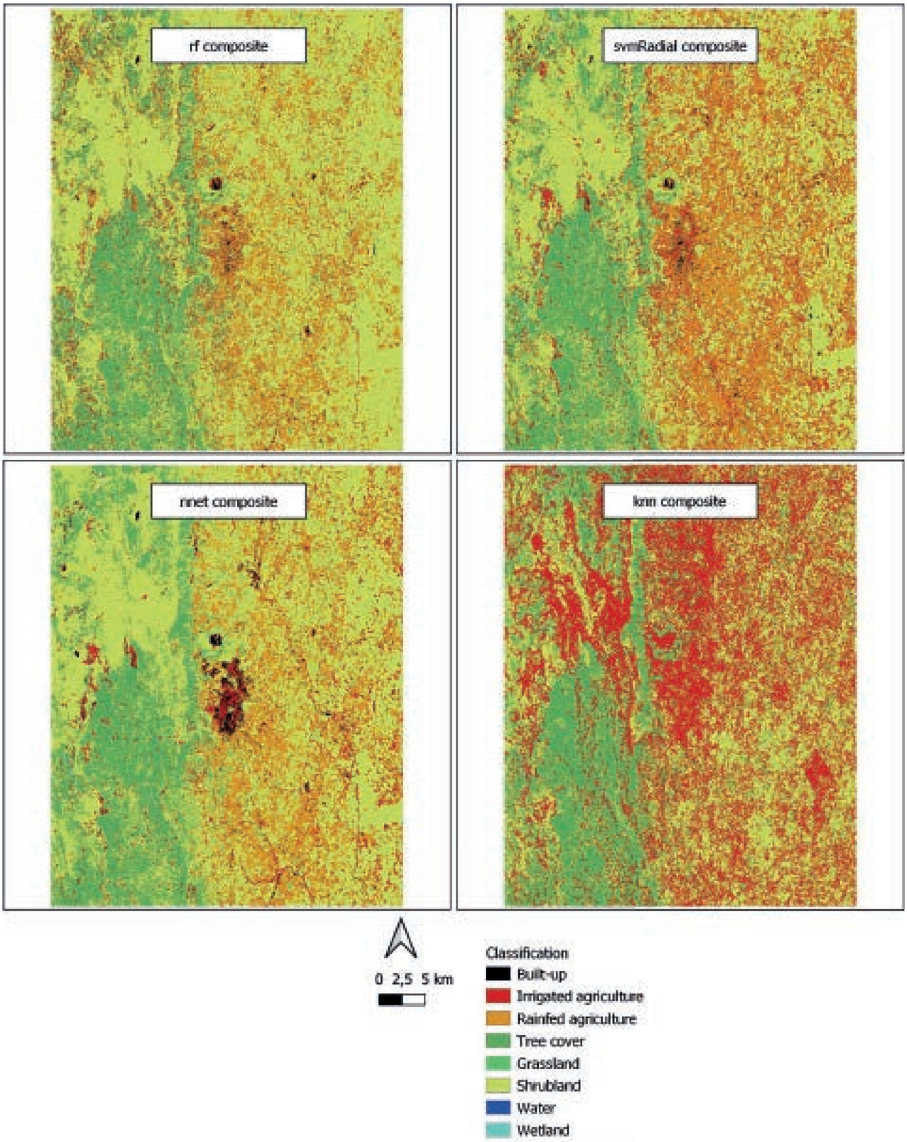
Comparison composite length - Manica 2m rf composite

	Reference class (sample counts)						Σ		Accuracy		
Map class	Bu	Dv	Ia	Lv	Ra	Wa	Map	Ref.	Producer's	User's	F-score
Bu	137	-	1	1	3	0	142	137	100.00 ± 0.00	96.48 ± 1.55	98.21 ± 0.80
Dv	-	2316	2	3	0	0	2321	2330	99.40 ± 0.16	99.78 ± 0.10	99.59 ± 0.09
Ia	-	13	312	9	6	0	340	329	94.83 ± 1.18	91.76 ± 1.49	93.27 ± 0.96
Lv	-	1	7	280	8	0	296	301	93.02 ± 1.40	94.59 ± 1.31	93.80 ± 0.96
Ra	0	0	7	8	221	0	236	238	92.86 ± 1.59	93.64 ± 1.59	93.25 ± 1.13
Wa	0	0	0	0	0	2446	2446	2446	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

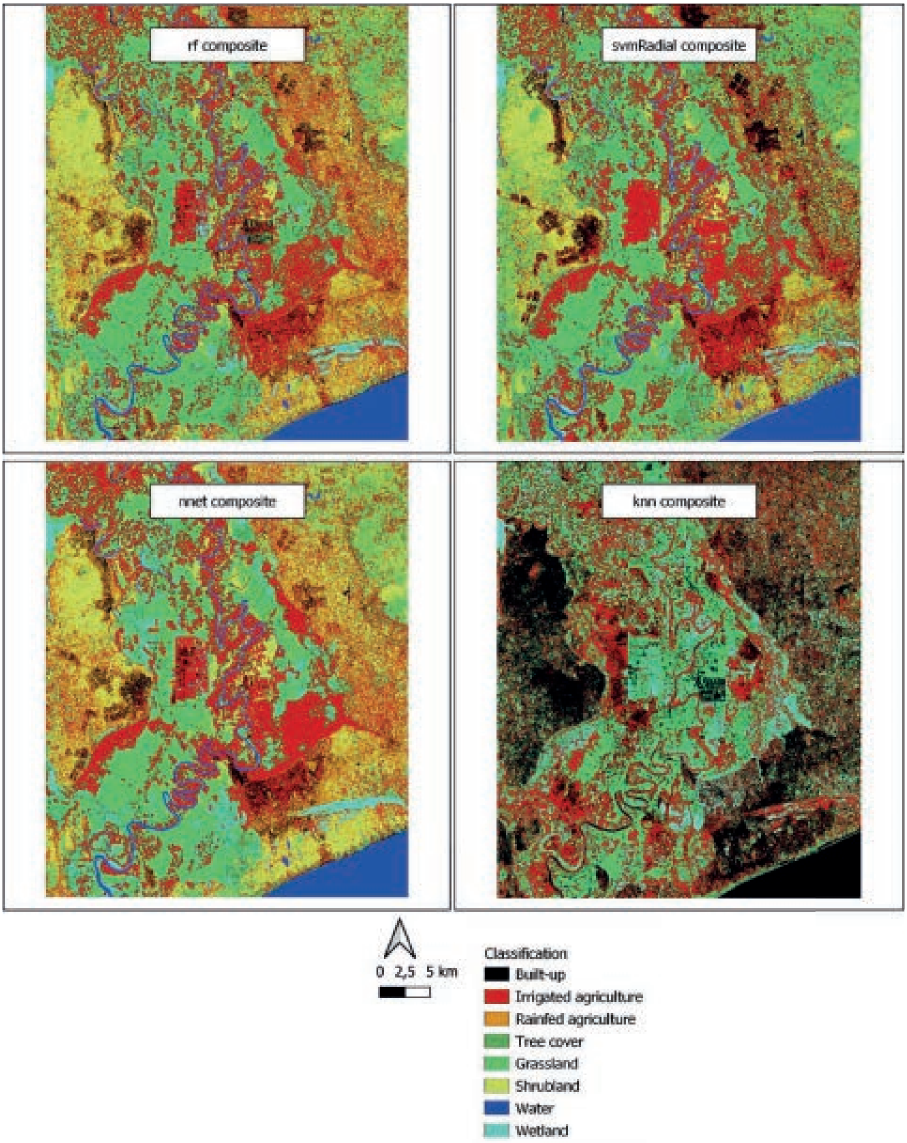
Overall accuracy = 98.81 ± 0.14

7.2. Annex 2: Classification maps

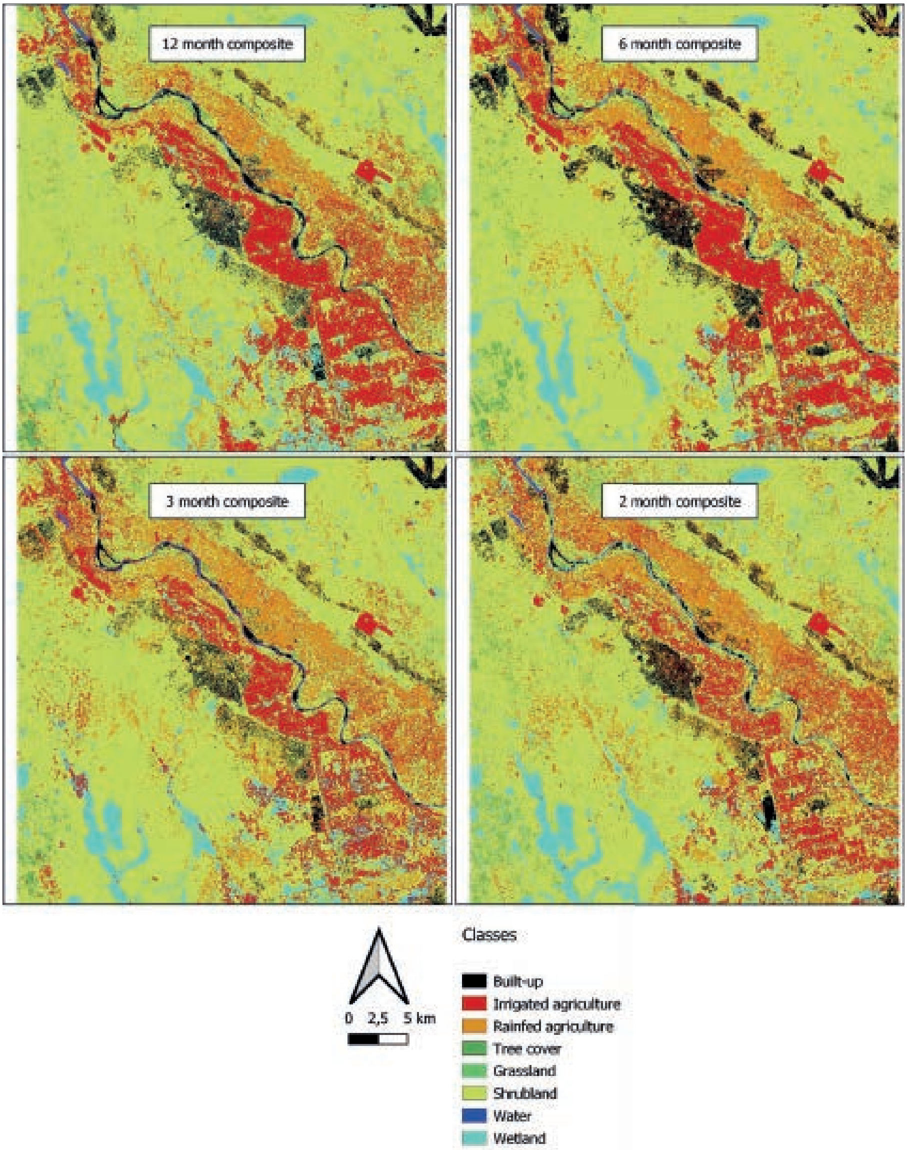
7.2.1. *Catandica*



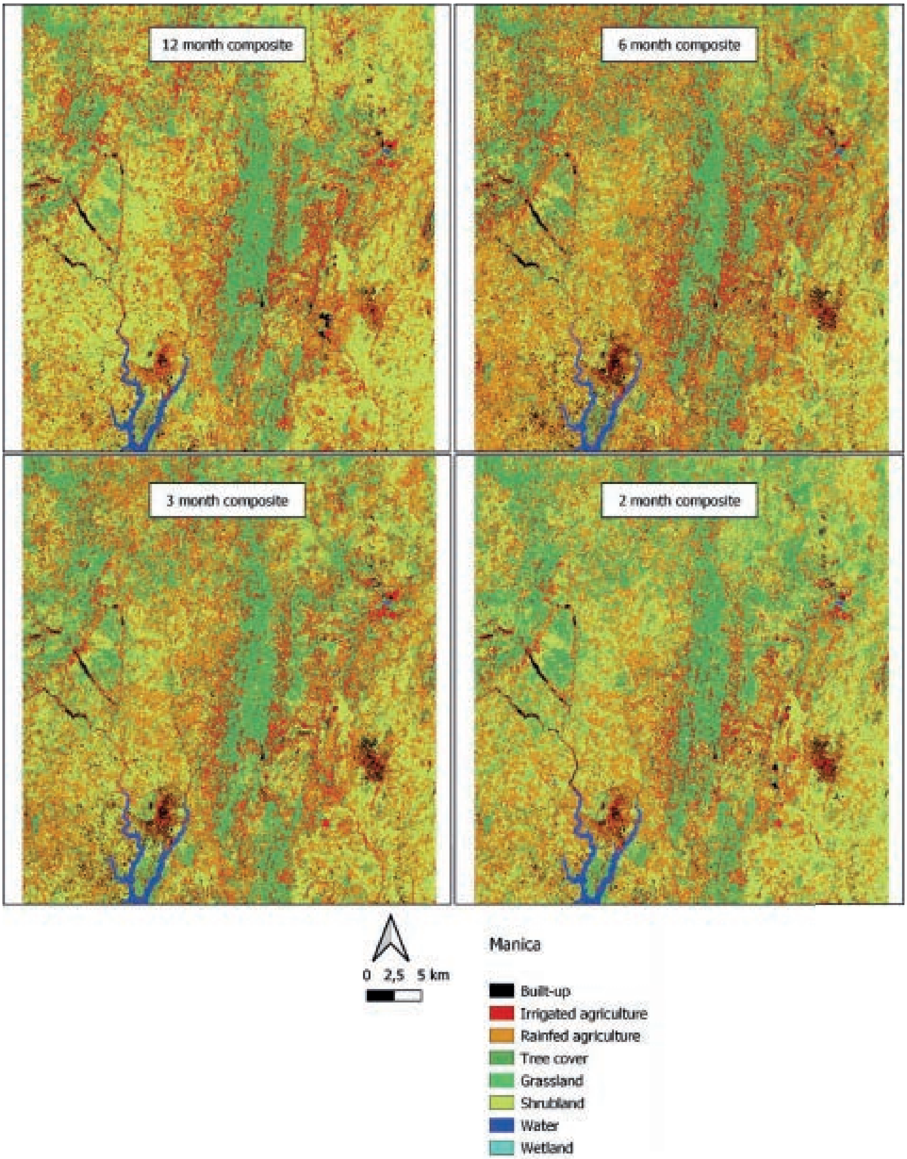
7.2.2. *Xai-Xai*


rf composite
svmRadial composite
nnet composite
knn composite
0 2,5 5 km
Classification
Built-up
Irrigated agriculture
Rainfed agriculture
Tree cover
Grassland
Shrubland
Water
Wetland

7.2.3. *Chokwe*


12 month composite
6 month composite
3 month composite
2 month composite
0 2,5 5 km
Classes
Built-up
Irrigated agriculture
Rainfed agriculture
Tree cover
Grassland
Shrubland
Water
Wetland

7.2.4. Manica


12 month composite
6 month composite
3 month composite
2 month composite
0 2,5 5 km
Manica
Built-up
Irrigated agriculture
Rainfed agriculture
Tree cover
Grassland
Shrubland
Water
Wetland

Chapter 4

Evaluating the effect of training data size and composition on the accuracy of smallholder irrigated agriculture mapping in Mozambique using remote sensing and machine learning algorithms

This chapter is published as:
Weitkamp, T., & Karimi, P. (2023).

Evaluating the Effect of Training Data Size and Composition on the Accuracy of Smallholder Irrigated Agriculture Mapping in Mozambique Using Remote Sensing and Machine Learning Algorithms.

Remote Sensing, 15(12), 3017. https://doi.org/10.3390/rs15123017

1. Abstract

Mapping smallholder irrigated agriculture in sub-Saharan Africa using remote sensing techniques is challenging due to its small and scattered areas and heterogenous cropping practices. A study was conducted to examine the impact of sample size and composition on the accuracy of classifying irrigated agriculture in Mozambique's Manica and Gaza provinces using three algorithms: random forest (RF), support vector machine (SVM), and artificial neural network (ANN). Four scenarios were considered, and the results showed that smaller datasets can achieve high and sufficient accuracies, regardless of their composition. However, the user and producer accuracies of irrigated agriculture do increase when the algorithms are trained with larger datasets.

The study also found that the composition of the training data is important, with too few or too many samples of the "irrigated agriculture" class decreasing overall accuracy. The algorithms' robustness depends on the training data's composition, with RF and SVM showing less decrease and spread in accuracies than ANN. The study concludes that the training data size and composition are more important for classification than the algorithms used. RF and SVM are more suitable for the task as they are more robust or less sensitive to outliers than the ANN. Overall, the study provides valuable insights into mapping smallholder irrigated agriculture in sub-Saharan Africa using remote sensing techniques.

2. Introduction

The size and composition of training samples are critical factors in remote sensing classification, as they can significantly impact classification accuracy. While sampling design is well-documented in the literature (Foody, 2009; Foody et al., 2006, 2016; Olofsson et al., 2014; Stehman & Foody, 2019), questions remain about the optimal number of samples required, their quality, and class imbalance (Collins et al., 2020; Mellor et al., 2015; Millard & Richardson, 2015). Class imbalance occurs when one or more classes is more abundant in the dataset than others, and since most machine learning classifiers try to decrease the overall error, the models are biased towards the majority class, leading to lower performances in classifying minority classes than majority classes (Ebrahimy et al., 2022). Generally, class imbalance can be dealt with through i) model-oriented solutions, where misclassifications are penalised, or where the algorithm focusses on a minority class, or ii) data-oriented solutions, where classes are balanced by over- or undersampling (Douzas et al., 2019).

Collecting a large number of quality training samples can be challenging due to limited time, access, or interpretability constraints. Practical issues and budget limitations can affect the sampling strategy, particularly in areas that are difficult to access, where rare land cover classes may be under-represented compared to more abundant classes (Mellor et al., 2015; C. A. Ramezan et al., 2021). Additionally, if data quality is a concern, selecting an algorithm that is less sensitive to such issues may be necessary. In the above cases it would be valuable to know how the sample size and composition affect the classification, and if additional samples are needed for increased accuracies. On the other hand, if a large sample size is already available, it may influence the choice of classifier.

These questions are even more relevant for mapping the extent monitoring irrigated agriculture. Especially smallholder irrigated agriculture is often inadequately represented in datasets and policies aimed at agricultural production and irrigation development, due to informal growth and lack of government or donor involvement (Beekman et al., 2014b; Veldwisch et al., 2019b; Venot et al., 2021; Woodhouse et al., 2017b). This results in a underrepresentation of smallholder irrigation in official statistics, even though smallholders provide most of the local food.

There are two general reasons for this underrepresentation. The first is the often a modernistic view of what constitutes irrigation by officials and data collectors (de Bont et al., 2019), in other words large scale systems. The second reason is that African smallholder agriculture is complex, with variability in field shape, cropping systems, and timing of agronomic activities (Bégué et al., 2018; Izzi et al., 2021; Veldwisch et al., 2019), often in areas that are hard to reach. Government officials and technicians that do not know about these areas will not visit them, fortifying the idea that there is no other irrigation than the large-scale systems (which are easier to reach and to recognize). Even if they do know about these systems, they might mislabel the very heterogeneous irrigated fields (i.e. many weeds) as natural vegetation.

To our knowledge, there have not been any studies yet that have investigated the effects of these biases in the training data set on classification results, and how choices made by the data collector result in changing accuracies. Choices could include oversampling irrigated agriculture because that is the class of interest, or being restricted in budget and only collecting a few samples. Ramezan et al., (2021) investigated the effects of sample size on different algorithms and we build on their ideas by including possible scenarios of how biased datasets can lead to misrepresentation.

There is ample literature on best practices regarding sampling strategies, however these are not always followed. Although training data (TD) is often assumed to be completely accurate, it almost always contains errors (Stehman & Foody, 2019). These errors can come from issues with the sample design and the collection process itself and can lead to significant inaccuracies in maps created using machine learning algorithms, which can negatively impact their usefulness and interpretation (Elmes et al. 2020). It is very likely that data collection efforts in sub-Saharan Africa (SSA) are biased towards classes of interest, or heavily underestimate rare classes. That is why the main objective of this study is to investigate how different training data sizes and compositions affect the classification results of irrigated agriculture in SSA, and what the trade-offs are between cost, time, and accuracy.

This research focuses on mapping smallholder irrigation in complex landscapes in two provinces of Mozambique and explores the effects of different training data sets on the classified extent of irrigated agriculture in four scenarios: 1) *Size* (same ratio, smaller dataset), 2) *Balance* (equal numbers per class), 3) *Imbalance* (over and under sampling irrigated agriculture), and 4) *Mislabelling* (assigning wrong class labels). To fully understand the specific effects of each type of noise source, this study uses three commonly used algorithms (RF, SVM, and ANN) in cropland mapping. This research aims to inform analysts on the effects of noise in TD on irrigated agriculture classification results.

3. Method

3.1. General method

The same training data (TD) that was used in Weitkamp et al., (2023) is used in this research, including the same satellite data; specifically, the 2x6-month composites will be used due to the acceptable trade-off between computing time and accuracies. Figure 1 shows the overview of the method and how the various scenarios (explained in section 2.5) are run for the three algorithms, random forest (RF), support vector machine (SVM), and artificial neural network (ANN).

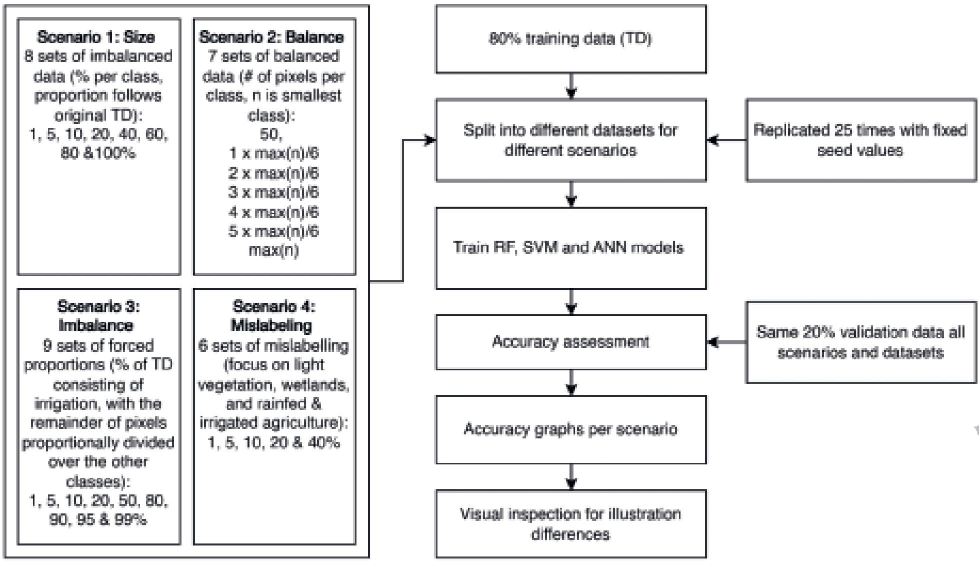


Figure 1 General overview of the methodology

3.2. Study area & RS data

In this study, we compare two provinces, both having two study areas of 40x40 km (Figure 2). The two provinces are different in climate and landscape, allowing for more comparisons between models. These study areas were chosen as they contain diverse landscapes such as dense forests, wetlands, grasslands, mountains, and agriculture.

Figure 2 The four study areas, from top to bottom: Catandica and Manica in Manica province; Chokwe and Xai-Xai in Gaza province.

The following land-cover classes were mapped for this analysis (Table 1):

Table 1 Class descriptions

Cropland irrigated	Croplands under management mainly during the dry season
Cropland rainfed	Croplands under management mainly during the wet season
Dense vegetation	Natural vegetation comprising mainly of trees and dense undergrowth.
Light vegetation	Natural vegetation comprising of mainly low shrubs, grasses, and some trees.
Grassland	Natural vegetation of primarily grass.
Wetland	Natural vegetation that is submerged part of the year (mainly during the rainy season and first part of the dry season).
Water	Water bodies and rivers.
Built-up area	Man-made surfaces and built-up areas, including bare areas such as sand (no vegetation).

Two types of remotely sensed data were used: optical (Sentinel 2) and SAR (Sentinel 1). Two composites of six months were made of the region, using Digital Earth Africa geomads (DEA, 2021) and median deviations (see Roberts, Dunn, and Mueller (2018) and Wellington and Renzullo (2021) for more information on these concepts). Further information on the indices can be found in Weitkamp et al., (2023). The specific scripts can be found on GitHub (https://github.com/TimonWeitkamp/training-data-size-and-composition)

3.3. Training and validation samples per scenario

Table 2 shows the number of polygons (and hectares) collected per class per study area in clustered random strategy, supplemented with some additional irrigated pixels (purposively sampled). During the simulations, we grouped the samples based on their province to increase the total number of training data per simulation.

Table 2 Polygon distribution and size (hectares) per area and class.

	Catandica		**Manica**		**Xai-Xai**		**Chokwe**	
	# polygons	**hectares**	**# polygons**	**hectares**	**# polygons**	**hectares**	**# polygons**	**hectares**
Cropland irrigated	45	16,4	58	10,2	157	38,3	68	166
Cropland rainfed	34	10,9	32	7	19	5,8	48	40,4
Tree cover	9	148	19	104	9	37,2	15	12,5
Shrubland	25	89,5	20	11,3	28	26	104	187
Grassland	0	0	0	0	52	111	0	0
Wetland	0	0	0	0	6	27	12	144
Water	0	0	9	113	9	42,6	5	17,2
Built-up area	10	3,4	10	5,6	10	18,1	10	11,5
Total	123	268,2	148	251,1	290	306	262	578,6

Of this data, the same 20% of the data per class (fixed seed number) was excluded from the training dataset intended for validation; hence each of the results is compared with the same validation data.

This paper investigates four aspects of training data (TD) errors resulting from various sources, focusing on irrigated agriculture. The following scenarios will be explored:

Scenario 1: Size *(same ratio, smaller dataset).* In this scenario, we investigate the relationship between the amount of training data (TD) and the model's accuracy. Specifically, we want to determine whether adding more TD in the same ratio always leads to better results or if similar results can be achieved with fewer data.

To do this, we used eight imbalanced data sets, each with a different proportion of the original training data. The data sets ranged in size from 1% to 100% of the original dataset, with increments of 1, 5, 10, 20, 40, 60, 80, and 100%. The pixel ratio for set 8 of both provinces is shown in Table 3.

Table 3 Number of pixels in set 8 per province (size dataset)

	Gaza	Manica
Class	set 8 (100%)	set 8 (100%)
Built-up area	2 849	1 064
Irrigated agriculture	19 601	3 260
Rainfed agriculture	4 798	2 540
Dense vegetation	6 111	22 185
Grassland	10 157	-
Light vegetation	20 386	9 782
Water	5 504	9 720
Wetland	16 582	-

Scenario 2: Balance *(equal numbers per class)*. In this aspect of the study, we will examine the effect of class balance in the training data on the classification results. Simple random sampling often results in class imbalance, where rare classes are under-represented in the training set due to their smaller area. In particular, we will investigate the impact of using larger, balanced datasets on the classification performance.

We used 7 sets of balanced data to achieve this, where each class has the same number of TD samples. The first set consists of 50 samples, and the remaining sets will be divided into six equal steps based on the class with the lowest abundance (i.e., the smallest class determines the step sizes). The specific sample sizes (in pixels) for each set are shown in Table 4.

Table 4 Number of pixels per set (balanced dataset)

	set 1	set 2	set 3	set 4	set 5	set 6	set 7
Gaza	50	508	966	1424	1882	2340	2798
Manica	50	225	400	575	750	925	1100

Scenario 3: Imbalance *(over and under-sampling irrigated agriculture)*. In this scenario, we aim to investigate the effect of class imbalance caused by purposive sampling on the classification performance. Specifically, we will simulate a scenario where the proportion of samples from the class "irrigated agriculture" is increased at the cost of other classes.

To do this, we created nine sets of data, each with a different proportion of "irrigated agriculture" samples. The proportions will be 1%, 5%, 10%, 20%, 50%, 80%, 90%, 95%, and 99%. To ensure that the same total number of training data is used in each set, the number of samples for the other classes were adjusted accordingly. The remaining training data were divided equally among the other classes, following the method described in Millard and Richardson (2015). The number of samples in each class for each set is summarized in Table 5.

Table 5 Number of pixels per set (imbalanced dataset)

	Class	set 1 (1%)	set 2 (5%)	set 3 (10%)	set 4 (20%)	set 5 (50%)	set 6 (80%)	set 7 (90%)	set 8 (95%)	set 9 (99%)
Gaza	Irrigated agriculture	202	1008	2015	4030	10076	16122	18137	19144	19950
	Rest of the classes (7)	2850	2735	2591	2303	1439	576	288	144	29
	Total	20152	20153	20152	20151	20149	20154	20153	20152	20153
Manica	Irrigated agriculture	54	268	535	1071	2677	4283	4819	5086	5300
	Rest of the classes (5)	1060	1017	964	857	535	214	107	54	11
	Total	5354	5353	5355	5356	5352	5353	5354	5356	5355

Scenario 4: Mislabelling *(assigning wrong class labels)*. In this study, we will examine the effect of mislabelling on the classification accuracy. In smallholder agriculture SSA, class labels can be misassigned due to the heterogeneous nature of the agriculture and the potential for errors or intentional mislabelling.

To simulate this scenario, we created five sets of data, each with a different proportion of mislabelled pixels. The proportions were 1%, 5%, 10%, 20%, and 40%. The focus will be on mislabelling classes that may be considered "border cases" that are likely to be confused rather than randomly selected classes, following Foody et al. (2016). These classes are irrigated agriculture, rainfed agriculture, and light vegetation. The number of misclassified pixels is shown in Table 6.

Table 6 Total number of pixels mislabelled per set for non-focus classes (irrigated and rainfed agriculture and light vegetation).

	set 1 (1%)	set 2 (5%)	set 3 (10%)	set 4 (20%)	set 5 (40%)
Gaza	860	4299	8599	17198	34396
Manica	486	2428	4855	9710	19420

3.4. Algorithm and cross-validation parameter tuning

We have used three different algorithms, namely radial support vector machines (SVM), random forests (RF), and artificial neural networks (ANN). For a description of the algorithms, we refer readers to Abdolrasol et al. (2021); Maxwell et al. (2018); Ramezan et al. (2021); Thanh Noi and Kappas (2017). We want to illustrate that the algorithms may interpret the data differently and lead to different classifications with different accuracies.

We used the *caret* package (Kuhn 2008), which uses the free statistical software tool R and allows for systematically comparing different algorithms and composites in a standardized method. We used *rf*, *svmRadial*, and *nnet* algorithms from *caret* for the random forest, support vector machine, and artificial neural network, respectively.

Cross-validation is a widely used method for evaluating the performance of machine learning algorithms and models. In cross-validation, the data is divided into multiple folds or subsets, typically of equal size. The algorithm is trained on one subset and tested on the other subsets, so each subset is used for testing exactly once. The algorithm's performance is then evaluated based on the average performance across all the folds.

Spatial K-fold cross-validation is a variation of the traditional cross-validation approach that considers the spatial relationships between the samples in the dataset (Meyer et al., 2018a). The spatial k-folds method divides the data into k subsets, with each subset consisting of samples that are spatially close to each other. This is particularly useful in remote sensing, where the spatial relationships between the samples are important in understanding the underlying patterns in the data. In this study, we used spatial k-fold cross-validation.

3.5. Classifications and replications

To ensure the accuracy and reliability of our models, we conducted 25 iterations of all steps for each of the three algorithms using the same seed numbers. By replicating the process, we could account for the variability in accuracies that may depend on the specific training data sets used in each run. This allowed us to evaluate the robustness and generalizability of the models and determine whether they were sensitive to specific training data points and seed numbers or whether they were more robust and generalizable to the study area.

We created various sample sizes and compositions by using random subsampling from the complete sample set, with different seed values. To decrease computation time, we used the caret::train() function and included all variables in the model rather than using forward feature selection of the variables.

Figure 3 displays the range of model parameter values per scenario, training data set, and province based on the overall accuracy. The range of values used by the same algorithms across different seed values and scenarios demonstrates the inherent randomness in the model results, even with the same training data. Some parameter values, such as the *mtry* value of 2 for RF and the *decay* and *size* values for ANN, consistently show higher preference across all datasets. However, *sigma* from SVM exhibits little overlap between the provinces and scenarios. These findings suggest that parameter tuning is highly recommended for SVM and ANN while less necessary for rf, as evident from the lack of clear patterns in the results – similar to what Phalke et al., (2020) also found.

Figure 3 Parameter values and how often a model uses that value per algorithm, per scenario (dataset).

3.6. Accuracy assessment

We calculated the overall accuracy and the user's and producer's accuracies using the same validation dataset for each iteration (Table 7).

Table 7 Sample sizes per class used for accuracy assessment.

	Gaza	Manica
Built-up area	668	252
Irrigated agriculture	4936	823
Rainfed agriculture	1227	607
Dense vegetation	1496	5577
Grassland	2536	-
Light vegetation	5132	2428
Water	1339	2452
Wetland	4165	-

4. Results

The four scenarios (Tables 3, 4, 5, and 6) were designed to demonstrate the impact of training data composition on accuracy, based on possible design and collection errors. Firstly, each scenario's mean overall accuracy per dataset is presented, separated by the province to account for varying climates and agricultural regions. Then, a closer examination of the classification of irrigated agriculture within each scenario is conducted, using the user and producer accuracies.

4.1. The overall accuracy of all scenarios

Figure 4 summarizes the mean overall accuracy of the three classification methods, per scenario and study area. In scenarios 1 (same class ratio, but smaller) and 2 (equal number of pixels per class), high accuracy plateaus of greater than 90% are achieved within the first two sets (5% of total and 508/225 pixels per class, respectively), with similar results across all algorithms. In scenario 3, which involves over and under-sampling of the "irrigated agriculture" class, the accuracy starts high and peaks at sets 3 and 4. However, depending on the algorithm used, it decreases to less than 30-60% in Gaza and 40-50% in Manica when more than three quarters of the dataset contains a single class. Scenario 4, which involves mislabelling, shows high accuracy with the first sets (1-5% mislabelling), particularly with the SVM algorithm remaining stable, while the other two algorithms drop by only five percentage points.

The overall accuracy is mainly affected by the majority classes and hides considerable variation of individual runs. Thus, we will also investigate the classification results of the irrigated agriculture class by using user and producer accuracies.

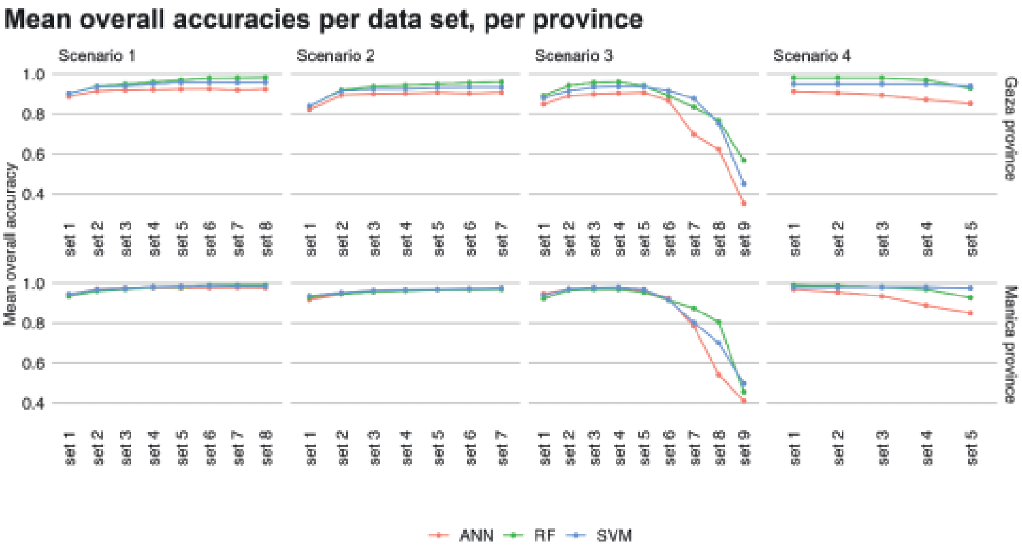


Figure 4 Mean overall accuracies per algorithm, dataset, province, for each scenario.

4.2. Class specific accuracies per scenario

4.2.1. Scenario 1: same ratio, smaller dataset

Figure 5 compares the accuracies of irrigated agriculture between Gaza and Manica using different algorithms, for scenario 1. Generally, larger datasets (set 8) show higher accuracies and less variation in values per dataset than smaller datasets, although there are still differences between the algorithms and study areas.

In Gaza, the more homogeneous study area, the RF algorithm has the lowest accuracy spread and the highest accuracy values, whereas the SVM and ANN have more spread and slightly lower accuracies. The three algorithms are quite stable, with set 2 already leading to comparable results as set 8, which is 10-20 times larger. For each algorithm, the user and producer accuracies are in the same range, indicating that “irrigated agriculture” (user), as well as other classes (producer), are accurately classified. The accuracies are also similar to the mean overall accuracies.

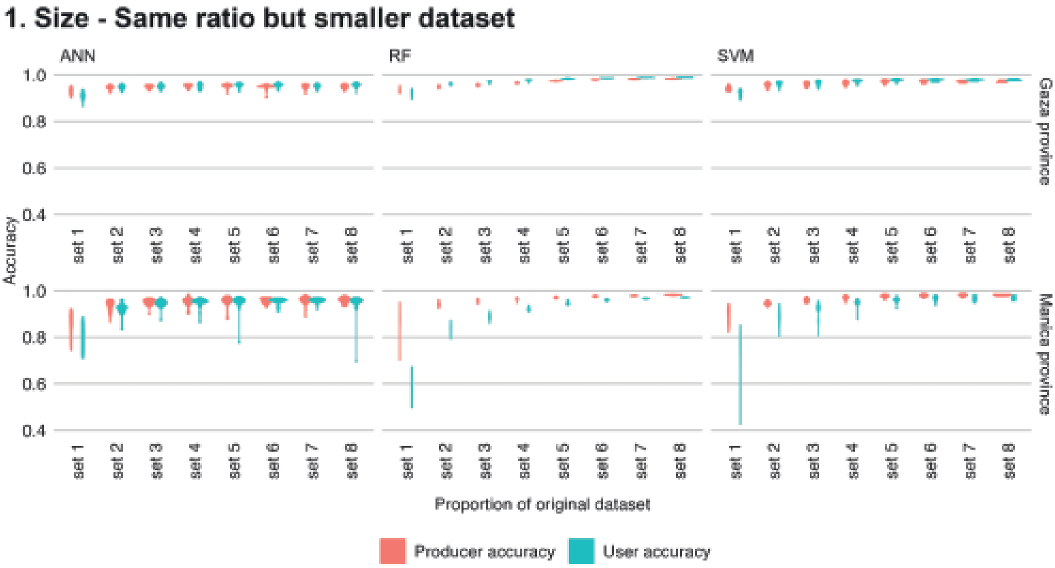


Figure 5 Distribution of user and producer accuracy irrigated agriculture for each algorithm and dataset, per province, for scenario 1: size.

In Manica, which is more heterogeneous, the user and producer accuracies start low and increase until a plateau of ~95% is reached after the fifth set with all algorithms. The most extensive spread in values can be found with ANN in all sets and both accuracies, followed by SVM in the user accuracy, whereas RF shows the least spread in values. Set 1 (the smallest dataset) has the lowest accuracies with the largest spread with all algorithms. However, ANN still has high accuracies (around 80%). It also reaches the plateau the fastest, suggesting that ANN performs well on smaller datasets, albeit with a larger spread, indicating sensitivity to the specific dataset used. The user accuracy is generally lower than the producer accuracy for RF and SVM, at least in the first few sets, indicating that these models were less able to identify “irrigated agriculture” (user), but better at identifying other classes (producer). This could be due to the models not being exposed to enough “irrigated agriculture” samples in the training phase or the models overfitting other classes, meaning they can classify those classes well but not the “irrigated agriculture” class. The producer’s accuracy is in line with the mean overall accuracy, whereas the user’s is less so.

4.2.2. Scenario 2: equal numbers per class

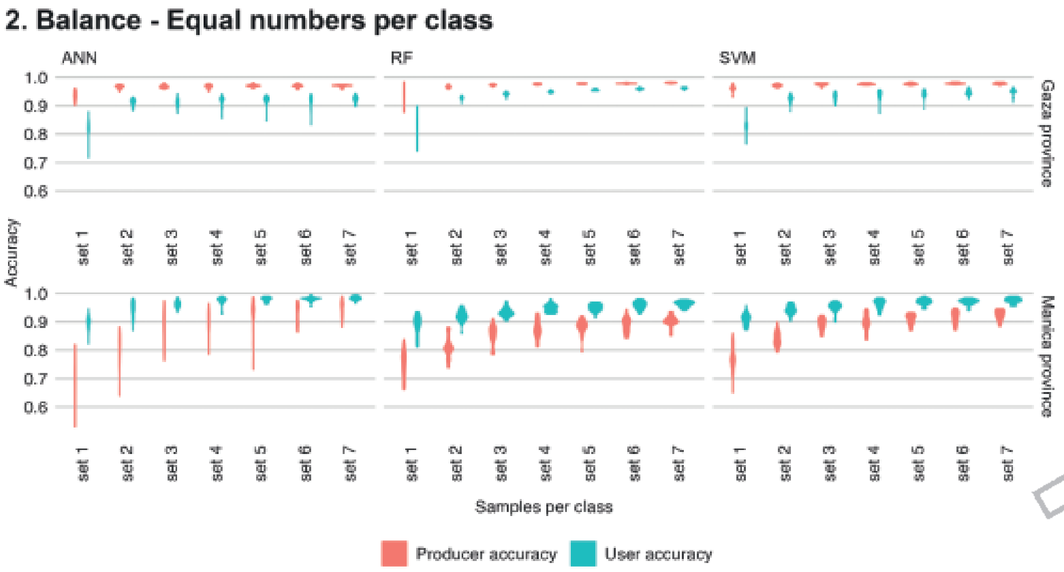


Figure 6 Distribution of user and producer accuracy irrigated agriculture for each algorithm and dataset, per province, for scenario 2: equal numbers per class.

Where the producer accuracy is higher than the user accuracy in Gaza, it is the other way around in Manica (Figure 6). In Gaza, this indicates that the models are not very good at identifying the class of interest (irrigated agriculture) to the user, but they are very good at identifying other classes. In Manica, the models are very good at identifying the class of interest (irrigated agriculture) to the user but not as good at identifying other classes.

In Gaza, most of the producer accuracy values are well above 95%, indicating that almost all the training data samples have been correctly classified. The user accuracies, although high, show more spread in values and remain lower (only the last sets reach 95%), indicating that there is a slight overestimation of irrigated agriculture, especially when the training data contains fewer irrigated agriculture pixels (first few sets). Excluding set 1, RF has the least spread in values, followed by SVM. ANN seems to have the most difficulty in consistent classifications, even as the total number of pixels increases.

In Manica, there is an overall increase in class-specific accuracies with an increasing sample size of irrigated agriculture with all three algorithms (Figure 5). The spread in accuracies in the models with the most irrigated agriculture pixels (set 7) is less than those with fewer samples (set 1), suggesting more robust classifications. However, there is not much difference between the last four sets. ANN shows the largest spread in producer accuracies between the algorithms and starts with the lowest accuracies, while RF and SVM show less spread. Although ANN showed the largest spread, it also achieved the highest accuracies (between 90-95%), followed by RF and SVM with slightly lower accuracies (85-95%). The user accuracies of the three algorithms are more similar and mostly above 90% accuracy, with ANN having the smallest (set 7) and largest (set 1) spread and the highest accuracies, followed by RF and SVM with slightly lower accuracies and larger spreads.

4.2.3. Scenario 3: over and under sampling

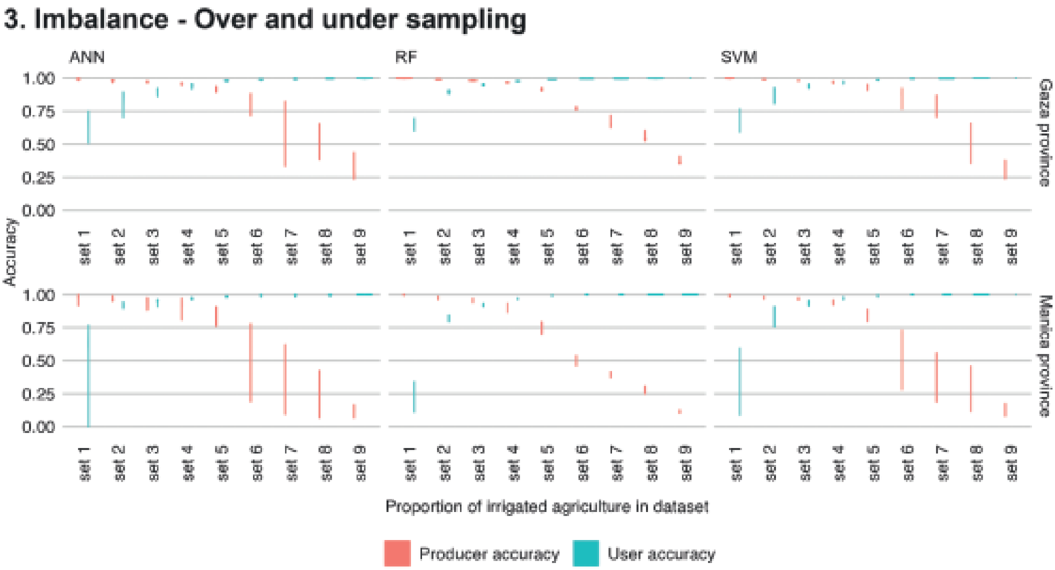


Figure 7 Distribution of user and producer accuracy irrigated agriculture for each algorithm and dataset, per province, for scenario 3: over and under sampling.

Scenario 3, as shown in Figure 7, reveals that the user and producer accuracies are similar around sets 3 and 4, which contain between 10-20% of the “irrigated agriculture” class. This composition is similar to that of the training dataset in Gaza and Manica, which is 22% and 6%, respectively. The producer accuracy remains high until set 4, after which it drops rapidly as the proportion of “irrigated agriculture” increases. The user accuracy is the opposite and increases until set 4, after which it reaches 100% accuracy. This is not surprising, as most of the map will be classified as “irrigated agriculture,” meaning the validation data will be correct for that class. The other classes will be less present in the later sets, resulting in a low producer accuracy.

The RF algorithm shows the least spread in both user and producer accuracy. ANN and SVM have larger spread in producer than user accuracy, and user accuracy spread is small after sets 2/3. Producer accuracy spread starts small but increases with each set for these two algorithms.

4.2.4. Scenario 4: mislabelling irrigated, rainfed, and light vegetation

Figure 8 Distribution of user and producer accuracy of irrigated agriculture for each algorithm and dataset, per province, for scenario 4: mislabelling.

Scenario 4 (Figure 8) reveals that in Gaza, the SVM algorithm's accuracies remain high in all five sets (over 95%), with only a slight decrease in accuracy and minimal spread in values. The RF algorithm follows this trend but dips slightly lower in set 5. ANN has the largest downward trend and the most spread in accuracy values.

In Manica, as seen in Gaza, the SVM algorithm performs best with stable and high (over 95%) accuracies. The RF algorithm starts high but drops to 75-85% accuracy in the last set, with slightly more spread in values. The ANN algorithm has the largest spread and a larger downward trend.

4.3. Visual inspection

In this section, we present a visualization of the level of agreement among models for classifying irrigated agriculture in the Chokwe area. The images depict areas with varying degrees of green and red, with darker shades indicating higher agreement or disagreement among models (referred to as agreement maps), respectively. Specifically, the darkest green shade corresponds to areas where 25 models agreed on the classification of the pixel as irrigated agriculture, while the dark red shade indicates a classification by only one model. In cases where no red or green shades are present, it means that the pixel was classified as a different class other than irrigated agriculture. We have chosen to display only the first and last sets per scenario to illustrate the extremes.

4.3.1. Scenario 1: same ratio, smaller dataset

Figure 9 presents a comparison between the results of set 1 (1% of the data) and set 8 (100% of the data) for scenario 1. Our analysis reveals that set 8 identifies a substantially higher amount of irrigated agriculture compared to set 1, particularly in the southern region of the Limpopo River, which encompasses the Chokwe Irrigation Scheme (CIS). In contrast, the northern bank consists of rainfed agriculture and farmer-led irrigation. Set 1 performs poorly in identifying irrigated agriculture in this region, except for areas near wetlands and a few clusters.

Furthermore, we observed differences in the performance of the algorithms. The artificial neural network (ANN) algorithm identified considerably less irrigated agriculture than the random forest (RF) and support vector machine (SVM) algorithms, which demonstrated similar performances. In particular, ANN severely underestimated the amount of irrigated agriculture in the northern bank, as well as within the CIS.

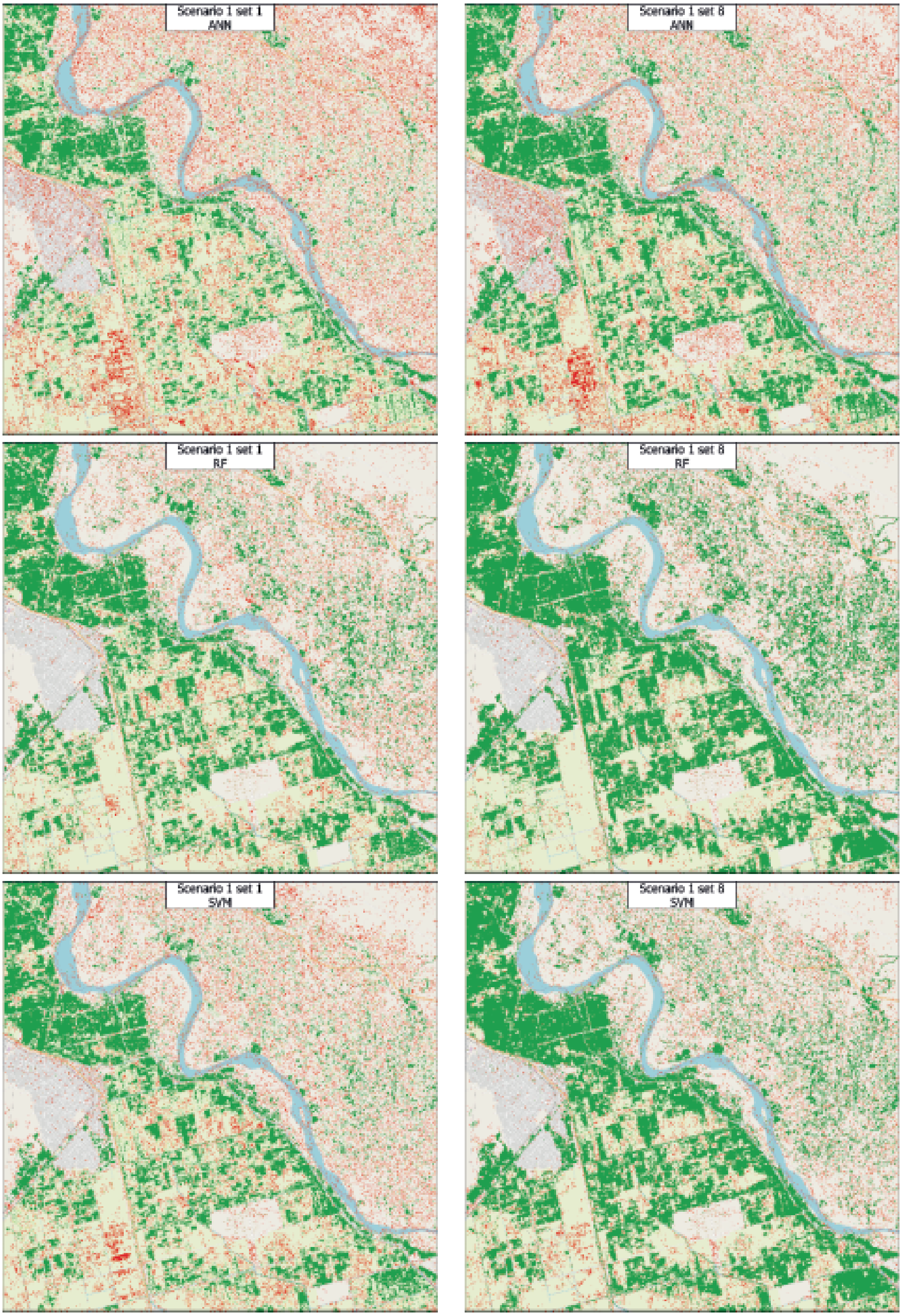


Figure 9 Scenario 1 agreement maps

4.3.2. Scenario 2: equal numbers per class

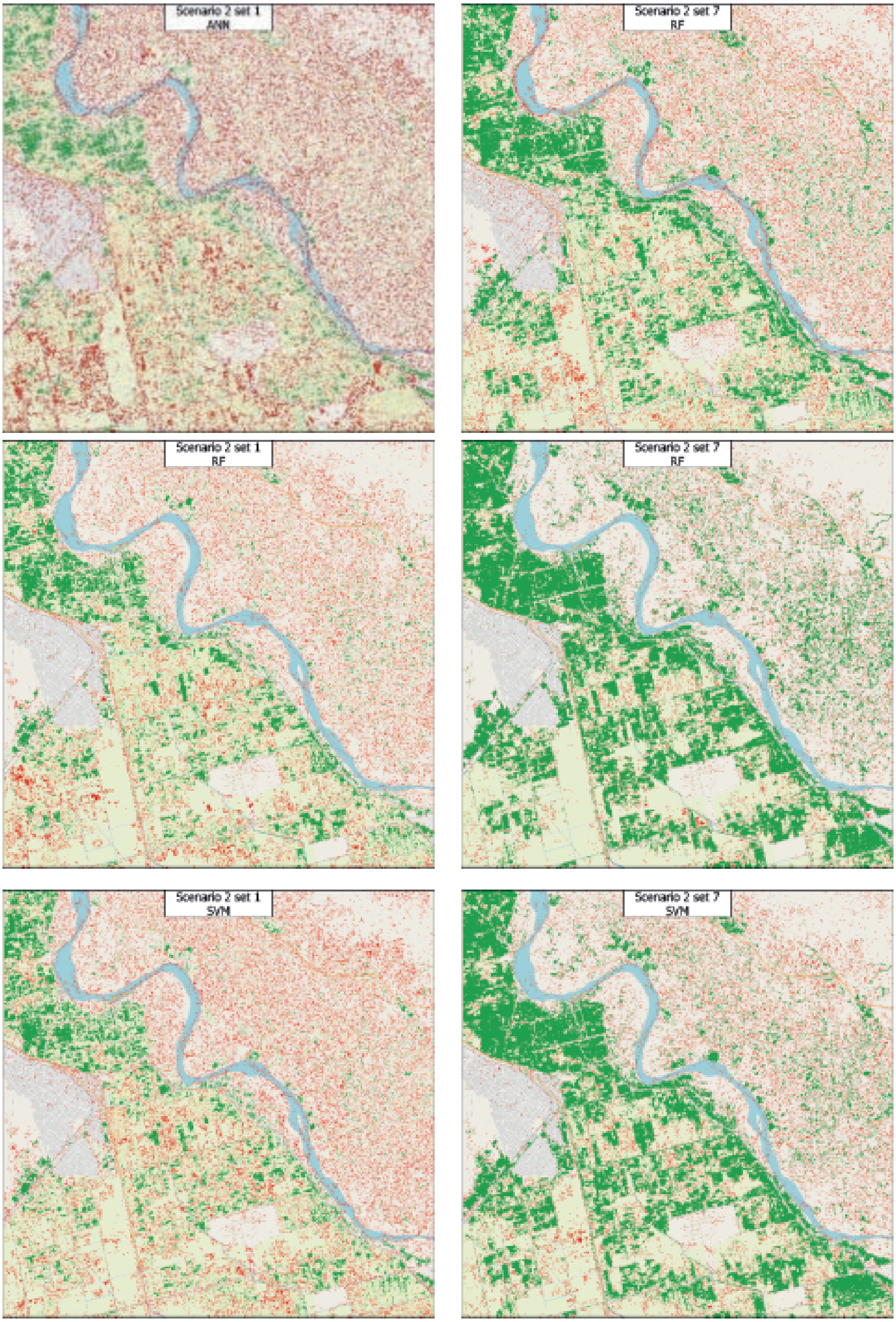


Figure 10 Scenario 2 agreement maps

Scenario 2, where each class has the same number of pixels, shows more significant differences between the smallest and largest datasets than scenario 1 (Figure 10). Set 1 underclassifies the CIS and shows limited irrigation agriculture on the northern bank. The red pixels, where only a few models classify irrigated agriculture, mostly correspond to individual trees or small groups of trees. In contrast, set 7 presents a more balanced map with fewer red areas and larger clusters of irrigated agriculture.

The RF and SVM maps are similar in both sets, while ANN shows fewer areas classified as irrigated agriculture, similar to scenario 1. Additionally, ANN misclassifies the natural vegetation on the Limpopo banks as irrigated agriculture in both sets.

4.3.3. Scenario 3: over and under sampling

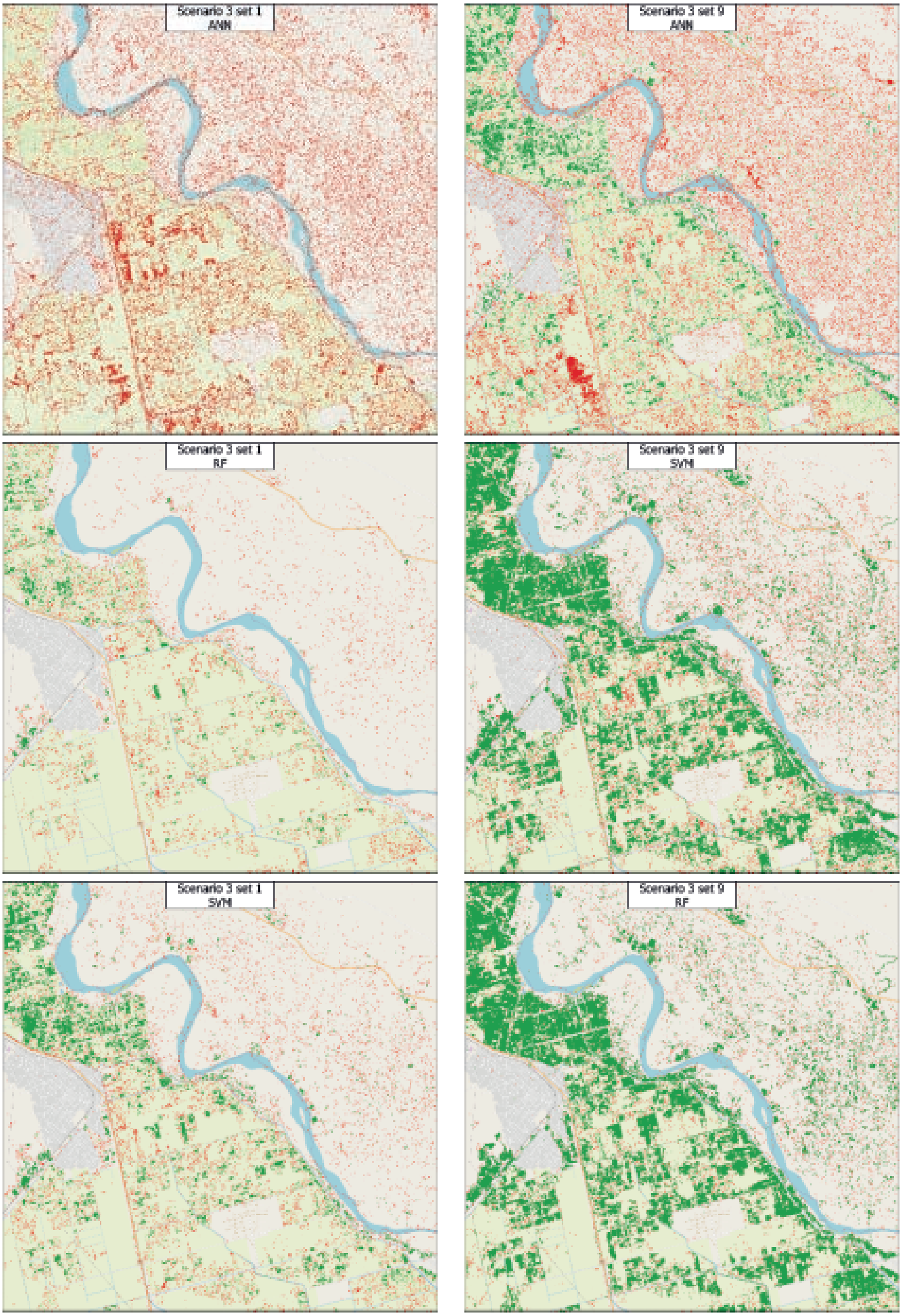


Figure 11 Scenario 3 agreement maps

Scenario 3 highlights the impact of over and under sampling of irrigated agriculture, where set 1 has only 1% of the pixels as irrigated agriculture, while set 9 has 99% (Figure 11). As expected, having very few training data for irrigated agriculture results in limited classification of that class, while having almost only class-specific training data leads to cleaner maps with fewer red areas on the north bank (at least for RF and SVM).

Comparing the algorithms, we observe that ANN classifies more irrigated agriculture in set 1 than the other two algorithms, but there is minimal agreement among the 25 models (no green areas present in set 1). Set 3 using ANN shows more irrigated agriculture, but still less than the other two algorithms. With fewer data (set 1), RF and SVM are less similar, but in set 9, they become more similar again.

4.3.4. Scenario 4: mislabelling irrigated, rainfed, and light vegetation

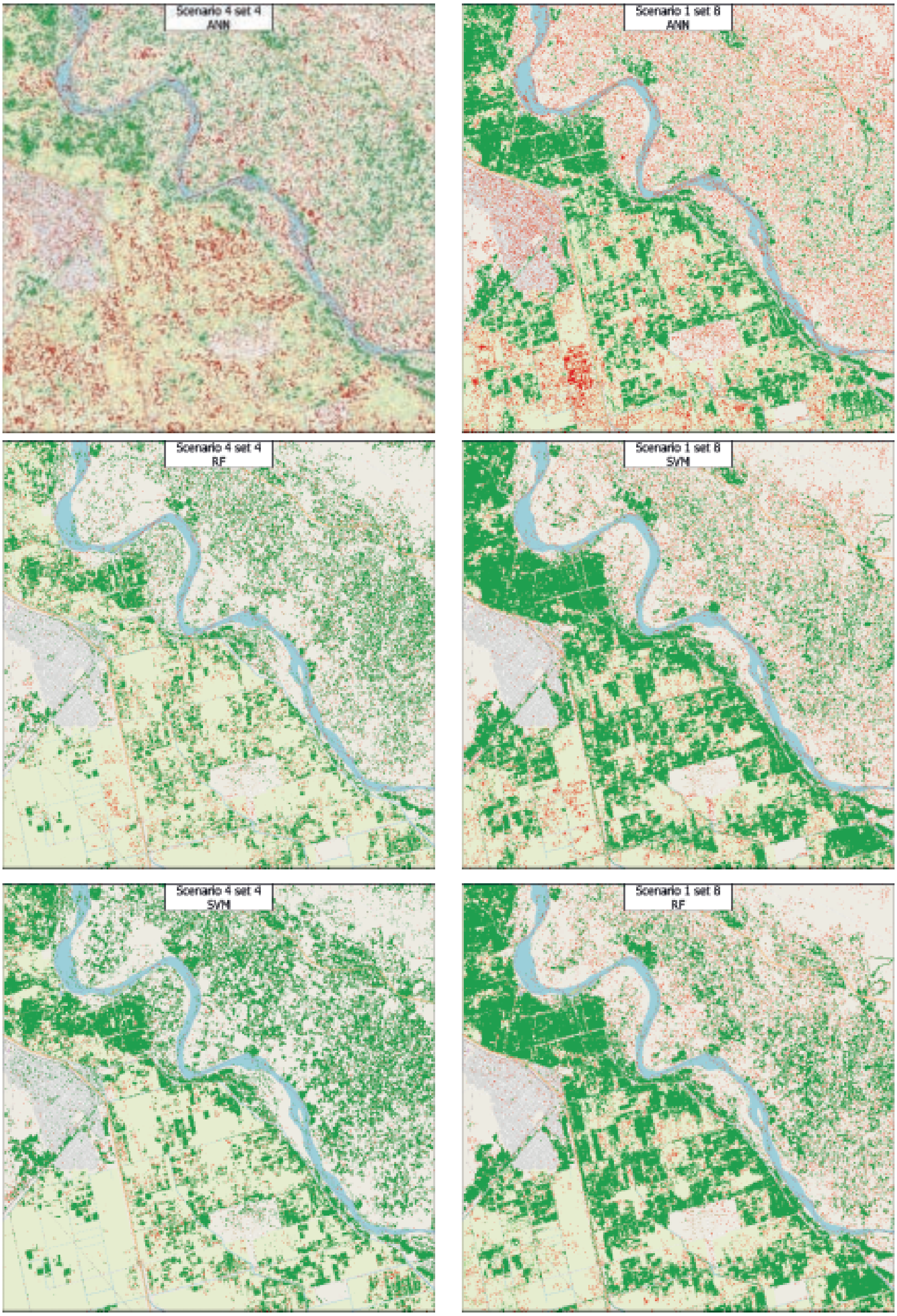


Figure 12 Scenario 4 agreement maps

In Figure 12, we compare Scenario 4 set 4 (with 40% misclassification) with scenario 1 set 8 (with 0% misclassification) for reference. Scenario 4 set 4 shows that more irrigated agriculture is classified on the north bank than the south bank, with all three algorithms, compared to the other scenarios. At the same time, there is less irrigated agriculture in the CIS, with more emphasis on heterogeneous areas for classifying irrigated agriculture.

As in all previous scenarios, the algorithm ANN classifies the least area as irrigated agriculture, followed by RF. The SVM algorithm classifies the most irrigated agriculture.

5. Discussion

The results of this study align with previous research by Ramezan et al., (2021), which found that larger sample sizes lead to improved classifier performance and that increasing the sample set size after a certain point did not substantially improve the classification accuracy. Scenarios 1 and 2 in this research show that larger datasets improve overall classification results, but not by much. This plateauing of overall accuracy is not unexpected, as when classifications reach very high overall accuracy, there is little potential for further increases. Our study is also in line with what Ramezan et al., (2021) found, in that user and producer accuracies continued to increase with larger sample sizes, indicating that larger sample sizes are still preferable to smaller sizes, even with similar overall accuracy results.

A large spread in accuracy means that the specific results depend more on the dataset that is used for that classification than other factors. For example, the SVM algorithm in Manica in Scenario 1 resulted in a user accuracy of just above 40%, but also 85%. By chance, any of the two could have become the final classification; if it was the 85% classification, one would think enough data is collected for the study, whereas the other sets show that higher accuracies are possible, with less spread in values. The lower spread in values also indicates a more stable model which can generalize more. It also means that the specific dataset used for the classification is less important, as similar results can be expected from any random subset, also seen in Section 3.3.

Scenario 1, where eight datasets ranged in size from 1% to 100% of the original dataset were used, shows that larger training datasets lead to the higher user and producer accuracy with less spread in values (Figure 5). The size of set 5 in Manica falls between sets 3 and 4 of Gaza (40% vs. 10-20%, respectively), which are also the sets after which the accuracies plateau in Gaza. This corresponds to ~1300 pixels of irrigated agriculture for Manica and ~1900-3900 for Gaza. This reinforces the statement that larger training data sets are preferable over smaller sets but that there is an optimum after which accuracies only marginally increase at the cost of more computing time and, effectively, more resources 'lost' collecting that data in the first place. To find out if enough data is collected for a classification of irrigated area, researchers and practitioners can use this *subsetting* method to evaluate if different iterations yield the same, stable results, or if additional resources should be put towards more field data collection.

Scenario 2 also examines the impact of data size on classification performance, but with equal numbers of samples per class, spread over seven sets. Similar to scenario 1, larger datasets generally result in higher user and producer accuracies (Figure 6). However, this scenario highlights differences in the performance of the classifiers in the two study areas. In scenario 1, the results of both study areas followed similar patterns but with different accuracy values. In this scenario, however, the user and producer accuracy trends are reversed, depending on the study area. In Gaza, the user's accuracy is consistently lower than the producer's, whereas in Manica, the user's accuracy is consistently higher than the producer's. Manica also shows a larger spread in values for both user and producer accuracy.

This trend reversal suggests that the models in Gaza are better able to classify the non-irrigated agriculture classes than the irrigated agriculture class, indicating a more generalized model. Conversely, the Manica models can better classify the irrigated agriculture class than the non-irrigated agriculture classes, indicating a less generalized model. As all classes have the same number of pixels per dataset within the same study area, the complexity of the landscape likely plays a role in this difference. The two provinces generally have different landscapes (flat vs mountain), climate (little vs lot rainfall) and consequently, different agricultural practices, with different field sizes (larger vs small) and shapes (regular vs irregular). It is worth noting that, even though Gaza has twice the number of pixels as Manica, sets 1 are the same size in both cases, and 3 of Gaza and 7 of Manica are similar in size. However, even for these sets with similar sizes, Gaza has higher producer accuracies, and Manica has higher user accuracies.

Scenario 3, where irrigated agriculture is vastly over and under-sampled in nine sets ranging from 1% to 99%, shows a peak in overall accuracy around sets 3 and 4 (Figure 7, 10% and 20% irrigated agriculture in the dataset). These two sets reflect the 'true' composition of the dataset, which was found in the field. When irrigated agriculture is underrepresented (sets 1 and 2, 1% and 5%), the overall accuracy is not much lower. This is because the other majority classes have a greater impact on the overall accuracy. As more irrigated agriculture is present in the training datasets (sets 5 to 9, 50-99%), the other classes decrease in size, and irrigated agriculture becomes the majority class. The high user accuracy indicates that any irrigated agriculture in the validation set is correctly classified (not surprising as all pixels are classified as such). However, the reverse is that the producer accuracy is extremely low (many of the pixels are wrongly classified as irrigated agriculture instead of a different class).

Scenario 4, where similar classes are mislabelled on purpose in 5 sets from 1% to 40% mislabelling, shows a decrease in overall accuracy (Figure 4) for ANN and only a minor decrease in the last set for RF. SVM does not seem to be affected, possibly because the support vectors used for distinguishing the different classes do not change much between the sets, indicating that SVM is less sensitive to data set compositions.

The user and producer accuracies (Figure 8) also show that SVM can handle this imbalance, perhaps because it uses the same support vectors to distinguish the different classes in all the sets. Adding more data will not help the algorithm, as that data is not near the separation planes between classes. RF is similarly stable, except for the last set, which also shows a larger spread in accuracy values. The user accuracy is also higher than the producer's, which comes from slowly oversampling irrigated agriculture (among other classes). The ANN has many difficulties with the changing compositions, as seen from the large spread in values and decreased accuracies. Overall, RF and SVM seem to handle this mislabelled data well.

The results of the study demonstrate the importance of the dataset and algorithm selection in accurately classifying irrigated agriculture in remote sensing data. Visual inspection reveals that different areas are classified as irrigated agriculture depending on the dataset and algorithm used. In some cases, the models prioritize farmer-led irrigated areas over more conventional large-scale irrigated areas, but the latter is generally classified more accurately. The amount of data used and the balance between classes also have a significant impact on the accuracy of classification, with too few data or imbalanced data resulting in underestimation of the extent of farmer-led irrigation, and too much noise resulting in overestimation. The RF and SVM algorithms are found to be more robust with noisy data than the ANN algorithm. Although the maps do not distinguish between farmer-led irrigation and large-scale irrigation, our knowledge of the area enables us to interpret the maps in terms of these different types of irrigation.

Generally, there are many oversampling and undersampling strategies which have not been tested. The focus of this study was not to find the best method to deal with imbalanced data, but to illustrate what imbalanced data does with the final results.

Overall, ANN showed high results but with a large spread in all scenarios and study areas. The RF and SVM showed results similar to each other, depending on the scenario's dataset and study area, which algorithm resulted in higher accuracies with lower spreads. Both are recommended for mapping irrigated agriculture. The large spread in ANN shows that it may be suitable for detecting irrigated agriculture, but only in certain circumstances - when there is much data (scenario 1 final sets), and the landscape is more homogeneous (Gaza, all scenarios). Nevertheless, the random chance of high or low accuracies is higher with ANN than with RF and SVM (i.e., larger spread), indicating that the specific dataset used in modelling is more important for ANN than the other two algorithms.

According to Maxwell et al., (2018), the training sample size and quality can have a greater impact on classification accuracy than the choice of algorithm. As a result, differences in accuracy between datasets within the same algorithm should be more pronounced than those between different algorithms. This is supported by scenarios 1, 2, and 3, where the algorithms show similar trends and values but exhibit greater variability within datasets. Scenario 3 demonstrates that user and producer accuracies may cross over, but the differences between datasets are still more significant than those between algorithms. However, scenario 4 is less conclusive since there is little variation in the high accuracies of the RF and SVM algorithms across all sets, with some variation in Manica. At the same time, ANN shows dissimilar trends and greater differences between sets compared to the other two algorithms.

6. Conclusion

The results of this study indicate that larger sample sizes generally lead to the higher user and producer accuracies. However, there is an optimum after which accuracies only marginally increase at the cost of more computing time and collection effort (Scenario 1). We also show that the models trained on Gaza were better at classification of all classes (i.e., a more generalized model) than in Manica (Scenario 2). In other words, the more homogeneous landscape of Gaza lead to models that could generally classify all classes, whereas models of the more heterogeneous Manica were overfitting towards irrigated agriculture, even though all classes had the same number of pixels in the training data sets. Scenarios 3 and 4 show that the field data collected should reflect the actual landscape composition and that class labels can bias towards heterogeneous areas (i.e., no oversampling of irrigated agriculture or mislabelling), and that random forest and support vector machine are more suitable for classifying irrigated agriculture than the artificial neural network, as they are less sensitive to the specific dataset.

This study provides valuable insights for practitioners and researchers mapping irrigated agriculture in sub-Saharan Africa by means of remote sensing techniques. It highlights the importance of carefully considering sample size and composition when collecting and using data. African smallholder agriculture is complex, with variability in field shape, cropping systems, and timing of agronomic activities. Based on this study, to accurately predict such smallholder irrigated agriculture, we recommend to:

- Ensure that training data represents the area being classified and includes sufficient samples to achieve high accuracy. This can be done best using a random sampling design. Although perfect data is desirable, models (RF and SVM) can tolerate some noise.

- Evaluate multiple algorithms when classifying data, as different algorithms may perform better or worse depending on the specific characteristics of the data being classified.

- Interpret classification results carefully, as accuracies alone may not correctly represent the classification performance. Visual inspection and further interpretation are needed to understand the results and potential limitations of the classification fully.

- Perform multiple simulations with different subsets of the data to estimate if the training data yields robust results (i.e., minimal variation in accuracies between sets), which can indicate that sufficient data has been collected.

Chapter 5

The generalisation of machine learning models for predicting irrigated agriculture in heterogeneous landscapes: an examination of model transferability in Mozambique

This chapter is published as:
T. Weitkamp and P. Karimi (2023)

The generalisation of machine learning models for predicting irrigated agriculture in heterogeneous landscapes: an examination of model transferability in Mozambique

International Journal of Applied Earth Observation and Geoinformation (*Under review*)

1. Abstract

Mapping smallholder irrigation in sub-Saharan Africa using remote sensing is challenging due to the complex nature of small-sized, irregularly shaped fields and the diversity in agronomic activities. A robust and comprehensive set of training data is, thus, a fundamental prerequisite for producing reliable and accurate maps. Collecting ground data in new areas is expensive and time-consuming, making it crucial to determine the extent to which models can be transferred between areas to save time and effort while improving prediction accuracy. This study explores the use of the "Area of Applicability" (AOA) concept for finetuning irrigated agriculture hotspot maps, particularly for transferred models. Spatial cross-validation and random forest algorithms are used to assess model performance and robustness. The research addresses the spatial-temporal transferability of machine learning models in remote sensing by assessing their performance across different locations. Various scenarios are considered, including simple model transfers between areas with different climates, training on a more varied dataset, and transferring to an unsurveyed area. The findings show that model transfer in complex landscapes remains challenging and that the AOA does not exclude as many areas as expected when comparing different scenario results to baseline results.

2. Introduction

Mapping smallholder irrigation in sub-Saharan Africa (SSA) presents challenges due to the nature of small, irregularly shaped fields with in-class variance, including inter- and mix-cropping systems and variability in the timing of agronomic activities such as planting, harvesting, and irrigation (Bey et al., 2020c; Nabil et al., 2020; Rufin et al., 2022; Weitkamp & Karimi, 2023). Collecting data in new areas of smallholder farming can be expensive, labour-intensive, and time-consuming, especially if a randomised sampling strategy is used and the road network density is low, which is often the case in rural Africa. As a result, it is important and relevant to explore new techniques that may help in transferring prediction models between areas to save time, costs, and effort while producing more accurate predictions over larger areas. Creating more impactful maps relies on the capacity of the models to generalise effectively and to omit areas with unreliable predictions.

The complex nature of smallholder agriculture and training models makes it difficult for trained remote sensing-based models to generalise well enough to accurately predict in other areas. Model generalisation refers to the ability of a trained model to perform well on new, unseen data that comes from areas that were not part of the training dataset. In other words, how well models perform when transferred to other areas. Collecting data through opportunistic field-sampling (which involves densely clustered training data) might result in reduced model generalizability and transferability (Ludwig et al., 2022). This translates to the model becoming overly specialised and having limited accuracy when making predictions for unfamiliar regions.

Poor generalizability can also result from using overly complex models that are too focused (i.e., overfit) on the specific environments they were trained on (Barbiero et al., 2020). The more predictors a model has, the more complex and specialised the relationships it learns will be. This makes the model more likely to overfit, meaning it performs well on the training data but poorly on new, unseen data. It also increases the chances of the model not accounting for predictor combinations in new locations (Ludwig et al., 2023). A proven strategy to prevent overfitting and enhance model generalisation is simplifying the model by removing predictors that don't significantly affect the outcome, such as through forward feature selection (Ludwig et al., 2022).

After training and transferring a model to new areas, its performance is often communicated through accuracy metrics such as the error matrices and its associated overall, user and producer accuracy. However, simply communicating the performance is not sufficient (Meyer & Pebesma, 2021). Additionally, validating a model using data collected within the same geographic extent to which it was trained does not offer a valid assessment of model generalisation performance to new geographic extents (Maxwell et al., 2021). This is because field samples used for model training are often not evenly distributed; they may be heavily clustered due to opportunistic field-sampling campaigns or biased towards "known" areas, and there may be areas without sufficient training data (Yates et al., 2018).

Another consideration, besides simplifying the model, is to verify the geographical extent to which a model can generalise and provide meaningful predictions for new instances of the problem. A model that is only able to deal with the data it was trained (i.e., overfitted) on is generally considered useless, regardless of its performance on the training dataset (Barbiero et al., 2020). When a model is transferred to a new geographical area, it assumes that the statistical relationships learned from the training data and predictor variables still hold. However, the new environment likely differs greatly from what was observed in the training data. This means that classes that are not present in the training data are being classified as another (wrong) class, and must be considered problematic (Meyer & Pebesma, 2021). In other words, the predictions of these locations are too uncertain to be considered for further action. Improved analysis and communication of uncertainties of spatial predictions is therefore needed (Meyer & Pebesma, 2021).

To effectively address uncertainties, it is essential to define the scope within which a prediction model can be confidently used. This can be accomplished through the utilisation of the "Area of Applicability" (AOA) concept, as proposed by Meyer & Pebesma (2021). Understanding the AOA becomes particularly important when generating predictions for diverse regions based on limited field data or when extrapolating across study areas where the model's suitability for the new context is ambiguous (Meyer & Pebesma, 2021). Consequently, instances where there is a high spatial model error and/or a restricted area of applicability serve as indicators of inadequate model generalization (Ludwig et al., 2023).

Model transfer, which is the ability of a model to generalise to new areas, is an active area of research, particularly in data-scarce regions. Traditional machine learning methods such as random forest and support vector machine, or a combination of the two, are often used in model transfer in the agricultural context (Gao et al., 2022; Li et al., 2020; Mills, 2008; Orynbaikyzy et al., 2022; Phalke & Özdoğan, 2018; Wang et al., 2019). These methods have been shown to be effective in many applications and have the advantage of being computationally efficient and easy to implement. To our knowledge, there are no studies that focus on model transfer and irrigated agriculture, let alone for SSA.

Deep learning (DL) models have also been shown to generalise well to new data. However, due to their high level of data abstraction of various deep learning methods (Nowakowski et al., 2021; Pires de Lima & Marfurt, 2020; Tong et al., 2020; Xu et al., 2020), we have chosen to use traditional machine learning methods because of their ease of use and interpretability. DL models are often complex and difficult to understand, making it challenging to identify errors or biases in the predictions. Additionally, deep learning methods require large amounts of training data, which can be challenging to obtain in data-scarce regions. While DL models may outperform traditional machine learning methods in some cases (e.g. Du et al., 2022; Zhang et al., 2020), we believe that traditional methods are more suitable for our study given our data and research questions.

African smallholder farmers are active, the new environment likely differs greatly from what was observed in the training data, resulting in the model underfitting the training data. To our knowledge, the AOA has not been applied to irrigation mapping before, or has it been applied in SSA. Furthermore, the AOA has not been integrated into previous model transfer studies due to its recent emergence. Given its novelty, we aim to pioneer its application in the field and specifically focus on its implementation in SSA.

When transferring models, particularly those involving numerous predictor variables, there is a concern of overfitting. This arises when the model encounters pixel values that correspond to a class it hasn't encountered before. This scenario is likely to occur in SSA, where smallholder farmers employ diverse irrigation methods across different regions. To address this issue when transferring models to focus on SSA's irrigation patterns, it becomes essential to utilise the AOA to exclude regions that the model hasn't been exposed to. Furthermore, simplifying models aids in creating more broadly applicable models that are suitable for transferability, prompting the incorporation of feature selection.

In this study, we investigate the extent to which transferring remote sensing models to new geographic areas with distinct land use, climate, and agricultural practices can be done, using feature selection and the AOA. Since creating new training data and models for these areas is often laborious and time-consuming, exploring the geographic generalisation of region-specific models can contribute considerably to early mapping exercises and research, particularly given the growing availability of remote sensing data.

3. Method

3.1. Study area

The research was carried out across four distinct regions in Mozambique: Chokwe and Xai-Xai in the Gaza province, and Manica and Catandica in the Manica province (Figure 2). These specific regions were selected due to their diverse agroecological features and the presence of both small-scale and large-scale irrigated agricultural systems. The case studies encompassed an approximate area of 40x40 km.

Within the Manica province, the landscape is characterised by mountains and is supported by small streams that serve as sources for irrigation. Farmers redirect the water into earthen canals known as "furrows" and employ techniques like sprinkler irrigation, small pumps, and bucket irrigation. The size of these systems is relatively smaller compared to those in the Gaza province, and they adapt based on water availability. During the dry season, horticultural crops receive irrigation, while maise is cultivated during the rainy season.

In the Gaza province, irrigation systems, both large and small in scale, are situated along the banks of the Limpopo River. Flood-based irrigation practices are common, and pumps are utilised to access elevated areas. In proximity to Xai-Xai, irrigated zones with shallow groundwater tables necessitate drainage post the rainy season. The agricultural focus during the irrigation season includes horticulture and maise, whereas rice and maise take precedence during the rainy season.

Figure 2 The four study areas in Mozambique, from top to bottom: Catandica, Manica (Manica province), Chokwe, and Xai-Xai (Gaza province).

3.2. Model training

Spatial feature selection ffs

Meyer et al., (2018) propose using forward feature selection along with a study-specific cross-validation strategy to choose predictor variables suitable for spatial prediction. The adoption of 'spatial feature selection' or *ffs*, which incorporates spatial cross-validation during selection, prioritises variables that improve the predictive performance in new geographical regions (Ludwig et al., 2023). Essentially, this means grouping pixels from the same polygon within the same training fold, thereby creating training and validation data clusters based on polygons.

AOA

While models are often assumed to be applicable across the entirety of the area of interest, there are instances where the suitability of the model in new environments can be unsuitable. Therefore, it becomes necessary to quantify the dissimilarity between predictors at new locations and those present in the training data. This Dissimilarity Index (DI) can be computed by determining the minimum distance to training data in the weighted predictor space, which is then normalised by the average distances among the training data (Meyer & Pebesma, 2021).

To pinpoint regions that significantly differ from the training data and thus cannot be reliably used for predictions, the derivation of an AOA necessitates setting a DI threshold. This threshold is determined by the *aoa* function in the CARET package (Kuhn, 2019), version 0.8.1 as used, which extracts the threshold from the training data by identifying the maximum dissimilarity among the training data after the removal of outliers through cross-validation. For specific details, we refer to (Meyer & Pebesma, 2021).

Classification

To investigate the potential transferability of machine learning models for predicting irrigated agriculture, various models will be trained and explored in two distinct scenarios:

- Scenario 1: Simple model transfers between two areas with different climates: assessing the extent to which a model trained on one area can be applied to another very different area, without modifications. This scenario includes a baseline model, in which a model is trained and applied to the same area.

- Scenario 2: Training on a more varied dataset and transfer to 'unseen' area: assessing the extent to which a model trained on three varying areas can be applied to a fourth, assuming it contains at least some of the relationships learned by the model.

For each scenario, the study will compare a transferred model (e.g. from Chokwe to Manica) with a locally-trained model (e.g. from Chokwe to Chokwe) and evaluate the transferability of the model, either in time or space. This will be done for two areas, Chokwe and Manica. At the same time, we will run each model using the *caret::train* and *caret::ffs* functions, the first uses all variables, the second uses feature selection to exclude variables. Each model is replicated 4 times with different seed values.

Additionally, each model will be repeated with five spatial cross-validations to enhance the model's robustness. This allows for a more accurate assessment of the model's performance and reduces the risk of overfitting due to spatial autocorrelation. Random forest (RF) will be used in each scenario.

Table 1 provides a detailed overview of each scenario and its respective objectives. In total, each model will be run 4 times for 'all variables (*train)*' and 'feature selection (*ffs*)' per scenario (six in total), resulting in 6 × 4 × 2 = 48 models (excluding the internal cross-validation etc). To effectively visualise this, we will make use of hotspot maps for irrigated agriculture, as well as the AOA. We will combine these to make final classification maps, in which the hotspot maps can also show negative values, indicating classified irrigation that is highly unlikely to be so. In other words, the AOA hotspot values (range 0 – 4) are subtracted from the irrigation hotspot values (range 0 – 4), potentially resulting in negative values if only one model classified irrigation, but four AOA models found the pixel outside of the acceptable range (1 – 4 = -3).

Table 1 Description of the different transferability scenarios explored in this study.

	Model	Train location	Test location	All variables	Feature selection
Scenario 1	1	Chokwe	Chokwe	A	B
	2	Manica (baseline)	Chokwe	A	B
	3	Manica	Manica	A	B
	4	Chokwe (baseline)	Manica	A	B
Scenario 2	5	Chokwe + Catandica + Xai-Xai	Manica	A	B
	6	Manica + Catandica + Xai-Xai	Chokwe	A	B

Within a test location, the different models are trained with the same training data sample size (determined by the baseline model's training data). The multiple location transfer models are trained on three different locations, each contributing one third of the total training data sample size. We have done this to take out the influence of training data size on model transferability, so that we could focus on the model training and AOA aspects.

All code and data can be found on the GitHub repository: https://github.com/TimonWeitkamp/model-transferability

Satellite data

Satellite data was collected within the Digital Earth Africa (DEA) 'sandbox', which provides access to Open Data Cube products in a Jupyter Notebook environment[2]. Sentinel- 2 geomedian products (a robust high-dimensional statistic like the normal median that maintains relationships between spectral bands, DEA, 2021; Roberts et al., 2018) were generated at 10-meter resolution for a 6-monthly composite, covering April – September 2020, corresponding to the dry season. Images with more than 30% cloud cover were filtered out.

From Sentinel-2 we calculated the Normalised Difference Vegetation Index (NDVI), Bare Soil Index (BSI), and Normalised Difference Water Index (NDWI), using the DEA indices package for the Sentinel-2 composites (Wellington & Renzullo, 2021). Three second-order statistics (Median Absolute Deviations (MADs)) were also calculated, which are change statistics based on the geomedian: the Euclidean (EMAD, based on Euclidean distance), Spectral (SMAD, based on cosine distance), and Bray-Curtis (BCMAD, based on Bray-Curtis dissimilarity) MADs (Roberts et al., 2018). Wellington & Renzullo (2021) used these change statistics, as well as a few of the indices in their classification of irrigated areas, with success. We used these indices and statistics to cover the different phases of croplands, from browning (BSI) to greening (NDVI), the NDWI for water detection, while the MADs are suitable for change detection, particularly for irrigation (Wellington & Renzullo, 2021).

All bands and indices were merged into one dataset, forming a 14-variable dataset (Table 2).

2 Sandbox link and explanation can be found on https://docs.digitalearthafrica.org/en/latest/sandbox/index.html

Table 2 Overview of variables per composite time-length

Group	Variable	Equation
Sentinel-2	Blue	
	Green	
	Red	
	Near Infrared (NIR)	
	Red-edge 1 (RE1)	
	Red-edge 2 (RE2)	
	Shortwave Infrared 1 (SWIR1)	
	Shortwave Infrared 2 (SWIR2)	
Indices S2	Normalised Difference Vegetation Index (NDVI)	(NIR - Red)/(NIR + Red)
	Normalised Difference Water Index (NDWI)	(NIR – SWIR1)/(NIR + SWIR1)
	Bare Soil Index (BSI)	((Red + SWIR1) - (NIR + Blue))/ ((Red + SWIR1) + (NIR + Blue))
	Chlorophyll index (CI)	(NIR / Red Edge 1) - 1
Temporal variation	3 MADS S2	See Roberts et al. (2018) and Wellington and Renzullo (2021) for more details on equations
Sentinel-1	VV	
	VH	
Indices S1	RVI	x VH / (VV + VH)

3.3. Accuracy assessment

We assessed the performance of the models by utilising metrics such as overall map accuracy, user accuracy, and producer accuracy. To evaluate these models, we employed a cross-validation approach. This involved dividing the training data into folds, and subsequently, the model with the highest performance was compared against 20% of the validation data. The outcomes for each model were then recorded in a confusion matrix. It is worth noting that overall accuracy can exhibit bias towards the most prevalent class within the training data. Therefore, it is also valuable to take into consideration both user accuracy and producer accuracy, as these metrics provide more detailed insights into the model's performance for specific classes.

The confusion matrices can be found on GitHub: https://github.com/TimonWeitkamp/model-transferability/tree/main/confusion_matrices

4. Results

In this study, we investigate the extent to which transferring remote sensing models to new geographic areas with distinct land use, climate, and agricultural practices can be done, using feature selection and the AOA. We will present the overall accuracy, and the user and producer accuracies for irrigated agriculture of the baseline study and the two transfers, for both the feature selection model and the model that uses all 14 variables. A visual inspection then follows for two case studies: Chokwe and Manica, where we take a closer look at the extent of irrigated agriculture.

4.1. Accuracies

Figure 3 shows the accuracies of each scenario and model; the producer and user accuracy are for irrigated agriculture. The accuracies are split over whether the variables are selected using *caret::ffs* or *caret::train*, and presented per validation location (Chokwe and Manica). The scenarios compared are the baseline scenario, multiple locations transfer, or single location transfer.

It is evident that the baseline scenarios yield the highest accuracies on all three metrics, for both locations (>80%). Both the multiple and single location transfer accuracies are low (below 50% and 30%, respectively) for Chokwe indicating unsuitable transfer. This means that the other three locations from which the training data comes have different landcover types or temporal patterns than Chokwe. A similar situation occurs for Manica, although here the overall accuracy is high in the transferred models (>75%). The user and producer accuracies of the multiple and single location transfer models remain low (also below 50% and 30%, respectively). The high overall accuracy indicates that other classes are likely better classified than irrigated agriculture, perhaps because those are prevalent compared to irrigated agriculture.

As all training data for the models was the same size per validation location, Figure 3 also shows that including multiple locations leads to (slightly) higher accuracies than when only a single location is used when using the *train* function. However, when using the *ffs* function, the specific dataset used (i.e., seed value) could lead to individual results of single location transfers to be higher than that of multiple location transfer models. Generally, the multiple locations have higher accuracies.

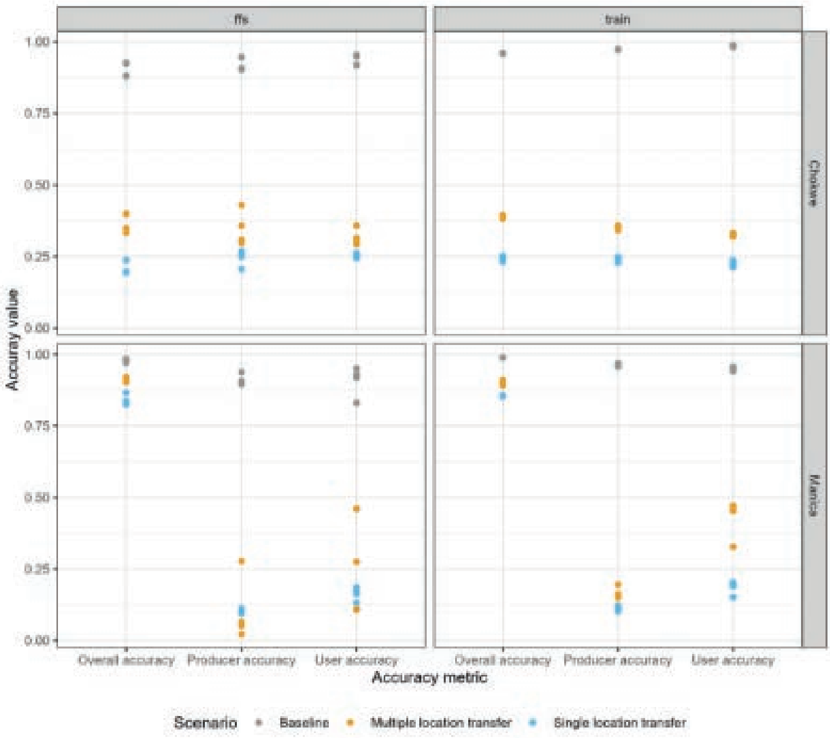


Figure 3 Accuracies per scenario and variable selection method. The multiple points of the same scenario show the accuracies per seed value (i.e., replication).

4.2. Visual inspection Chokwe

4.2.1. AOA maps

We start by showing the hotspot maps of the AOA, where a value of 4 indicates that all four AOA models predicted that those areas fall outside of the acceptable range (Figure 4). The top row of the figure shows the AOA maps of *ffs* and the bottom row shows *train* results.

If we compare six outputs, we can see two noticeable things. The first is that the three *train* AOA maps are all more similar to each other than the three of *ffs*. The *train* maps almost exclusively show value 4 and on the same locations (mainly the Limpopo River, roads, and wetlands) and extent. The second point is that the *ffs* maps show a wider range of values and where these can be found. The baseline map has the largest AOA, followed by the multiple location model which shows similar locations of exclusion, but geared towards the urban areas, mostly with value 1. The single location transfer, i.e., the Manica model applied

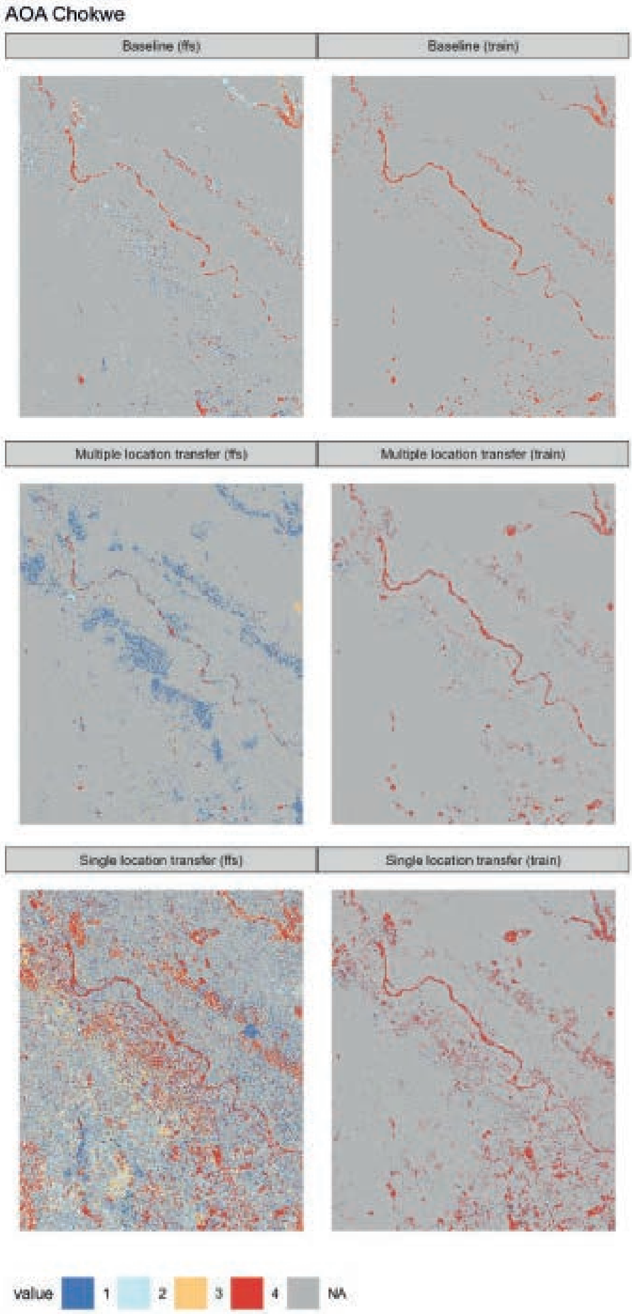


Figure 4 AOA hotspot map of Chokwe. The values in the legend show how many models classified a pixel falling outside of the acceptable range. A value of 4 indicates that all four AOA models predicted that those areas fall outside of the acceptable range.

in Chokwe, shows the same urban areas but now with value 4, as well as more exclusion throughout the whole study area.

Comparing the scenarios shows that the two baseline AOA maps are relatively similar to each other. However, in line with the previous comparison of *ffs* with *train*, the multiple and single location models are very dissimilar. Looking at this, we might expect to see similar irrigation maps for the three scenarios when using *train* and more diverging maps when using *ffs.*

4.2.2. Irrigation classification maps

Figure 5 shows the irrigation classification maps for the different scenarios and variable selection methods. The baselines show clear areas of agriculture with value 4, especially the *train* models seem certain. The two maps are not completely the same, but similar patterns can be found in where irrigation occurs. The *ffs* models show more variation in hotspot values, showing more areas of 1 and 2 than the *train* models. The two transfer scenarios show irrigated agriculture on the whole map, with different degrees of confidence. The two *train* models show value 4 almost everywhere, where the *ffs* models at least show these areas as values 1 and 2. This indicates that although the accuracies of *ffs* in Figure 3 are not very different than those of the *train* models, the *ffs* model seems more likely. Nevertheless, the four maps of the transfer scenarios vastly overestimate the extent, as we show in section 3.4.

Next, we will look at what is actually classified as irrigated agriculture by the transfer models. All transfer models and feature selection methods misclassify vast areas of light vegetation, in addition to classifying irrigation.

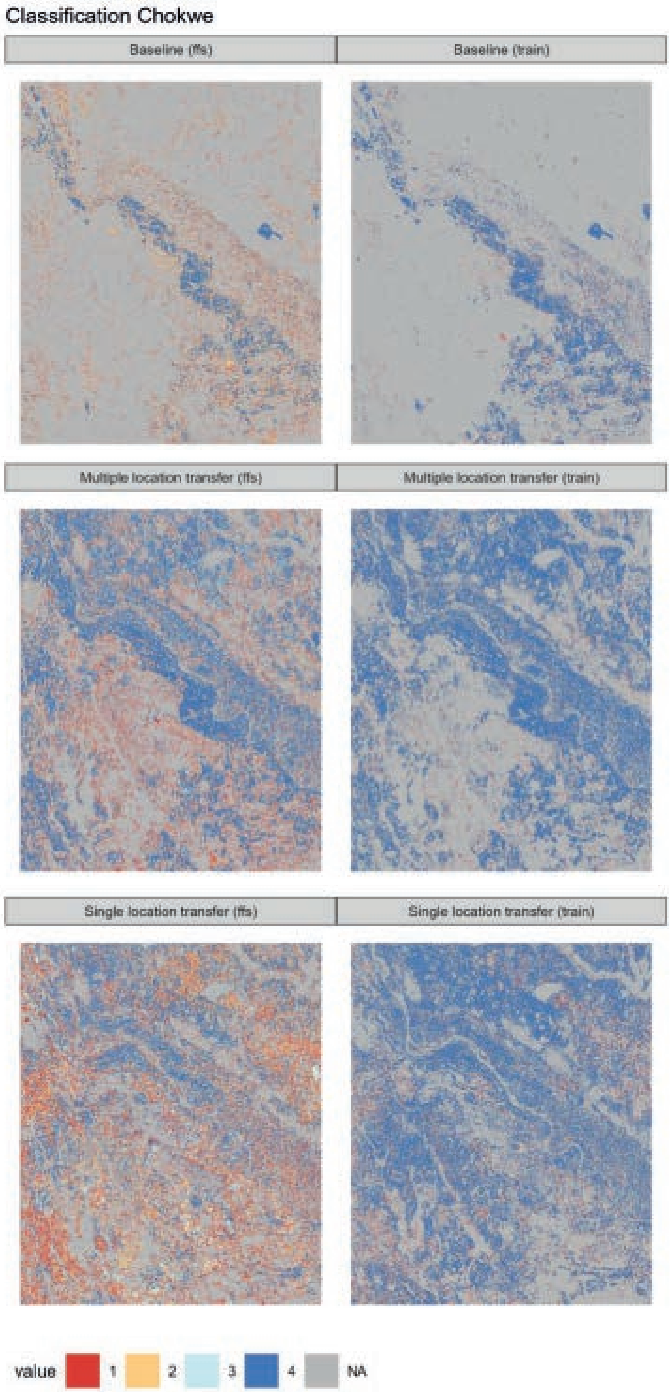


Figure 5 Irrigation classification hotspot map

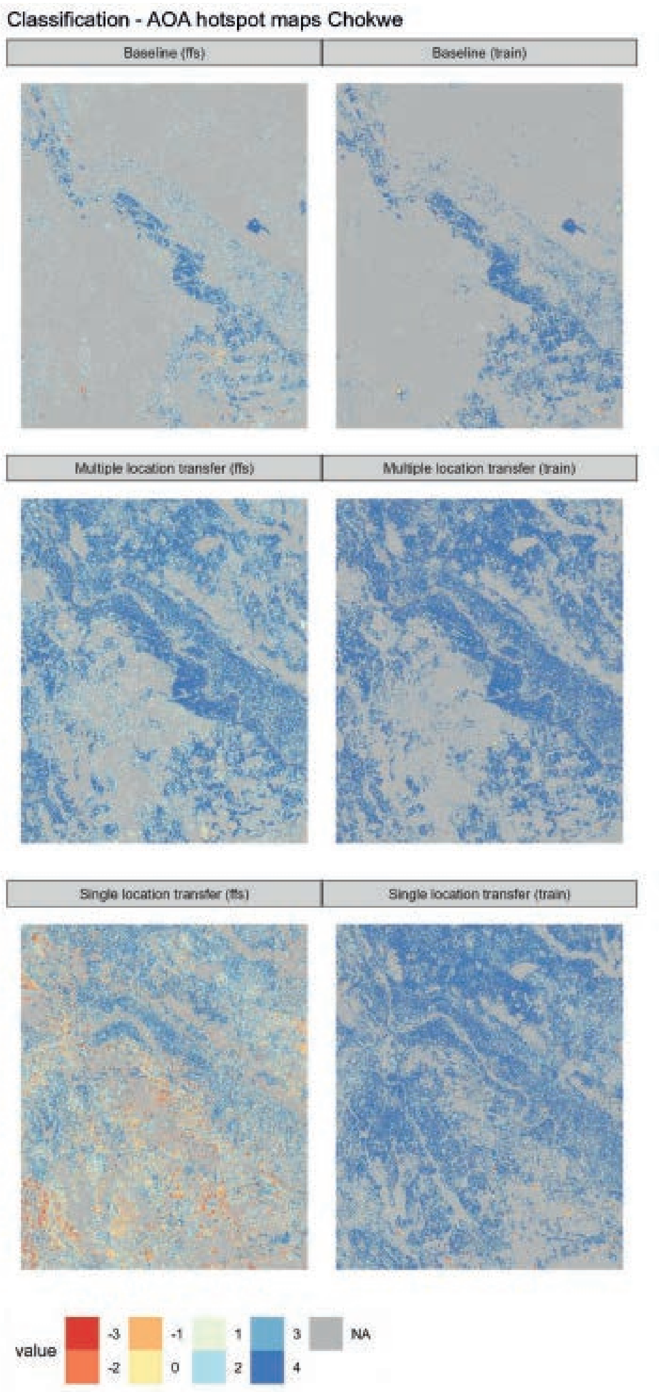


Figure 6 Combined hotspot map of AOA and irrigated agriculture.

4.2.3. Combined

Figure 6 shows the combined map of the previous AOA and irrigation hotspot maps (Figure 4 and Figure 5). The hotspot values of the AOA are subtracted from the irrigated agriculture hotspots to subtract the uncertainty from the classification and resulting in pixel values of -3 to +4. A pixel value below 2 can be considered unreliable, and -3 means one model classified irrigation, and 4 AOA models found that pixel not belonging in the acceptable DI range, irrespectable of the class that pixel was classified as

Considering the new maps in Figure 6, we mainly see changes in the *ffs* maps, especially the single transfer map. This final map has little overlap with the baseline map and mostly classifies rainfed agriculture as irrigation. The other *ffs* map and the two *train* maps seem similar but vastly overestimate areas of irrigated agriculture. The AOA did not take away much uncertainty of the classified maps.

4.3. Visual inspection Manica

4.3.1. AOA maps

Figure 7 shows the AOA hotspot maps for Manica for the different scenarios and feature selection methods. The baseline maps show pixels falling outside of the range throughout the study area. Furthermore, the lake in the south has value 4 in all transfer models, indicating that the water classes of the other areas are dissimilar to the one in Manica. The two *train* transfer maps are similar to each other, and other than the lake itself, also similar to the baseline map. The *ffs* transfer maps are dissimilar to each other, in pixels falling outside of the acceptle DI range and their value.

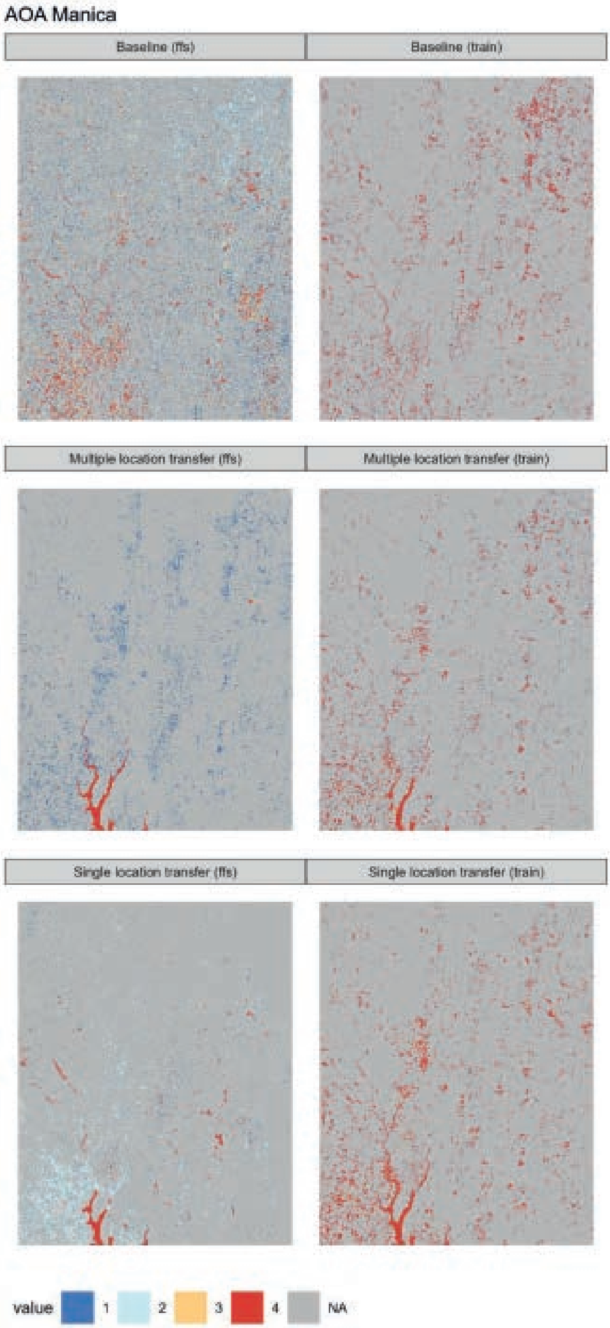


Figure 7 AOA hotspot maps of Manica. The values in the legend show how many models classified a pixel falling outside of the acceptable range. A value of 4 indicates that all four AOA models predicted that those areas fall outside of the acceptable range.

4.3.2. *Irrigation classification maps*

Figure 8 shows the irrigation hotspot maps for the different scenarios. Where the three *train* maps of Chokwe mostly showed value 4 pixels, that is only the case for two of the three maps in Manica. The multiple location transfer map is different, with mainly value 1. Irrigated agriculture is also classified all over the study area, and with different agreements in each scenario. The *ffs* maps mainly show hotspot values of 1 and 2, again spread out over the study area, with a few clusters of value 4.

Next, we will look at what is actually classified as irrigated agriculture by the transfer models. Starting with the multiple location, we see that irrigated agriculture is correctly classified, but mostly trees along the streams and small wetlands are (wrongly) classified as irrigation, for both feature collection methods. The single location transfer models follow the same trends.

4.3.3. *Combined*

Figure 9 shows the combined AOA minus irrigation hotspot maps, which mostly show positive values of 2 or higher. Especially the *train* maps show values of 3 and 4. The *train* maps classify much more irrigated agriculture than the *ffs* maps in each scenario, and the multiple location transfer map shows the least irrigated agriculture. Both transfer scenarios also show less irrigated agriculture than the baseline study, which is the other way around from what we saw in Chokwe.

This indicates that irrigated agriculture in the other three areas shows different spectral patterns than in Manica.

4.4. Hectares of irrigated agriculture

Figure 10 shows the number of hectares classified in the combined hotspot maps, per hotspot value. It shows that not a lot of irrigated area was classified that fell outside of the AOA, otherwise the hectares belonging to values -3 to -1 would be larger. The figure also shows that the *train* maps consistently showed more irrigated agriculture in the 4-value category, whereas the *ffs* models classified more hectares in the 1 to 3 value categories. This shows that the *ffs* models are more uncertain than the *train* models, which we also saw in the visual inspection.

In Chokwe, both of the transferred maps overestimate how much irrigated agriculture there is (class 4), compared to the baseline. They also classify the 2 and 3 value class, but less extreme. In Manica, the baseline map shows the highest estimated hectares, meaning the transferred models underestimate the extent in comparison.

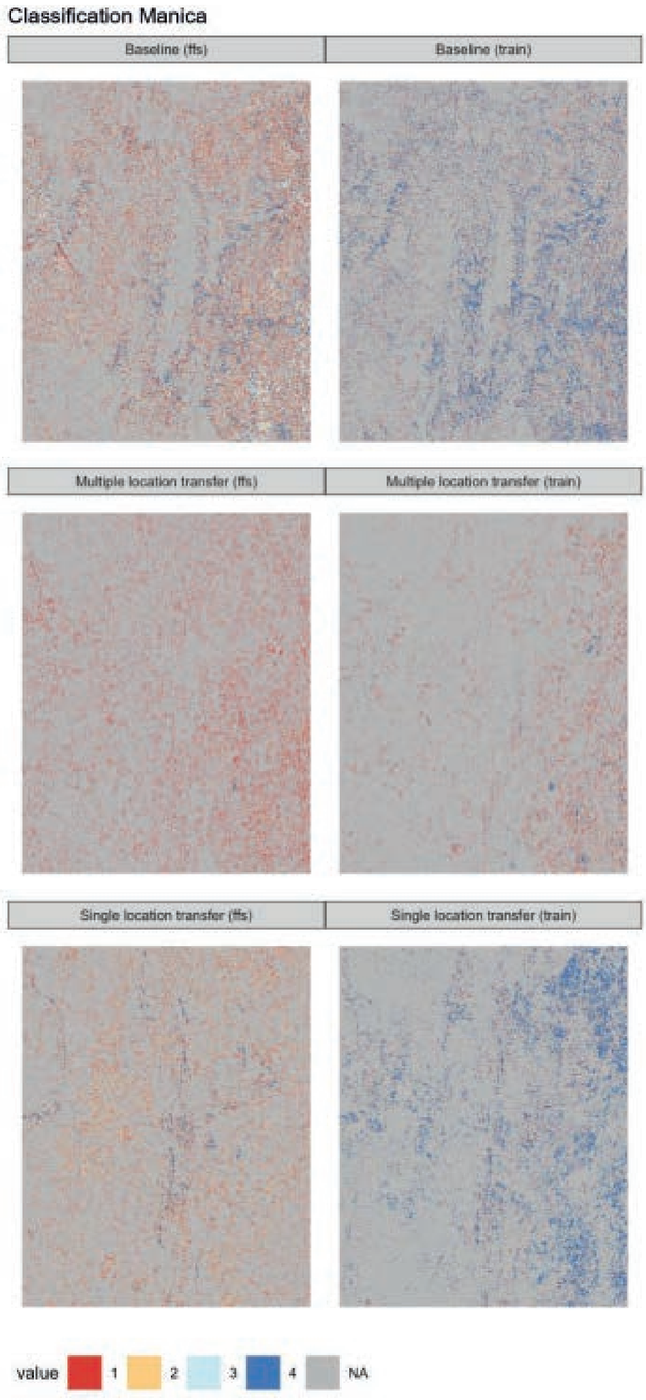


Figure 8 Irrigation classification maps

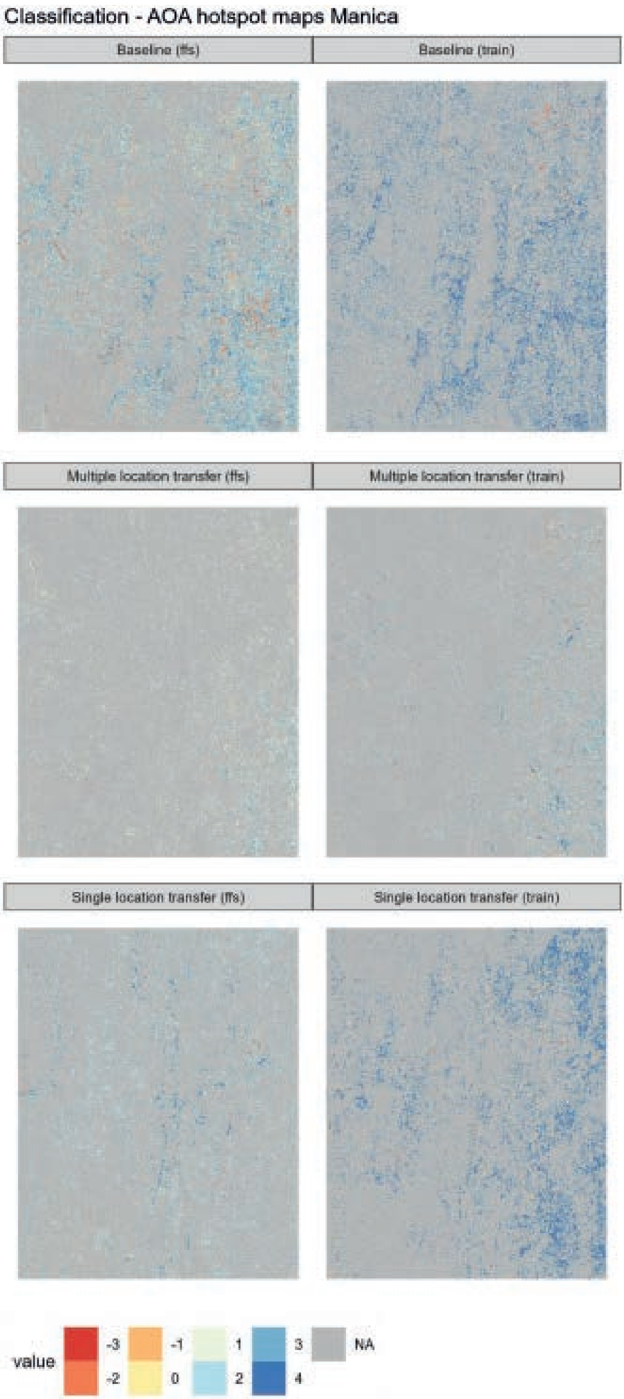


Figure 9 Combined AOA and irrigation map

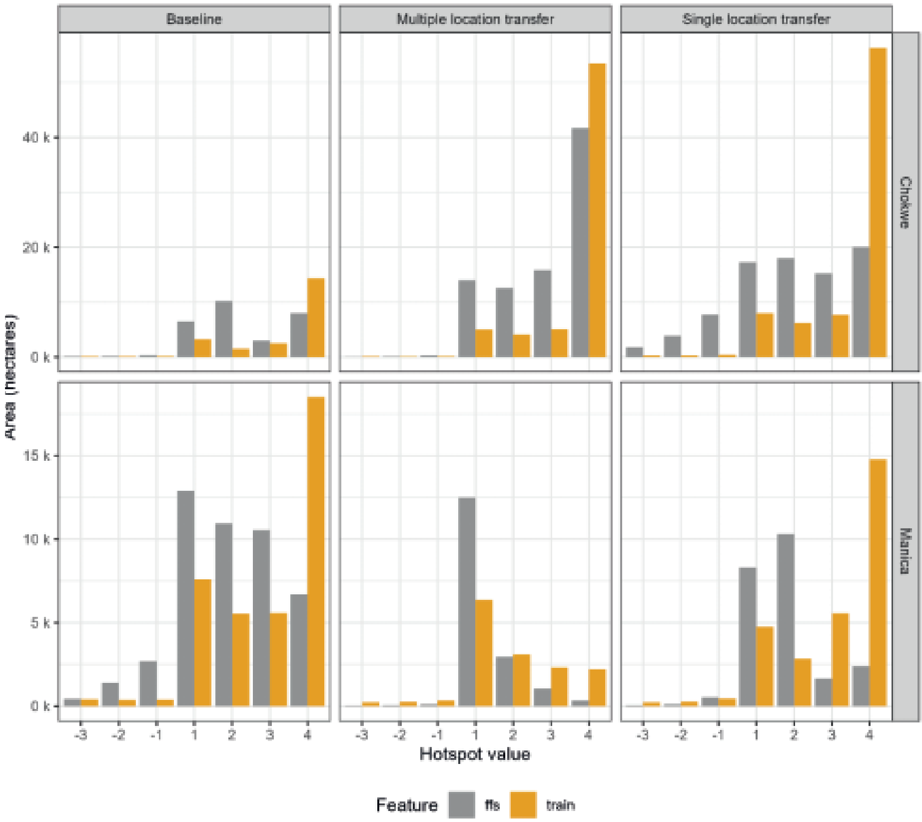


Figure 10 Hectares per hotspot value per scenario

5. Discussion

This study explored three aspects, namely transferring RS-based models between different geographical regions and how feature selection and the Area of Application (AOA) could be used to make maps with higher accuracies and less uncertainties.

Starting with the transferring of models, we explored two scenarios: training a model on data from three different areas (multiple location transfer) and training a model on data from one area (single location transfer). We did this twice, for two study areas, Chokwe that has a more homogeneous landscape and is drier than Manica, which has a more heterogeneous landscape and is wetter. The two other areas that were used for training the multiple location transfer are similar in climate to the other two areas but have slightly different landscape characteristics.

The results show that compared to the baseline maps, both transfer models under-performed, with the single location models performing slightly worse. There can be a few reasons for this. From other studies (Weitkamp et al., 2023) we know that irrigated agriculture can successfully be classified in these areas, hence the training data should be sufficient. This is also confirmed with the baseline studies, which show accuracies of over 75%. Yet the difference is the overall number of variables used. In the other study, we used data of the rain season in addition to the dry season, which we used in this study. We hypothesise that if more variables were present, the single location transfer model would not be affected much in its training, as it might be overfit for the trained location, and hence would still misclassify much of the area of the new location. However, the multiple location model will have had more instances to learn from the statistical relationships of the same landcovers but under different circumstances. We expect that included more predictor variables in the first place might lead to better results for the multiple location transfer scenarios.

We also compared the performances of the models when using feature selection or using all predictor variables. Using the hotspots maps, we see that the train models are more confident in their classifications, which we can derive from the vast areas of value 4 and the low number of hectares for the other hotspot values. The *ffs* models also show many value 4 areas, but lower, and they show more hectares for the values 3 and 2. The *ffs* models are simpler, usually only two to five variables were used (not shown in results), compared to the 14 of the *train* models, with only slightly lower performances.

It is difficult to say if the *train* models overfit to the new locations more so than the *ffs* models, as both vastly overestimated the extent of irrigated agriculture, compared to the baselines. But considering that the *ffs* models show more nuance in their hotspot values, we might consider this as a proxy for generalisation. This aspect needs to be explored some more in future studies – in other words, can the hotspot maps be used as proxy for how well models generalise their assumptions?

Finally, from the overestimated maps (compared to the baselines) we know that the predictions of some locations are too uncertain to be considered for further action, for example to base decision-making on. These areas can be delineated though the AOA, in principle. Yet, the AOA maps miss out on much of these areas, in all transfer scenarios, areas, and variable selection methods. This is surprising, as we expected that, at least for the single location transfer model, many locations would fall outside of the acceptable dissimilarity index (DI) range. This threshold is determined automatically, we left the settings to their defaults. From these findings, we can speculate that the landcovers found in Chokwe and Manica may belong to different classes, spectrally speaking they are very similar, but at a hydrological different scale. For example, the transfer maps applied to Chokwe classified light vegetation as irrigated agriculture, whilst the transfer maps applied to Manica classified dense vegetation as irrigated agriculture. Seemingly, both classes have similar spectral responses as irrigated agriculture in other areas.

Van Passel et al. (2020) conducted a study where they trained models on landscapes of both uniform and diverse compositions. Their goal was to investigate whether models trained on diverse landscapes would outperform those trained on uniform ones when applied to new environments, given that the former capture a broader range of environmental variations, as suggested by previous research. Surprisingly, their findings did not align with this hypothesis. Contrary to expectations, models trained on arid landscapes and then transferred to wetter settings exhibited better performance compared to the reverse scenario. This unexpected outcome had also been observed by Tsalyuk et al. (2017).

Although the differences in user and producer accuracies of irrigated agriculture are small, the models trained in Manica (wetter) and applied to Chokwe (drier) performed slightly better. This is not in line with the findings of (Tsalyuk et al., 2017; Van Passel et al., 2020). However, the overall accuracy of the Chokwe model (and applied to Manica) is high, over 75%. This indicates that other classes are correctly classified and are likely more present, such as dense vegetation (see error matrices on GitHub). We speculate that irrigated agriculture is more difficult to classify in heterogeneous, wet landscapes, where other classes are very similar. When those models are applied to drier areas, it becomes 'easier' to distinguish irrigated agriculture from the surrounding dry areas. However, the classification maps of Chokwe show that irrigated agriculture was classified over the whole study area. Note that the single transfer *ffs* models mostly classified some rainfed areas in the north of the study area and a few irrigated areas, whereas the *train* models classified the majority of the study area as irrigated agriculture.

Ludwig et al., (2023) summarise that a high spatial model error and/or a small area of applicability is an indication of poor model generalisation. This would mean that our models, which have hight spatial model error but a large area of applicability, hint at generalisability. Yet the visual inspection clearly shows the maps overestimate irrigated agriculture. We do not think it is the method of calculating the DI or the threshold, but the simplicity of the models in the first place, combined with the complexity of the landscape. In their study, (Ludwig et al., 2022) also compare models with 12 and 5 variables (also with a similar methodology), but found more meaningful AOA results, but in a less complex landscape. To explore this hypothesis, a future study could include tens of variables to analyse if the AOA better reflects the expected results.

Limitations to the study

We used random forest in this study, but other results may have been achieved if we used other algorithms, such as support vector machine. From our experience (Weitkamp et al., 2023; Weitkamp & Karimi, 2023), these two algorithms performed similarly, hence we only used one.

As noted earlier, the models may have been too simple in the first place (14 variables in total). Although this was enough for the baseline maps, properly training models with the intent to transfer them may require more predictor variables.

5

We also designed the transfer models to have the same training data size as the baseline model, but with different compositions. Training the models on larger datasets also gives the models more opportunities to learn different statistical relationships, potentially improving the generalisability.

The combined use of the predictor variables and training data size may have been a limitation for properly calculating the AOA. We have not explored this aspect, but we expect this to have some influence on size of the AOA.

6. Conclusion

In this research we explored the possibilities of transferring models and using the area of application (AOA) to exclude pixels from the classification that are likely misclassified. We combined the classified maps with the AOA maps to update hotspots maps, which show where irrigated agriculture can likely be found.

We found that the transferred models overestimated the extent of irrigated agriculture, whether three separate locations are used for training, or a single location – the multiple location models only had a marginally higher accuracies than the single location models.

Additionally, we set out to understand how the AOA would change for the different scenarios and found that in our case, the AOA insufficiently masks out misclassified pixels. A possible explanation could be the training data size or number of predictor variables, which can be increased in a future study.

We still believe that the combination of irrigation hotspots and AOA seems like a valid strategy to further highlight the likely misclassified irrigated areas, at least when assessing the baseline maps. However, more research is needed to determine the minimum requirements of the AOA models in terms of landscape, training data size, and predictor variables.

Predicting where irrigated agriculture takes place with higher certainties gives better insights in spatial and temporal trends of irrigation. It also allows policy makers to make better decisions on where interventions may be required and where farmers manage by themselves. Finally, the maps and derived statistics give a more realistic view of what is happening in the field, minimising the need for extensive field visits.

Chapter 6

General discussion

Timon Weitkamp

This research investigated the mapping of the spatial-temporal extent of irrigated agriculture in sub-Saharan Africa (SSA) using remote sensing (RS) data and how modelling choices influence these maps. The mapping of irrigated agriculture holds considerable relevance for supporting irrigation development, monitoring water use, and tracking changes in land cover. Nonetheless, there are several challenges associated with mapping smallholder irrigation. The highly diverse and ever-changing environments where smallholder farmers irrigate make it challenging to differentiate between classes with similar spectral behaviours. Moreover, the presence of small, irregularly shaped fields, inter- and mixed-cropping systems, and the variability in the timing of agricultural practices like planting, harvesting, and irrigation further contribute to the complexity of the task. Additionally, mapping irrigated agriculture requires knowledge of the surrounding land covers to distinguish them. Ground data was collected from various land covers in four areas of Mozambique, covering a wide range of landscape and agricultural practices. I used it to train machine learning models about the relationships between land covers and their spectral properties.

This research investigates how the classification models are sensitive to data inputs, which depend on human choices. The primary objective is to identify unconscious and undesired influences on irrigation mapping, report and reflect on them, and, where necessary, avoid them. This research demonstrates how remote sensing-based mapping of irrigated agriculture is sensitive to the many methodological choices that often remain hidden or implicit.

My research addressed four key research questions (RQs) related to the process of irrigation classification:

- **RQ 1**: How have recent RS-based irrigation mapping projects in SSA consciously and unconsciously defined and classified irrigated agriculture, and how do these choices impact irrigation mapping?

- **RQ 2**: How does the selection of algorithms and composite lengths influence the accuracy of predicting irrigated agriculture in various landscapes and cropping systems?

- **RQ 3**: How does the size and composition of training data impact the accuracy of predicting irrigated agriculture in diverse landscapes and cropping systems?

- **RQ 4**: What approaches can enable the successful application of models trained on one area to other areas, minimising the need for extensive field data collection?

The following section will briefly reflect on these questions.

1. Insights and implications of irrigation mapping

Mapping irrigated agriculture based on remote sensing data is a complex process involving mapping vegetation and knowing how to differentiate between irrigated, rainfed, and non-crop vegetation. Determining how farmers manage the fields, whether it is held privately or publicly, individually or collectively, or through what development process it came about (farmer-led irrigation development or through external intervention) requires information beyond satellite imagery. While RS-derived maps may show small and fragmented irrigated areas, factors like farmers' access to markets, inputs, and finance play a crucial role in understanding the nature of irrigation practices. This non-satellite information necessitates on-the-ground interviews and a deep understanding of local management practices. As a result, a map of irrigated agriculture often provides approximations rather than definitive conclusions, and it is especially difficult to do justice to small-scale, individual, and migrating irrigation practices.

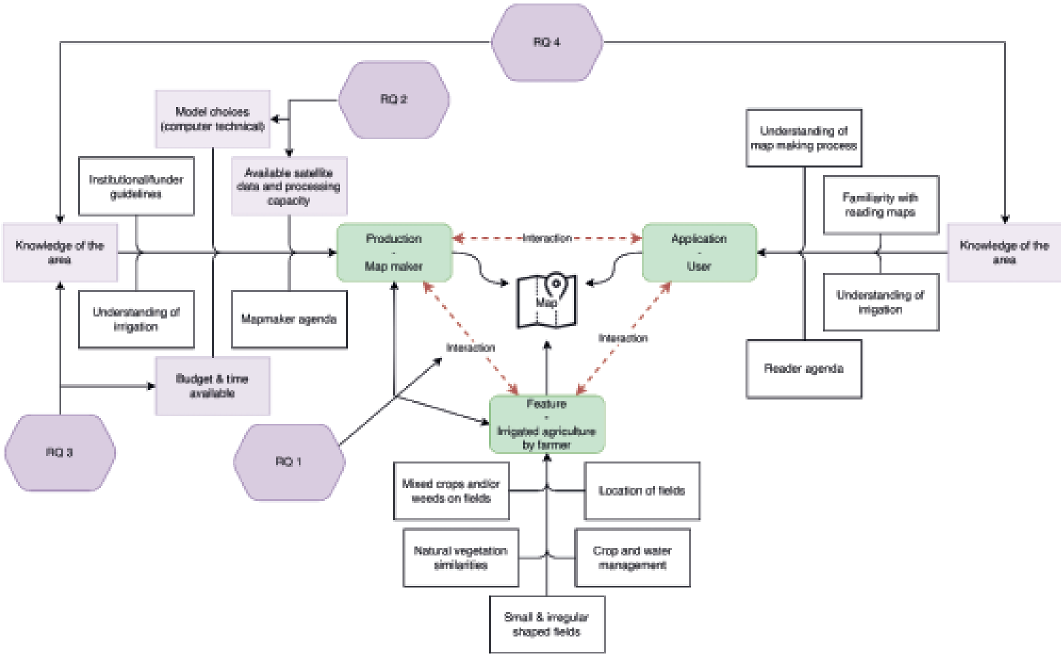


Figure 1 Schematic overview of the many aspects to consider when making and reading a map of irrigated agriculture.

Figure 1 (introduced in Chapter 1) illustrates the various elements involved in the creation of a map, including the individuals responsible for its production (*Production side*), the intended users (*Application side*), and the subject being mapped, which in this case is irrigated agriculture (*Feature side*). I re-introduce it here to illustrate the various elements that play a role in what ends on the map and summarise how the four research questions work together to inspect some of these elements.

1.1. Choices in the classification process.

Maps of irrigated agriculture in SSA vary highly in the extent and location of where they situate irrigation. It is not always fully clear from where these differences emerge. Therefore, it is essential to approach new studies in this field with a critical mindset, recognising the value of exploring alternative methods and conducting context-specific investigations (Maxwell et al., 2018; Ramezan et al., 2021).

In Chapter 2, I analyse recent academic journal articles published since the launch of the Sentinel satellites in 2015. Initially, I intended to conduct a content analysis of the modelling choices and their implications across these studies. However, it became apparent that there was a lack of reporting on these choices, making reproducing their methods impossible,

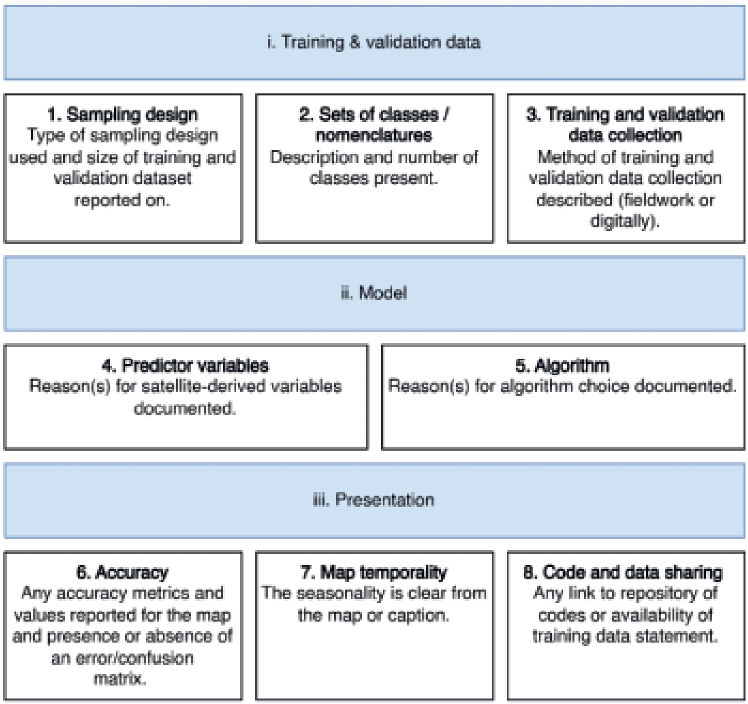


Figure 2 Framework overview containing the eight steps divided over three elements.

and specific steps had to be estimated or guessed. Without sufficient reporting on the methodological choices, the map maker masks their assumptions and intentions (Braun, 2021). I developed a framework to explicitly assess modelling choices, covering eight steps that all classification studies typically go through (**RQ 1**), if not explicitly, then implicitly or by default. The framework (Figure 2) allows me and others to evaluate the reproducibility of results across different studies. I analysed each article to determine which modelling choices and steps the authors reported on and categorised them into three levels: fully reported, partially reported, and not reported. Among the eight steps analysed, steps two, three, and seven (nomenclatures, field data collection, and map seasonality) were reported the most. In contrast, steps one, four and eight (the sampling design, algorithm adequacy, and data and code sharing) received the least attention. None of the articles was complete on every step, and no single article covered all the steps comprehensively.

The results of Chapter 2 highlight that the absence of transparent choices hampers the accurate evaluation of irrigated agriculture's extent, particularly smallholder irrigation, and can affect mapping accuracy significantly. Making these choices explicit not only aids in the evaluation of maps but also allows for the sharing and reusing of relevant components, fostering transparency and collaboration in remote sensing studies.

Furthermore, sharing the elements used in the classification process openly is essential. It enables remote sensing scientists to assess the reliability of new methods and modelling techniques. While authors typically go through the first seven steps of the classification framework, they often leave some steps undocumented in the final publication. Understanding the reasons behind this behaviour requires further investigation through surveys and interviews with remote sensing scientists.

The framework can also serve as a self-assessment tool, ensuring data, models, and code are included before publication. Journals, funders, and institutions can make it a publication prerequisite, offering recognition for exemplary practices. Thorough documentation in irrigation mapping aids targeted government support, minimises resource waste, and fosters global collaboration, advancing irrigation practices and benefiting agriculture and food security.

1.2. Algorithms and composite lengths.

Widely used algorithms in irrigation classification include random forest (RF), support vector machine (SVM), artificial neural networks (ANN), and k-nearest neighbours (k-NN). These algorithms are trained using field data and satellite data, often in the form of composites. Composites are commonly used to generate cloud-free and spatially consistent images from satellite time series by aggregating summary measures from the time series, such as the mean pixel value. Creating monthly, seasonal, or annual composites can effectively capture vegetation phenology.

In Chapter 3, I examined how different algorithms and composite lengths (Steps 4 & 5, Table 1) affect the accuracy of predicting irrigated agriculture in Mozambique (**RQ 2**). Specifically, I evaluate how four classifiers (RF, SVM, ANN, and k-NN) and four composite lengths (1 × 12-monthly, 2 × 6-monthly, 4 × 3-monthly, and 6 × 2-monthly) classify irrigated agriculture. I present the results using "agreement maps" that illustrate the consensus among the models regarding the classification of an area as irrigated agriculture or non-irrigated. These maps highlight the presence of core areas of irrigated agriculture, known as hotspots, which show a high level of certainty. Surrounding these hotspots is an uncertainty zone where the models show less agreement. These maps can combine the strengths of multiple models and reduce the possibility of false positives (areas incorrectly classified as irrigated agriculture). This method is unique as it focuses on a specific class distribution in the area and classification certainty. My analysis, including 16 models, revealed that the composite length and algorithm choice substantially influence the results. Therefore, it is crucial to integrate the findings from various models to address model-specific biases. These findings call for three key recommendations.

- The algorithm selection strongly affects the accuracy of remote sensing-based models for identifying irrigated agriculture. I observed that ANN, SVMs, and RF all performed effectively in classifying irrigated areas. However, there was no single "best" algorithm. I recommend using at least two algorithms to address the landscape's heterogeneous and homogeneous characteristics adequately.

- The composite length is essential in accurately identifying irrigated agriculture in diverse landscapes. I found shorter composites more suitable for complex and heterogeneous landscapes. On the other hand, longer composites are sufficient for more uniform landscapes. Promising options, such as 6-month and 3-month composites, offer reduced computation time and data size advantages while still achieving high classification accuracy.

- My analysis demonstrates that combining models with different composite lengths and algorithms into agreement maps improves the accuracy of identifying irrigated agriculture. These agreement maps provide valuable insights that aid in decision-making processes and assist in prioritising targeted field surveys or management decisions. Additionally, identifying irrigation hotspots through these maps helps decision-makers better understand irrigation dynamics, leading to more informed and effective actions.

1.3. Training data size and composition.

Chapter 4 centres on the impact of training sample size and composition (Steps 1-3, Table 1) on the accuracy of RS classification for mapping smallholder irrigated agriculture in SSA (**RQ 3**). In particular, the optimal number of samples, their quality, and the class imbalance issue are investigated. Models were made on the province scale (i.e., combining data from the two study areas per province) to include more variety in the training data. Generally, the landscape in Gaza province is more homogeneous, and the agricultural fields are more regular and larger than in Manica, where the landscape is more complex, and fields follow the contours of the topography more. Manica also has more rainfall, increasing the complexity of distinguishing cropland from other vegetation classes.

Collecting extensive and high-quality training samples presents difficulties due to limitations in time, access, and interpretability. As a result, class imbalance, where certain classes are more abundant in the training data, can lead to challenges in accurately classifying minority classes. The sample size can affect the choice of algorithm, as some algorithms require a larger dataset than others. These challenges are particularly relevant in the context of smallholder irrigated agriculture, as it is often underrepresented in datasets.

In addition to the dataset's size, training data biases can affect classification outcomes. These biases can arise from limited local knowledge, mislabelling, and the human aspect of interpretation. While some studies have explored the effects of sample size (e.g., Elmes et al., 2020; Ramezan et al., 2021), I found no studies that have investigated the impact of these biases in the training data set on classification results and how choices made by the data collector result in changing accuracies.

The scenarios explored in Chapter 4 show that larger sample sizes generally improve user and producer accuracies; these are class-specific accuracies that indicate whether that class is over- or underestimated. However, there is a point of diminishing returns where further increases in sample size only marginally increase accuracy and require more resources (Scenario 1). In Scenario 2, models trained on data from Gaza province (drier and more homogeneous) perform better overall but not so on irrigated agriculture, indicating a more generalised model. The model trained on data from Manica (wetter and more heterogenous) favoured irrigated agriculture more than other classes, overestimating the extent of the class (i.e., overfitting). In other words, the Gaza model could better predict all classes without much preference towards single classes. In contrast, the Manica model favoured irrigated agriculture more than other classes. Scenarios 3 and 4 highlight the importance of collecting representative field data and using suitable algorithms, such as RF and SVM, which are less sensitive to specific dataset characteristics than the ANN.

I replicated the simulations 25 times, allowing me to deduce that a considerable range of accuracies indicates that particular outcomes rely more on the dataset used for classification than other factors like the algorithm or variables. This variability implies that the model's stability and ability to generalise are compromised when there is a wider distribution of accuracy values.

Agreement maps show the influence of training data biases on classifying farmer-led and conventional irrigation. In the region surrounding Chokwe, small-scale irrigation developed by smallholder farmers can be mainly found on the north side of the Limpopo River, whereas conventional irrigation is more prevalent on the south side. The agreement maps revealed that, in most scenarios, the areas with small-scale irrigation were consistently underrepresented when using data containing more irrigated agriculture, in contrast to the conventional irrigation areas.

Based on the above findings, the following conclusions emerge:

- The quality of training data is one of the most determining factors in successfully mapping smallholder irrigation. Therefore, training data must reflect the target area and include an adequate number of samples for high accuracy, preferably using a random sampling design. Some noise in the data can be tolerated by models such as RF and SVM.

- There is no single algorithm that provides the best results in all circumstances. Given the data's specific characteristics, evaluating multiple algorithms is needed to find the best performer.

- Accuracy values alone may not fully represent classification performance. Visual inspection and further analysis are necessary to understand the results and their limitations comprehensively.

- Multiple replications with different data subsets are needed to assess the training data's robustness. Slight variations in accuracies between replications indicate sufficient training data collection.

1.4. Model transferability and generalisability

Chapter 5 investigates whether transferring models between regions can improve model performance and save resources compared to collecting new data (**RQ 4**). I explored three key aspects: the transfer of RS-based models between different geographical regions, the impact of feature selection on model accuracy, and the use of the Area of Application (AOA) to improve map reliability. The study focused on two distinct study areas, Chokwe and Manica, with differing landscapes and climates.

Firstly, the study examined the transferability of RS-based models. I investigated two scenarios: multiple location transfer, where I trained models on data from three different areas, and single location transfer, where I trained models on one area. Both transfer models underperformed compared to baseline maps, with single-location models performing slightly worse. This could be attributed to the limited number of predictor variables used in the study, as including more variables might have improved the multiple location transfer models' ability to adapt to diverse landscapes.

Additionally, the study compared model performances with and without feature selection. The simpler models with feature selection exhibited more confidence in their classifications. However, the differences were marginal compared to the models that use all variables.

The study also revealed that the predictions of some areas were too uncertain for practical decision-making. I expected the AOA maps to identify these uncertain areas. However, they did not do so effectively across all transfer scenarios and variable selection methods. This led to speculation that different land cover classes in Chokwe and Manica may be similar, spectrally speaking. For example, dense vegetation in one area might resemble irrigated agriculture in another, complicating the classification process.

Notably, despite having high spatial model errors, the models in this study displayed a large area of applicability, suggesting a degree of generalizability. However, a visual inspection of the maps revealed a consistent overestimation of irrigated agriculture. This overestimation might be linked to the models' simplicity, limited predictor variables, and the inherent complexity of the landscape. Future investigations could explore how incorporating more predictor variables might impact AOA results.

Despite these challenges, combining irrigation hotspots and AOA remains a promising strategy to refine the accuracy of predictive maps and support informed decision-making. Additional research is necessary to ascertain the optimal prerequisites for AOA models, encompassing landscape complexity, training data size, and predictor variables, to facilitate more precise predictions of irrigated agriculture. These improved predictions can give valuable insights into irrigation trends, aid policy formulation, and reduce the necessity for extensive field visits.

2. Reflections and directions for further research

2.1. Mapping smallholder irrigation on larger scales

While I used the four regions of Mozambique as case studies, the conclusions drawn from this research apply to the broader irrigation sector in Mozambique and the rest of Africa. To cover a wide range of climatological, geographical and societal diversity, I intentionally modelled four distinct regions to capture as much variety as possible and reflect the broader African context. However, it is essential to note that small-scale irrigation practices developed by farmers themselves exist other than the furrows in mountainous areas and petrol pumps and canals in flatter regions of Mozambique, such as spate irrigation in Sudan (Fujihara et al., 2020), urban irrigation in Ghana (Drechsel & Keraita, 2014) or wetland reclamation in Malawi (Veldman, 2012). Although anecdotal, many more examples can be found, and multiple forms of irrigation exist in the same area and country.

Consequently, agricultural areas under these other practices may not be "seen" by the models used in this thesis, highlighting the need for further exploration beyond the scope of this research. As highlighted in Chapter 5, there is a trade-off between developing a model that accurately identifies irrigation in a small area versus one that can generalise well enough to be applied to larger regions.

Although the models may not apply to larger areas, the methods and concepts can be. A next step in developing this approach is using them on a regional or country-wide scale, which will likely lead to different challenges than the ones explored on a smaller scale, such as managing large datasets, cloud computing, and how to be sure all irrigation practices are included.

2.2. More computational expensive methods

Taking a practical approach, my thesis focused on relatively simple methods that do not require extensive computational resources. While the results were satisfactory, they may not represent the best possible outcome. One avenue not extensively explored in this research involves more advanced classification techniques, including deep learning, object-based methods, and complex time series analysis. These methods often require abundant training data, which was not readily available for this study. To fully harness the potential of these approaches, extensive field data campaigns would be needed to gather the necessary training samples and explore cloud computing for its analysis. Such endeavours present promising avenues for future research. They could enhance the accuracy and robustness of identifying irrigated agriculture in a more advanced and computationally intensive manner.

2.3. Other spatial resolution

When considering further research avenues, exploring the potential benefits and drawbacks of utilising different satellite data sources is important. While the Sentinel data used in this study offered open-source accessibility and satisfactory resolution, incorporating higher resolution imagery, such as from Planet data, could provide other advantages. The improved level of detail from higher resolution imagery enables better identification and differentiation of crop types, leading to more accurate mapping of irrigated agriculture. However, it is essential to acknowledge the associated challenges, including increased costs and greater data processing requirements, particularly for larger areas.

2.4. Other temporal resolution and variables

In Chapter 3, I examined different composite lengths spanning 12 months (1 × 12-monthly, 2 × 6-monthly, 4 × 3-monthly, and 6 × 2-monthly). The analysis revealed that not all variables and months held the same level of importance, and the final model did not even use some. When a model does not use a particular variable, it suggests that the information provided by that variable does not significantly contribute to the model's ability to differentiate between different classes or predict outcomes accurately. Mostly, the end and start of the dry season were most important for identifying irrigated agriculture. Further research could explore the incorporation of specific periods, such as focusing on 1) the months during the dry season, 2) a combination of the end of the dry season (as farmers prepare for the rainy season) and the start of the dry season, or 3) the peak irrigation month(s). By narrowing the focus to these specific periods, generating similar or even better maps may be possible while utilising fewer satellite data.

2.5. Expanding on framework

Although the preceding sections have predominantly focused on technological aspects, there is an opportunity for additional research relating to the framework to make modelling choices explicit, as outlined in Chapter 2. One potential avenue for further exploration involves conducting interviews with authors to investigate their decision-making process and assess their level of consciousness regarding the subjective nature of their choices. Such interviews could offer valuable insights into the factors influencing their methodology, enriching our understanding of the subjectivity inherent in the mapping process. Additionally, it would be worthwhile to further explore the incorporation of additional classification steps to enhance the accuracy and refinement of the mapping process.

2.6. Interaction between farmer and map user

Maps can influence how we perceive and interact with our surroundings. That means that maps inherently possess a political dimension and do not simply mirror reality and hold significant power in shaping perceptions and influencing the division and allocation of space (Bennett et al., 2022).

Critical RS (Bennett et al., 2022) is emerging as a field of study that takes a step back to evaluate remote sensing and considers both who uses satellites for observation and those being observed. It asks questions such as: Who determines who can access the data? Are data being used to directly overcome ground access that would otherwise be unfeasible for political or other reasons? Is this study reinforcing or dismantling an existing power structure? (Luna-Cruz, 2021)

A further research direction is on applying the critical RS concept to mapping smallholder irrigated agriculture and investigating how farmers are affected by those irrigation maps. The study could focus on how policy based on maps affects opportunities for different farmers (e.g., their differentiated access to inputs, markets, knowledge, and funding), if they get a voice in local politics or what happens to their and their fields' privacy. Considering all the conscious and unconscious biases explored in this study (Figure 1), officials and technicians may acknowledge the numbers from the maps and the practices initiated by smallholder farmers. However, they may also reject them to avoid legitimising their practices or justify measures to limit farmer-led irrigation development due to perceived adverse effects on water resources (Venot et al., 2021).

3. Concluding remarks

The remote sensing-based land use/land cover (LULC) classification field has become more accessible and widespread due to various factors. These include the availability of open-source software like QGIS and R, open data policies by organisations such as Landsat, MODIS, and Sentinel, and the emergence of cloud computing platforms like Google Earth Engine and Digital Earth Africa. Additionally, online tutorials and platforms like GitHub have made RS techniques more accessible and widely adopted. This accessibility has empowered individuals and smaller groups who previously lacked the resources to engage in mapping activities. However, by now, it is clear that the diversity of methods and (research) objectives used in creating these maps poses a challenge: it is not always straightforward what methods to use or not, what to report on, and extrapolating the results to other cases.

The question of the added value of RS arises, particularly because field data are still indispensable in the context of smallholder agriculture, characterised by its inherent heterogeneity. Considering the intrinsic uncertainty and inaccuracies, one might question the rationale behind investing considerable efforts in making maps with remote sensing data if field visits are still required for model training and validation. However, RS does offer several advantages despite the need for field data.

Firstly, it provides broader spatial coverage, allowing for identifying and monitoring trends and patterns over large areas that would be challenging to achieve solely through field visits. It is a cost-effective means for obtaining a general overview of the agricultural landscape based on initially collected data.

Secondly, remote sensing can aid in prioritising field visits and optimising resource allocation. By identifying potential hotspots or areas of interest through maps and satellite imagery, field visits can be targeted to specific locations where further data collection, validation, and policy refinement are needed. This approach helps to make the most efficient use of limited resources, maximising the impact of interventions.

Finally, remote sensing can assist in providing historical data and long-term monitoring, enabling the analysis of changes and trends over time. This information is valuable for understanding the dynamics of agricultural systems, identifying drivers of change, and informing evidence-based decision-making. By combining remote sensing with ground-truthing data, models and algorithms can be refined and improved, leading to more accurate and reliable predictions and assessments.

Ultimately, RS-based maps are just tools, and it is up to the people who use them to determine how they will be used. Policymakers can make maps of irrigated agriculture and use them as they see fit. However, the ease with which maps can be made nowadays means that farmers can also get a voice by making alternative maps using equally scientifically sound principles that reflect the perspectives and interests of the local community, challenging dominant narratives and power structures.

References

Abdolrasol, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., & Milad, A. (2021). Artificial Neural Networks Based Optimization Techniques: A Review. *Electronics*, *10*(21), Article 21. https://doi.org/10.3390/electronics10212689

Abera, D., Kibret, K., & Beyene, S. (2019). Tempo-spatial land use/cover change in Zeway, Ketar and Bulbula sub-basins, Central Rift Valley of Ethiopia. *Lakes & Reservoirs: Research & Management*, *24*(1), 76–92. https://doi.org/10/gk8kd6

Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. B. S. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, *7*(4), 221–248. https://doi.org/10.1089/big.2018.0175

Abubakar, G. A., Wang, K., Shahtahamssebi, A., Xue, X., Belete, M., Gudo, A. J. A., Mohamed Shuka, K. A., & Gan, M. (2020). Mapping Maize Fields by Using Multi-Temporal Sentinel-1A and Sentinel-2A Images in Makarfi, Northern Nigeria, Africa. *Sustainability*, *12*(6), Article 6. https://doi.org/10/gkzh5s

Adams, W. M., & Carter, R. C. (1987). Small-scale irrigation in sub-Saharan Africa. *Progress in Physical Geography: Earth and Environment*, *11*(1), 1–27. https://doi.org/10.1177/030913338701100101

African Union. (2020). *Framework for Irrigation Development and Agricultural Water Management in Africa*. African Union. https://au.int/sites/default/files/documents/38632-doc-framework_for_irrigation_development_english.pdf

Ajaz, A., Karimi, P., Cai, X., De Fraiture, C., & Akhter, M. S. (2019). Statistical Data Collection Methodologies of Irrigated Areas and Their Limitations: A Review. *Irrigation and Drainage*, *68*(4), 702–713. https://doi.org/10/gk464q

Aredehey, G., Mezgebu, A., & Girma, A. (2018). Land-use land-cover classification analysis of Giba catchment using hyper temporal MODIS NDVI satellite images. *International Journal of Remote Sensing*, *39*(3), 810–821. https://doi.org/10/gk8kfb

ARSET - Fundamentals of Remote Sensing | NASA Applied Sciences. (n.d.). Retrieved 13 June 2023, from http://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-remote-sensing

Barbiero, P., Squillero, G., & Tonda, A. (2020). *Modeling Generalization in Machine Learning: A Methodological and Computational Study* (arXiv:2006.15680). arXiv. http://arxiv.org/abs/2006.15680

Baudoux, L., Inglada, J., & Mallet, C. (2021). Toward a Yearly Country-Scale CORINE Land-Cover Map without Using Images: A Map Translation Approach. *Remote Sensing*, *13*(6), Article 6. https://doi.org/10.3390/rs13061060

Beekman, W., Veldwisch, G., & Bolding, A. (2014a). Identifying the potential for irrigation development in Mozambique: Capitalizing on the drivers behind farmer-led irrigation expansion. *Physics and Chemistry of the Earth, Parts A/B/C*, *76*, 54–63.

Beekman, W., Veldwisch, G. J., & Bolding, A. (2014b). Identifying the potential for irrigation development in Mozambique: Capitalizing on the drivers behind farmer-led irrigation expansion. *Physics and Chemistry of the Earth, Parts A/B/C*, *76–78*, 54–63. https://doi.org/10/gk8kfc

Bégué, A., Arvor, D., Bellon, B., Betbeder, J., de Abelleyra, D., P. D. Ferraz, R., Lebourgeois, V., Lelong, C., Simões, M., & R. Verón, S. (2018). Remote Sensing and Cropping Practices: A Review. *Remote Sensing*, *10*(2), 99. https://doi.org/10/gczn8x

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10/f8ndk8

Bennett, M. M., Chen, J. K., Alvarez León, L. F., & Gleason, C. J. (2022). The politics of pixels: A review and agenda for critical remote sensing. *Progress in Human Geography*, *46*(3), 729–752. https://doi.org/10.1177/03091325221074691

Bey, A., Jetimane, J., Lisboa, S. N., Ribeiro, N., Sitoe, A., & Meyfroidt, P. (2020a). Mapping smallholder and large-scale cropland dynamics with a flexible classification system and pixel-based composites in an emerging frontier of Mozambique. *Remote Sensing of Environment*, *239*, 111611. https://doi.org/10.1016/j.rse.2019.111611

Bey, A., Jetimane, J., Lisboa, S. N., Ribeiro, N., Sitoe, A., & Meyfroidt, P. (2020b). Mapping smallholder and large-scale cropland dynamics with a flexible classification system and pixel-based composites in an emerging frontier of Mozambique. *Remote Sensing of Environment*, *239*, 111611. https://doi.org/10/gk8kff

Bey, A., Jetimane, J., Lisboa, S. N., Ribeiro, N., Sitoe, A., & Meyfroidt, P. (2020c). Mapping smallholder and large-scale cropland dynamics with a flexible classification system and pixel-based composites in an emerging frontier of Mozambique. *Remote Sensing of Environment*, *239*, 111611. https://doi.org/10.1016/j.rse.2019.111611

Bjornlund, V., Bjornlund, H., & Van Rooyen, A. F. (2020). Exploring the factors causing the poor performance of most irrigation schemes in post-independence sub-Saharan Africa. *International Journal of Water Resources Development*, *36*(sup1), S54–S101. https://doi.org/10.1080/07900627.2020.1808448

Bofana, J., Zhang, M., Nabil, M., Wu, B., Tian, F., Liu, W., Zeng, H., Zhang, N., Nangombe, S. S., Cipriano, S. A., Phiri, E., Mushore, T. D., Kaluba, P., Mashonjowa, E., & Moyo, C. (2020). Comparison of Different Cropland Classification Methods under Diversified Agroecological Conditions in the Zambezi River Basin. *Remote Sensing*, *12*(13), 2096. https://doi.org/10/ghfn2v

Bolding, A., Manzungu, E., & Zaag, P. van der (Eds.). (1996). *The practice of smallholder irrigation: Case studies from Zimbabwe*. University of Zimbabwe Publications.

Brandt, M., Bäumler, R., & Samimi, C. (2009). Agricultural suitability of dune system and Limpopo Basin soils near Xai-Xai, Mozambique. *South African Journal of Plant and Soil*, *26*(4), 206–212. https://doi.org/10.1080/02571862.2009.10639956

Braun, A. C. (2021). More accurate less meaningful? A critical physical geographer's reflection on interpreting remote sensing land-use analyses. *Progress in Physical Geography: Earth and Environment*, *45*(5), 706–735. https://doi.org/10.1177/0309133321991814

Cai, X., Magidi, J., Nhamo, L., & van Koppen, B. (2017). *Mapping irrigated areas in the Limpopo Province, South Africa*. International Water Management Institute (IWMI). https://doi.org/10.5337/2017.205

Campbell, J. B., & Wynne, R. H. (2011). *Introduction to Remote Sensing, Fifth Edition* (5th edition). The Guilford Press.

Castilla, G. (2016). We Must all Pay More Attention to Rigor in Accuracy Assessment: Additional Comment to "The Improvement of Land Cover Classification by Thermal Remote Sensing". Remote Sens. 2015, 7, 8368–8390. *Remote Sensing*, *8*(4), Article 4. https://doi.org/10.3390/rs8040288

Collins, L., McCarthy, G., Mellor, A., Newell, G., & Smith, L. (2020). Training data requirements for fire severity mapping using Landsat imagery and random forest. *Remote Sensing of Environment*, *245*, 111839. https://doi.org/10.1016/j.rse.2020.111839

Comber, A., Fisher, P., & Wadsworth, R. (2004). Integrating land-cover data with different ontologies: Identifying change from inconsistency. *International Journal of Geographical Information Science*, *18*(7), 691–708. https://doi.org/10.1080/13658810410001705316

Comber, A., Fisher, P., & Wadsworth, R. (2005). What is Land Cover? *Environment and Planning B: Planning and Design*, *32*(2), 199–209. https://doi.org/10/cw6pbk

De Bont, C., Komakech, H. C., & Veldwisch, G. J. (2019). Neither modern nor traditional: Farmer-led irrigation development in Kilimanjaro Region, Tanzania. *World Development*, *116*, 15–27. https://doi.org/10.1016/j.worlddev.2018.11.018

De Bont, C., Liebrand, J., Veldwisch, G. J., & Woodhouse, P. (2019). Modernisation and African farmer-led irrigation development: Ideology, policies and practices. *Water Alternatives*, *12*(1), 107–128.

de Bont, C., Liebrand, J., Veldwisch, G. J., & Woodhouse, P. (2019). Modernisation and African Farmer-Led Irrigation Development: Ideology, Policies and Practices. *Water Alternatives*, *12*(1), 23.

de Bont, C., & Veldwisch, G. J. (2020). State Engagement with Farmer-led Irrigation Development: Symbolic Irrigation Modernisation and Disturbed Development Trajectories in Tanzania. *The Journal of Development Studies*, *56*(12), 2154–2168. https://doi.org/10/gjdtf3

de Fraiture, C., & Giordano, M. (2014). Small private irrigation: A thriving but overlooked sector. *Agricultural Water Management*, *131*, 167–174.

DEA. (2021). *DEA GeoMAD*. https://docs.digitalearthafrica.org/en/latest/data_specs/GeoMAD_specs.html#Triple-Median-Absolute-Deviations-(MADs)

Douzas, G., Bacao, F., Fonseca, J., & Khudinyan, M. (2019). Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the

Geometric SMOTE Algorithm. *Remote Sensing*, *11*(24), Article 24. https://doi.org/10.3390/rs11243040

Drechsel, P., & Keraita, B. (2014). *Irrigated urban vegetable production in Ghana: Characteristics, benefits and risk mitigation* (2nd ed). International Water Management Institute (IWMI). https://doi.org/10.5337/2014.219

Du, M., Huang, J., Wei, P., Yang, L., Chai, D., Peng, D., Sha, J., Sun, W., & Huang, R. (2022). Dynamic Mapping of Paddy Rice Using Multi-Temporal Landsat Data Based on a Deep Semantic Segmentation Model. *Agronomy*, *12*(7), Article 7. https://doi.org/10.3390/agronomy12071583

Duker, A. E. C., Cambaza, C., Saveca, P., Ponguane, S., Mawoyo, T. A., Hulshof, M., Nkomo, L., Hussey, S., Van den Pol, B., Vuik, R., Stigter, T., & van der Zaag, P. (2020). Using nature-based water storage for smallholder irrigated agriculture in African drylands: Lessons from frugal innovation pilots in Mozambique and Zimbabwe. *Environmental Science & Policy*, *107*, 1–6. https://doi.org/10/gjkwhx

Duker, A. E. C., Maseko, S., Moyo, M. A., Karimba, B. M., Bolding, A., Prasad, P., De Fraiture, C., & Van Der Zaag, P. (2023). The Changing Faces of Farmer-Led Irrigation: Lessons from Dynamic Irrigation Trajectories in Kenya and Zimbabwe. *The Journal of Development Studies*, 1–20. https://doi.org/10.1080/00220388.2023.2204176

Ebrahimy, H., Mirbagheri, B., Matkan, A. A., & Azadbakht, M. (2022). Effectiveness of the integration of data balancing techniques and tree-based ensemble machine learning algorithms for spatially-explicit land cover accuracy prediction. *Remote Sensing Applications: Society and Environment*, *27*, 100785. https://doi.org/10.1016/j.rsase.2022.100785

Eckert, S. (2016). *Large-and small-scale cropland classification on the foothills of mount Kenya based on SPOT-5 Take 5 data time series* (rayyan-310812505). European Space Agency, (Special Publication) ESA SP, Centre for Development and Environment, University of Bern, Hallerstrasse 10, Bern, 3012, Switzerland. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84988457278&partnerID=40&md5=753107a29cb07f4cec10f02564adfdf2

Eckert, S., Kiteme, B., Njuguna, E., & Zaehringer, J. (2017). Agricultural Expansion and Intensification in the Foothills of Mount Kenya: A Landscape Perspective. *Remote Sensing*, *9*(8), 784. https://doi.org/10/gg3kbv

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J., Fishgold, L., Friedl, M., Jain, M., Kohli, D., Laso Bayas, J., Lunga, D., McCarty, J., Pontius, R., Reinmann, A., Rogan, J., Song, L., Stoynova, H., Ye, S., Yi, Z.-F., & Estes, L. (2020). Accounting for Training Data Error in Machine Learning Applied to Earth Observations. *Remote Sensing*, *12*(6), 1034. https://doi.org/10/ggx5w3

Elwan, E., Le Page, M., Jarlan, L., Baghdadi, N., Brocca, L., Modanesi, S., Dari, J., Quintana Seguí, P., & Zribi, M. (2022). Irrigation Mapping on Two Contrasted Climatic Contexts

Using Sentinel-1 and Sentinel-2 Data. *Water*, *14*(5), Article 5. https://doi.org/10.3390/w14050804

Espey, J. (2019). Sustainable development will falter without data. *Nature*, *571*(7765), 299–299. https://doi.org/10.1038/d41586-019-02139-w

Exner, A., Bartels, L. E., Windhaber, M., Fritz, S., See, L., Politti, E., & Hochleithner, S. (2015). Constructing landscapes of value: Capitalist investment for the acquisition of marginal or unused land—The case of Tanzania. *Land Use Policy*, *42*, 652–663. https://doi.org/10/f6t66f

Foody, G. (2009). Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing*, *30*(20), 5273–5291. https://doi.org/10/dr5xdk

Foody, G. (2021). Impacts of ignorance on the accuracy of image classification and thematic mapping. *Remote Sensing of Environment*, *259*, 112367. https://doi.org/10/gjn8wf

Foody, G. M. (2020). Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment*, *239*, 111630. https://doi.org/10/ghzkkq

Foody, G., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, *104*(1), 1–14. https://doi.org/10/bfpdp4

Foody, G., Pal, M., Rocchini, D., Garzon-Lopez, C., & Bastin, L. (2016). The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS International Journal of Geo-Information*, *5*(11), 199. https://doi.org/10/f9g2ss

Fujihara, Y., Tanakamaru, H., Tada, A., Ahmed Adam, B. M., & Eltaib Elamin, K. A. (2020). Analysis of cropping patterns in Sudan's Gash Spate Irrigation System using Landsat 8 images. *Journal of Arid Environments*, *173*, 104044. https://doi.org/10/gk8kf7

Gao, Q., Zribi, M., Escorihuela, M., Baghdadi, N., & Segui, P. (2018). Irrigation Mapping Using Sentinel-1 Time Series at Field Scale. *Remote Sensing*, *10*(9), 1495. https://doi.org/10/gfnvvx

Gao, Z., Guo, D., Ryu, D., & Western, A. W. (2022). Enhancing the Accuracy and Temporal Transferability of Irrigated Cropping Field Classification Using Optical Remote Sensing Imagery. *Remote Sensing*, *14*(4), Article 4. https://doi.org/10.3390/rs14040997

Gella, G. W., Bijker, W., & Belgiu, M. (2021). Mapping crop types in complex farming areas using SAR imagery with dynamic time warping. *ISPRS Journal of Photogrammetry and Remote Sensing*, *175*, 171–183. https://doi.org/10/gjh2bs

Ghebreamlak, A., Tanakamaru, H., Tada, A., Ahmed Adam, B., & Elamin, K. (2018). Satellite-Based Mapping of Cultivated Area in Gash Delta Spate Irrigation System, Sudan. *Remote Sensing*, *10*(2), 186. https://doi.org/10/gfkdcp

Gitelson, A. A., Viña, A., Ciganda, V., Rundquist, D. C., & Arkebauer, T. J. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters*, *32*(8). https://doi.org/10.1029/2005GL022688

Gumbo, A. D., Kapangaziwiri, E., Chikoore, H., Pienaar, H., & Mathivha, F. (2021). Assessing water resources availability in headwater sub-catchments of Pungwe River Basin in a changing climate. *Journal of Hydrology: Regional Studies*, *35*, 100827. https://doi.org/10.1016/j.ejrh.2021.100827

Harmon, G., Jepson, W., & Lefore, N. (2023). Farmer-led irrigation development in sub-Saharan Africa. *Wiley Interdisciplinary Reviews: Water*, e1631. https://doi.org/10.1002/wat2.1631

Harrison, E. (2018). Engineering change? The idea of 'the scheme' in African irrigation. *World Development*, *111*, 246–255. https://doi.org/10.1016/j.worlddev.2018.06.028

Hasenbein, K., Abdel-Rahman, E. M., Adan, M., Gachoki, S. M., King'ori, E., Dubois, T., & Landmann, T. (2022). Availability of Sentinel-2-based time-series observations: Which vegetation phenology-based metrics perform best for mapping farming systems in complex landscapes? *Remote Sensing Letters*, *13*(7), 695–707. https://doi.org/10.1080/2150704X.2022.2068985

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10/bbcpxk

Higginbottom, T. P., Adhikari, R., Dimova, R., Redicker, S., & Foster, T. (2021). Performance of large-scale irrigation projects in sub-Saharan Africa. *Nature Sustainability*, *4*(6), Article 6. https://doi.org/10.1038/s41893-020-00670-7

Homer, C., Dewitz, J., Jin, S., Xian, G., Costello, C., Danielson, P., Gass, L., Funk, M., Wickham, J., Stehman, S. V., Auch, R., & Riitters, K. (2020). Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database. *ISPRS Journal of Photogrammetry and Remote Sensing*, *162*, 184–199. https://doi.org/10.1016/j.isprsjprs.2020.02.019

Hornun, S. T., & Bolwig, S. (2020). *The Growth of Small-Scale Irrigation in Kenya: The Role of Private Firms in Technology Diffusion.* UNEP DTU Partnership. https://backend.orbit.dtu.dk/ws/portalfiles/portal/223530158/The_Growth_of_Small_Scale_Irrigation_in_Kenya1.pdf

Hounkonnou, D., Kossou, D., Kuyper, T., Leeuwis, C., Nederlof, S., RÃling, N., Sakyi-Dawson, O., & Huis, A. (2012). An innovation systems approach to institutional change: Smallholder development in West Africa. *Agricultural Systems*, *108*. https://doi.org/10/f3xknn

Izzi, G., Denison, J., & Veldwisch, G. J. (2021). *The farmer-led irrigation development guide: A what, why and how-to for intervention design.* World Bank.

Jennewein, J. S., Lamb, B. T., Hively, W. D., Thieme, A., Thapa, R., Goldsmith, A., & Mirsky, S. B. (2022). Integration of Satellite-Based Optical and Synthetic Aperture Radar Imagery to Estimate Winter Cover Crop Performance in Cereal Grasses. *Remote Sensing*, *14*(9), Article 9. https://doi.org/10.3390/rs14092077

Johannsen, C. J., & Daughtry, C. S. T. (2009). Surface Reference Data Collection. In T. A. Warner, M. D. Nellis, & G. Foody, *The SAGE Handbook of Remote Sensing* (pp. 244–256). SAGE Publications, Inc. https://doi.org/10.4135/9780857021052.n17

Kajisa, K., & Payongayong, E. (2011). Potential of and constraints to the rice Green Revolution in Mozambique: A case study of the Chokwe irrigation scheme. *Food Policy*, *36*(5), 615–626. https://doi.org/10.1016/j.foodpol.2011.07.002

Khatami, R., Mountrakis, G., & Stehman, S. V. (2016). A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, *177*, 89–100. https://doi.org/10/f8gwsn

Khatami, R., Southworth, J., Muir, C., Caughlin, T., Ayana, A. N., Brown, D. G., Liao, C., & Agrawal, A. (2020). Operational Large-Area Land-Cover Mapping: An Ethiopia Case Study. *Remote Sensing*, *12*(6), Article 6. https://doi.org/10/gpbxqv

Knauer, K., Gessner, U., Fensholt, R., Forkuor, G., & Kuenzer, C. (2017). Monitoring Agricultural Expansion in Burkina Faso over 14 Years with 30 m Resolution Time Series: The Role of Population Growth and Implications for the Environment. *Remote Sensing*, *9*(2), 132. https://doi.org/10/gg3r7b

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M. (2019). *The caret Package*. https://topepo.github.io/caret/

Kumar, M., Phukon, S. N., Paygude, A. C., Tyagi, K., & Singh, H. (2022). Mapping Phenological Functional Types (PhFT) in the Indian Eastern Himalayas using machine learning algorithm in Google Earth Engine. *Computers & Geosciences*, *158*, 104982. https://doi.org/10.1016/j.cageo.2021.104982

Landmann, T., Eidmann, D., Cornish, N., Franke, J., & Siebert, S. (2019). Optimizing harmonics from Landsat time series data: The case of mapping rainfed and irrigated agriculture in Zimbabwe. *Remote Sensing Letters*, *10*(11), 1038–1046. https://doi.org/10/gk8kgq

Lawrence, R. L., & Moran, C. J. (2015). The AmericaView classification methods accuracy comparison project: A rigorous approach for model selection. *Remote Sensing of Environment*, *170*, 115–120. https://doi.org/10/f7xm72

Lebourgeois, V., Dupuy, S., Vintrou, É., Ameline, M., Butler, S., & Bégué, A. (2017). A Combined Random Forest and OBIA Classification Scheme for Mapping Smallholder Agriculture at Different Nomenclature Levels Using Multisource Data (Simulated Sentinel-2 Time Series, VHRS and DEM). *Remote Sensing*, *9*(3), 259. https://doi.org/10/gk8kgr

Li, Q., Qiu, C., Ma, L., Schmitt, M., & Zhu, X. X. (2020). Mapping the Land Cover of Africa at 10 m Resolution from Multi-Source Remote Sensing Data with Google Earth Engine. *Remote Sensing*, *12*(4), Article 4. https://doi.org/10.3390/rs12040602

Ludwig, M., Bahlmann, J., Pebesma, E., & Meyer, H. (2022). Developing transferable spatial prediction models: A case study of Satellite based landcover mapping. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B3-2022*, 135–141. https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-135-2022

Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., & Meyer, H. (2023). Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography*, *n/a*(n/a). https://doi.org/10.1111/geb.13635

Luna-Cruz, Y. (2021, December 15). *Critical Remote Sensing for the geosciences*. AGU Fall Meeting 2021. https://agu.confex.com/agu/fm21/meetingapp.cgi/Paper/830716

Magidi, J., Nhamo, L., Mpandeli, S., & Mabhaudhi, T. (2021). Application of the Random Forest Classifier to Map Irrigated Areas Using Google Earth Engine. *Remote Sensing*, *13*(5), Article 5. https://doi.org/10/gk66dj

Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J. M., McNairn, H., & Rao, Y. S. (2020). Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data. *Remote Sensing of Environment*, *247*, 111954. https://doi.org/10.1016/j.rse.2020.111954

Marín Del Valle, T., & Jiang, P. (2022). Comparison of common classification strategies for large-scale vegetation mapping over the Google Earth Engine platform. *International Journal of Applied Earth Observation and Geoinformation*, *115*, 103092. https://doi.org/10.1016/j.jag.2022.103092

Massari, C., Modanesi, S., Dari, J., Gruber, A., De Lannoy, G. J. M., Girotto, M., Quintana-Seguí, P., Le Page, M., Jarlan, L., Zribi, M., Ouaadi, N., Vreugdenhil, M., Zappa, L., Dorigo, W., Wagner, W., Brombacher, J., Pelgrum, H., Jaquot, P., Freeman, V., … Brocca, L. (2021). A Review of Irrigation Information Retrievals from Space and Their Utility for Users. *Remote Sensing*, *13*(20), Article 20. https://doi.org/10.3390/rs13204112

Maxwell, A. E., Sharma, M., Kite, J. S., Donaldson, K. A., Maynard, S. M., & Malay, C. M. (2021). Assessing the Generalization of Machine Learning-Based Slope Failure Prediction to New Geographic Extents. *ISPRS International Journal of Geo-Information*, *10*(5), 293. https://doi.org/10.3390/ijgi10050293

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10/gfghjc

Mbaabu, P. R., Ng, W.-T., Schaffner, U., Gichaba, M., Olago, D., Choge, S., Oriaso, S., & Eckert, S. (2019). Spatial Evolution of Prosopis Invasion and its Effects on LULC and Livelihoods in Baringo, Kenya. *Remote Sensing*, *11*(10), 1217. https://doi.org/10/gk8kg6

Meier, J., & Mauser, W. (2023). Irrigation Mapping at Different Spatial Scales: Areal Change with Resolution Explained by Landscape Metrics. *Remote Sensing*, *15*(2), Article 2. https://doi.org/10.3390/rs15020315

Mellor, A., Boukir, S., Haywood, A., & Jones, S. (2015). 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*, 155–168. https://doi.org/10.1016/j.isprsjprs.2015.03.014

Melsen, L. A., Torfs, P. J. J. F., Uijlenhoet, R., & Teuling, A. J. (2017). Comment on "Most computational hydrology is not reproducible, so is it really science?" by Christopher Hutton et al.: REPRODUCING COMPUTATIONAL STUDIES. *Water Resources Research*, *53*(3), 2568–2569. https://doi.org/10/gctq7j

Melsen, L. A., Vos, J., & Boelens, R. (2018). What is the role of the model in socio-hydrology? Discussion of "Prediction in a socio-hydrological world". *Hydrological Sciences Journal*, *63*(9), 1435–1443. https://doi.org/10/ghp56w

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633. https://doi.org/10/gkbtkx

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018a). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10/gc2tsg

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018b). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10/gc2tsg

Millard, K., & Richardson, M. (2015). On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing*, *7*(7), Article 7. https://doi.org/10/gf7ccz

Mills, H. (2008). *Analysis of the transferability of support vector machines for vegetation classification*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10/bq3jpc

Morales-Barquero, L., Lyons, M. B., Phinn, S. R., & Roelfsema, C. M. (2019). Trends in Remote Sensing Accuracy Assessment Approaches in the Context of Natural Resources. *Remote Sensing*, *11*(19), Article 19. https://doi.org/10.3390/rs11192305

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10/frc3mg

Msigwa, A., Komakech, H. C., Verbeiren, B., Salvadore, E., Hessels, T., Weerasinghe, I., & Griensven, A. van. (2019). Accounting for Seasonal Land Use Dynamics to Improve Estimation of Agricultural Irrigation Water Withdrawals. *Water*, *11*(12), 2471. https://doi.org/10/gk8khg

Muluneh, A., Tadesse, T., & Girma, R. (2022). Assessing potential land suitable for surface irrigation using GIS and AHP techniques in the Rift Valley Lakes Basin, Ethiopia. *Sustainable Water Resources Management*, *8*(2), 46. https://doi.org/10.1007/s40899-022-00632-1

Nabil, M., Zhang, M., Bofana, J., Wu, B., Stein, A., Dong, T., Zeng, H., & Shang, J. (2020). Assessing factors impacting the spatial discrepancy of remote sensing based cropland products: A case study in Africa. *International Journal of Applied Earth Observation and Geoinformation*, *85*, 102010. https://doi.org/10.1016/j.jag.2019.102010

Nalepa, R. A., & Bauer, D. M. (2012). Marginal lands: The role of remote sensing in constructing landscapes for agrofuel development. *The Journal of Peasant Studies*, *39*(2), 403–422. https://doi.org/10/gm6m4k

Nalepa, R. A., Short Gianotti, A. G., & Bauer, D. M. (2017). Marginal land and the global land rush: A spatial exploration of contested lands and state-directed development in contemporary Ethiopia. *Geoforum*, *82*, 237–251. https://doi.org/10/gbnmd7

Namara, R. E., Hope, L., Sarpong, E. O., De Fraiture, C., & Owusu, D. (2014). Adoption patterns and constraints pertaining to small-scale water lifting technologies in Ghana. *Agricultural Water Management*, *131*, 194–203. https://doi.org/10/f5j9b6

Nhamo, L., Ebrahim, G. Y., Mabhaudhi, T., Mpandeli, S., Magombeyi, M., Chitakira, M., Magidi, J., & Sibanda, M. (2020). An assessment of groundwater use in irrigated agriculture using multi-spectral remote sensing. *Physics and Chemistry of the Earth, Parts A/B/C*, *115*, 102810. https://doi.org/10/gkq8tw

Nkoka, F., Veldwisch, G., & Bolding, J. (2014a). Organisational modalities of farmer-led irrigation development in Tsangano District, Mozambique. *Water Alternatives*, *7*(2), 414–433.

Nkoka, F., Veldwisch, G. J., & Bolding, A. (2014b). *Organisational Modalities of Farmer-led Irrigation Development in Tsangano District, Mozambique*. 7(2), 20.

Nowakowski, A., Mrziglod, J., Spiller, D., Bonifacio, R., Ferrari, I., Mathieu, P. P., Garcia-Herranz, M., & Kim, D.-H. (2021). Crop type mapping by using transfer learning. *International Journal of Applied Earth Observation and Geoinformation*, *98*, 102313. https://doi.org/10.1016/j.jag.2021.102313

ODK. (2023). *ODK Docs*. https://docs.getodk.org/getting-started/

ODK collect. (2022). *ODK Collect*. https://docs.getodk.org/

Olofsson, P., Foody, G., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10/f55gjk

Orynbaikyzy, A., Gessner, U., & Conrad, C. (2022). Spatial Transferability of Random Forest Models for Crop Type Classification Using Sentinel-1 and Sentinel-2. *Remote Sensing*, *14*(6), Article 6. https://doi.org/10.3390/rs14061493

Ouattara, B., Forkuor, G., Zoungrana, B. J. B., Dimobe, K., Danumah, J., Saley, B., & Tondoh, J. E. (2020). Crops monitoring and yield estimation using sentinel products in semiarid smallholder irrigation schemes. *International Journal of Remote Sensing*, *41*(17), 6527–6549. https://doi.org/10.1080/01431161.2020.1739355

Ozdogan, M., Yang, Y., Allez, G., & Cervantes, C. (2010). Remote Sensing of Irrigated Agriculture: Opportunities and Challenges. *Remote Sensing*, *2*(9), 2274–2304. https://doi.org/10/dttc3d

Pflugmacher, D. (2022). *mapac: Map accuracy and area estimation* (0.11) [Computer software]. https://pages.cms.hu-berlin.de/pflugmad/mapac/

Phalke, A. R., & Özdoğan, M. (2018). Large area cropland extent mapping with Landsat data and a generalized classifier. *Remote Sensing of Environment*, *219*, 180–195. https://doi.org/10.1016/j.rse.2018.09.025

Phalke, A. R., Özdoğan, M., Thenkabail, P. S., Erickson, T., Gorelick, N., Yadav, K., & Congalton, R. G. (2020). Mapping croplands of Europe, Middle East, Russia, and Central Asia using Landsat, Random Forest, and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, *167*, 104–122. https://doi.org/10.1016/j.isprsjprs.2020.06.022

Pires de Lima, R., & Marfurt, K. (2020). Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sensing*, *12*(1), Article 1. https://doi.org/10.3390/rs12010086

Ramezan, C. A., Warner, T. A., Maxwell, A. E., & Price, B. S. (2021). Effects of Training Set Size on Supervised Machine-Learning Land-Cover Classification of Large-Area High-Resolution Remotely Sensed Data. *Remote Sensing*, *13*(3), Article 3. https://doi.org/10/gn38dp

Ramezan, C., Warner, T., & Maxwell, A. (2019). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing*, *11*(2), 185. https://doi.org/10/gk8kd5

Redicker, S., Dimova, R., & Foster, T. (2022). Synthesising evidence on irrigation scheme performance in West Africa. *Journal of Hydrology*, *610*, 127919. https://doi.org/10.1016/j.jhydrol.2022.127919

Roberts, D., Dunn, B., & Mueller, N. (2018). Open Data Cube Products Using High-Dimensional Statistics of Time Series. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 8647–8650. https://doi.org/10/gk8khv

Roberts, D., Mueller, N., & Mcintyre, A. (2017). High-Dimensional Pixel Composites From Earth Observation Time Series. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(11), 6254–6264. https://doi.org/10/gcgqn3

Rufin, P., Bey, A., Picoli, M., & Meyfroidt, P. (2022). Large-area mapping of active cropland and short-term fallows in smallholder landscapes using PlanetScope data. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102937. https://doi.org/10.1016/j.jag.2022.102937

Salmon, J. M., Friedl, M. A., Frolking, S., Wisser, D., & Douglas, E. M. (2015). Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data. *International Journal of Applied Earth Observation and Geoinformation*, *38*, 321–334. https://doi.org/10/f67sjj

Sedano, F., Molini, V., & Azad, M. A. K. (2019). A mapping framework to characterize land use in the sudan-sahel region from dense stacks of landsat data. *Remote Sensing*, *11*(6). Scopus. https://doi.org/10.3390/RS11060648

Segarra, J., Buchaillot, M. L., Araus, J. L., & Kefauver, S. C. (2020). Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications. *Agronomy*, *10*(5), Article 5. https://doi.org/10.3390/agronomy10050641

Shao, G., Tang, L., & Liao, J. (2019). Overselling overall map accuracy misinforms about research reliability. *Landscape Ecology*, *34*(11), 2487–2492. https://doi.org/10/ggvtj4

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10/gkfsb5

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., & James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources. *Scientific Data*, *6*(1), 190030. https://doi.org/10.1038/sdata.2019.30

Statista. (2022, July 21). *World population by continent 1800-2100*. Statista. https://www-statista-com.ezproxy.library.wur.nl/statistics/997040/world-population-by-continent-1950-2020/

Stehman, S. V., & Czaplewski, R. L. (1998). Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sensing of Environment*, *64*(3), 331–344. https://doi.org/10.1016/S0034-4257(98)00010-8

Stehman, S. V., & Foody, G. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, *231*, 111199. https://doi.org/10/gf7n7m

Stehman, S. V., & Wickham, J. (2020). A guide for evaluating and reporting map data quality: Affirming Shao et al. "Overselling overall map accuracy misinforms about research reliability". *Landscape Ecology*, *35*(6), 1263–1267. https://doi.org/10/gk8mhf

Tena, Mwaanga, & Nguvulu. (2019). Impact of Land Use/Land Cover Change on Hydrological Components in Chongwe River Catchment. *Sustainability*, *11*(22), 6415. https://doi.org/10/gk8kgg

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*, *18*(2), 18. https://doi.org/10/gf55h8

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, *237*, 111322. https://doi.org/10.1016/j.rse.2019.111322

Traoré, F., Bonkoungou, J., Compaoré, J., Kouadio, L., Wellens, J., Hallot, E., & Tychon, B. (2019). Using Multi-Temporal Landsat Images and Support Vector Machine to Assess the Changes in Agricultural Irrigated Areas in the Mogtedo Region, Burkina Faso. *Remote Sensing*, *11*(12), 1442. https://doi.org/10/gk8kjb

Tsalyuk, M., Kelly, M., & Getz, W. M. (2017). Improving the prediction of African savanna vegetation variables using time series of MODIS products. *ISPRS Journal of Photogrammetry and Remote Sensing*, *131*, 77–91. https://doi.org/10.1016/j.isprsjprs.2017.07.012

Van Passel, J., De Keersmaecker, W., & Somers, B. (2020). Monitoring Woody Cover Dynamics in Tropical Dry Forest Ecosystems Using Sentinel-2 Satellite Imagery. *Remote Sensing*, *12*(8), Article 8. https://doi.org/10.3390/rs12081276

van Rooyen, A. F., Ramshaw, P., Moyo, M., Stirzaker, R., & Bjornlund, H. (2017). Theory and application of Agricultural Innovation Platforms for improved irrigation scheme management in Southern Africa. *International Journal of Water Resources Development*, *33*(5), 804–823. https://doi.org/10.1080/07900627.2017.1321530

Veldman, R. (2012). *Dynamics in wetland agriculture: Competition and desiccation: A case study of household dynamics, agricultural intensification and competition for resources in wetland agriculture in Badwa dambo, Malawi* [MSc, Wageningen University and Research]. https://edepot.wur.nl/197176

Veldwisch, G. J., Bolding, A., & Wester, P. (2009). Sand in the Engine: The Travails of an Irrigated Rice Scheme in Bwanje Valley, Malawi. *The Journal of Development Studies*, *45*(2), 197–226. https://doi.org/10/c9fn9w

Veldwisch, G. J., Venot, J.-P., Woodhouse, P., Komakech, H. C., & Brockington, D. (2019a). Re-introducing politics in African farmer-led irrigation development: Introduction to a special issue. *Water Alternatives*, *12*(1), 1–12.

Veldwisch, G. J., Venot, J.-P., Woodhouse, P., Komakech, H. C., & Brockington, D. (2019b). *Re-introducing Politics in African Farmer-Led Irrigation Development: Introduction to a Special Issue*. *12*(1), 12.

Venot, J.-P., Bowers, S., Brockington, D., Komakech, H., Ryan, C., Veldwisch, G. J., & Woodhouse, P. (2021). *Below the Radar: Data, Narratives and the Politics of Irrigation in Sub-Saharan Africa*. *14*(2), 27.

Vogels, M. F. A., de Jong, S. M., Sterk, G., & Addink, E. A. (2019). Mapping irrigated agriculture in complex landscapes using SPOT6 imagery and object-based image analysis – A case study in the Central Rift Valley, Ethiopia –. *International Journal of Applied Earth Observation and Geoinformation*, *75*, 118–129. https://doi.org/10/gk8kjd

Vogels, M. F. A., de Jong, S., Sterk, G., Douma, H., & Addink, E. (2019). Spatio-Temporal Patterns of Smallholder Irrigated Agriculture in the Horn of Africa Using GEOBIA and Sentinel-2 Imagery. *Remote Sensing*, *11*(2), 143. https://doi.org/10/gk8kjf

Wang, S., Azzari, G., & Lobell, D. B. (2019). Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. *Remote Sensing of*

Environment, *222*, 303–317. https://doi.org/10.1016/j.rse.2018.12.026

Weemstra, H., Oord, A. L., de Boer, F. S., & Beekman, P. W. (2014). Baseflow prediction in a data-scarce catchment with Inselberg topography, Central Mozambique. *Physics and Chemistry of the Earth, Parts A/B/C*, *76–78*, 16–27. https://doi.org/10.1016/j.pce.2014.09.005

Weitkamp, T., & Beekman, W. (2022). *Technical report: Remote sensing methodology to measure irrigated agriculture in Mozambique* (Technical Report FASIMO report 3A). Resilience BV.

Weitkamp, T., & Karimi, P. (2023). Evaluating the Effect of Training Data Size and Composition on the Accuracy of Smallholder Irrigated Agriculture Mapping in Mozambique Using Remote Sensing and Machine Learning Algorithms. *Remote Sensing*, *15*(12), Article 12. https://doi.org/10.3390/rs15123017

Weitkamp, T., Veldwisch, G. J., Karimi, P., & de Fraiture, C. (2023). Mapping irrigated agriculture in fragmented landscapes of sub-Saharan Africa: An examination of algorithm and composite length effectiveness. *International Journal of Applied Earth Observation and Geoinformation*, *122*, 103418. https://doi.org/10.1016/j.jag.2023.103418

Wellington, M. J., & Renzullo, L. J. (2021). High-Dimensional Satellite Image Compositing and Statistics for Enhanced Irrigated Crop Mapping. *Remote Sensing*, *13*(7), Article 7. https://doi.org/10/gmfbrm

Wiggins, S., & Lankford, B. (2019). *Farmer-led irrigation in sub-Saharan Africa: Synthesis of current understandings*.

Woodcock, C. E., & Gopal, S. (2000). Fuzzy set theory and thematic maps: Accuracy assessment and area estimation. *International Journal of Geographical Information Science*, *14*(2), 153–172. https://doi.org/10/cmkvng

Woodhouse, P., Veldwisch, G. J., Venot, J.-P., Brockington, D., Komakech, H., & Manjichi, Â. (2017a). African farmer-led irrigation development: Re-framing agricultural policy and investment? *The Journal of Peasant Studies*, *44*(1), 213–233.

Woodhouse, P., Veldwisch, G. J., Venot, J.-P., Brockington, D., Komakech, H., & Manjichi, Â. (2017b). African farmer-led irrigation development: Re-framing agricultural policy and investment? *The Journal of Peasant Studies*, *44*(1), 213–233. https://doi.org/10/gk8kjm

Xie, Y., Lark, T. J., Brown, J. F., & Gibbs, H. K. (2019). Mapping irrigated cropland extent across the conterminous United States at 30 m resolution using a semi-automatic training approach on Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, *155*, 136–149. https://doi.org/10/ghgc6p

Xiong, J., Thenkabail, P. S., Gumma, M. K., Teluguntla, P., Poehnelt, J., Congalton, R. G., Yadav, K., & Thau, D. (2017). Automated cropland mapping of continental Africa using Google Earth Engine cloud computing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *126*, 225–244. https://doi.org/10/f949js

Xu, J., Zhu, Y., Zhong, R., Lin, Z., Xu, J., Jiang, H., Huang, J., Li, H., & Lin, T. (2020). DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sensing of Environment*, *247*, 111946. https://doi.org/10.1016/j.rse.2020.111946

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding Challenges in the Transferability of Ecological Models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Yu, L., Liang, L., Wang, J., Zhao, Y., Cheng, Q., Hu, L., Liu, S., Yu, L., Wang, X., Zhu, P., Li, X., Xu, Y., Li, C., Fu, W., Li, X., Li, W., Liu, C., Cong, N., Zhang, H., … Gong, P. (2014). Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *International Journal of Remote Sensing*, *35*(13), 4573–4588. https://doi.org/10/gfghjg

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., & Arino, O. (2022). *ESA WorldCover 10 m 2021 v200* (Version v200) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.7254221

Zhang, W., Liu, H., Wu, W., Zhan, L., & Wei, J. (2020). Mapping Rice Paddy Based on Machine Learning with Sentinel-2 Multi-Temporal Data: Model Comparison and Transferability. *Remote Sensing*, *12*(10), Article 10. https://doi.org/10.3390/rs12101620

Appendices

– Summary –

– Samenvatting –

– List of publications –

– Acknowledgements –

– SENSE Education Certificate –

Summary

In recent years, there has been a renewed interest in irrigation in sub-Saharan Africa (SSA) due to the need for agricultural development and food security. Expanding irrigation is necessary to meet the region's food requirements with the projected population growth. Smallholder farmers have long been driving irrigated agriculture in SSA for a long time through farmer-led irrigation development (FLID). Farmers have independently initiated, operated, and maintained irrigation infrastructure, often focusing on high-value cash crops to improve their income. However, FLID often goes unnoticed by official institutions due to its fragmented nature and the technical bias in defining irrigation. The small scale and heterogeneity of FLID make it challenging to accurately count and report official statistics. Moreover, the practices of smallholder farmers are sometimes considered inferior or irrelevant compared to "modern" irrigation technologies.

Similar challenges arise when mapping with remote sensing (RS) due to the complex and diverse nature of these systems. Several factors contribute to the difficulty in accurately measuring and classifying irrigated agriculture using satellite sensors. These factors include the similarity in spectral signatures between different land cover classes, mixed spectral signatures within the same land cover class, complex shapes and arrangements of fields, and subjective definitions of irrigation.

Despite these challenges, RS offers several advantages for mapping irrigated agriculture. It provides wide spatial coverage, allows monitoring of temporal and spatial trends, and assists in prioritizing field visits. RS data can be consistently analysed over time and is easily accessible. Different classes of irrigated agriculture can be distinguished by considering factors such as the timing of image acquisition, variations in vegetation colour, and notable changes.

This thesis aims to examine the production of remote sensing maps and their ability to depict irrigated agriculture. While remote sensing cannot directly measure farmer-led irrigation, it can capture the diverse and dispersed nature of small-scale irrigated agriculture, which requires interpretation through fieldwork and local expertise. The research identifies and addresses potential challenges in mapping irrigated agriculture in SSA using remote sensing data.

The research uses four case studies in Mozambique, specifically Chokwe, Xai-Xai, Manica, and Catandica, chosen for their diverse agroecological characteristics and the presence of both small-scale and large-scale irrigated agriculture.

In Chapter 2, I look at common RS classification steps that all mapping studies go through. I developed a framework to explicitly address and assess modelling choices, covering seven steps that all classification studies typically go through. The framework aims to evaluate the reproducibility of results across different studies. The primary results highlight two key findings. Firstly, the study demonstrates and systematizes the impact of different choices on the classification process. Secondly, it reveals a concerning culture of insufficient reporting on eight crucial choices. The lack of reporting in these eight domains suggests a potential lack of awareness among map makers regarding the significance of their methodological choices in accurately defining the extent of irrigated agriculture and reproducibility. Consequently, the produced maps likely underreport the full extent of irrigated agriculture, especially that of smallholder farmers.

In Chapter 3, I examined how different algorithms and composite lengths affect the accuracy of predicting irrigated agriculture in Mozambique. Composites are commonly used to generate cloud-free and spatially consistent images from satellite time series by aggregating summary measures from the time series, such as the mean pixel value. Creating composites on a monthly, seasonal, or annual basis can effectively capture vegetation phenology. Specifically, I evaluated how four classifiers (the random forest (RF), support vector machine (SVM), artificial neural networks (ANN), and k-nearest neighbours (k-NN)) and four composite lengths (1 × 12-monthly, 2 × 6-monthly, 4 × 3-monthly, and 6 × 2-monthly) classified irrigated agriculture. I present the results using “agreement maps” that illustrate the consensus among the models regarding the classification of an area as irrigated agriculture or non-irrigated. These maps highlight the presence of core areas of irrigated agriculture, known as hotspots, which exhibit a high level of certainty. Surrounding these hotspots is an uncertainty zone where the models exhibit less agreement. These maps can combine the strengths of multiple models and reduce the possibility of false positives (areas incorrectly classified as irrigated agriculture).

I found that artificial ANN, SVM, and RF all performed effectively in classifying irrigated areas. However, there was no single “best” algorithm. For complex and heterogeneous landscapes, shorter composites are found to be more suitable. Conversely, longer composites are sufficient for more uniform landscapes. Promising options, such as 6-month and 3-month composites, offer advantages in reduced computation time and data size while still achieving high classification accuracy. My analysis demonstrates that combining models with different composite lengths and algorithms into agreement maps improves the accuracy of identifying irrigated agriculture.

Chapter 4 centres on the impact of training sample size and composition on the accuracy of RS classification for mapping smallholder irrigated agriculture in SSA. In particular, I investigate the optimal number of samples, their quality, and the class imbalance issue. Collecting extensive and high-quality training samples presents difficulties due to limitations in time, access and interpretability. As a result, class imbalance, where certain classes are more abundant in the training data, can lead to challenges in accurately classifying minority classes. The available sample size can affect the choice of algorithm, as some algorithms require a larger dataset than others. These challenges are particularly relevant in the context of smallholder irrigated agriculture, as it is often underrepresented in datasets and policies. In addition to the dataset's size, training data biases can affect classification outcomes. These biases can arise from limited local knowledge, mislabelling, and the human aspect of interpretation.

The various explored scenarios of Chapter 4 show that larger sample sizes generally improve user and producer accuracies; these are class-specific accuracies that can be used to show if that class is over- or underestimated. However, there is a point of diminishing returns where further increases in sample size only marginally increase accuracy and require more resources. The study also reveals that models trained on Gaza perform better overall, indicating a more generalized model compared to the overfitting observed in Manica; in other words, the Gaza model was better able to predict all classes without much preference towards single classes. In contrast, the Manica model favoured irrigated agriculture more than other classes. Other scenarios highlight the importance of collecting representative field data and using suitable algorithms, such as RF and SVM, which are less sensitive to specific dataset characteristics compared to the ANN.

Chapter 5 investigates whether transferring models between regions can improve model performance and save resources compared to collecting new data. I hypothesize that targeted data collection is necessary in the new area since the relationships between spectral responses and land covers learned in one area may not apply due to variations in weather conditions, landscapes, and farming practices. Instead of random data collection, I focused on identifying areas with high prediction errors to guide targeted data collection efforts.

Various models were trained on data from different scenarios to investigate the potential transferability of machine learning models for predicting irrigated agriculture. The study found that simple transfers of models were not effective in correctly classifying new areas due to insufficient training data. However, incorporating more diverse data from multiple regions improved the classification performance. Unsurprisingly, the best results were achieved when using only data from the target area, excluding data from other areas.

To conclude, the field of remote sensing-based land use/land cover classifications has been democratised due to various factors, including the availability of open-source software like

QGIS and R, open data policies by organizations such as Landsat, MODIS, and Sentinel, as well as the emergence of cloud computing platforms like Google Earth Engine and Digital Earth Africa. Additionally, online tutorials and platforms such as GitHub have made RS techniques more accessible and widely adopted. This accessibility has empowered individuals and smaller groups who previously lacked the resources to engage in mapping activities. However, the diversity of methods and (research) objectives used in creating these maps poses a challenge: it is not always straightforward what methods to use or not, what to report on, and extrapolating the results to other cases. The results of this research have implications for documenting and reporting of methods and choices, presenting irrigated agriculture through maps, and showing how easy it is to manipulate those maps with slight tweaks to models.

Samenvatting

De afgelopen jaren is er hernieuwde interesse ontstaan in irrigatie in Afrika ten zuiden van de Sahara (*SSA – sub-Saharan Africa*) vanwege de groeiende noodzaak van landbouwontwikkeling en voedselzekerheid. De uitbreiding van irrigatie is cruciaal om te kunnen voldoen aan de verwachte bevolkingsgroei en de groeiende voedselbehoefte in de regio. Kleinschalige boeren spelen al lange tijd een leidende rol in geïrrigeerde landbouw in SSA, voornamelijk door middel van door boeren geleide irrigatieontwikkeling (*FLID – Farmer-Led Irrigation Development*). Deze boeren hebben zelf het initiatief genomen voor het opzetten, exploiteren en onderhouden van irrigatie-infrastructuur, vaak gericht op hoogwaardige gewassen om hun inkomens te verbeteren. Ondanks deze belangrijke bijdrage wordt FLID vaak over het hoofd gezien door officiële instanties vanwege de versnipperde aard en technologische vooringenomenheid bij de definitie van irrigatie. De kleinschaligheid en diversiteit van FLID maken het uitdagend om nauwkeurige statistieken te verzamelen en te rapporteren. Bovendien worden de praktijken van kleinschalige boeren soms als minderwaardig of irrelevant beschouwd in vergelijking met "moderne" irrigatietechnologieën.

Vergelijkbare uitdagingen doen zich voor bij het gebruik van *remote sensing* (RS) voor het in kaart brengen van geïrrigeerde gebieden, vanwege de complexe en diverse aard van deze agrarische systemen. Diverse factoren bemoeilijken het nauwkeurig meten en classificeren van geïrrigeerde landbouw met behulp van satellietsensoren. Enkele van deze factoren zijn de gelijkenis in spectrale signalen tussen verschillende klassen van landbedekking, gemengde spectrale signalen binnen dezelfde klassen van landbedekking, ingewikkelde vormen en indelingen van velden, en subjectieve definities van irrigatie.

Desondanks biedt RS diverse voordelen voor het in kaart brengen van geïrrigeerde landbouw. Het levert een uitgebreide ruimtelijke dekking, maakt de monitoring van zowel temporele als ruimtelijke trends mogelijk, en ondersteunt bij het prioriteren van veldbezoeken. RS-gegevens kunnen consistent worden geanalyseerd in de loop van de tijd en zijn gemakkelijk toegankelijk. Door rekening te houden met factoren zoals het tijdstip van beeldopname, variaties in de kleur van de vegetatie en opvallende veranderingen, kunnen verschillende klassen van geïrrigeerde landbouw worden onderscheiden.

Deze dissertatie onderzoekt de productie van kaarten met RS en hun vermogen om geïrrigeerde landbouw weer te geven. Hoewel RS niet direct FLID kan meten, kan het wel de diverse en verspreide aard van kleinschalige geïrrigeerde landbouw vastleggen, waarvoor interpretatie vereist is door middel van veldwerk en lokale expertise. Het onderzoek identificeert en behandelt potentiële uitdagingen bij het in kaart brengen van geïrrigeerde landbouw in SSA met behulp van RS gegevens.

Deze studie maakt gebruik van vier casestudies in Mozambique, namelijk Chokwe, Xai-Xai, Manica en Catandica. Deze locaties zijn geselecteerd vanwege hun gevarieerde agro-ecologische kenmerken en de aanwezigheid van zowel kleinschalige als grootschalige geïrrigeerde landbouw.

In hoofdstuk 2 onderzoek ik de gemeenschappelijke stappen die alle karteringsstudies doorgaan bij de classificatie van RS kaarten. Ik heb een raamwerk ontwikkeld dat expliciet ingaat op en beoordeelt welke keuzes in modellen worden gemaakt, bestaande uit zeven stappen die typisch zijn voor classificatiestudies. Het doel van dit raamwerk is om de reproduceerbaarheid van resultaten tussen verschillende studies te evalueren. De belangrijkste bevindingen benadrukken twee cruciale punten. Ten eerste toont de studie aan hoe verschillende keuzes het classificatieproces beïnvloeden en systematiseert deze impact. Ten tweede brengt het een zorgwekkend gebrek aan rapportage aan het licht met betrekking tot acht essentiële keuzes. Het ontbreken van gedetailleerde verslaglegging over deze acht gebieden duidt op mogelijk onvoldoende bewustzijn bij de makers van de kaarten over het belang van hun methodologische keuzes voor een nauwkeurige definitie van de omvang van geïrrigeerde landbouw en reproduceerbaarheid. Hierdoor is het waarschijnlijk dat de geproduceerde kaarten een onderschatting geven van de volledige omvang van geïrrigeerde landbouw, met name die van kleine boeren.

In hoofdstuk 3 heb ik onderzocht hoe verschillende algoritmen en samengestelde lengtes de nauwkeurigheid beïnvloeden bij het voorspellen van geïrrigeerde landbouw in Mozambique. Composieten worden vaak gebruikt om wolkenvrije en ruimtelijk consistente beelden te genereren uit satelliettijdreeksen door samenvattende maten uit de tijdreeksen te aggregeren, zoals de gemiddelde pixelwaarde. Het maken van composieten op maand-, seizoens- of jaarbasis kan de vegetatiefenologie effectief vastleggen. In het bijzonder heb ik geëvalueerd hoe vier classificeerders (het random forest (RF), de support vector machine (SVM), kunstmatige neurale netwerken (ANN) en k-nearest neighbours (k-NN)) en vier samengestelde lengtes (1 × 12-maandelijks, 2 × 6-maandelijks, 4 × 3-maandelijks en 6 × 2-maandelijks) geïrrigeerde landbouw classificeerden. De resultaten worden gepresenteerd aan de hand van "akkoordkaarten" (*agreement maps*) die de consensus tussen de modellen illustreren met betrekking tot de classificatie van een gebied als geïrrigeerde landbouw of niet-geïrrigeerd. Deze kaarten benadrukken de aanwezigheid van kerngebieden van geïrrigeerde landbouw, bekend als hotspots, die een hoge mate van zekerheid vertonen. Rondom deze hotspots bevindt zich een onzekerheidszone waar de modellen minder overeenstemming vertonen. Deze kaarten kunnen de sterke punten van meerdere modellen combineren en de kans op fout-positieven (gebieden die ten onrechte geclassificeerd zijn als geïrrigeerde landbouw) verkleinen.

Ik ontdekte dat ANN, SVM en RF allemaal effectief presteerden bij het classificeren van geïrrigeerde gebieden. Er was echter geen enkel “beste” algoritme. Voor complexe en heterogene landschappen bleken kortere composieten geschikter te zijn, terwijl omgekeerd langere composieten voldoende waren voor meer uniforme landschappen. Veelbelovende opties, zoals 6-maands en 3-maands composieten, boden voordelen in termen van kortere rekentijd en gegevensomvang, terwijl nog steeds een hoge classificatienauwkeurigheid werd behouden. Mijn analyse toont aan dat het combineren van modellen met verschillende samengestelde lengtes en algoritmes in akkoordkaarten de nauwkeurigheid van het identificeren van geïrrigeerde landbouw verbetert.

Hoofdstuk 4 richt zich op de invloed van de omvang en samenstelling van trainingssteekproeven op de nauwkeurigheid van RS -classificatie voor het in kaart brengen van kleinschalige geïrrigeerde landbouw in SSA. Specifiek onderzoek ik het optimale aantal data, hun kwaliteit en het probleem van ongelijk verdeelde klassen. Het vergaren van uitgebreide trainingsdata van hoge kwaliteit is uitdagend vanwege beperkingen in tijd, toegang en interpreteerbaarheid. Hierdoor kan klasse-onbalans, waarbij bepaalde klassen meer vertegenwoordigd zijn in de trainingsgegevens, problemen veroorzaken bij het nauwkeurig classificeren van minderheidsklassen. De beschikbare steekproefgrootte kan de keuze van het algoritme beïnvloeden, aangezien sommige algoritmen een grotere dataset vereisen dan andere. Deze uitdagingen zijn vooral relevant in de context van kleinschalige geïrrigeerde landbouw, omdat deze vaak ondervertegenwoordigd is in datasets en beleidsdocumenten. Naast de grootte van de dataset kunnen vertekeningen in de trainingsgegevens de resultaten van de classificatie beïnvloeden. Deze vertekeningen kunnen voortkomen uit beperkte lokale kennis, onjuiste labels en het menselijke aspect van interpretatie.

De verschillende onderzochte scenario’s in hoofdstuk 4 laten zien dat over het algemeen grotere steekproeven de gebruikers- en producentennauwkeurigheid verbeteren, waarbij deze nauwkeurigheden klasse specifiek zijn en kunnen worden gebruikt om aan te tonen of een klasse over- of onderschat is. Er is echter een punt van afnemende meeropbrengst waar verdere verhogingen van de steekproefomvang de nauwkeurigheid slechts marginaal verbeteren en meer middelen vereisen. Het onderzoek toont ook aan dat modellen getraind op de regio Gaza over het algemeen beter presteren, wat duidt op een meer gegeneraliseerd model in vergelijking met de overaanpassing die werd waargenomen in de regio Manica. Met andere woorden, het Gaza-model was beter in staat om alle klassen zonder voorkeur voor afzonderlijke klassen te voorspellen. Het Manica-model gaf daarentegen meer de voorkeur aan geïrrigeerde landbouw dan aan andere klassen. Andere scenario’s benadrukken het belang van het verzamelen van representatieve veldgegevens en het gebruik van geschikte algoritmen, zoals RF en SVM, die minder gevoelig zijn voor specifieke kenmerken van de dataset in vergelijking met het ANN.

Hoofdstuk 5 onderzoekt of het overbrengen van modellen tussen regio's de modelprestaties kan verbeteren en middelen kan besparen in vergelijking met het verzamelen van nieuwe gegevens. Ik veronderstel dat gerichte datacollectie noodzakelijk is in het nieuwe gebied, aangezien de relaties tussen spectrale responsen en bodembedekkingen die in het ene gebied zijn geleerd, mogelijk niet van toepassing zijn vanwege variaties in weersomstandigheden, landschappen en landbouwpraktijken. In plaats van willekeurige gegevensverzameling heb ik me gericht op het identificeren van gebieden met hoge voorspellingsfouten om gerichte gegevensverzameling te sturen.

Verschillende modellen werden getraind op gegevens uit verschillende scenario's om de potentiële overdraagbaarheid van machinaal leren modellen voor het voorspellen van geïrrigeerde landbouw te onderzoeken. Uit het onderzoek bleek dat eenvoudige overdrachten van modellen niet effectief waren in het correct classificeren van nieuwe gebieden vanwege onvoldoende trainingsgegevens. Het toevoegen van meer diverse gegevens uit meerdere regio's verbeterde echter de classificatieprestaties. Het was dan ook geen verrassing dat de beste resultaten werden behaald wanneer alleen gegevens uit het doelgebied werden gebruikt, met uitsluiting van gegevens uit andere gebieden.

Concluderend kan worden gesteld dat het veld van op RS gebaseerde classificaties van landgebruik en bodembedekking gedemocratiseerd is door verschillende factoren. Onder andere de beschikbaarheid van open source software zoals QGIS en R, het open data beleid van organisaties zoals Landsat, MODIS en Sentinel, en de opkomst van cloud computing platforms zoals Google Earth Engine en Digital Earth Africa hebben hieraan bijgedragen. Daarnaast hebben online tutorials en platforms zoals GitHub RS-technieken toegankelijker gemaakt en op grote schaal verspreid. Deze toegankelijkheid heeft individuen en kleinere groepen die voorheen niet over de middelen beschikten om aan karteringsactiviteiten deel te nemen, meer mogelijkheden geboden.

De diversiteit aan methoden en onderzoeksdoelen die worden gebruikt bij het maken van deze kaarten vormt echter een uitdaging. Het is niet altijd duidelijk welke methoden wel of niet gebruikt moeten worden, waarover gerapporteerd moet worden en hoe de resultaten geëxtrapoleerd kunnen worden naar andere gevallen. De resultaten van dit onderzoek hebben implicaties voor het documenteren en rapporteren van methoden en keuzes, het presenteren van geïrrigeerde landbouw door middel van kaarten, en het laten zien hoe gemakkelijk het is om die kaarten te manipuleren met kleine aanpassingen aan modellen.

List of publications

Weitkamp, T., Veldwisch, G. J., Karimi, P., & de Fraiture, C. (2023). Mapping irrigated agriculture in fragmented landscapes of sub-Saharan Africa: An examination of algorithm and composite length effectiveness. International Journal of Applied Earth Observation and Geoinformation, 122, 103418. https://doi.org/10.1016/j.jag.2023.103418

Weitkamp, T., & Karimi, P. (2023). Evaluating the Effect of Training Data Size and Composition on the Accuracy of Smallholder Irrigated Agriculture Mapping in Mozambique Using Remote Sensing and Machine Learning Algorithms. Remote Sensing, 15(12), 3017. https://doi.org/10.3390/rs15123017

T. Weitkamp and P. Karimi (2023). The generalisation of machine learning models for predicting irrigated agriculture in heterogeneous landscapes: an examination of model transferability in Mozambique. International Journal of Applied Earth Observation and Geoinformation (*Under review*)

Acknowledgements

I did not take this PhD journey alone; many people have helped me along the way in different ways which made the journey much more enjoyable than if I did it alone. In no particular order, I want to thank you all!

I want to start by thanking my promotor: Prof. Charlotte de Fraiture. Thanks for allowing me to continue my journey in the form of a PhD on a topic that motivates me a lot. I have learned much from questioning and digging deeper into a topic, as well as from your experiences with irrigated agriculture. I am also grateful for my co-promotor Gert Jan Veldwisch. As we both worked in Wageningen, you were always available for a quick discussion over a coffee or a walk over campus or the uiterwaarden. Your experiences in Mozambique helped me get additional insights for my own study. I also want to thank my co-promotor, Poolad Karimi. During this PhD journey, new opportunities arose for you in Washington, meaning most of our discussions happened online. Nevertheless, your remote sensing experience helped me focus, as I quickly found out that the opportunities seem endless. I enjoyed our discussions in which I also got a glimpse of what happens 'in the real world'. I would like to thank all three of you for teaching me the skills of critical thinking and writing structured and scientifically, as well as the many facets of smallholder irrigation and mapping it with satellite data.

This research would not have been possible if it was not for the FASIMO project and my many team members from INIR, ISPM and ISPG: Mario, Andre, Wilson, Emilio, Lateiro, Lasse, Hélder, Eugénio, Elton, Daniel, Moisês, Nascimento, Zivale, Banze. Your involvement allowed me to test ideas and concepts, learn from on-the-ground experiences and get a glimpse of how remote sensing can be applied and how it is perceived outside of the academic world.

I also want to thank my many colleagues spread around the world, Gundana, Nelson, Geronimo, Katrien, Olga, Jelle, Júbia, Cremilde, Piloto, Namurove, Simone; Thijs, Thijs, Thijs (you fight who is who ;)), Elna, Eveline, Annemieke, Abigail, Emily, Zoé, Frederique; Titus (¡¡PhD!!), Susanna, Miriam, Anniek, Sander, and of course the four+1 people who keep it all running smoothly: Wouter, Gert Jan, Guido, Joep, and Daniel. Your experiences and stories kept me motivated to get it finished in 4 years, and you gave me the trust and energy to explore remote sensing in different projects and services. Although I did a PhD, the external PhD construction allowed me to apply what I learned into practice, learning more about real-world requirements and putting me back on the ground with both feet when stuff got too technical. Many thanks also to my former and current (PhD) colleagues at the Water Resources Management group for the coffee discussions during breaks and writing retreats, allowing me to put my study and journey into perspective: Matthijs, Rozemarijn, Lena, Gerlo, Wahib, Zerihun, Xiulu, Melle, Lena, Hanan, and staff members whom most I have known since the start of my academic career in 2011; Pieter, Henk, Petra, Bert, Harm, Jeroen, Gerardo, Jaime, Boukje, Maartje, and Natalja.

I would not have come far if it was not for friends and family. You got me to translate the deep-level work to more *jip and janneke* language, allowing me to get new insights as well. Thanks *bill batties*, Peter, Jasper, and Bas, we started our university journey together in the bachelor BIL in 2011 and we have followed similar paths and interests ever since. We know how to keep each other motivated. I also want to thank Dominus: Julian, Michal, Michael, Ian, Alex, Jelle & Peter. We also started together in 2011 and experienced a lot. Our talks on *warmptepompen* and *zonnepanelen* kept me focused on the journey after my PhD. Work has to be balanced with fun – which I did a lot with box lacrosse. I want to thank you for giving me a place and time to vent over the last few years: Lars, Hidde, Iepe, Davie and the Dutch team I've played with since the world championships in 2019. Another way to vent is with games and good food, and for this, I want to thank my *buren* Tom and Annemerel; thanks for the pre-, during-and post-lockdown dinner and game nights.

As I'm nearing the end, I want to thank my family for their support as well; Tirza, Yvonne, Gert, Manuela, Lizetta, Jeroen, Britt, Jeff, Ria, Gerard, Tonny, Karin, Karel, Myrte, Linus, Nardy and Olaf. Although most of what I did was probably magic to you, you still had a listening ear. And where would I have been without the unquestioned support of Nala, who let me have many breaks to play but also to walk and think about ideas. The support team expanded with Pip and later on with Pumba, also giving me ample opportunities to give them attention instead of my own work.

Last but certainly not least, a heartfelt appreciation to Kim for steadfastly supporting me over the past four years, holding down the fort during my data collection in Mozambique, and providing the space for me to delve into this meaningful topic. I aspire to reciprocate the same opportunity and support in return.

Sense Education Certificate

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

D I P L O M A

for specialised PhD training

The Netherlands research school for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Timon Karel Bernard Weitkamp

born on the 2nd of February 1993 in Monze, Zambia

has successfully fulfilled all requirements of the
educational PhD programme of SENSE.

Wageningen, 2 February 2024

Chair of the SENSE board

Prof. dr. Martin Wassen

The SENSE Director

Prof. Philipp Pattberg

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Timon Karel Bernard Weitkamp** has successfully fulfilled all requirements of the educational PhD programme of SENSE with a work load of 38.2 EC, including the following activities:

SENSE PhD Courses

- Environmental research in context (2020)
- Research in context activity: Training on mapping smallholder irrigation in Mozambique (2022)

Other PhD and Advanced MSc Courses

- Basic statistics, PE&RC and SENSE (2020)
- Critical thinking and argumentation, Wageningen Graduate Schools (2020)
- Scientific writing, Wageningen Graduate Schools (2021)
- Bayesian statistics, PE&RC and SENSE (2020)
- Uncertainty Propagation in Spatial Environmental Modelling, PE&RC (2020)
- Introduction to machine learning, PE&RC (2021)

Management and Didactic Skills Training

- Supervising 2 MSc students with thesis titled 'Mapping irrigated agriculture in a complex landscape 'and 'What is the potential of using Sentinel-1 and Sentinel-2 data to map farmer-led irrigated agriculture with machine learning?' (2021)
- Supervising BSc student with thesis titled 'The potential of providing location specific feedback on crop growth factors of sugarcane using Sentinel 2 remote sensing data'(2021)

Oral Presentations

- *Field data collection for remote sensing-based mapping.* FASIMO, March 2020, Chokwe & Chimoio, Mozambique
- *SmartCane.* Kick-Start project finalisation for European Space Agency and Netherlands Space Office. May 2021, online.
- *Maize monitoring from space.* Netherlands Space Office project evaluation, May 2022, The Hague, Netherlands.
- *How does crop monitoring work for you? SASPEN project, June 2022, Wageningen, The Netherlands*

SENSE coordinator PhD education

Dr. ir. Peter Vermeulen

The research described in this thesis was financially supported by the International Development Research Centre (IDRS), grant number (PROJECT ID) 109039 and Resilience B.V.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design: Printservice Ede
Layout and printing: Printservice Ede