**Masterarbeit**
Zur Erlangung des akademischen Grades Master of Science Global Change Geography

# Mapping cropped area in the Rabi season of Andhra Pradesh between 2017 and 2019 using training data generated through the agtech platform Plantix

eingereicht von: Lara Schmitt

Gutachter: Dr. Philippe Rufin
Prof. Dr. Patrick Hostert

Eingereicht am Geographischen Institut der Humboldt-Universität zu Berlin am: 28.02.2020

## Contents

## List of figures



## List of tables

## List of abbreviations

| | |
|---|---|
| ANN | Artificial neural network |
| AP | Andhra Pradesh |
| API | Application programming interface |
| APSDPS | Andhra Pradesh State Development Planning Society |
| AOI | Area of interest |
| CI | Confidence interval |
| DES | Directorate of Economics and Statistics |
| DLR | German Aerospace Center |
| DNN | Deep neural network |
| FNN | Feedforward neural network |
| G | Gatherix |
| GEE | Google Earth Engine |
| GDP | Gross domestic product |
| GSDP | Gross state domestic product |
| GUF | Global Urban Footprint |
| HRI | High resolution imagery |
| ICAR | Indian Council of Agricultural Research |
| IMD | India Meteorological Department |
| ISMR | Indian summer monsoon rainfall |
| Jan | January |
| JRC | Joint Research Centre - European Commission |
| OA | Overall accuracy |
| OSM | Open Street Map |
| PEAT | Progressive Environmental and Agricultural Technologies |
| P | Plantix |
| PA | Producer's accuracy |
| SWM | South west monsoon |
| STM | Spectral-temporal metric |
| UIDAI | Unique Identification Authority of India |
| USGS | United States Geological Survey |
| UA | User's accuracy |
| VHR | Very high resolution |

# Mapping cropped area in the Rabi season of Andhra Pradesh between 2017 and 2019 using training data generated through the agtech platform Plantix

**Abstract**:

The agricultural sector of India represents an essential element in the challenge of ensuring global agricultural production and the farmer's livelihoods under changing weather patterns and growing population. About 70% of India's rural households still depend primarily on agriculture for their livelihood, with 82% being small and marginal farmers. However, detailed knowledge about the crop production of smallholders in India is rare. In order to fill these existing knowledge gaps, remote-sensing based information on cropped area in the different agricultural seasons could be a valuable auxiliary resource. The objective of this study was to map stable cropped area and its share of rice cropped area in India's federal state of Andhra Pradesh for the Rabi (dry) season of the years 2017 - 2019. Accounting for the large spatial and temporal variability of the onset of the Rabi season in Andhra Pradesh, a restricted timeframe (1 January to 31 March) was applied in order to cover the common main growing period of Rabi crops in Andhra Pradesh. Traditional methods to collect the required ground truth data are expensive and time-consuming. The potential to overcome these disadvantages of traditional data collection methods could lie in crowdsourced data collection methods as they provide large amounts of data at comparatively low cost. As ground truth data for cropped area crowdsourced data provided by the agricultural-technology startup PEAT was utilized. The company receives large amounts of geotagged and timestamped image data of several crop varieties on a daily basis through the automatic image recognition feature of their smartphone application Plantix. To limit the PEAT data to a suitable reference dataset, the size of the raw dataset was substantially reduced by applying several spatial and thematic filters. Nine different subsets of the PEAT data were generated in order to conduct a sensitivity analysis. Ancillary data for the remaining classes forming non-cropped area was obtained through several global datasets for woody canopy (Hansen et al. 2013), water (Pekel et al. 2016) and urban areas (Felbier et al. 2014) as well as manually collected in Google Earth Pro for unsown cropland, bare land and areas of sparse natural vegetation. I used Landsat 7 and Landsat 8 imagery to compute spectral-temporal metrics as basis for the classification utilizing Random Forest classifiers. The result for aggregated stable cropped area within the selected timeframe (5.7 Mha) was not in agreement with officially reported numbers for Rabi cropped area in 2017 - 2019 (1.7 - 2.06 Mha) revealing the unsuitable approach to map stable Rabi cropped area in the entire study region. The determined shares of rice cropped area were in better agreement with the official statistics with a deviation of +7.1%. The accuracy assessment of the individual maps shows the appropriate usability of extensively filtered crowdsourced PEAT data as ground truth data for cropped area with user's accuracies exceeding > 90%.

**Keywords:** agriculture; India; Andhra Pradesh; drought; remote sensing; spectral-temporal-metrics; crowdsourced data; deep neural network; Plantix; Progressive Environmental and Agricultural Technologies; PEAT; smallholder farmers

## 1 Introduction

One of the world's biggest concerns is that climate change will endanger global agricultural and food production. How big a threat is climate change to the climate-sensitive and economically important sector of agriculture in developing countries? How well will farmers be able to adapt to the consequences of climate change?

Smallholder farms play a crucial role in this issue as they make up 84% of the world's 570 million farms (Lowder et al. 2016). About two-thirds of the developing world's 3 billion rural population live in small farm households with land plots less than 2 hectares (Rapsomanikis 2015). They are responsible for more than half of the food calories produced globally (Samberg et al. 2016). Smallholder farmers are particularly vulnerable to changing environmental conditions induced by climate change because they usually do not have access to appropriate technologies to mitigate vulnerability such as crop insurance or capital for improved seed stock (Lobell et al. 2008).

About 24% of the world's farms are located in India making its agriculture sector an essential element

in the challenge of ensuring global food security under changing weather patterns and growing population (Lowder et al. 2016). In the past five years, India has experienced below normal monsoon rainfall levels and at the same time an increase in extreme rainfall events. The small and marginal farmers have particularly experienced crop loss to due erratic rainfall and drought-like situations (Basha 2018). A comprehensive understanding of the agricultural dynamics in India in response to changing climatic conditions is therefore critical for ensuring the farmer's livelihood as well as ensuring global food production. To be able to conduct prospective policy interventions and develop strategies for smallholder farmers to cope with the changing environmental conditions, it is of high importance to gain more detailed knowledge about cropping patterns of smallholder farmers.

Having the objective of filling the existing knowledge gaps on smallholder crop production in India, remote-sensing based information could be a valuable auxiliary resource. Remote sensing data is offered at low-cost, covering large areas. More importantly, it can be continuously updated in comparison to survey data. Various studies have been conducted aiming at identifying cropping practices such as crop type and cropping intensity over the last several decades. A variety of studies have utilized high temporal-resolution data like MODIS to assess crop type and cropping intensity based on crop phenologies (Galford et al. 2008; Chang et al. 2007; Löw et al. 2018; Shao et al. 2016). Using the unique temporal signature of crops is a powerful technique given the fact that it is typically difficult to differentiate crops based on solely spectral signatures. For instance, rice can be accurately distinguished from other monsoon crops in South Asia using the unique phenological signature of field flooding and rice transplanting in the beginning of the growing season (Sakamoto et al. 2009). While the studies utilizing MODIS data were able to assess crop type and cropping intensity with high classification accuracies, these studies have been only applied in regions with large farm plots (Jain et al. 2013). In order to capture smallholder cropland extent, remote sensing imagery with a higher spatial resolution needs to be utilized. Landsat images having a spatial resolution of 30 m is more similar in size to single smallholder plot. However, while Landsat imagery offers higher spatial resolution over MODIS, it is possible that Landsat data lacks the necessary temporal resolution for cropped area mapping.

In order to make use of remote-sensing based information, suitable reference data is needed to accurately map agricultural land use. Traditional methods of ground truth data collection are expensive and time-consuming. The potential to overcome these disadvantages of traditional data collection methods could lie in crowdsourced data collection methods. They provide large amounts of data at comparatively low cost. In recent years, crowdsourced data has gained importance as a consequence of the interactivity of Web 2.0 and the fast proliferation of GPS enabled mobile smartphones all over the world. Numerous research fields have utilized crowdsourced data in recent years such as public health (Paul and Drezde 2012), environmental data acquisition (Fienen and Lowry 2012), clinical research (Chandler and Shapiro 2016), protective area management (Levin et al. 2017; Walden-Schreiner et al. 2018) and transportation planning (Jestico et al. 2016). More recently, crowdsourced data has also gained importance in the remote sensing context. Various studies explored the potential of the integration of crowdsourced data in remote sensing applications. Salk et al. (2016) evaluated the quality of volunteer's contributions who assessed 165,000 satellite images for the presence of cropland. The different facets of how crowdsourcing and citizen science impact upon the validation, use and enhancement of observations from satellites products and services was further investigated by Mazumdar et al. (2017). Johnson and Iizuka (2016) explored the potential of training data automatically extracted from OpenStreepMap. Fritz et al. (2017) derived global land cover and land use data from the Geo-Wiki crowdsourcing platform in order to make them available for the scientific community to be utilized as reference data for global satellite-derived products. Further very recent studies made first attempts to utilize crowdsourced data in the remote sensing context. Lesiv et al. (2019) presented an unique approach to quantify and map agricultural field size globally using crowdsourced information collected through a campaign where participants visually interpreted 130,000 very high resolution satellite imagery using the Geo-Wiki platform. Panteras and Cervone (2018) utilized Twitter data to enhance the temporal resolution of satellite-based flood extent generation. Herfort et al. (2019) revealed that crowdsourcing combined with deep learning outperforms existing remote-sensing based approaches to map human settlements.

In this study, I intended use crowdsourced data provided by the agricultural-technology startup 'Progressive Environmental and Agricultural Technologies' (PEAT) as ground truth data for cropped area in order to map stable cropped area in the Rabi (dry) season, lasting from November to April, in India's federal state of Andhra Pradesh. PEAT receives large amounts of geotagged and timestamped crowdsourced data on a daily basis through their Android smartphone application Plantix. The application, which is primarily used by Indian farmers, provides an automatic image recognition feature for crop diseases based on a deep neural network. The outstanding benefit of the PEAT dataset is that it provides continuously updated information opening the chance for continuous monitoring of cropped area extent of different agricultural seasons. However, the use of crowdsourced data poses various challenges as it contains a considerable amount of noise, usually represents a biased sample and often lacks the quality standards of traditional data collection measures (Senaratne et al. 2017). To limit the PEAT data to a suitable reference dataset, I reduced the size of the raw dataset substantially by applying several spatial and thematic filters. Ancillary data for the remaining classes forming non-cropped area was obtained through several global datasets for woody canopy (Hansen et al. 2013), water (Pekel et al. 2016) and urban areas (Felbier et al. 2014) as well as manually collected in Google Earth Pro for unsown cropland, bareland and areas of sparse natural vegetation.
Until now, any studies have been conducted aiming at mapping cropland extent of the Rabi season in Andhra Pradesh. Only the study from Jain et al. (2016) utilizing the MODIS Enhanced Vegetation Index mapped the India annual winter cropped area (growing season from October to March). The mapping of Rabi cropped area in Andhra Pradesh constitutes a complex objective as the onset of the agricultural seasons in Andhra Pradesh differ spatially as the sowing dates and therefore harvesting dates of the Kharif (monsoon) crops are directly connected to the onset of the monsoon rains which arrive gradually in the different region of Andhra Pradesh. The time period of the Kharif harvest then in turn determines the sowing dates of the following Rabi crops in Andhra Pradesh. Thus, having a north-south extension of 729 km and covering different climatic subregions, the sowing dates of the rainfed Rabi crops in Andhra Pradesh vary in a range of several weeks. However, a considerable share (36%) of the cultivated area in Andhra Pradesh is irrigated. Thus, the different access to irrigation water potentially determines the distribution of Rabi cropped area in Andhra Pradesh to a great extent.
My approach to capture the Rabi cropped area in Andhra Pradesh is to include the years 2017 - 2019 as being covered by PEAT data and to limit the timeframe from 1 January to 31 March for each year in order to cover the main growing period of the Rabi crops. The remotely sensed imagery used is Landsat 7 and Landsat 8 imagery. The classification results for the cropped area per year are aggregated to obtain the stable cropped area within the selected timeframe assuming this approach captures the actual Rabi cropped area.
Rice represents the major food crop grown in Andhra Pradesh. The rice cropped area in the Rabi season accounts for about 40% of the total Rabi cropped area (Government of Andhra Pradesh 2019) and therefore constitutes the predominant Plantix submission crop variety during the season. Hence, I further examined the usability of rice Plantix submissions as ground truth data in order to map rice cropped area within the previously classified cropped area. Various previous efforts have been made in order to map time series of paddy rice extent using Landsat time using different approaches, e.g. combined with phenology-based algorithms for mapping rice expansion in China exceeding overall accuracies of 95% (Dong et al. 2015) or for the Mekong delta across multiple growing seasons aiming at differentiating between single-, double-, and triple cropped fields (Kontgis et al. 2015). However, this study aims at mapping paddy rice extent for the narrow time frame of just one agricultural season.

My specific research questions are:

1. Is the PEAT database a valid source of ground truth data for binary cropped area mapping and binary rice/non-rice cropped area mapping?
2. Is mapping of stable Rabi cropped area in Andhra Pradesh feasible utilizing the aggregated recurrent cropped area within the timeframe 1 January - 31 March of the years 2017 - 2019?
3. What is the share of rice cropped area of the Rabi cropped area in Andhra Pradesh?

## 2 Study Area

### *2.1 General geographic, administrative and climate characteristics*

My study area is the federal state of Andhra Pradesh where the Plantix application received the largest number of submissions in the Rabi season between the years 2017 and 2019 (PEAT 2020). It lies between 12°41' and 19°07' N latitude and 77° and 84°40'E longitude at the southeastern coast of India with a 974 m long coastline bordering the Bay of Bengal (Figure 1). It is bordered by Telangana, Chhattisgarh, and Orissa in the north, Tamil Nadu to the south and Karnataka to the west. Two major rivers, the Godavari and the Krishna run across the state. Andhra Pradesh is India's seventh largest state covering an area of 162,968 km² (Ministry of Commerce and Industry 2019). Until 2014, Andhra Pradesh comprised a larger area with the region of Telangana belonging to the state of Andhra Pradesh. On 2 June 2014 the state was split. The north-western portion comprising 40% of the former area was carved out to create the new state Telangana after prolonged protests by residents of Telangana who felt the under-developed region had long been neglected by the government of Andhra Pradesh (Telangana State Portal 2019). The population of the present-day Andhra Pradesh is 53.61 million; the population density is 308 persons/km² (UIDAI 2019). The state can be divided in the two main regions of Rayalaseema and Coastal Andhra Pradesh. Andhra Pradesh consists of 13 official districts with population sizes varying between 2,699,471 (Srikakulam) and 4,083,31 (Anantapur). Nine of districts are located in Coastal Andhra Pradesh and the other four in Rayalaseema (Figure 1).

The climate is semi-arid tropical with temperatures varying from 15° to 45 °C. The average temperature in the cooler months of December and January is 24 °C, in the summer months of May and June the average temperature is 33°C (APSDPS 2019). Most of the precipitation falls within July to September (68.5% of total annual rainfall) during the south-west monsoon (summer monsoon). The monsoon covers the entire state and provides 66% of the annual rainfall for most of the districts. The north-east monsoon (winter monsoon) which brings precipitation only to the southern part of India's peninsula usually enters Andhra Pradesh in October and last until December. It provides 24% of the Andhra Pradesh's annual precipitation, however, only its southern districts receive rainfall from the north-east monsoon (DES 2019).

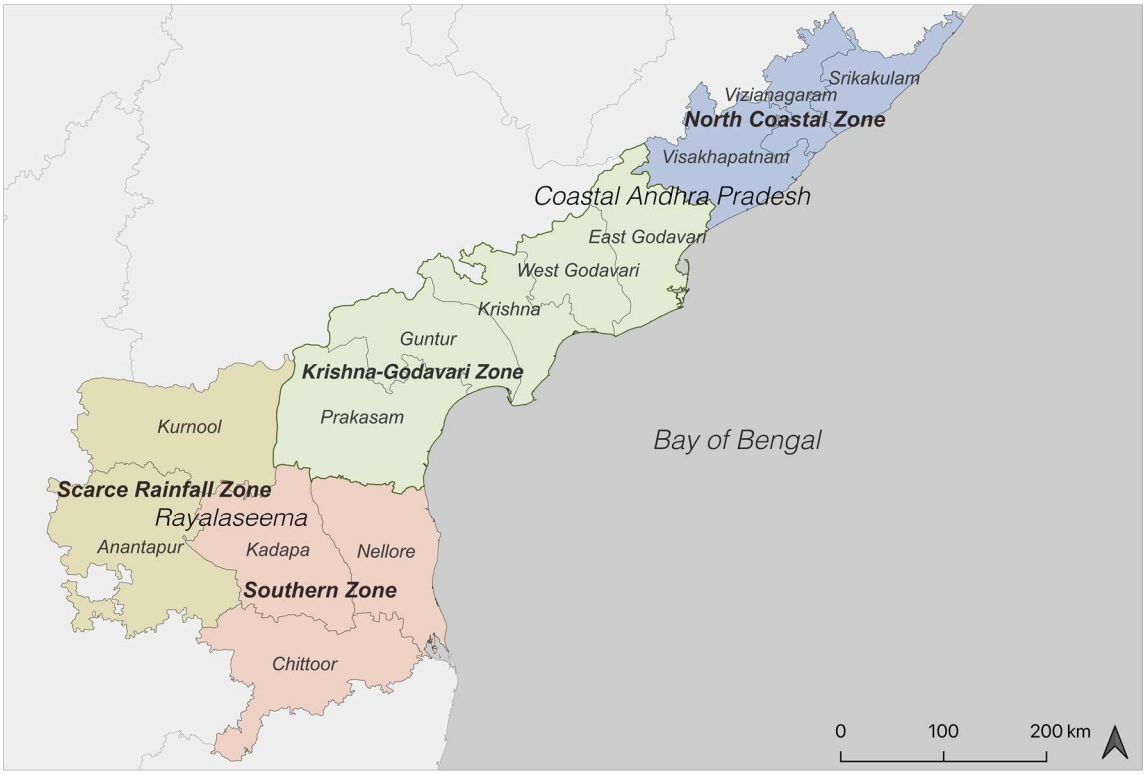



Figure 1: Map of Andhra Pradesh, showing the division in Rayalaseema and Coastal Andhra Pradesh, agro-climatic zones and federal districts. Source: Department of Agriculture, Andhra Pradesh.

## 2.2 *Agricultural characteristics*

Andhra Pradesh is one of the most important agrarian states in India. The state is an agro-based economy with agriculture and allied sectors contributing more than 29% of the Gross State Domestic Product (GSDP) as against 17% of India's entire GDP (Ministry of Commerce and Industry 2019). In 2017, the Ministry of Agriculture reported that 38.1% of Andhra Pradesh's area is under cultivation, while another 22.9% is forest (Forest survey of India 2017).The major food crop grown in Andhra Pradesh is rice (Figure 2), constituting for approximately 77% of the total food grain production which amounts to approximately 7% of the total state GDP (DES 2018).

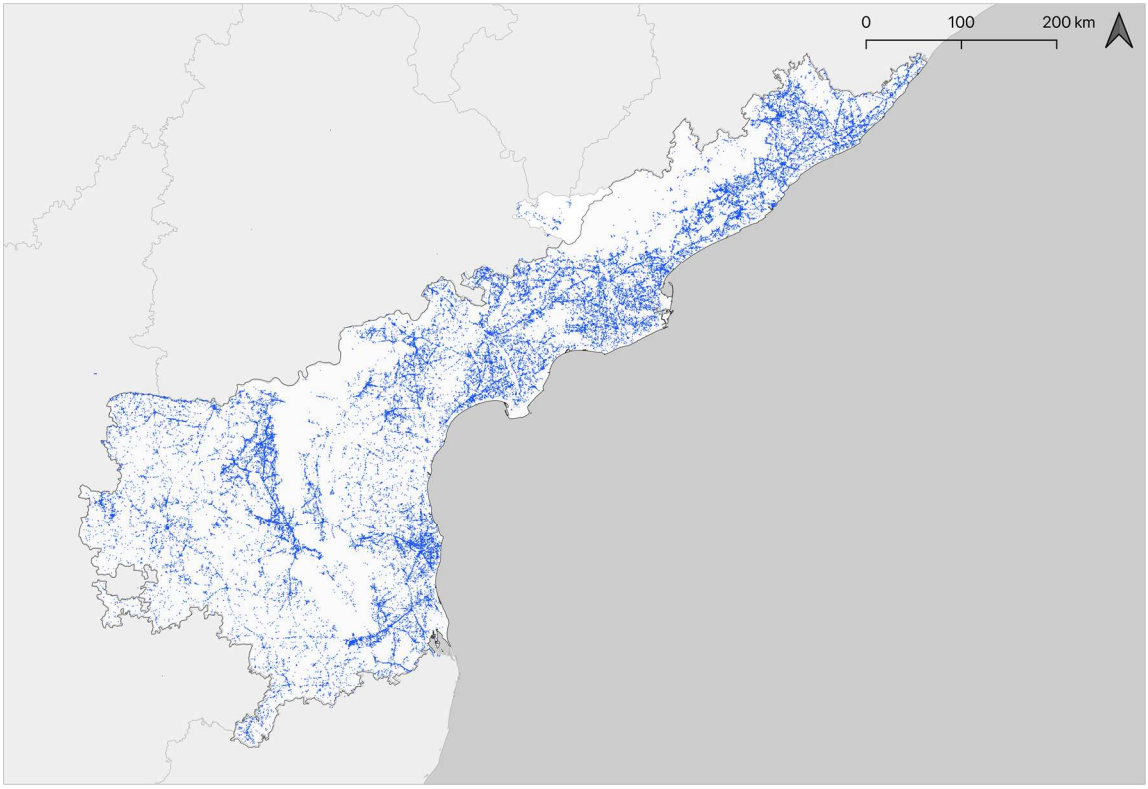



Figure 2: Distribution of Plantix Submissions (2017-2019) in Andhra Pradesh detected as rice by the DNN.
Source: PEAT 2020.

Like most parts of India, Andhra Pradesh has two main cropping seasons: Kharif, lasting from June to November and Rabi, lasting from November to April. The terms originate from Arabic language where Kharif means autumn and Rabi means spring. Kharif crops are grown with rainwater from the southwest monsoon and the Kharif represents the major growing season. The Kharif crops are usually sown with the beginning of the first monsoon rains around June and harvested in September or October. The Rabi crops are sown around November and harvesting happens in springtime (March - April) as indicated by the name. The Rabi crops in Andhra Pradesh are grown mainly with rainwater that has percolated into the ground or are irrigated. Solely the farms in the southern parts of Andhra Pradesh benefit from additional precipitation during the Rabi season from the north-east monsoon (October-December). Paddy rice, cotton and peanut account for over 70% of Andhra Pradesh's cropped area in Kharif, while paddy rice, chickpea and black gram are the dominant Rabi crops (Ministry of Agriculture and Farmers' Welfare 2017).

Andhra Pradesh can be categorized into four agro-climatic zones (Table 1) based on the precipitation amount as well as type and topography of the soils (Department of Agriculture 2015).

Table 1: Agro-climatic zones in Andhra Pradesh with respective districts, precipitation during south west monsoon, temperature, soil types and crops grown. Source: Department of Agriculture.

| Name | Districts | PPT | Temp. | Soil Type | Crops Grown |
|---|---|---|---|---|---|
| **North Coastal Zone** | Srikakulam, Vizianagaram, Visakhapatnam, uplands of East Godavari | SWM: 1000-1100 mm | Min: 26-27 °C<br>Max: 33-36 °C | Red soils with clay base | Paddy rice, groundnut, jowar (sorghum), bajra (pearl millet), raw jute (jute & mesta), sesame, black gram, horticultural crops |
| **Krishna-Godavari Zone** | East & west Godavari, Krishna, Guntur, contiguous areas of Khammam, Nalgonda and Prakasam | SWM: 800-1100 mm | Min: 23-24 °C<br>Max: 32-36 °C | Deltaic alluvium, red soils with clay, red loams, coastal sands & saline soils | Paddy rice, groundnut, jowar, bajra, tobacco, cotton, capsicum, sugarcane, horticultural crops |
| **Southern Zone** | Nellore, Chittoor, southern parts of Prakasam and Kadapa, eastern parts of Anantapur | SWM: 700-1100 mm | Min: 23-25 °C<br>Max: 33-46 °C | Red loamy soils, shallow to moderately deep | Paddy rice, groundnut, cotton, sugarcane, millets, horticultural crops |
| **Scare Rainfall Zone** | Kurnool, Anantapur, west Prakasam, north Kadapa | SWM: 500-700 mm | Min: 24-30 °C<br>Max: 32-36 °C | Red earths with loamy soils, red sandy soils | Cotton, korra (foxtail millet), jowar, other millets, groundnut, legumes, paddy rice |

Andhra Pradesh has one of the largest irrigated areas in India. The gross irrigated area is 6.28 Mha which is 36% of the cultivated area. It accounts for nearly 7.3% of the total irrigation area in India. The major source of irrigation water is groundwater, with nearly 49% of the net irrigation coming from wells and tube wells. The rest of the irrigation water is derived from sources such as canal-based systems and tanks. The increase in irrigated agriculture has been essential to the state's economic development and poverty reduction. The irrigation expansion is considered to have played a major role in the state's fast agricultural growth over the last three decades assuring agricultural profitability in the dry regime. Most of the state's area is underlain by hard rock aquifers with very poor storage. Thus, most regions of Andhra Pradesh do not provide an advantageous environment for intensive use of groundwater resources. However, limited access to surface irrigation systems makes farmers resort to well irrigation through open wells (Prasuna et al. 2018).

### *2.3 Threat of increased drought frequency*

Andhra Pradesh's climate, as the entire Asian monsoon climate, is significantly dominated by the Indian Summer Monsoon Rainfall (ISMR). Its importance for agricultural production, water availability, and food security is well-documented (Wahl and Morrill 2010). The ISMR averaged over the entire of India is remarkably steady from year to year, with a variation coefficient of only 10%. However, even these small variations strongly affect agricultural production (Mishra et al. 2012; Vinnarasi and Dhanya 2016). The production of the Kharif crops directly depends on the ISMR, whereas the production of the Rabi crops depends on the soil moisture availability, which in turn depends entirely on the ISMR in most regions of India (Prasanna 2014). While any significant deviation from the ISMR average, even a well-predicted one, will negatively affect the economy, the impact of climate change further exacerbates these effects. Numerous studies conclude that climate change has a significant effect on the interannual and multidecadal variability of the monsoon - despite having regard to its year-to-year variability as well as the El Niño-monsoon relationship to climate change (Dash et al. 2011; Ghosh et al. 2012; Singh et al. 2014). Observations reveal that ISMR, particularly over central and north India, has experienced a statistically significant weakening since the 1950s (Naidu et al. 2015; Quesada et al. 2017). Simultaneously, the frequency and intensity of monsoon droughts (Roxy and Chaithra 2018) as well as the frequency of heavy and extreme rains (Singh et al. 2019) have increased. In recent years, a delay in the onset of the monsoon over India has been observed (Loo et al. 2015; Sahana et al. 2015).

In semi-arid regions such as Andhra Pradesh, droughts are a natural occurrence. However, as their frequency and intensity are expected to increase with climate change, this threatens agricultural production and thus, the economic and social situation of the farmers. In recent years, several regions of

Andhra Pradesh have been afflicted by severely to extremely dry conditions during the Rabi season. The rainfall statistics for the crop year 2015 - 2016 still showed satisfactory seasonal conditions with an overall deficit of -3% (912.5 mm as against the normal of 966.0 mm) being in the normal deviation range. In detail, the rainfall deficit during the south west monsoon (June to September) was -5.9% over normal and -3.0% during the north east monsoon (October to December) (Department of Agriculture 2016). However, the following three years were characterized by alarmingly dry conditions. The Department of Agriculture reported an overall deficit of 29.9% over normal rainfall during the crop year 2016 - 2017 with a -4% deviation during the south west monsoon and -71.2% deviation during the north east monsoon. In 2017 - 2018 the rainfall deviation during the south west monsoon was 2% and -40% during the north east monsoon (Department of Agriculture 2018). In 2018-2019 the reported overall rainfall deficit was -32% divided up in -18% deviation during the south west monsoon and -58% during the north east monsoon (Department of Agriculture 2019). In particular, farmers in the south-eastern states of Andhra Pradesh who do not have access to irrigation water, hence rely on winter rainfalls during the north east monsoon, experienced severe drought conditions due to the described high rainfall deficits in the past three Rabi seasons. However, the drought also aggravates the aforementioned problem of diminishing groundwater levels as the precipitation amount is not sufficient to replenish the groundwater reservoirs.

### *2.4 Current situation of smallholder farmers*

The major proportion, approximately 80% of Andhra Pradesh's farmers are small and marginal farmers (Basha 2018). Their present economic situation and social situation is concerning. In the last two decades, an increased frequency of suicides among farmers has been documented. Several studies investigated the reasons which are driving farmers to commit suicide. According to Vaddiraju (2004) the reasons are economic and ecological. The economic causes are fluctuating prices and inadequate marketability of the yields and the dependency on private moneylenders and non-institutional credits that the farmers could not pay back after a failure of yields. Another study conducted by Tada (2004) found that the lack of minimum support price and its execution as well as the lack of crop insurance schemes are further reasons. Overall, the main issue is that the fixed market rent for leased land has to be paid to the landholder in many parts of Andhra Pradesh regardless of whether the tenant was able to achieve a succesful harvest (Tada 2004). The ecological reasons are an increase in the frequency of droughts and the shift to intensive farming. The shift from multiple cropping to mono-cropping of crops such as cotton and capsicum increased the vulnerability of farmers. Moreover, falling groundwater levels in many parts of the country have led to a hydrological crisis being a subset of the agricultural-ecological challenges Andhra Pradesh's farmers are facing today (Vaddiraju 2004). A World Bank report from 2008 containing a case-study about climate variability in two arid sectors of Andhra Pradesh, Anantapur and Chittoor, reinforces this assessment of an amplifying ecological problem in Andhra Pradesh. Based on their evaluation, the negative effects of climate change, especially increased frequency of droughts and floods may lead to substantial decline in farm income in the districts. They identified the farmers of the districts as highly vulnerable to further climate variability because of their limited adaptive capacity, due to low incomes and restricted alternative employment opportunities (World Bank 2008). The fact that several districts of Andhra Pradesh have experienced a high frequency of droughts in recent years (IMD 2019) aggravates the current critical situation of the small-scale farmers and decreases their resilience.

## 3 Material and Methods

### *3.1 Data cropped area*

#### 3.1.1 Crowdsourced data generated through the smartphone app Plantix

Crowdsourcing data collection consists of building datasets with the help of a large group of people. With the proliferation of GPS enabled mobile smartphones crowd-contributed data has become a powerful source for several research fields (Paul and Drezde 2012; Fienen and Lowry 2012; Jestico et al. 2016). Traditional methods of data collection are expensive, time consuming, and lack spatial and temporal detail whereas the outstanding benefit of crowdsourced data is that it can be

collected at comparatively low cost.

The agricultural-technology startup PEAT receives large amounts of crowdsourced geotagged and timestamped data through the automatic image recognition feature of their Android smartphone application Plantix. PEAT launched its application in 2016 with the goal of empowering farmers and growers worldwide to fight crop shortfalls due to plant diseases. Plantix offers besides the automatic image recognition of crop diseases, a crop advisory, a community feature and a weather forecast. The major user base of the application (72% of all users worldwide) lies in India (Figure 3). The remaining users are mainly located in Bangladesh (5%), Pakistan (5%) and Egypt (4%). Plantix has been installed 10,3 million times by October 2019 and its monthly active userbase consist of roughly one million users per month. PEATS's database currently consists of over 20 million image submissions. About 75% of all Plantix submissions are submitted with geolocation information (PEAT 2020).

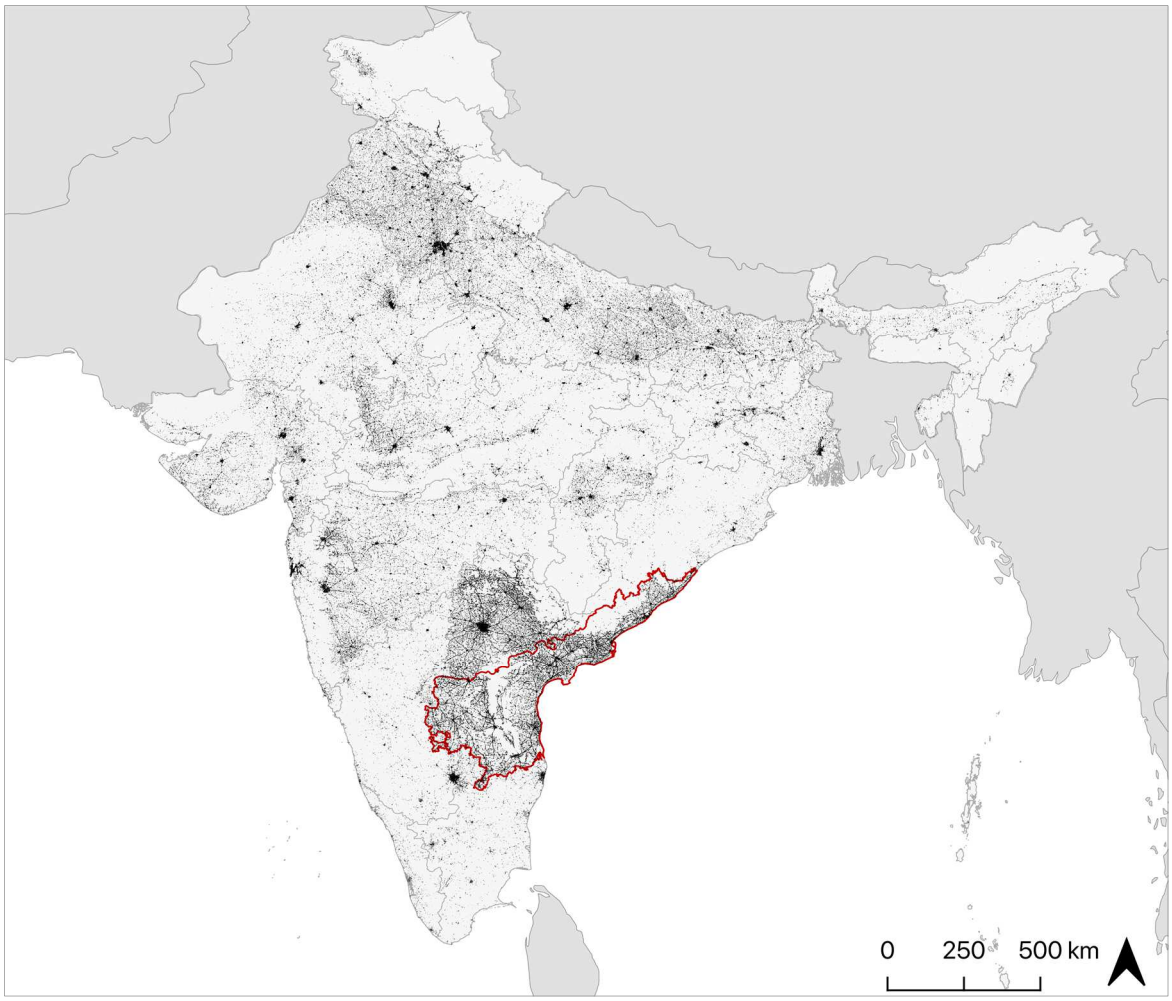



Figure 3: Distribution of geotagged Plantix and Gatherix submissions in India during the Rabi season 2017 - 2018 and location of Andhra Pradesh (blue). A total of 1,065,898 images were received in the timespan from 1 November 2017 to 31 March 2018 in India. Notable is the high concentration of submissions around urban areas. Source: PEAT 2020.

The key feature of the application, the automatic image recognition, is based on a deep artificial neural network. PEAT uses a Convolutional Neural Network (CNN) which is a class of deep neural networks (DNNs) commonly applied to analyze visual imagery (Sladojevic et al. 2016). (For a detailed explanation of the development and functioning of DNNs, see Appendix 1). The datasets needed to train the network are fed by the crowdsourced data gathered through the Plantix application. By using the feature, the image's rights of use are transferred to PEAT, so that the image can be included in their image database. The learning model PEAT applies to train the DNN is a supervised learning strategy, i.e., plant experts label a subset of the received data per crop and disease that fulfills certain quality criteria to build the training datasets. Based on these expert labels, the neural network is trained to predict the crop type as well as the disease pattern of the Plantix submissions. The quality

and size of the training datasets increased with a growing number of image submissions over the past three years and so did the accuracy of PEATs neural network. In October 2019, the overall accuracy of the neural network's detection results for crop types was 95.09% with 53 classes being trained on. The overall accuracy for the disease net trained on 486 disease patterns, pests and nutrient deficiencies was 82.50% (PEAT 2019).
In its starting phase, PEAT employed workers to build up the initial datasets. These field surveyors took photos of various crops and diseases and uploaded them to the database through an application named Gatherix which was developed by PEAT for this purpose. By now, the enhancement of the training data sets as well as the extension of crop varieties and diseases that are supported in the smartphone application is nearly solely done by simply using the crowdsourced Plantix data. Further to PEAT's own benefit, their crowdsourced geotagged and timestamped data is potentially a promising data source for other fields. In this study I examined whether the data is suitable as ground truth for remote sensing analyses.

### 3.1.2 Plantix & Gatherix submissions from Andhra Pradesh

Andhra Pradesh was the first state in which Plantix was officially launched in 2016, therefore the number of images received from Andhra Pradesh is among the highest (Figure 3). In peak times the submissions of plant images per day were between 2,000 and 2,500 (Figure 4). Between the launch of Plantix in 2016 and October 2019, PEAT received 1,007,460 submissions with an attached geolocation and timestamp from Andhra Pradesh. In this study, I only used a fraction of the Rabi submissions, more precisely, I queried all Plantix and Gatherix submissions received between 1 January and 31 March from Andhra Pradesh for the years 2017 to 2019 from PEAT's database. The queried attributes were image ID, timestamp, latitude, longitude, location accuracy, app name, DNN variety result, DNN variety similarity for the most recent net from October 2019, object net result, user ID, plant ID and predicate if the image was uploaded from the smartphone gallery.

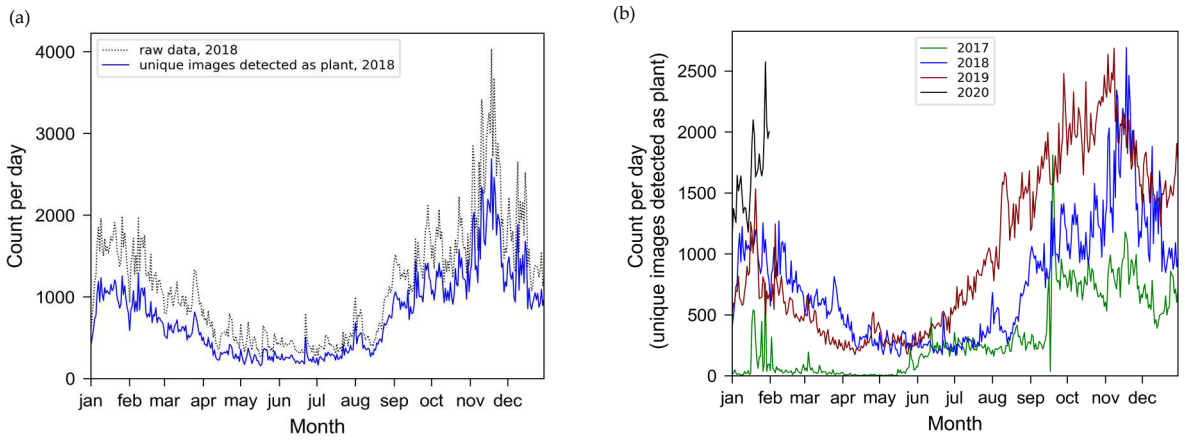



Figure 4: Temporal distribution and amount of Plantix and Gatherix submissions in Andhra Pradesh.(a) Submissions per day in Andhra Pradesh from 1 January to 31 December 2018. The gray graph shows the raw data count per day; the blue graph shows the data count after removing image duplicates and filtering for only images that were detected as plant. A share of 68% of all submissions in 2018 in Andhra Pradesh was detected as plant. (b) Filtered submission (as in (a)) per day from 1 January to 31 December for the years 2017-2019 and 1 January to 31 January 2020. Source: PEAT 2020.

### 3.1.3 Plantix & Gatherix data filtering

The raw data count containing all Plantix and Gatherix submissions for the study area and the selected timespan was 328,754 in total (47,154 for 2017, 182,120 for 2018 and 99,480 for 2019). To retrieve a suitable training data set, I applied several filter steps (Figure 5). The first coarse filter step comprised the removal of all submissions that were detected by the DNN as a non-plant object. The second, coarse filter step was composed of the exclusion of multiple submissions from the same latitude-longitude combination. Then, all submissions that were uploaded from the user's smartphone gallery were removed from the datasets from 2018 and 2019 which included the attribute introduced in August 2017. Afterwards, I applied two spatial filters in order to retrieve images that were taken in

agricultural fields and not for instance from urban hobby gardeners. To do so, I filtered out all points that lay in urban and suburban areas and points that were in proximity to major roads. As a mask for the urban filter I used the Global Urban Footprint (GUF) Layer (DLR 2014) with a dilation buffer of 30 m around urban pixels. A comprehensive road vector layer was obtained from OSM (OpenStreetMap contributors 2019). After a visual assessment of which road types were crucial to include, I selected primary, secondary and residential road objects from the layer. Roads located in rural areas with small widths such as tertiary roads and paths were omitted. The selected road vector objects were enclosed with an 18 m buffer relating to the center line of the road. A thematic filter was applied to the data set excluding tree varieties from the dataset such as mango, citrus and papaya as well as ornamental varieties and additional varieties PEAT's DNN was not trained on. The following filter step comprised the removal of images that were detected as the corresponding crop variety below a certain threshold. In order to gain knowledge about the distribution and accuracy of different thresholds, I examined a subset of all images of major non-tree variety images that were detected as crops in Andhra Pradesh within our selected timespan (21 in total). I defined accuracy ranges in 10%-steps starting from 50%. For each crop variety, I selected a random sample of 40 images per accuracy range. Subsequently, I visually determined the crop variety for every image in order to identify the threshold from which the similarity index is high enough to give a correct crop variety result. Two thresholds, ≥ 50 % and ≥ 80 %, were applied as filters. Finally, for both threshold subsets, I filtered submissions that were uploaded with varying location accuracies (≤ 100 m, ≤ 30 m, ≤ 10 m).
For the exploration and the filtering of the raw PEAT data I used the Python programming language (Python Software Foundation) and the geographic information system software QGIS (QGIS Development Team 2019).

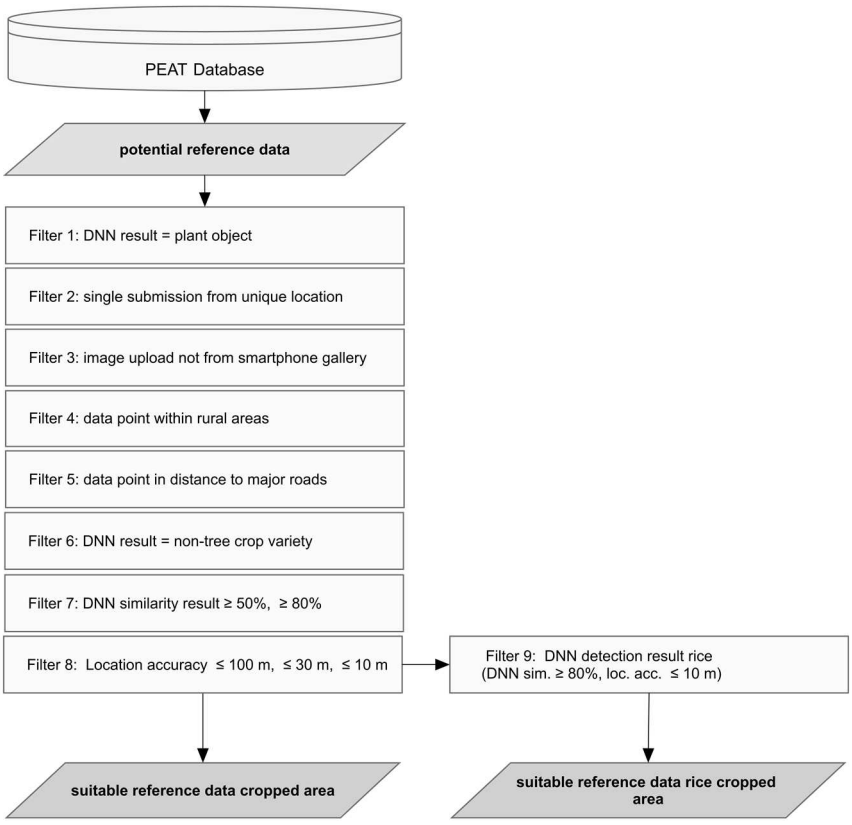



Figure 5: Filter steps applied to PEAT data.

### 3.1.4 Location accuracy of Plantix & Gatherix submissions

The overall location accuracy of the PEAT data is low due to the limited GPS capability of smartphones. The location estimate in smartphones mainly depends on cellular connectivity and the availability of WiFi networks. Since the valid Plantix & Gatherix submissions are made from rural areas

where cell towers have high distances between them and mostly no WiFi networks are available, the location estimate of the device is solely dependent on the GPS signal. This leads to generally low location accuracies from submissions from rural areas. Besides this general issue, several factors further influence the distribution of the location accuracy values among the PEAT data. In 2017, PEAT used the 'best location' estimate strategy that was available for Android devices at that time. This strategy is based on manually defining and accessing the location providers, e.g. GPS, Network (WiFi and cell tower). The best location estimate is then built by evaluating the location estimates of the different location providers. In August 2017, the Plantix developers changed the strategy for accessing the smartphone's location when a fused location provider became available for Android devices. The fused location provider is a location application programming interface (API) in Google Play services that combines different signals to provide the location information, e.g. location fixes can be shared among different applications. The developers of an application can specify the required quality of the API service (Google Developers n. d.). Since an exact location accuracy is not relevant for PEAT's own data exploration purposes and to allow the app to be more battery-efficient, the Plantix developers decided to not request the most accurate data available but the best accuracy possible with no additional power consumption. In urban areas, the fused location provider can achieve high location accuracies due to the broad coverage of WiFi networks and cell towers. In rural areas, cell towers have higher distances between them and mostly no WiFi networks are available. Here, the location estimate is solely dependent on the GPS signal leading to substantially lower accuracy values.

A further issue leading to low location accuracies among the PEAT data is due to a shift in the strategy for subscribing to location updates. Executing a location request usually needs several iterations until a certain accuracy is achieved. The Plantix application version from 2017 and the Gatherix application stay subscribed to these location updates whereas the Plantix versions that were installed or updated after August 2017 unsubscribes from location updates if the provided location lies within a 200 m radius of the previous received location.

### 3.1.5 Composition of subsets

I generated ten different subsets of the PEAT dataset; nine subsets for the sensitivity analysis of the cropped area mapping and one 'sub-subset' for the rice/non-rice crop mapping. Six subsets were compiled using varying location accuracy thresholds as well as varying DNN similarity thresholds; one subset contained only the major crop varieties represented in the PEAT data from Andhra Pradesh. One subset contained only data from 2017 (assuring submissions with 'best estimate' location request) and 2019 (submissions with 'fused location provider' request), respectively. The data counts of the subsets were adjusted to comparable sizes. The specific filter attributes applied to the subsets and the according subset sizes are listed in Table 2. The distribution of the data points for subset 6 and 8 are displayed in Figure 6.

Table 2: Subsets of the PEAT data (P = Plantix, G = Gatherix).

| | **Subset specifications** | **Size** |
|---|---|---|
| (1) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 50%, location accuracy ≤ 100m | 3200 |
| (2) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 50%, location accuracy ≤ 30m | 3200 |
| (3) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 50%, location accuracy ≤ 10m | 3200 |
| (4) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 80%, location accuracy ≤ 100m | 3200 |
| (5) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 80%, location accuracy ≤ 30m | 3200 |
| (6) | P & G submissions from 2017 – 2019, all non-tree crop varieties, DNN similarity ≥ 80%, location accuracy ≤ 10m | 3200 |
| (7) | Major crop varieties: P & G submissions from 2017 – 2019, non-tree crop varieties with ≥ 1000 submissions after filter step 6, DNN similarity ≥ 80%, location accuracy ≤ 10m | 2800 |
| (8) | 'Best estimate location strategy': P & G submissions from 2017, all non-tree crop varieties DNN similarity ≥ 80%, location accuracy ≤ 10m | 1200 |
| (9) | 'Fused location provider': P Submissions from 2019, all non-tree crop varieties DNN similarity ≥ 80%, location accuracy ≤ 10m | 1200 |
| (10) | Subset of (8): P & G submissions detected as rice by the DNN (n = 800) & P & G submissions detected as non-rice crop by the DNN (n = 800) | 1600 |

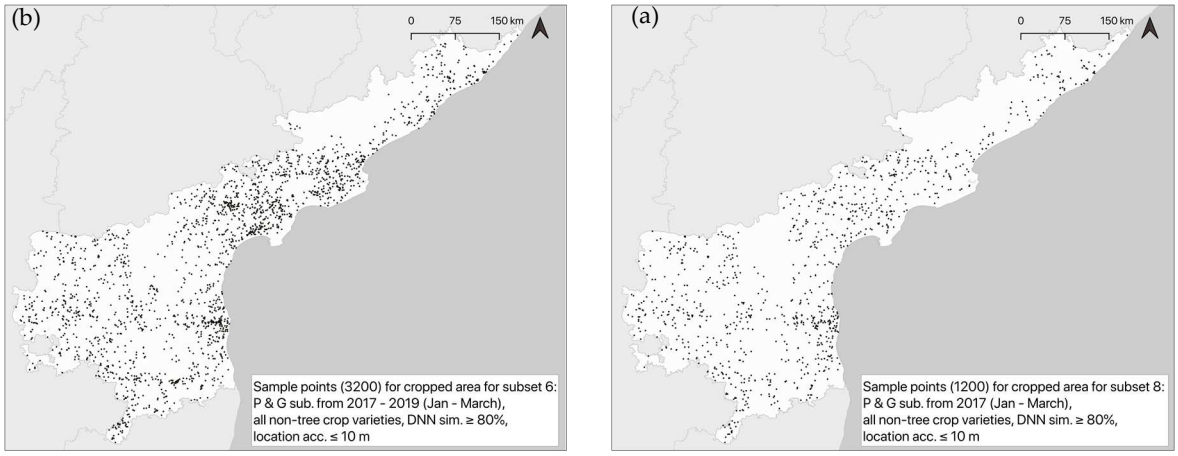



Figure 6: Distribution of PEAT subset data points. (a) Subset 6. (b) Subset 8. Source: PEAT 2020.

### *3.2 Data non-cropped area*

Data for urban areas, water, woody canopy, unsown cropland/bare land/areas of sparse vegetation (hereinafter collectively referred to as non-cropped area) was retrieved using several global datasets which are available in the Google Earth Engine (GEE) public data catalog as well as manually collected in Google Earth Pro for the latter.

As sampling area for urban training points I used the GUF layer (DLR 2014) which I obtained from the German Aerospace Center. The GUF project aims at the worldwide mapping of human settlements with unparalleled spatial resolution of 0.4 arcsecs (~ 12 m). The German Aerospace Center processed a total of 180,000 scenes from the radar satellites TerraSAR-X and TanDEM-X scenes from the years 2010 - 2013 to create the GUF layer. The layer consists of a binary, thematic raster dataset with all pixels containing data either categorized as built-up areas or non-built up areas. A built-up area is defined as a region featuring building structures with a vertical component (Felbier et al. 2014). After cropping the layer to the extent of Andhra Pradesh, I masked all urban pixels and eroded 30m of the mask edges in order to get only reliable urban areas as sampling extent. A random sample of points across the filtered extent was generated.

As sampling extent for water reference points, I used the Global Surface Water Mapping Layers version 1.1 from the European Commission's Joint Research Centre. The dataset contains maps of the location and temporal distribution of surface water. It was generated using Landsat 5, 7 and 8 scenes acquired between 16 March 1984 and 31 December 2018. An expert system was used to classify each pixel individually into water or non-water (Pekel et al. 2016). The product consists of 1 image containing 7 bands showing different facets of the spatial and temporal distribution of surface water in the timespan. Regions where water has never been detected are masked. I applied filters on the recurrence band and seasonality band to retrieve only valid water areas for our analysis. Recurrence is a measurement of the degree of inter-annual variability in the presence of water. The measurement describes how frequently water returned from one year to another within a water period. A water period is defined as the time period from the first month in the first year in which water is observed to the last month of the last year in which water is observed of the entire 35-year period. The threshold on the recurrence band was set to 99. The seasonality band provides information concerning the intra-annual behavior of water surfaces for a single year (2018) showing the number of months water was present. All pixels were masked that did not reach a value of 12 in the seasonality layer. By finally combining the two masks, the extent of seasonal water bodies as well as paddy rice fields were removed (Figure 7a). Thereafter, I eroded the filtered surface water extent to increase the reliability of the sampling area. I sampled random points across the extent according to the number of training points for urban areas.

Sampling areas for woody canopy were obtained through the Hansen Global Forest Change version 1.6 dataset (updated version produced with data up to 2018). The global dataset contains the results from an extensive time-series analysis of Landsat 7 images in characterizing forest extent and change.

First, I generated a classified image containing four tree cover classes out of the 'tree cover 2000' band. The band contains pixel values ranging from 0 to 100 indicating the tree cover canopy for the year 2000, defined as canopy closure for all vegetation taller than 5 m. The classes were set to tree cover levels of 20% starting with 20% (Figure 7b). Pixels with less than 20% tree cover were excluded lying in between the forest definition of the FAO and the definition of Hansen et al. (2013). According to the FAO (2015), forest is defined as 'land with a canopy cover of more than 10%'; Hansen et al. (2013) define forest as land with 25% or greater canopy closure. Our second processing step comprised using the 'loss' band to mask forest loss areas in our classified image. Hansen et al. (2013) define forest loss as stand-replacement disturbance, i.e. a change from a forest to non-forest state. Subsequently, I eroded the edges of the forest pixels by 30 m and created a stratified random sample across the four classes.

Data points for the collective class of unsown cropland/bare land/sparsely naturally vegetated areas were collected manually using Google Earth Pro. I collected training polygons in areas where very high-resolution (VHR) imagery was available in the software for the relevant timeframes of the three years. For each year, I collected 800 reference points capturing different agricultural regions of Andhra Pradesh. The inspection of an unsown cropland point in single VHR image (from a single point of time within the timespan from 1 Jan to 31 March) did not provide enough reliable information to determine if the plot at the point location was indeed uncropped for the entire timespan. Therefore, I further examined NDVI time series for collected points for unsown area previously filtered out by setting a maximum NDVI value threshold of > 0.1.

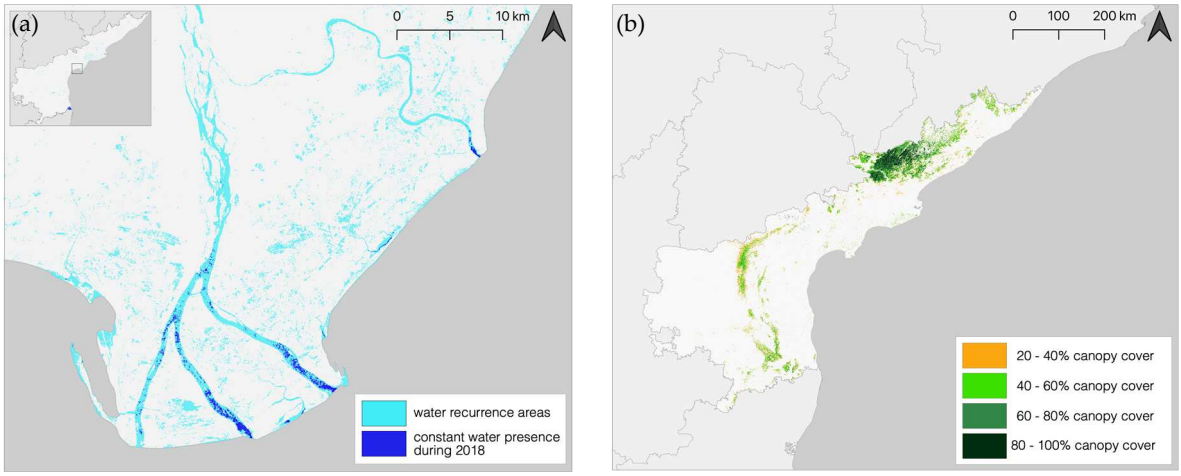



Figure 7: Defining sample areas for non-cropped area from global datasets. (a) JRC Surface Water Mapping Layers. Water occurrence throughout the year (where water occurred for a minimum of one month) and filtered water areas (where water was present for a duration of 12 months). (b) Classified woody canopy gradient based on Hansen et al. (2013).

Finally, the location points for urban areas, woody canopy, water (three different random sample sizes per class: 800, 700 and 300) and the location points for unsown cropland/bare land/ areas of sparse natural vegetation (according number of points) were combined resulting in dataset sizes equal to the PEAT datasets (Table 3).

Table 3: Sizes of feature subsets (name abbreviations according to PEAT subsets specifications, Table 2).

| | | **Subset size per class** | | | | | **Total size** |
|---|---|---|---|---|---|---|---|
| | | **Cropped** | **Non-cropped** | | | | |
| | | | Water | Urban | Woody veg. | Unsown/ Spars. veg. | |
| (1) | DNN ≥ 50%, acc. ≤ 100m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (2) | DNN ≥ 50%, acc. ≤ 30m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (3) | DNN ≥ 50%, acc. ≤ 10m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (4) | DNN ≥ 80%, acc. ≤ 100m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (5) | DNN ≥ 80%, acc. ≤ 30m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (6) | DNN ≥ 80%, acc. ≤ 10m | 3200 | 800 | 800 | 800 | 800 | **6400** |
| (7) | Major crops (DNN ≥ 80%, acc. ≤ 10m) | 2800 | 700 | 700 | 700 | 700 | **5600** |
| (8) | Best loc. estimate (DNN ≥ 80%, acc. ≤ 10m) | 1200 | 300 | 300 | 300 | 300 | **2400** |
| (9) | Fused loc. prov. (DNN ≥ 80%, acc. ≤ 10m) | 1200 | 300 | 300 | 300 | 300 | **2400** |

## *3.3 Remotely sensed image data and image processing*

As remotely sensed input data I used the Landsat 7 and Landsat 8 image collections that were available in the GEE data catalog. Correspondingly, the processing of the image data was also done in the GEE.

### 3.3.1 Google Earth Engine

The Google Earth Engine is a cloud-based platform for planetary-scale geospatial analysis. It consists of a high-performance, intrinsically parallel computation service and a multi-petabyte data catalog with analysis-ready data.

Due to its considerable capabilities as a parallel computation service, the GEE has been used in various studies covering different disciplines (Hansen et al. 2013; Dong et al. 2016; Pekel et al. 2016). While a traditional computing environment usually starts computing the pixels as soon as an expression is processed, the GEE works with a different approach. It delays computing the output pixels until it has calculated the context in which output is required. For instance, if the output is meant to be displayed on an interactive map, then the zoom level and the view bounds of the map determines the projection and resolution of the output in a dynamic way. Otherwise, if the result is needed as an input to another computation, then that computation can inquire an adequate projection, resolution, and filter bounds for the pixels. Using this information, input data is automatically resampled and re-projected on the fly. Therefore, it is possible to visualize results in a very short time in the Earth Engine which is advantageous for spotting issues in one's analysis. Thus, the GEE promotes an interactive and iterative approach to data exploration and algorithm development.

Another benefit of using the GEE is that the user does not have to deal with the details of working in a parallel processing environment. The system handles and hides almost every facet of how a computation is managed. The decisions about resource allocation, parallelism, data distribution, and retries are entirely made by the system. However, the liberation from these details also means that the user is unable to influence them (Gorelick et al. 2017).

Being a shared computing resource, limits on the maximum duration of requests in the GEE (currently 270 s), the total number of simultaneous requests per user (40) and the number of simultaneous executions of certain expensive operations such as spatial aggregations (25) are imposed (Gorelick et al. 2017).

### 3.3.2 Landsat data

The Landsat series of satellite missions has collected imagery of the Earth's surface since 1972, making it the longest running uninterrupted Earth observation program. It provides an unprecedented record of the dynamics and status of Earth. Due to a policy change in 2008, all archived and new Landsat data kept by the United States Geological Survey (USGS) have been made available free of charge over the internet to any user (Woodcock et al. 2008). The main value of the Landsat program is its long-term record of observations. Before the opening of the archive to the public, change mapping was limited to a great extent by the paid data access. Thus, analyses were restricted either to large areas over coarse time steps or short time steps over small areas (Wulder et al. 2012). Since the opening of the archive various long-term and large area analyses have been conducted respectively (Patrick Griffiths et al. 2014; Song et al. 2017; Rufin et al. 2019).

Currently, the Landsat 7 and Landsat 8 satellites are in orbit. The spatial resolution provided by the satellites for the bands utilized in this study (visible, near infrared and shortwave spectrum) is 30 m. Landsat 7 was launched on 15 April 1999. The satellite has the Enhanced Thematic Mapper (ETM+) sensor on board, an improved version of the Thematic Mapper instruments that were carried by the Landsat 4 and Landsat 5 satellites. Landsat 7 data is forwarded as 8-bit images with 256 grey levels (USGS 2019c). Due to the Scan Line Corrector (SLC) failure on 31 May 2003, the sensor has acquired and delivered data with gaps since that time. After processing the corrupted scenes, the remaining number of pixels is 78%. However, this data is still among the most geometrically and radiometrically accurate civilian satellite data in the world (Storey 2003). Landsat 8 was launched on 11 February 2013 and carries two sperate sensors: The Operational Land Imager (OLI) and the Thermal Infrared

Sensor (TIRS). OLI gathers data with a radiometric precision over a 12-bit dynamic range. This converts into 4096 possible gray levels, compared with only 256 gray levels in Landsat 7-bit instruments. Both satellites orbit the Earth in a sun-synchronous, near-polar orbit, at an altitude of 705 km with a 16-day repeat cycle. The satellite orbits are offset to allow 8-day repeat coverage of any Landsat scene area on the globe. The conjointly acquisition capacity of Landsat 7 and Landsat 8 is currently around 1,500 scenes per day. Data from the satellites is acquired on the Worldwide Reference System-2 (WRS-2) path/row system with swath overlap varying from minimum 7% at the Equator to maximum 85% at the poles. The size of a Landsat 7 or Landsat 8 scene is 170 km x 185 km (USGS 2019b).

#### 3.3.3 Clear sky pixel observation count

All Landsat 7 ETM+ and Landsat 8 OLI Collection 1 Level 1 Tier 1 Surface Reflectance images that were available in the GEE catalog for the selected timespan per year were utilized in order to compute clear sky observation counts per pixel. A total of 182 (2107), 184 (2018), and 187 (2019) images were included covering Andhra Pradesh's area spanning over 18 Landsat tiles. Based on the three clear sky observation images, I generated a no data mask.

#### 3.3.4 Spectral-temporal metrics

Band-wise spectral-temporal metrics (STMs) per selected timespan per year were generated using all available clear sky observations. In order to remove double observations in the areas where Landsat tiles overlap, daily mosaics were computed. The STMs I computed were mean, median, $10^{th}$ percentile, $90^{th}$ percentile and standard deviation of all reflectance values. In total, the five STMs for each band summed up to 30 features. Each of the STM datasets was masked with the no data mask I created in 3.4.2. I extracted the year specific STMs at the point locations of the nine reference data sets resulting in 27 different feature subsets.

### *3.4 Classification*

#### 3.4.1 Classification models for cropped area mapping

The feature subsets were applied in Random Forest classification models. Random Forest is an ensemble learning method, i.e., a method that generates many classifiers and combines their results. The idea behind this decision tree based algorithm was first developed by Breiman (2001): a large number of relatively uncorrelated models (trees) can produce ensemble predictors that are more accurate than any of the individual predictions. Specifically, the final prediction to classify a new instance in a Random Forest is conducted by using the majority vote of the individual trees. The Random Forest algorithm represents an advanced, more robust version of bagged decision trees. The bagging (bootstrap aggregation) method comprises the procedure of drawing different training subsets randomly with replacement from the entire training set. In a Random Forest, N bootstrap samples are drawn usually from two-thirds of the training dataset. Each training subset is used as an input to base learners whereas the remaining training samples are used to estimate each tree's classification error, leading to an aggregated Out-of-Bag (OOB) error. The improvement of Random Forests over bagged decision trees is the mode with which the decision rules are calculated at each tree node. Instead of splitting each tree node using the best split among all variables as it is done in bagged decision trees, the Random Forest algorithm splits a node by using the best among a subset of predictors which are randomly chosen at that node. This additional randomness in the model minimizes the correlation between the classifiers in the ensemble.
Advantages of the Random Forests are its computational speed and that Random Forests classifiers are not sensitive to noise or overfitting. Moreover, it provides variable importance measures and the OOB error as an internal measure of accuracy. Finally, its simple parametrization: The main parameter needed are the number of trees and the number of variables used to split each node. After conducting an evaluation of the Out-of-Bag errors for four of the different training data sets (cf. OOB error plots in Appendix 5), I set the number of trees to 200 for all models and the variables per split to the square root of all features. The classified maps were exported using the projected coordinate reference system WGS 84/Andhra Pradesh which uses the WGS 84 geographic 2D CRS as its base

coordinate reference system and the Andhra Pradesh National Spatial Framework Lambert Conic Conformal (2SP) as its projection.

#### 3.4.2 Classification strategy for rice/non-rice crop mapping

Based on the cropped area maps of the best performing model from 3.4.1, I generated 'active cropland' masks which I applied to the STM datasets. The cropped STM datasets were used as a basis for classification of rice/non-rice crops within the cropped area for the respective years. The parametrization of the applied Random Forest classifier (Appendix 5d) and the export projection was done analog to 3.4.1.

### *3.5 Validation data & accuracy assessment for cropped area mapping*

#### 3.5.1 Sampling design

The sampling design defines how to select the subset of the map forming the basis for the accuracy assessment. Following Olofsson et al.'s (2014) recommendation for the sampling design for land cover maps, I used of a stratified random sampling approach since the use of simple systematic sampling can result in having difficulties capturing rare classes. Stratification is the division of the AOI into smaller areas (strata), in which each assessment unit is assigned to a single stratum. The strata must be mutually exclusive and inclusive of the AOI, with neither areas that are in multiple strata classes nor omitted from the strata. Since I had to ensure to be able to derive estimations of cropped area in different sub-national regions as well as ensuring the sufficient representation of small classes, stratification was done in two steps. The first stratification step was done by region (Andhra Pradesh's agro-climatic zones) and the second step by map subclasses. Based on the assumption of user's accuracies of 90% for all subclasses and targeting a standard error of the overall accuracy of 1%, I determined that an overall sample size of 900 was required. As I had to take account of adequate sample allocation for each of the regions, I further defined specified allocation schemes per agro-climatic zone based on their area size. Equal sample size of the map strata favors estimation of user's accuracy (UA), while proportional allocation usually results in smaller standard errors for producer's accuracy (PA) and overall accuracy (OA) (Olofsson et al. 2014). Accounting for both needs, I used a sample allocation lying between equal and proportional allocation ensuring a minimum sample size of 50 per stratum. Due to this adjustment the overall sampling size summed up to a number of 1550.

#### 3.5.2 Response design

The response design defines the mode of determining whether the map and the reference data agree. The following four features were chosen as parameters for the response design. First, as spatial assessment unit the pixel was chosen. The second feature of the response design are the sources of the reference data. According to Olofsson et al. (2014) the reference data and/or the process with which the reference data is acquired has to be of higher accuracy than the process to create the map or the process to create the reference classification has be more accurate than the map classification. My reference data consisted of high-resolution imagery from Google Earth and maximum NDVI composites. The maximum NDVI composites were generated in the GEE using Landsat 7 and Landsat 8 imagery for each year, covering the timespan from 1 January to 31 March. I created a stack of the maximum NDVI bands from each year and masked no data pixels. To retrieve manageable file sizes, I exported the image stack in 13 tiles, each covering one district of Andhra Pradesh. My strategy for labelling the reference points consisted of observing the validation points simultaneously in Google Earth and in the maximum NDVI composites. The very high-resolution imagery was used to determine whether a validation point lies in an agricultural field while the maximum NDVI data revealed whether the validation point lies in a pixel that contained vegetation within the timespan from 1 January to 31 March. I used the EO Time Series Viewer QGIS Plugin (Jakimow et al. 2019) to simultaneously visualize the maximum NDVI images in order to clearly identify changes in the reference data within the selected timespan of the years 2017 - 2019. Only reference points that showed

no significant change in maximum NDVI within the timespan of all three years were included. Reference points for which I was not able to determine the class labels with high confidence were omitted. In a second step, I verified all validation points for the classes cropped area and unsown/bare land/sparsely vegetated area by looking at NDVI time series spanning from Nov 2016 to Oct 2019 (cf. Appendix 12).

#### 3.5.3 Analysis

For each classification model the prediction for the validation points of the five subclasses was performed first for the whole AOI using the entire reference points set, and second for the four agro-climatic zones using only the reference points that were collected within the according region. I generated error matrices for all feature subsets/region combinations in order to assess commission and omission errors. Lastly, I calculated area-adjusted class accuracies as well as area-adjusted overall accuracies for all feature subsets/region combinations. The binary classified maps of the best performing classification model for the three Rabi seasons were exported in order to visualize the stable cropped area in the timespan from 1 Jan to 31 March in the years 2017-2019 with an RGB-overlay.

### *3.6 Accuracy assessment for rice/non-rice crop mapping*

Due to the lack of independent validation data for rice cropped areas, I solely used the OOB error to measure the prediction error of the applied random forest classification. Consequently, I was not able to compute area-adjusted class accuracies for classified the rice/non-rice crop maps.

The complete workflow of my study is presented in Figure 8.

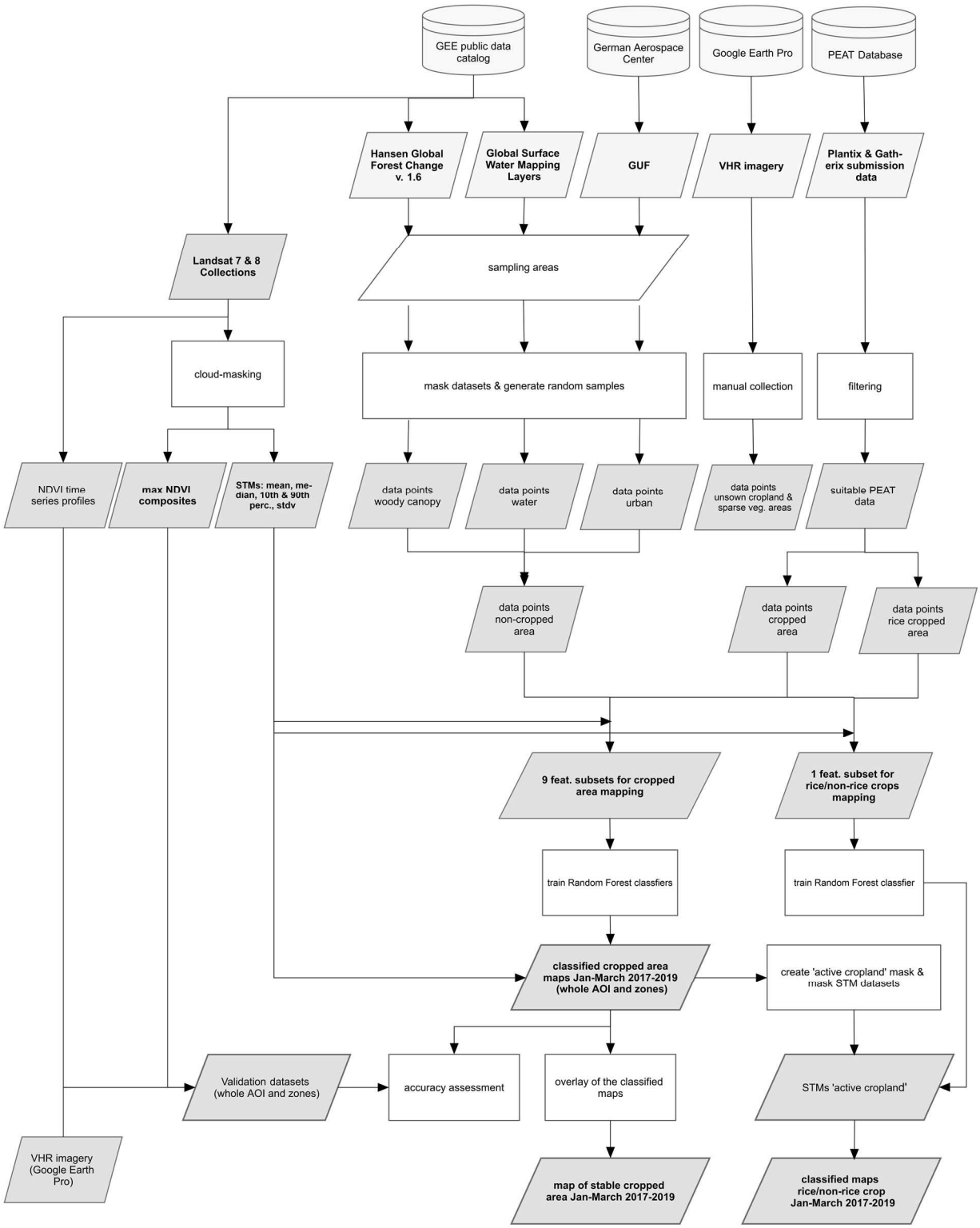



Figure 8: Workflow overview. Data points for non-cropped area were retrieved from the Google Earth Engine data catalog for water (JRC Surface Water Mapping Layers) and woody canopy (Hansen Global Forest Change v. 1.6). The DLR Global Urban Footprint was used as sampling area for urban data points. Data points for unsown cropland/bare land/sparse vegetated areas were manually collected using VHR imagery in Google Earth Pro. Data for cropped area was obtained from PEAT's database for the timespan 1 Jan – 31 March of the years 2017 - 2019. The PEAT data was filtered in several steps (see detailed steps in Figure 5) to retrieve suitable reference points. As remotely sensed image data Landsat 7 and Landsat 8 scenes spanning from 1 Jan - 31 March 2017 -2019 were used. Based on this data, maximum NDVI composites were calculated as validation data sources and five STMs were calculated, providing the mapping basis. Nine training data subsets for the cropped area mapping were composed by extracting the STMs at the point locations for the different subclasses; one training data subset was created for the rice/non-rice mapping. I trained Random Forest classification models in order to generate classified maps for the three timespans for the whole AOI as well as for AP's four agro-climatic zones. The accuracy assessment of the classified maps was conducted using the maximum NDVI composites as well as VHR imagery and NDVI time series profiles as reference data. To identify the stable cropped area of the three years, I created a map overlay. As basis for the rice/non-rice crops mapping, I created 'active cropland' masks out of the classified cropped area maps. The rice/non-rice classification was applied within the previous as cropped area classified regions.

# 4 Results

## *4.1 PEAT data filtering*

I reduced the PEAT data count from a total of 328,754 submissions to 5,785 (1.7% of the raw data count) for the filter combination of DNN similarity ≥ 50% and location accuracy ≤ 100 m and to 3,329 (1.0% of the raw data count) for the filter combination of DNN similarity ≥ 80% and location accuracy ≤ 10m respectively. Table 4 shows the data count reduction corresponding to each filter step as well as the share of the raw data count remaining at each filter step. The largest share of data (72.2% of the raw data count) was eliminated through the first two filter steps comprising the removal of non-plant submission and the removal of multiple submissions from the same transmitted location. The following filters up to the 7th filter step resulted in a reduction of the data count to 33,235 (DNN similarity ≥ 50%) representing 10.1% of the total count and 28,472 (DNN similarity ≥ 80%) representing 8.6% of the total count.

Table 4: Data count and share of raw data count per applied filter step. (The equivalent table showing the data count reduction only for Plantix data is presented in Appendix 2).

| **Filter step** | **TOTAL** | **share of raw data count** | ***Gatherix share/ f.dataset*** | **2017** | **share of raw data count** | ***Gatherix share/ f.dataset*** | **2018** | **share of raw data count** | ***Gatherix share/ f.dataset*** | **2019** | **share of raw data count** | ***Gatherix share/ f.dataset*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw submissions count for AOI & timespan | **328,754** | | *35.1%* | **47,154** | | *40.2%* | **182,120** | | *36.8%* | **99,480** | | *29.4%* |
| 1) Removal non-plant submissions | **254,202** | **77.3%** | *41.9%* | 34,268 | 72.7% | *37.5%* | 137,440 | 75.5% | *47.3%* | 82,494 | 82.9% | *34.9%* |
| 2) Removal multiple submissions | | | | | | | | | | | | |
| (1) from the same crop at the same location | **110,009** | **33.4%** | 2.6% | 4,927 | 10.4% | *4.6%* | 63,917 | 35.1% | *2.9%* | 41,165 | 41.3% | 2-3% |
| (2) from different crops from the same location | **91,604** | **27.8%** | 1.6% | 4,054 | 8,6% | *3.4%* | 53,955 | 29.6% | *1.8%* | 33,595 | 33.7% | 1.4% |
| 3) Removal of submissions uploaded from smartphone gallery | **77,199** | **23.5%** | 2.1% | - | - | | 44,482 | 24.4% | *2.5%* | 28,663 | 28.8% | 2.3% |
| 4) Removal of submissions located in urban areas | **53,447** | **16.3%** | 2.6% | 2,891 | 6.1% | *3.8%* | 29,202 | 16.0% | *3.1%* | 21,354 | 21.5% | 1.7% |
| 5) Removal of points in proximity to major roads | **51,531** | **15.6%** | 2.6% | 2,769 | 5.8% | *3.8%* | 28,117 | 15.4% | *3.1%* | 20,645 | 20.7% | 1.6% |
| 6) Removal of tree varieties | **37,956** | **11.5%** | 3.1% | 2,415 | 5.1% | *2.9%* | 20,283 | 11.1% | *4.1%* | 15,258 | 15.3% | 2.1% |
| 7) DNN similarity | | | | | | | | | | | | |
| ≥ 50% | **33,235** | **10.1%** | 3.6% | 2,255 | 4.7% | *3.1%* | 17,343 | 9.5% | *4.8%* | 13,647 | 7.5% | 2..2% |
| ≥ 80% | **28,472** | **8.6%** | 4.1% | 2,019 | 4.3% | *3.5%* | 13,647 | 7.5% | *4.7%* | 11,855 | 6.5% | 2.3% |
| 8) location acc. | | | | | | | | | | | | |
| DNN ≥ 50% & | | | | | | | | | | | | |
| location acc. ≤100 m | **5,785** | **1.7%** | 1.6% | 1,907 | 4.0% | *2.3%* | 2,577 | 1.4% | *26.1%* | 1,301 | 0.7% | 15.2% |
| location acc. ≤ 30 m | **4,927** | **1.4%** | 1.8% | 1,854 | 3.9% | *2.1%* | 2,062 | 1.1% | *32.4%* | 1,011 | 0.6% | 18.2% |
| location acc. ≤ 10 m | **3,654** | **1.1%** | 2.0% | 1,585 | 3.4% | *1.3%* | 1,345 | 0.7% | *40.4%* | 724 | 0.4% | 21.7% |
| DNN ≥ 80% & | | | | | | | | | | | | |
| location acc. ≤100 m | **5,115** | **1.5%** | 1.0% | 1,703 | 3.6% | *2.5%* | 2,259 | 2.2% | *29.8%* | 1,153 | 0.6% | 17.1% |
| location acc. ≤ 30 m | **4,420** | **1.3%** | 2.0% | 1,656 | 3.5% | *2.4%* | 1,851 | 1.0% | *36.1%* | 931 | 0.5% | 21.7% |
| location acc. ≤ 10 m | **3,329** | **1.0%** | 2.2% | 1,422 | 3.0% | *1.5%* | 913 | 0.5% | *23.8%* | 668 | 0.4% | 44.0% |

The DNN similarity thresholds were set at 50% and 80% as a result of the conducted DNN accuracy score assessment. I was able to visually classify a total of 3,574 out of 4,200 viewed plant images, assuring a minimum of 30 images per range step. The overall results per range step as well as the share of images contained in each range step are shown in Table 5. In the range step of 90-100% containing 71.81% of the data, a share of 96.31% of the tested images were correctly classified. In the lowest range step of 50-60% less than 70% of the images were correctly classified by PEAT's deep neural network. However, it should be noted that the image count in the lower ranges is considerably low. Only 15.02% of all images were detected with a DNN similarity score of less than 80% whereas the major share (85.08%) of all images were classified with a DNN similarity of more than 80%. A comprehensive overview of the results per crop variety is depicted in Appendix 3.

Table 5: Overall result of accuracy range assessment. A total of 21 crop varieties was tested; per accuracy range step a minimum of 30 images per crop were tested.

| | **Range 50-60%** | *Data share* | **Range 60-70%** | *Data share* | **Range 70-80%** | *Data share* | **Range 80-90%** | *Data share* | **Range 90-100%** | *Data share* |
|---|---|---|---|---|---|---|---|---|---|---|
| **Average** | **69.36%** | *4.66%* | **77.70%** | *4.75%* | **83.21%** | *5.61%* | **88.83%** | *13.27%* | **96.31%** | *71.81%* |

The last filter step comprised the removal of all submissions that were transmitted to PEAT's database with a location accuracy of ≥ 100 m reducing the data count to a great extent. The previous data count (33,235 for DNN ≥50%; 28,472 for DNN ≥ 80%) was decreased by 82.6% respective 82.1%. The distribution of the recorded location accuracy of the PEAT data for plant images from unique locations (complies with filter step 3) is shown in Figure 9. Due to the sparse coverage of cell towers and WiFi network in rural areas the location accuracy could not be determined more exactly than within a radius of > 1000 m in most cases. A share of 60.2% of all submissions transmitted with a valid location accuracy lie in the accuracy range of > 1000 m while only 23.4% of all submissions (after filter step 3) lie in the accuracy range > 100 m (Figure 9a).

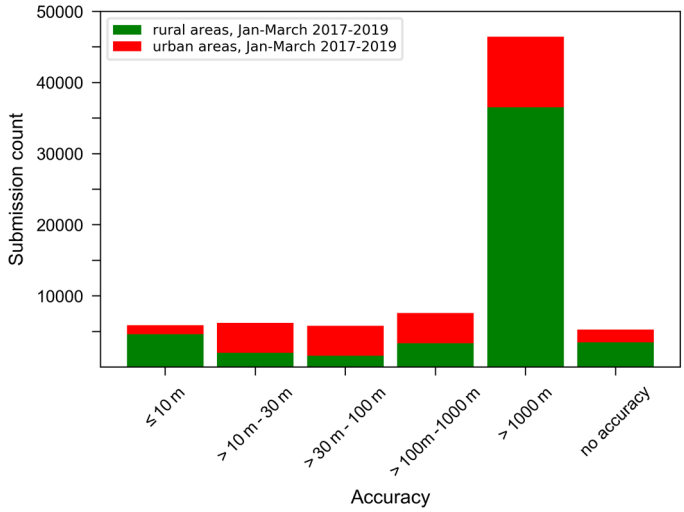



Figure 9: Distribution of the transmitted location accuracy among the PEAT data (after filter step 3 was applied: plant images from unique location) showing the shares of submissions received from rural areas and from urban areas. 60.1% of all submissions were transmitted with a location accuracy of > 1000 m, the major share of this bar (78%) was received from rural areas. 23.4% of all submissions after filter step 3 were submissions in the accuracy range steps > 100 m. For 6.8% of the submissions, no horizontal accuracy was captured by the location manager. Source: PEAT 2020.

The share of the total submissions (after filter step 3) from rural areas in the accuracy range < 100 m (Figure 9b) is 10.6%; 5.9% of the submissions lie in the fine-grained accuracy range of 1 m - 10 m. The change to the fused location provider combined with the abortion of requesting location updates if the previous location lies within a 200 m radius led to the fact that that the overall accuracy of Plantix submissions was reduced after 2017 (Figure 9c and 9d). The accuracy values of submissions received from the Gatherix application whose location strategy was not changed to the fused location provider were still mostly distributed in the range of 1 m to 10 m in 2019.

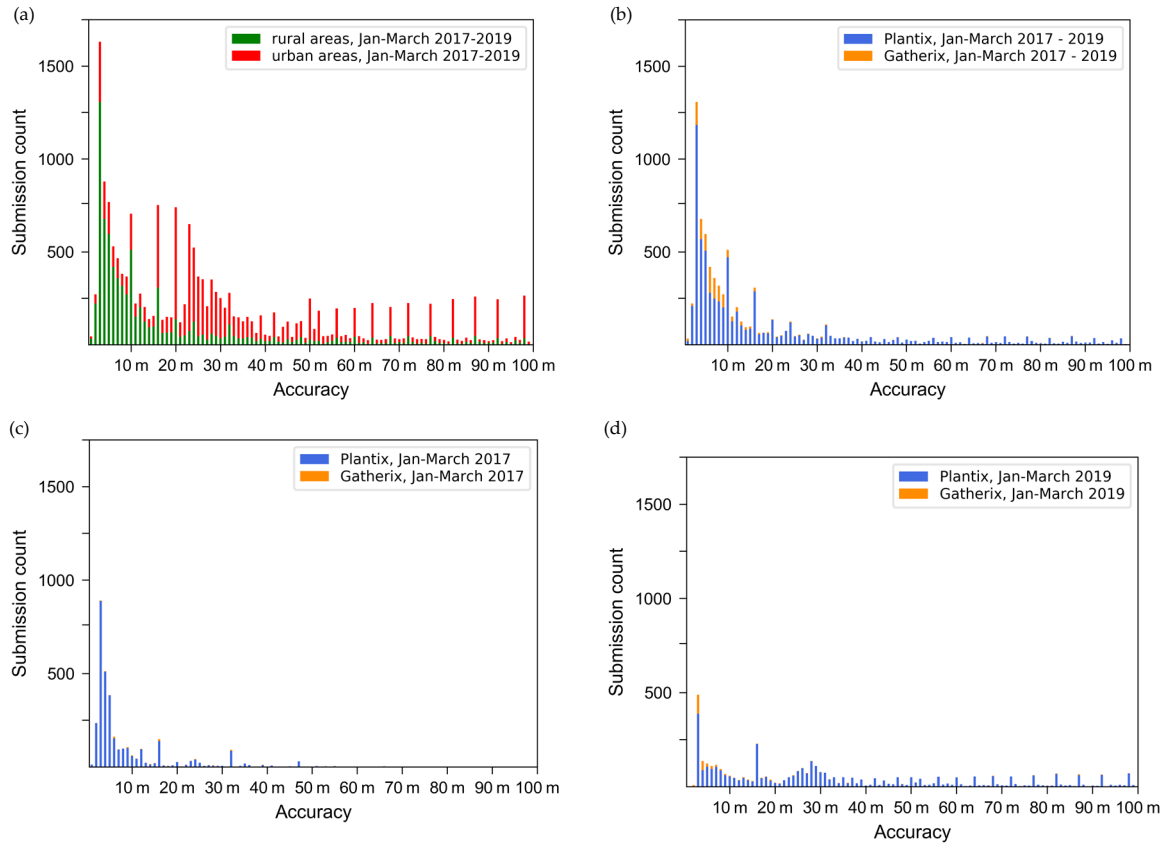



Figure 10: Distribution of the transmitted location accuracy in the range from 1 m to 100 m (after filter step 3 was applied). (a) Submissions divided into submissions from rural and urban areas. 45.5% of the submissions in the accuracy range < 100 m were received from rural areas. (b) Submissions only from rural areas, showing Plantix and Gatherix shares. (c) Submissions only from rural areas from 2017, showing Plantix and Gatherix shares. The major share of the distributions is concentrated in the range of 1 m – 10 m. (d) Submissions only from rural areas from 2019, showing Plantix and Gatherix shares. The accuracy of the Plantix data was determined via the fused location provider whereas the Gatherix location strategy was still the determined via the 'best estimate' location strategy (like all data from 2017). The peaks (in decreasing intervals as accuracy increases) and the fact that no Plantix submissions reach accuracy values > 3 m most likely refer to the specifications of the fused location provider (which are not traceable since Google Play Services are private APIs). Source: PEAT 2020.

### 4.2 *Clear sky observation density*

I recorded an average of 8.76 clear sky observations per pixel during the selected time spans. The minimum was zero observation (2017, 2019) respectively one observation (2018) and the maximum was 21 (2017, 2018) respectively 22 (2019). The visualization of the clear sky pixel observation count for 2019 (Figure 11) shows that there is data scarcity in the central-western region of Andhra Pradesh which is also the case for the two previous years. The percentage of pixels that have less than five clear sky pixel observations and therefore impair the meaning of the STMs that were calculated for these pixels are 4.4% (2019), 2.63% (2018) and 3.28% (2017) (see Table 6).

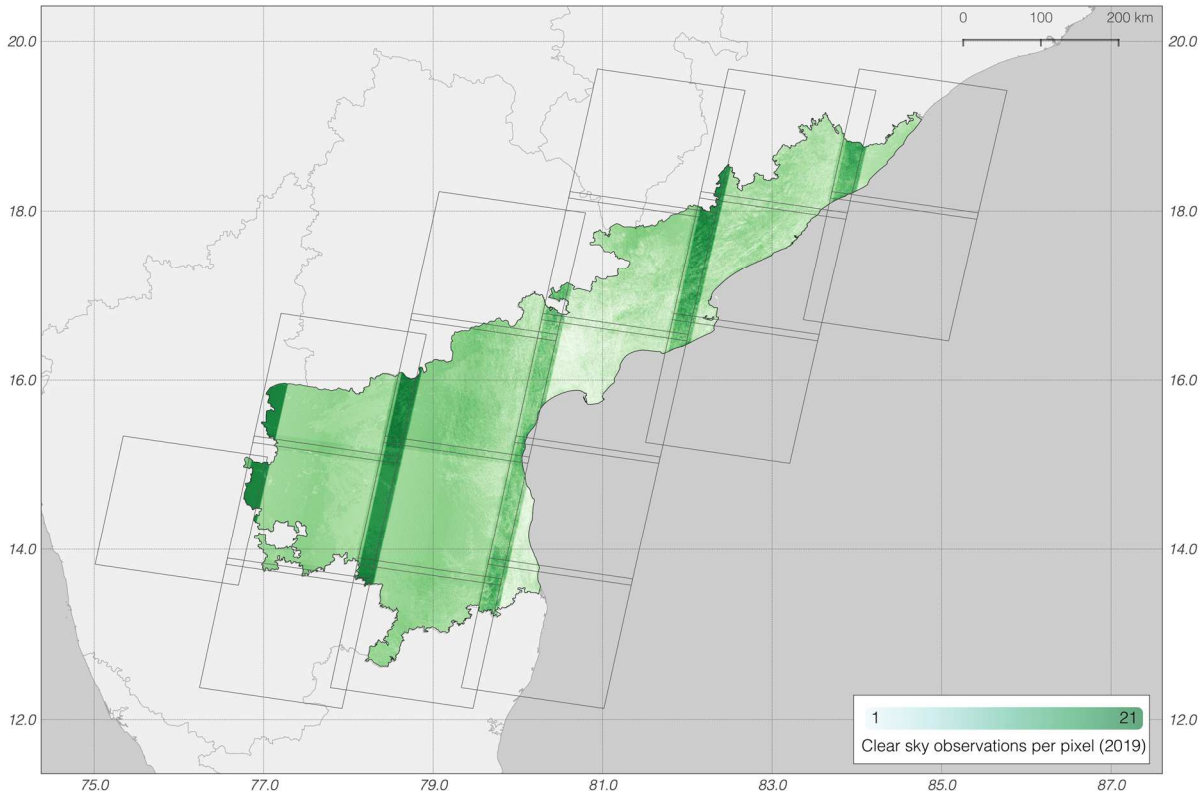



Figure 11: Study area and clear sky observation count per pixel (Landsat 7 and Landsat 8) for Andhra Pradesh acquired during the time period from 1 January to 31 March 2019. Overlay with the 18 Landsat WRS-2 scenes covering Andhra Pradesh. The clear sky observation maps for 2017 and 2019 show a similar pattern with data scarcity in the central-western area of Andhra Pradesh.

Table 6: Landsat clear sky observation statistics for 1 Jan - 31 March for the years 2017 - 2019.

| Year | Mean CSO count | Max CSO count | Area with ≥ 3 CSO | Area with ≥ 5 CSO | No Data |
|---|---|---|---|---|---|
| 2017 | 8.37 | 21 | 99.79% | 96.72% | $3.12^{e-7}$ % |
| 2018 | 8.72 | 21 | 99.89% | 97.37% | 0% |
| 2019 | 9.24 | 22 | 99.52% | 95.60% | $1.02^{e-13}$ % |

### *4.3 Mapping results*

The RGB-overlay of the classified binary maps (Figure 12) classified on the best performing model (model 8) depicts the areas of Andhra Pradesh where the agricultural fields were cropped in all three years between 1 Jan and 31 March (white), non-cropped areas (black) and areas where the plots where cropped just within one of the years (red, green, blue) respective two years (yellow, magenta, cyan).

Figure 13 shows the combined classification result of the best performing model for the binary cropped area mapping and the result of the rice classification result (within the cropped area) for the Rabi season 2018.

Figure 14 shows image details of the RGB-Overlay in comparison with VHR imagery.

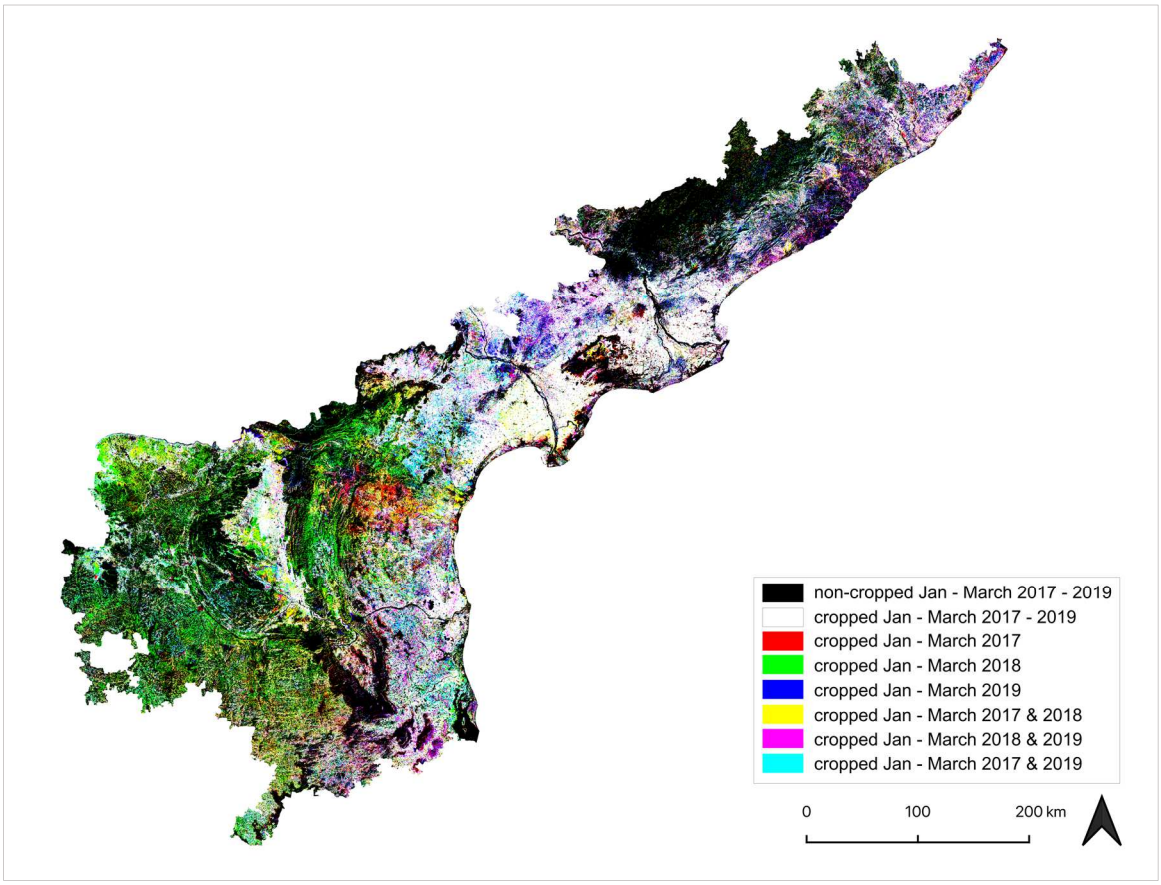



Figure 12: RGB-Overlay map (of model 8). Stable cropped areas are depicted in white, stable non-cropped areas in black.

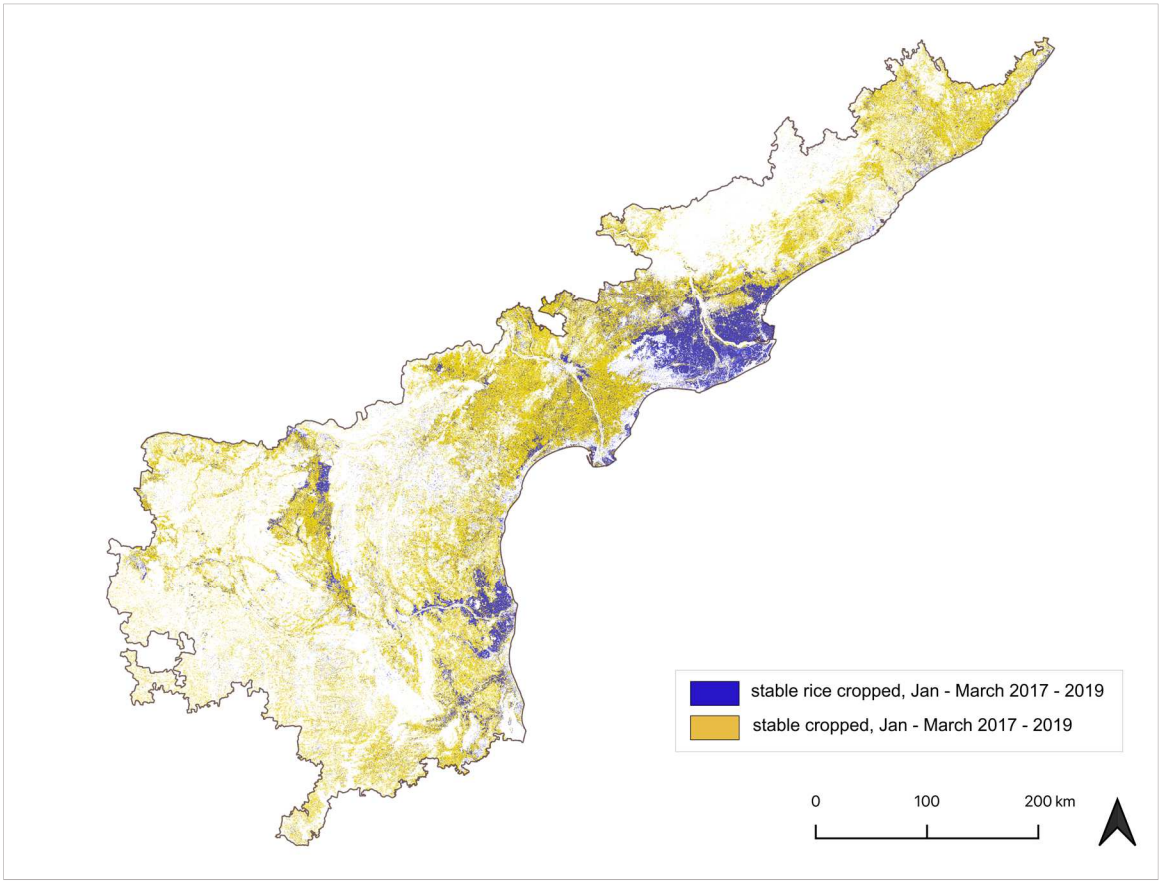
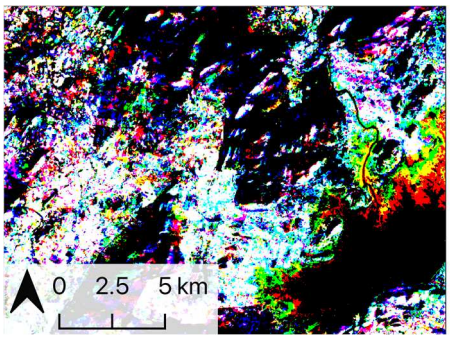
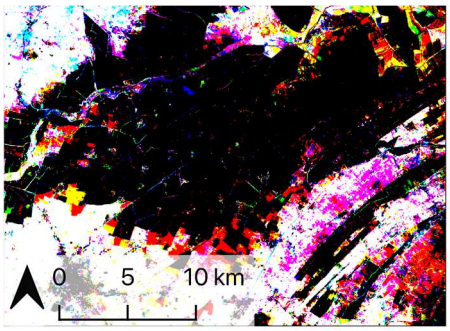


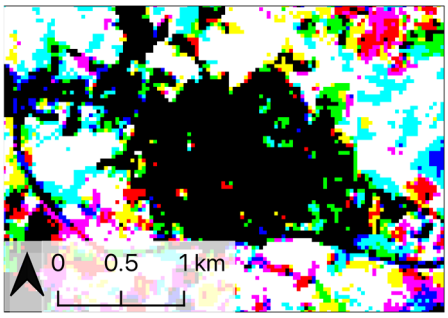

Figure 13: Classification result for stable cropped area for the timespan 1 Jan to 31 March during the years 2017 - 2019 showing the distribution of stable rice cropped area within the cropped area.



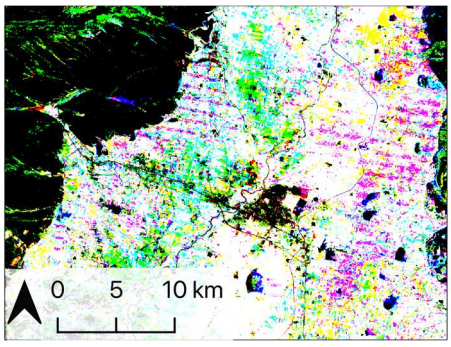

a) Forest patches and water body with non stable and stable cropped plots.




b) Aquaculture farms surrounded by predominantly stable cropped plots.




c) Urban area surrounded by mostly stable cropped plots.




d) Bare soil (left) and village (middle) with predominantly stable cropped plots.

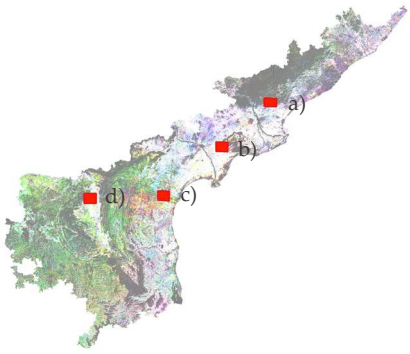




Figure 14: RGB (Figure 12) map details and corresponding VHR imagery.

## 4.4 *Classification accuracies*

### 4.4.1 Classification accuracies for cropped area maps of the entire study region

The area-adjusted OAs for the classified maps of the entire AOI lay between 84.75% [±1.65%] and 86.87% [±1.35%] for the STM dataset covering the timespan Jan - March 2017, between 89.41 [±1.20%] and 92.04 [±1.33%] for Jan - March 2018 and between 88.39 [±1.20%] and 88.76 [±1.16%] for Jan - March 2019 (cf. detailed depiction of all OAs in Appendix 6). Figure 15 shows the class-wise PAs and UAs for all models for the classification of the 2018 STM dataset. All models exceeded PAs and UAs higher than 85% for the classes cropped, water and woody canopy. The UAs for urban class range between 83% and 91%, the PAs between 71% and 80%. PAs for the unsown/bare land/sparsely vegetation class lay between 84% and 93%; UAs lay between 56% and 81%, indicating to a general overestimation of this class. The overall best result shows feature subset 8 (best location estimate strategy/DNN ≥ 80%, location acc. ≤ 10m); the worst result shows feature subset 9 (fused location provider/DNN ≥ 80%, location acc. ≤ 10m). The corresponding confusion matrices for feature subset 8 and 9 are shown in Appendix 7.

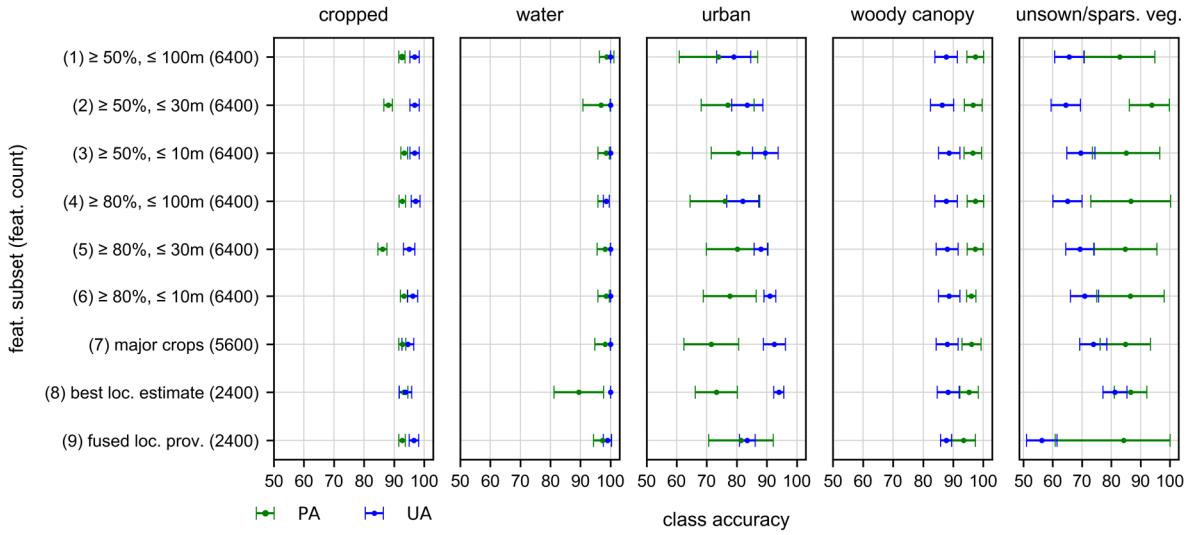



Figure 15: Area adjusted PAs and UAs accuracy for the Jan - March 2018 classification of the 9 different feature subsets. The 95% confidence intervals are depicted by the error bars.

### 4.4.2 Classification accuracies for cropped area maps of the agro-climatic zones

The area-adjusted OA of the two best performing models for the classified maps of the different climatic zones exceed 90% in either one of the models for the North Coastal Zone, Krishna-Godavari Zone and Scarce Rainfall Zone; the area-adjusted OAs for the Southern Zone are considerably lower ranging between 72.69% and 78.67% (Table 7).

Table 7: Area-adjusted OA for model 6 and model 8 for agro-climatic zones.

| Agro-climatic zone | Feature subset | 2017 OA [95% CI] | 2018 OA [95% CI] | 2019 OA [95% CI] |
|---|---|---|---|---|
| **North Coastal Zone** | (6) DNN ≥ 80%, loc. acc. ≤ 10m (n = 6400) | 89.67% [±3.54] | 93.46% [±2.78] | 94.45% [±2.78] |
| | (8) best estimate loc. strategy (n = 2400) | 92.80% [±2.70] | 87.78% [±3.76] | 93.81% [±2.83] |
| **Krishna-Godavari Zone** | (6) DNN ≥ 80%, loc. acc. ≤ 10m (n = 6400) | 90.49% [±2.48] | 98.17% [±0.72] | 96.78% [±0.70] |
| | (8) best estimate loc. strategy (n = 2400) | 91.94% [±2.11] | 92.24% [±1.05] | 96.61% [±1.01] |
| **Scare Rainfall Zone** | (6) DNN ≥ 80%, loc. acc. ≤ 10m (n = 6400) | 87.52% [±1.12] | 91.96% [±2.85] | 85.24% [±1.73] |
| | (8) best estimate loc. strategy (n = 2400) | 91.04 [±1.36] | 94.14% [±1.22] | 90.22% [±1.30] |
| **Southern Zone** | (6) DNN ≥ 80%, loc. acc. ≤ 10m (n = 6400) | 77.19% [±3.87] | 78.57% [±4.31] | 78.49% [±3.72] |
| | (8) best estimate loc. strategy (n = 2400) | 77.42% [±3.97] | 72.69% [±5.16] | 74.67% [±4.22] |

Figure 16 shows the class-wise PAs and UAs of the two best performing feature subsets 6 and 8 for the different agro-climatic zones for the classification of the 2018 STM dataset (cf. Appendix 8 for the confusion matrices for feature subset 8).

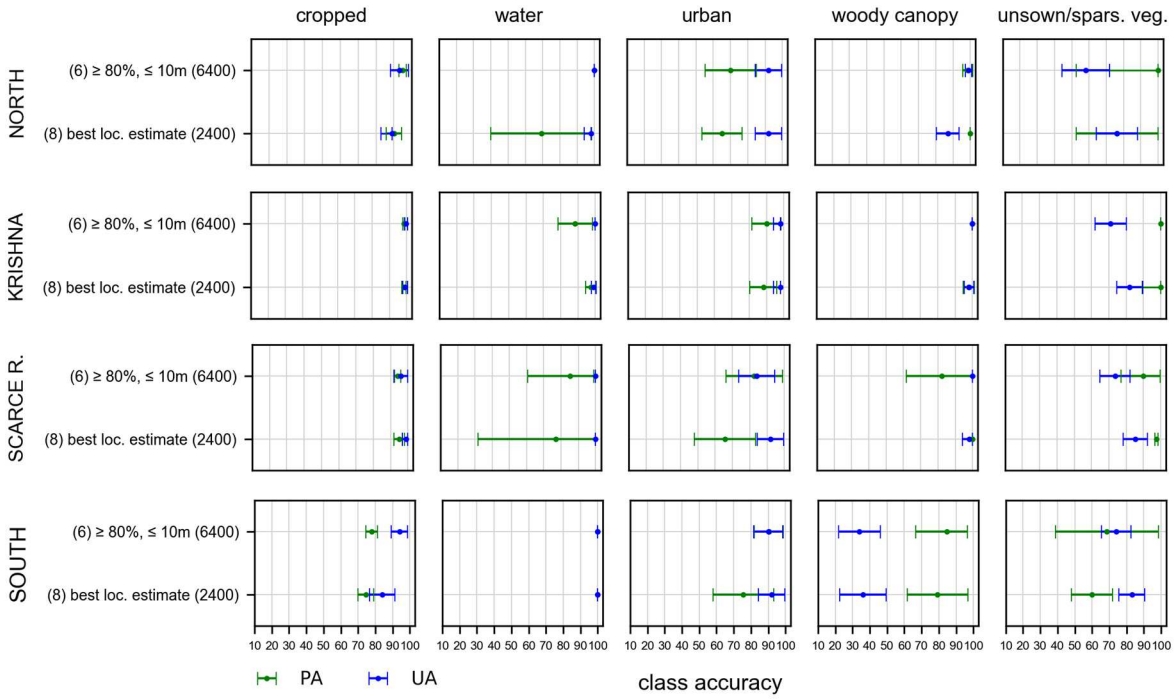



Figure 16: Area adjusted producer's and user's accuracy for the Jan - March 2018 classification (feature subset 6 and 8) of the four agro-climatic zones. (Validation was conducted using the zone-specific validation datasets). The 95% confidence intervals are depicted by the error bars.

### *4.5 Stable cropped area estimates*

The aggregated map estimate (feature subset 8 utilized) for stable cropped area within the time period 1 Jan - March 31 for the years 2017 - 2019 lies at 5.7 Mha (35.04% of total AOI) with the 95% CIs for the cropped area estimates per year ranging between ±1.44 and ±1.54 (cf. Appendix 9). Thereby, the map estimates are substantially higher than the officially reported numbers for Rabi cropped area which are 1.72 Mha (10.58%) for Rabi 2017, 2.06 Mha (10.93%) for Rabi 2018 and 1.78 Mha (10.93%) for Rabi 2019 (Government of Andhra Pradesh 2019).

The detailed map estimates for the stable cropped area per agro-climatic zone (with the 95% CIs for the cropped area estimates per year ranging between ±1.11 and ±5.31, cf. Appendix 10) indicate that the Krishna-Godavari Zone has the greatest stable cropped area share of its total area within the selected timespan in the years 2017 - 2019 (Table 8). However, regarding the accuracy of the map estimates the partially large range of the confidence intervals has to be noted.

Table 8: Stable cropped area map estimates for 1 Jan - 31 March 2017 - 2019 (feature subset 8) for Andhra Pradesh's agro-climatic zones.

| Agro-climatic zone | Map estimate for stable cropped area Jan – March 2017 – 2019 [Mha] | Share of total zone area [%] |
|---|---|---|
| **North Coastal Zone** | 0.80 | 21.67% |
| **Krishna-Godavari Zone** | 2.72 | 45.88% |
| **Scare Rainfall Zone** | 0.78 | 21.08% |
| **Southern Zone** | 1.41 | 32.15% |

### *4.6 Rice cropped area estimates*

Table 9 shows the map estimates (feature subset 8 utilized) for rice cropped area (not area-adjusted) in the entire study regions for the timeframes 1 Jan - 31 March for the years 2017-2019. The RF accuracies lay between 0.71% and 0.75%. Observable is a slight increase in rice cropped area in the map estimates of 6.1% from 2017 to 2019. The comparison of the map estimate for 2017 with the official reported numbers available for the Rabi 2017 rice cropped area shows high agreement with the map estimate being only 7.1% higher than the official reported number.

Table 9: Map area estimates of rice cropped area for 1 Jan - 31 March of the years 2017 - 2019 and comparison with official reported number for Rabi 2017 rice cropped area (not available for 2018 and 2019).

| | **AP rice cropped area map estimate Jan – March 2017 [Mha]** | **AP rice cropped area map estimate Jan – March 2018 [Mha]** | **AP rice cropped area map estimate Jan – March 2019 [Mha]** |
|---|---|---|---|
| **Map area estimate (feat. subset 8)** | 1.64 | 1.70 | 1.74 |
| ***Rabi rice cropped area (Government of Andhra Pradesh 2019)*** | *1.52* | *n.a.* | *n.a* |

In Table 10 the map estimates for rice cropped area are presented per agro-climatic zone, likewise with a comparison to the official reported numbers for Rabi 2017 cropped area. For the Krishna-Godavari Zone the map estimates are approximately in line with the official reported numbers. This also applies to the map estimates for the Scarce Rainfall Zone. However, the estimates for the Southern Zone and particularly for the North Coastal differ remarkably from the officially reported numbers.

Table 10: Map area estimates of rice cropped area per agro-climatic zone for 1 Jan - 31 March of the years 2017 - 2019 and comparison with official reported number for Rabi 2017 rice cropped area (not available for 2018 and 2019).

| **Agro-climatic zone** | **cropped rice area time frame 2017** | | **Rice cropped area map estimate Jan - March 2018 [Mha]** | **Rice cropped area map estimate Jan – March 2019 [Mha]** |
|---|---|---|---|---|
| | ***Rabi 2017 rice cropped area [Mha] (Government of Andhra Pradesh 2019)*** | **map estimate Jan – March 2017 [Mha]** | | |
| **North Coastal Zone** | *0.43* | 0.12 | 0.13 | 0.15 |
| **Krishna-Godavari Z.** | *0.85* | 0.96 | 0.92 | 1.08 |
| **Scare Rainfall Zone** | *0.09* | 0.15 | 0.18 | 0.13 |
| **Southern Zone** | *0.24* | 0.38 | 0.46 | 0.37 |

## 5 Discussion

The presented study ascertained the usability of crowdsourced data generated through the ag-tech platform Plantix as ground truth data for cropped area mapping with the prerequisite of the application of extensive filtering. The raw data count was drastically reduced; only 1% of the initial count remained as suitable reference data for the mapping purpose. The use of the filtered PEAT data as ground truth for cropped area resulted in user's and producer's accuracies exceeding over 90%. The objective of mapping stable Rabi cropped area in Andhra Pradesh utilizing the aggregated recurrent cropped area within the timeframe 1 January - 31 March of the years 2017 - 2019 was not reached due to the high temporal and spatial variability of cropping cycles in the study region. Therefore, the map estimates differ considerably to the official reported numbers. The mapping of exclusive rice cropped area within the selected timeframe, however, showed good agreement with the official reported numbers for rice cropped area in Andhra Pradesh's Rabi season.

### *5.1 Plantix data as ground truth for cropped area and rice cropped area mapping*

Crowdsourced data represents a promising resource as ground truth for remote sensing purposes, especially considering that the generation and collection of crowdsourced data will most likely increase in the near future. However, the use of crowdsourced data poses several challenges. First, a large dataset not being collected with a specific study design contains a considerable amount of noise as well as duplicate entries (Gao et al. 2016). Second, crowdsourced data usually represents a biased sample due to the often uneven geographical distribution of the crowd and the data might represent only a fraction of the sample as e.g., smartphones are more often used by people at an age below 40

years (Jestico et al. 2016). Third, crowdsourced data usually lacks the quality standards of traditional geographic data collection measures such as positional accuracy, temporal accuracy and thematic accuracy (Senaratne et al. 2017). In most cases, the crowdsourced data cannot be used directly to gain usable information. Various pre-processing steps and data mining steps are required to prepare the data (Barbier et al. 2012). In the case of my study, the application of the multitude of necessary filter steps to retrieve a suitable reference data set reduced the raw data count drastically. A large proportion (22.3%) of the received submissions were photos of non-plant objects. In the following, the raw data count was reduced substantially by removing multiple submissions from the same location and submissions that were uploaded from the smartphone gallery. The remaining submissions representing the 'noise-free' data constituted 23.5% of the initial data count. The following applied filter steps were specific to the intended use as ground truth for cropped area mapping. Especially the removal of submissions made from urban areas was crucial as the geographic distribution of the raw Plantix data is heavily concentrated in urban areas (cf. Figure 3), likely due to smartphone ownership, internet access and knowledge of the app. The fundamental filter step in order to retrieve a subset out of the PEAT dataset being suitable for the application as ground truth, however, was the final filter step of including only submissions with a high location accuracy.

The assessment of the DNN similarity score shows the overall high accuracy of PEAT's deep neural network regarding crop type classification (cf. Appendix 3). If an image was falsely classified by PEAT's DNN, then the confusion was mostly in between crops with similar leaf morphology (e.g. cereal grains or legumes) and only sparse confusion with ornamental plants. Furthermore, the shares of images in the lower accuracy ranges are very low for most crop varieties. This is reflected in the sensitivity analysis as the threshold of DNN similarity does not have a remarkable effect on the model's accuracy. Thus, for the purpose of cropped area mapping the low DNN similarity threshold of ≥ 50% was sufficient to ensure that the photo was taken of a crop variety.

The model performance of the feature subsets with the different location accuracy thresholds of ≤ 100 m, ≤ 30 m, and ≤ 10 m increased with higher location accuracy. It is remarkable that the model performance did not improve with a higher feature count. The best performing feature subset is subset 8 containing only Plantix & Gatherix submissions from 2017 (with DNN similarity ≥ 50% and location accuracy ≤ 10 m) representing the only feature subset that contained only submissions with the 'best estimate' location request. The worst performing feature subset is subset 9 containing only Plantix submissions from 2019 (with DNN similarity ≥ 50% and location accuracy ≤ 10 m) representing the only feature subset that contained only submissions with the current location request strategy of the fused location provider. This leads to the assumption that the 'best estimate' location strategy that was applied in the Plantix app until August 2017 generates considerably better accuracy results than the current location strategy. The reason might be that the 'best estimate' location strategy continuously refines its location request results while the fused location provider aborts location updates if the new location lies within a 200 m radius to the previous location. Assuming that the location accuracy of the Plantix submissions diminished due to PEAT's change from the best estimate location strategy to the fused location provider as indicated by the strong disparity between the user's accuracies of the according feature subsets 8 and 9, this entails an overall lower location accuracy reliability of Plantix submissions received after August 2017. This issue exemplifies the general problem of using crowdsourced data that is not intentionally collected for a research purpose. The parameters of the crowdsourced data can change over time, e.g., in this case due to an improvement for the user of the smartphone application with a more battery-efficient location strategy. Changing parameters leads to potential inconsistencies and deterioration of the crowdsourced data quality whereas traditional data collection methods are designed for a certain purpose and usually only improve.

I manually reviewed the location of all datapoints from feature subset 8 to ascertain its reliability. With 96% of the points being verified as lying in agricultural fields, the PEAT submissions with a location accuracy ≤ 10 m located in rural areas are of great value as ground truth for the mapping of cropped area. Thus, the collection of crowdsourced Plantix data holds a noteworthy potential for remote sensing-based agricultural mapping purposes for study areas where little or no ground truth data collection has been conducted so far. However, those regions currently constitute unfavorable conditions for acquiring valid crop submissions with high GPS accuracy in large quantities. On the

one hand, PEAT receives invalid submissions due to farmer's low digital literacy as well as their unfamiliarity with the Plantix app, and on the other hand, the location accuracy of the uploaded image's coordinates can vary considerably due to the limited availability of cellular data connection or WiFi connection for cellular positioning in rural areas. This leads to the issues that Plantix was misused in about 22% of the submissions utilized in this study and, even if the image information of the submission was useful, the low GPS accuracy made the submissions unusable for the mapping purpose in the majority of cases.
Notwithstanding the described challenges the PEAT data imposes when used in a remote sensing context, the data still holds a remarkable potential as the Plantix userbase is steadily growing (given the prerequisite that the startup can establish itself on the market). However, the transferability of my study to other study regions is currently limited to the few federal states of India wherefrom PEAT received submissions in sufficiently large quantities.

### *5.2 Mapping Rabi cropped area and Rabi rice cropped area in Andhra Pradesh*

Due to the high temporal and spatial variability of the onset of the Rabi season in Andhra Pradesh, the mapping of Rabi cropped area represents a challenge. My approach was to select the timeframe from 1 January to 31 March to cover the main growing period of the Rabi crops and to calculate the stable cropped area within that timeframe assuming this approach would capture the actual Rabi cropped area omitting plots that show a late Kharif harvest or plant material that was left on the fields after the Kharif harvest (cf. Appendix 8.1 and Appendix 13). However, the large difference (+3.9 Mha) between my map estimate for stable cropped area and the official reported numbers for Rabi cropped area from the government of Andhra Pradesh indicates that my chosen approach did not fulfill the objective to map the stable Rabi cropped area in Andhra Pradesh for the years 2017 - 2019. Likewise, the comparison with my mapping results for cropped area for the timespan from January to March 2017 with the study of Jain et al.'s (2016) aiming at mapping the annual winter cropped area in India from 2001 - 2016 shows a high discrepancy (cf. Appendix 14).
The comparison of my rice cropped area estimates for 2017 with the official statistics, however, shows higher agreement (+0.12 Mha). The timing for irrigated paddy rice cropping appears to be more distinctive (cf. Appendix 8.3) being independent from precipitation. The rice cropped area in the Rabi season of Andhra Pradesh (40% of the total Rabi cropped area) is entirely irrigated (Reddy and Motkuri 2013). In line with this, my mapping results for Rabi rice-cropped are in good agreement with the study of Ambika et al. (2016) who mapped high resolution irrigated area in India for 2000 to 2015. Thus, the precise mapping of irrigated rice cropped area in Andhra Pradesh's seasons utilizing Plantix data as ground truth represents a feasible research objective for further studies. However, the mapping of rainfed Rabi cropped area in Andhra Pradesh will remain a challenge in further research due to the described high temporal and spatial variability of the onset of the Rabi season.

### *5.3 Limitations and uncertainties*

Despite following a well-thought-out study design, uncertainties as well as limitations remain regarding the sampling design, the utilized data and the methodology.
The collective class of unsown cropland/bare land/sparsely vegetated areas has been shown to be overestimated in the accuracy assessment. A variety of factors could have contributed to the commission error of the class. First, the manual selection of the reference points for the class presents a potential source of errors. My limited knowledge about the regional cropping practices and their appearance in satellite imagery could have led to misclassifications of the training data as well as the validation data. Second, the class included the three subclasses of unsown cropland, bare land and shrubs. Particularly, defining a separate class for unsown cropland/ fallow fields could have been meaningful. Only limited attention has been devoted to the distinguishing between actively cropped fields and fallowed fields within agricultural lands (Tong et al. 2020). Fourth, large bare land areas have been misclassified as urban areas in the Scarce Rainfall Zone as Appendix 11 reveals.
The use of the three global datasets (Hansen et al. 2013; Pekel et al. 2016; DLR 2014) as sampling areas for non-cropped reference points also constitutes possible sources of errors. As global datasets they are not customized for the specification of the spectral signatures of these classes in Andhra Pradesh.

However, through the stringent definition of restricted sample areas, I obtained reliable sampling areas with a high probability for the classes of urban areas, woody canopy and water bodies.
The reliability of the accuracy values of the classification of the rice/non-rice cropped areas is very limited. The maps were not validated with an independent reference data set; merely the OOB error was used as accuracy measure. Due to the lack of an independent reference data set I was further not able to compute adjusted area probabilities.
The assumption of the deterioration of the Plantix data quality in regard of location accuracy due to the change to the fused location provider cannot be verified as the specification of the fused location provider as being part of the private Google Play Services API are not possible to examine.

Regarding the predominant small plot sizes in Andhra Pradesh of an average size of 1.15 ha (Department of Agriculture of India 2019), using Landsat image data having a resolution of 30 m was not favorable for the objective of the study. Remotely sensed image data from a satellite providing a higher resolution (e.g. Sentinel 2 with 10 m resolution) would have been preferable to avoid capturing a multitude of mixed pixels and therefore would have most probably led to more accurate area statistics results. However, at the starting point of time of this study, the Sentinel-2 top of atmosphere products from the European Space agency were not as geometrically accurate as Landsat data. The geolocation accuracy of Sentinel tiles was stated to be misregistered by more than 10 m pixel (Yan et al. 2018). Furthermore, the cloud/cloud shadow mask of the Level 2A Sentinel product as key element to enable an automatic processing of Sentinel 2 data has been stated as currently being not reliable for the use in time-series analysis (Baetens et al. 2019). Sentinel-2 cloud detections are unsatisfactory as low altitude clouds might not be detected in the cirrus band and bright land surfaces such as built-up structures happen to be misclassified as clouds when only spectral information is considered (Frantz et al. 2018). Therefore, the use of Sentinel-2 data in this study would have required pre-processing steps in order to correct for possible geometric inaccuracies as well as for pixels misclassified as clouds. Since this would have exceeded the scope of this study being a master thesis, I decided to utilize the Landsat Tier 1 product. The Landsat Tier 1 product 'includes Level-1 Precision and Terrain corrected data that have well-characterized radiometry and are inter-calibrated across the different Landsat instruments' (USGS 2019a). Furthermore, it includes the Quality Assessment Band enabling users to create cloud masks. These characteristics make Landsat data stable and reliable for pixel-level time-series analysis without additional pre-processing steps.

## 6 Conclusion

The study demonstrates that the crowdsourced PEAT data can be utilized as ground truth data for cropped area mapping as well as rice cropped area mapping if being filtered comprehensively. The fact that the Plantix submissions hold information on the crop variety, and moreover information on occurring crop diseases represents a large potential for the utilization of Plantix data for further crop type mapping as well as potentially even the mapping and monitoring of crop virus spreading. However, further studies aiming at the inclusion of Plantix data are currently limited to few regions in India where the present main Plantix userbase is located.
Noteworthy is that the amount of potential suitable Plantix data was eventually drastically reduced due to the low overall location accuracy of the Plantix submissions. Thus, the use of Plantix data for remote sensing purposes is restricted to areas with sufficient cellular connection. Unfortunately, areas where crowdsourced data collection methods would be very advantageous due to the lack of adequate survey data on cropped area, most likely do not fulfill this prerequisite. Nevertheless, in the face of the ongoing fast proliferation of smartphones and relating thereto, the universally growing mobile coverage, the use of crowdsourced data containing location information transmitted via mobile devices will most likely gain more importance in remote sensing contexts in the near future. As rice is the major food crop grown in India and therefore accounts for the largest share of Plantix submissions (20% of all submissions, cf. Appendix 4), further valuable research concerning rice cropped area using the PEAT data could be conducted in the remote sensing context, e.g., the examination of the distribution and possible expansion of irrigated rice area in Andhra Pradesh or in the further states of India where Plantix has a high userbase (cf. Figure 3).

The large spatial and temporal variability of the onset of the Rabi season in Andhra Pradesh presents a major challenge for the mapping of Rabi cropped area in Andhra Pradesh. The substantial result of this study is that defining one restricted time frame for the entire study area constitutes an unsuitable approach for Rabi cropped area mapping. This leads to the conclusion that a more sophisticated approach has to be considered in further studies. In order to accurately map Rabi cropped area in Andhra Pradesh, a detailed analysis of the spatial and temporal variation of the sowing dates across the state needs to be conducted in order to demarcate the different occurring Rabi cropping schedules. Subsequently, the applied timeframe needs to be adjusted according to the respective area. Additionally, regarding the small plot sizes in Andhra Pradesh, further studies should use remotely sensed image data of higher resolution.

In closing, this study represents the first approach of using crowdsourced data generated through the agtech platform Plantix as ground truth for cropland mapping. The depiction of the various required filter steps demonstrates the multiple previously recognized general challenges of working with crowdsourced data in the specific application of Plantix data in the remote sensing context. These valuable insights can be included in further studies on cropland and crop type mapping in areas where the smartphone application has a sufficiently active userbase.

## References


Amara J, Bouaziz B, Algergaway A (2017) A Deep Learning-based Approach for Banana Leaf Diseases Classification. In: Mitschang B (ed) BTW 2017 - Workshopband, Lecture Notes in Informatics, Gesellschaft für Informatik, Bonn

Ambika AK, Wardlow B, Mishra V (2016) Remotely sensed high resolution irrigated area mapping in India for 2000 to 2015. Scientific data 3

APSDPS (2019) Temperature and Humidity. Andhra Pradesh State Development Planning Society. Planning Department, Government of Andhra Pradesh. https://www.apsdps.ap.gov.in/pages/spatial_maps/temperature.html. Accessed 14 November 2019

Baetens L, Desjardins C, Hagolle O (2019) Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure. Remote Sensing

Barbier G, Zafarani R, Gao H, Fung G, Liu H (2012) Maximizing benefits from crowdsourced data. Computational and Mathematical Organization Theory 18

Basha PC (2018) Farmers Suicide in India - Causes and Remedies. International Journal of Research in Economics and Social Sciences 8

Breimann L (2001) Random Forests. Machine Learning 45

Chandler J, Shapiro D (2016) Conducting Clinical Research Using Crowdsourced Convenience Samples. Annual review of clinical psychology 12

Chang J, Hansen MC, Pittman K, Carroll M, DiMiceli C (2007) Corn and Soybean Mapping in the United States Using MODIS Time-Series Data Sets. Agronomy Journal 99

Collobert R, Weston J (2008) A unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In: Association for Computing Machinery (ed) Proceedings of the 25th International Conference on Machine Learning, Helsinki

Dash SK, Nair AA, Kulkarni MA, Mohanty UC (2011) Characteristic changes in the long and short spells of different rain intensities in India. Theoretical and Applied Climatology 105

Department of Agriculture (2015) Agro and sub Agro-Climatic Zones. https://farmech.dac.gov.in/FarmerGuide/AP/index1.html. Accessed 28 November 2019

Department of Agriculture (2016) Annual Administrative Report 2015-16. http://www.apagrisnet.gov.in/2018/Admin/Annual%20Administrative%20Report%202015-16%20Final.pdf. Accessed 30 November 2019

Department of Agriculture (2017) Annual Administrative Report 2016-17. http://www.apagrisnet.gov.in/2018/Admin/Annual%20Administrative%20Report%202016-17.pdf. Accessed 28 November 2019

Department of Agriculture (2018) Season and Crop Coverage Report Rabi 2017-18 for the week ending 03-01-2018. Government of Andhra Pradesh. http://www.apagrisnet.gov.in/2018/weekly/January/weekly_report_(Rabi)_12_03-01-18.pdf. Accessed 30 November 2019

Department of Agriculture (2019) Season and Crop Coverage Report. Rabi 2018-19 up to the week ending 20/02/2019. Government of Andhra Pradesh. http://www.apagrisnet.gov.in/2019/weekly/Febraury/weekly_report_(Rabi)_19_20-02-19.pdf. Accessed 28 November 2019

Department of Agriculture of India (2019) Agriculture Census 2015-16. All India Report on Number and Area of Operational Holdings

DES (2018) Second Advance Estimate of Production of Foodgrains for 2018-19. Directorate of Economics & Statistics. Directorate of Economics and Statistics, Agricultural Statistics Division. Department of Agriculture, Cooperation and Farmer's Welfare. http://agricoop.gov.in/sites/default/files/2ndADVEST201819_E.pdf. Accessed 28 November 2019

DES (2019) Monsoon Periods in Andhra Pradesh. Directorate of Economics & Statistics. Directorate of Economics and Statistics. Department of Agriculture, Cooperation and Farmer's Welfare. http://www.apsdps.gov.in/drought/directorate_of_economics_and_statistics.pdf

Dietterich TG (1997) Machine-Learning Research. Four Current Directions. American Association for Artificial Intelligence

DLR (2014) Global Urban Footprint. German Aerospace Center. https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-9628/16557_read-40454/

Dong J, Xiao X, Kou W, Qin Y, Zhang G, Li L, Jin C, Zhou Y, Wang J, Biradar C, Liu J, Moore B (2015) Tracking the dynamics of paddy rice planting area in 1986–2010 through time series Landsat images and phenology-based algorithms. Remote Sensing of Environment 160

Dong J, Xiao X, Menarguez MA, Zhang G, Qin Y, Thau D, Biradar C, Moore B (2016) Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine. Remote Sensing of Environment 185

Dyrmann M, Jørgensen RN, Midtiby HS (2017) RoboWeedSupport - Detection of weed locations in leaf occluded cereal crops using a fully convolutional neural network. Advances in Animal Biosciences 8

FAO (2015) FRA 2015 Terms and Definitions. Forest Resources Assessment Working Paper 180. Food and Agricultural Organization of the United Nations.

Felbier A, Esch T, Heldens W, Marconcini M, Zeidler J, Roth A, Klotz M, Wurm M, Taubenbock H (2014) The global urban footprint — Processing status and cross comparison to existing human settlement products. Conference Paper. IEEE Geoscience and Remote 13.07.2014 - 18.07.2014

Fienen MN, Lowry CS (2012) Social.Water—A crowdsourcing tool for environmental data acquisition. Computers & Geosciences 49

Forest survey of India (2017) India State of Forest Report (ISFR). Ministry of Environment, Forest & Climate Change, Government of India 15

Frantz D, Haß E, Uhl A, Stoffels J, Hill J (2018) Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. Remote Sensing of Environment 215:471–481

Fritz S, See L, Perger C, McCallum I, Schill C, Schepaschenko D, Duerauer M, Karner M, Dresel C, Laso-Bayas J-C, Lesiv M, Moorthy I, Salk CF, Danylo O, Sturn T, Albrecht F, You L, Kraxner F, Obersteiner M (2017) A global dataset of crowdsourced land cover and land use reference data. Scientific data 4

Galford GL, Mustard JF, Melillo J, Gendrin A, Cerri CC, Cerri CEP (2008) Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil. Remote Sensing of Environment 112

Gao J, Li Q, Zhao B, Fan W, Han J (2016) Mining Reliable Information from Passively and Actively Crowdsourced Data. In: Krishnapuram B, Shah M, Smola A, Aggarwal C, Shen D, Rastogi R (ed) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. ACM Press, New York, New York, USA

Ghosh S, Das D, Kao S-C, Ganguly AR (2012) Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes. Nature Clim Change 2:86–91

Google Developers (n. d.) Simple, battery-efficient location API for Android. https://developers.google.com/location-context/fused-location-provider/. Accessed 28 January 2020

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202

Government of Andhra Pradesh (2019) Season and crop coverage report. Rabi 2018-19 up to the week ending 16/01/2019.

Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JRG (2013) High-resolution global maps of 21st-century forest cover change. Science 342

Herfort B, Li H, Fendrich S, Lautenbach S, Zipf A (2019) Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning. Remote Sensing 11:1799

Ienco D, Gaetano R, Dupaquier C, Maurel P (2017) Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. IEEE Geoscience and Remote Sensing Letters 14

IMD (2019) Drought Monitoring. India Meteorological Department. http://www.imdpune.gov.in/hydrology/Drought_Monitoring.html

Jain M, Mondal P, DeFries RS, Small C, Galford GL (2013) Mapping cropping intensity of smallholder farms: A comparison of methods using multiple sensors. Remote Sensing of Environment 134

Jain M, Mondal P, Galford GL, Fiske G, DeFries RS (2016) India Annual Winter Cropped Area, 2001-2016. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)

Jakimow B, van der Linden S, Thiel F, Hostert P (2019) EO TIme Series Viewer - A QGIS plugin to explore Earth Observation Time Series Data. Abstract for oral presentation at QGIS User and Developer Conference, A Coruna, Spain, 4-10 March 2019

Jestico B, Nelson T, Winters M (2016) Mapping ridership using crowdsourced cycling data. Journal of Transport Geography 52

Johnson BA, Iizuka K (2016) Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. Applied Geography 67

Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 147

Kontgis C, Schneider A, Ozdogan M (2015) Mapping rice paddy extent and intensification in the Vietnamese Mekong River Delta with dense time stacks of Landsat data. Remote Sensing of Environment 169

Kussul N, Lavreniuk M, Skakun S, Shelestov A (2017) Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. IEEE Geoscience and Remote Sensing Letters 14

Lee SH, Chan CS, Wilkin P, Remagnino P (2015) Deep-plant: Plant identification with convolutional neural networks. 2015 IEEE International Conference on Image Processing, Quebec City

Lesiv M, Laso Bayas JC, See L, Duerauer M, Dahlia D, Durando N, Hazarika R, Kumar Sahariah P, Vakolyuk M'y, Blyshchyk V, Bilous A, Perez-Hoyos A, Gengler S, Prestele R, Bilous S, Akhtar IUH, Singha K, Choudhury SB, Chetri T, Malek Ž, Bungnamei K, Saikia A, Sahariah D, Narzary W, Danylo O, Sturn T, Karner M, McCallum I, Schepaschenko D, Moltchanova E, Fraisl D, Moorthy I, Fritz S (2019) Estimating the global distribution of field size using crowdsourcing. Global change biology 25

Levin N, Lechner AM, Brown G (2017) An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. Applied Geography 79

Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP, Naylor RL (2008) Prioritizing climate change adaptation needs for food security in 2030. Science 319

Loo YY, Billa L, Singh A (2015) Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia. Geoscience Frontiers 6

Löw F, Prishchepov A, Waldner F, Dubovyk O, Akramkhanov A, Biradar C, Lamers J (2018) Mapping Cropland Abandonment in the Aral Sea Basin with MODIS Time Series. Remote Sensing 10

Lowder SK, Skoet J, Raney T (2016) The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide. World Development 87

Lu H, Fu X, Liu C, Li L-g, He Y-x, Li N-w (2017) Cultivated land information extraction in UAV imagery based on deep convolutional neural network and transfer learning. Journal of Mountain Science 14

Luus FPS, Salmon BP, van den Bergh F, Maharaj BTJ (2015) Multiview Deep Learning for Land-Use Classification. IEEE Geoscience and Remote Sensing Letters 12

Mazumdar S, Wrigley S, Ciravegna F (2017) Citizen Science and Crowdsourcing for Earth Observations: An Analysis of Stakeholder Opinions on the Present and Future. Remote Sensing 9

McCulloch WS, Pitts WH (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. The bulletin of mathematical biophysics 5

Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, Francon O, Raju B, Shahrzad H, Navruzyan A, Duffy N, Hodjat B (2019) Evolving Deep Neural Networks. In: Kozma R, Alippi C, Choe Y, Morabito f (ed) Artificial Intelligence in the Age of Neural Networks and Brain Computing

Milioto A, Lottes P, Stachniss C (2017) Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science IV-2/W3

Ministry of Agriculture and Farmers' Welfare (2017) District-wise, season-wise crop production statistics. https://data.gov.in/catalog/district-wise-season-wise-crop-production-statistics?filters%5Bfield_catalog_reference%5D=87631&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc. Accessed 10 November 2019

Ministry of Commerce and Industry (2019) Industrial Development & Economic Growth in Andhra Pradesh. https://www.ibef.org/states/andhra-pradesh.aspx. Accessed 14 November 2019

Mishra V, Smoliak BV, Lettenmaier DP, Wallace JM (2012) A prominent pattern of year-to-year variability in Indian Summer Monsoon Rainfall. Proceedings of the National Academy of Sciences of the United States of America 109

Mohanty SP, Hughes DP, Salathé M (2016) Using Deep Learning for Image-Based Plant Disease Detection. Frontiers in plant science 7

Naidu CV, Dharma Raju A, Satyanarayana GC, Kumar PV, Chranjeevi G, Suchitra P (2015) An observational evidence of decrease in Indian summer monsoon rainfall in the recent three decades of global warming era. Global and Planetary Change 127

Olofsson P, Foody GM, Herold M, Stehman SV, Woodcock CE, Wulder MA (2014) Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148

OpenStreetMap contributors (2019). https://planet.openstreetmap.org

Panteras G, Cervone G (2018) Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring. International Journal of Remote Sensing 39

Patrick Griffiths, Tobias Kuemmerle, Matthias Baumann, Volker C. Radeloff, Ioan V. Abrudan, Juraj Lieskovsky, Catalina Munteanu, Katarzyna Ostapowicz, Patrick Hostert (2014) Forest disturbances, forest recovery, and changes in forest types across the Carpathian ecoregion from 1985 to 2010 based on Landsat image composites. Remote Sensing of Environment 151

Paul MJ, Drezde M (2012) A Model for Mining Public Health Topics from Twitter. Health

PEAT (2020) Progressive Environmental and Agricultural Technologies. http://peat.technology/

Pekel J-F, Cottam A, Gorelick N, Belward AS (2016) High-resolution mapping of global surface water and its long-term changes. Nature 540

Piccinini G (2004) The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity". Synthese 141

Prasanna V (2014) Impact of monsoon rainfall on the total foodgrain yield over India. Journal of Earth System Science 123

Prasuna V, Suneetha B, Madhavi K, Haritha GS, Ramakrishna Murthy GR (2018) Irrigation status, issues and management in Andhra Pradesh. Journal of Pharmacognosy and Phytochemistry

Python Software Foundation Python Language Reference, version 3.7.4. Available at http://www.python.org

QGIS Development Team (2019) QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

Quesada B, Devaraju N, Noblet-Ducoudré N de, Arneth A (2017) Reduction of monsoon rainfall in response to past and future land use and land cover changes. Geophysical Research Letters 44

Rapsomanikis G (2015) The economic lives of smallholder farmers. An analysis based on household data from nine countries. Food and Agricultural Organization of the United Nations

Reddy DN, Motkuri V (2013) SRI Cultivation in Andhra Pradesh: Achievements, Problems and Implications for GHGs and Work. Conference Paper. http://www.apagrisnet.gov.in/eindex.php

Reyes AK, Caicedo JC, Camargo JE (2015) Fine-tuning Deep Convolutional Networks for Plant Recognition. CLEF (Working Notes). Toulouse

Rojas R (1996) Fast Learning Algorithms. In: Rojas (ed) Neural Networks. Springer Berlin Heidelberg, Berlin, Heidelberg

Rosenblatt F (1958) The perceptron: A probalistic model for information storage and organization in the brain. Psychological Review 65

Roxy MK, Chaithra ST (2018) Impacts of Climate Change on the Indian Summer Monsoon. In: Ministry of Environment, Forest and Climate Change (ed) Climate Change and Water Resources in India

Rufin P, Frantz D, Ernst S, Rabe A, Griffiths P, Özdoğan M, Hostert P (2019) Mapping Cropping Practices on a National Scale Using Intra-Annual Landsat Time Series Binning. Remote Sensing 11:232

Sahana AS, Ghosh S, Ganguly A, Murtugudde R (2015) Shift in Indian summer monsoon onset during 1976/1977. Environmental Research Letters 10

Sakamoto T, van Phung C, Kotera A, Nguyen KD, Yokozawa M (2009) Analysis of rapid expansion of inland aquaculture and triple rice-cropping areas in a coastal area of the Vietnamese Mekong Delta using MODIS time-series imagery. Landscape and Urban Planning 92

Salk CF, Sturn T, See L, Fritz S, Perger C (2016) Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. International Journal of Digital Earth 9:410–426

Samberg LH, Gerber JS, Ramankutty N, Herrero M, West PC (2016) Subnational distribution of average farm size and smallholder contributions to global food production. Environmental Research Letters 11

Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M (2017) A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science 31

Shao Y, Lunetta RS, Wheeler B, Iiames JS, Campbell JB (2016) An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data. Remote Sensing of Environment 174

Singh D, Ghosh S, Roxy MK, McDermid S (2019) Indian summer monsoon: Extreme events, historical changes, and role of anthropogenic forcings. Wiley Interdisciplinary Reviews: Climate Change 10

Singh D, Tsiang M, Rajaratnam B, Diffenbaugh NS (2014) Observed changes in extreme wet and dry spells during the South Asian summer monsoon season. Nature Climate Change 4

Sladojevic S, Arsenovic M, Anderla A, Culibrk D, Stefanovic D (2016) Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification. Computational intelligence and neuroscience

Song X-P, Potapov PV, Krylov A, King L, Di Bella CM, Hudson A, Khan A, Adusei B, Stehman SV, Hansen MC (2017) National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. Remote Sensing of Environment 190

Storey JC (2003) Thematic Mapper Bumper Mode Scan Mirror Correction Model Algorithm Description. Scan Line Corrector Theoretical Basis. Version 1.1.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. In: Institute of Electrical and Electronics Engineers (ed) 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Tada PR (2004) Distress and the Deceased. Farmers Suicides in Andhra Pradesh. A Plausible Solution. The IUP Journal of Applied Economics 3

Telangana State Portal (2019) History. https://www.telangana.gov.in/About/History. Accessed 28 November 2019

Tong X, Brandt M, Hiernaux P, Herrmann S, Rasmussen LV, Rasmussen K, Tian F, Tagesson T, Zhang W, Fensholt R (2020) The forgotten land use class: Mapping of fallow fields across the Sahel using Sentinel-2. Remote Sensing of Environment 239

UIDAI (2019) State Wise Total Population (projected 2019). Unique Identification Authority of India. Government of India

USGS (2019a) Landsat Collection 1. https://www.usgs.gov/land-resources/nli/landsat/landsat-collection-1?qt-science_support_page_related_con=1#qt-science_support_page_related_con

USGS (2019b) Landsat Missions. Landsat 8. https://www.usgs.gov/land-resources/nli/landsat/landsat-8?qt-science_support_page_related_con=0#qt-science_support_page_related_con. Accessed 4 November 2019

USGS (2019c) Landsat Missions. Landsat 7. https://www.usgs.gov/land-resources/nli/landsat/landsat-7?qt-science_support_page_related_con=0#qt-science_support_page_related_con. Accessed 4 November 2019

Vaddiraju AK (2004) Farmers' Suicides and Rural Institutions in Andhra Pradesh. Governance and Policy Spaces (GAPS) Project Centre for Economic and Social Studies

Vinnarasi R, Dhanya CT (2016) Changing characteristics of extreme wet and dry spells of Indian monsoon rainfall. Journal of Geophysical Research: Atmospheres 121

Wahl ER, Morrill C (2010) Towards Understanding and Predicting Monsoon Patterns. Science 328

Walden-Schreiner C, Leung Y-F, Tateosian L (2018) Digital footprints: Incorporating crowdsourced geographic information for protected area management. Applied Geography 90

Woodcock CE, Allen R, Anderson M, Belward A, Bindschadler R, Cohen W, Gao F, Goward SN, Helder D, Helmer E, Nemani R, Oreopoulos L, Schott J, Thenkabail PS, Vermote EF, Vogelmann J, Wulder MA, Wynne R (2008) Free access to Landsat imagery. Science 320

World Bank (2008) Climate Change Impacts in Drought and Flood Affected Areas: Case Studies in India

Wulder MA, Masek JG, Cohen WB, Loveland TR, Woodcock CE (2012) Opening the archive: How free data has enabled the science and monitoring promise of Landsat. Remote Sensing of Environment 122

Yan L, Roy DP, Li Z, Zhang HK, Huang H (2018) Sentinel-2A multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology. Remote Sensing of Environment 215

## Appendix

### Appendix 1: Background about deep artificial neural networks

Artificial neural networks (ANNs) are computing systems that are inspired by information processing and distributed communication nodes of biological neural networks. The beginnings of research in neural networks date back to the 1940s with McCulloch and Pitts' (1943) first mathematical model of an artificial neuron widely seen as the pivotal publication triggering research in neural networks (Piccinini 2004). The first pioneering research in machine learning using simple algorithms started in the 1950s. Rosenblatt introduced the first simplest implementation of a neural network which consisted of a single artificial neuron in 1958. In the following decade, research on machine learning was focused on using logical, knowledge-based approaches. However, after the pioneering work of Rosenblatt and others, no efficient learning algorithm for multilayer neural network was known. This led to a pessimism around machine learning effectiveness and a stagnation of research in this field during the 1970s. The rediscovery of the backpropagation (backward propagation of errors) algorithm in the 1980s, together with the development of alternative network topologies initiated a resurgence in machine learning research (Rojas 1996). The following decade was characterized by a shift to a data-driven approach and an enormous increase in machine-learning research. Computer programs were developed to analyze large amounts of data and draw conclusions ('learn') from the results (Dietterich 1997). Since the 2000s large databases i.e., Big Data have become readily available as a result of the falling cost of large data storage and the increasing ease of collecting data over networks. Simultaneously, the amount of computing power had further increased by a considerable degree. This paved the way for a scaling up of machine learning systems. Moreover, besides the successful scaling up, the systems have become more powerful. Certain ideas that did not work before, can now be realized with millionfold more computing power and data (Miikkulainen et al. 2019). In the last decade, deep learning has become feasible, leading to the integration of machine learning in many widely used software devices and applications. Deep learning neural networks (DNNs) have improved state of the art substantially in various fields like computer vision and language processing (Collobert and Weston 2008; Szegedy et al. 2016). DNNs are typically Feed Forward Networks (FFNNs) in which the information flows only forward in the network, thus no feedback connections are applied. Deep learning (DL) refers to systems that use sophisticated mathematical modeling to process data in complex ways. The crucial characteristic of DNNs is that these networks are able to classify and order information in ways that go beyond simply input/output protocols. They are distinguished from the more trivial shallow neural networks by their depth, i.e., the number of node layers through which data must pass in a multistep pattern recognition process. Earlier neural network versions were composed of one input and one output layer, and at most one hidden layer in between. Deep learning is defined by using more than one hidden layer. Throughout the hidden layers a feature hierarchy with several levels of abstraction is applied. Each layer of nodes trains on a distinct set of features based on the previous layer's output, i.e., higher-level learned features are composed of lower-level features. To achieve an acceptable level of accuracy, DNN models need to be trained on large datasets (Miikkulainen et al. 2019).

Deep learning applied in agricultural contexts has become popular in recent years. The most frequent fields of application in agriculture are identification of weeds (Dyrmann et al. 2017; Milioto et al. 2017), land cover classification (Luus et al. 2015; Lu et al. 2017; Ienco et al. 2017), plant disease detection (Mohanty et al. 2016; Amara et al. 2017), plant recognition (Reyes et al. 2015; Lee et al. 2015) and crop type classification (Kussul et al. 2017) with the large majority of the publications dealing with image classification targeting crops (Kamilaris and Prenafeta-Boldú 2018).

## Appendix 2: Plantix data filtering counts.

| Filter step | TOTAL | share of raw data count | 2017 | share of raw data count | 2018 | share of raw data count | 2019 | share of raw data count |
|---|---|---|---|---|---|---|---|---|
| Raw submissions count for AOI & timespan | **213,394** | | **28,172** | | **115,023** | | **70,199** | |
| 1) Removal non-plant submissions | **147,554** | **69.1%** | 21,414 | 76.0% | 72,430 | 63.0% | 53,710 | 76.5% |
| 2) Removal multiple submissions | | | | | | | | |
| (1) from the same crop at the same location | **107,109** | **50.2%** | 4,700 | 16.7% | 62,203 | 54.1% | 40,206 | 57.3% |
| (2) from the same location | **89,996** | **42.2%** | 3,913 | 13.9% | 52962 | 46.0% | 33,121 | 47.2% |
| 3) Removal of submissions uploaded from smartphone gallery | **75,591** | **35.4%** | - | - | 43,489 | 37.8% | 28,189 | 40.2% |
| 4) Removal of submissions located in urban areas | **52,055** | **24.4%** | 2,781 | 9.9% | 28,285 | 24.6% | 20,989 | 29.9% |
| 5) Removal points nearby roads | **50,202** | **23.5%** | 2,663 | 9.5% | 27,238 | 23.7% | 20,301 | 28.9% |
| 6) Removal of tree varieties | **36,758** | **17.2%** | 2,345 | 8.3% | 19,452 | 16.9% | 14,961 | 21.3% |
| 7) DNN similarity | | | | | | | | |
| ≥ 50% | **32,047** | **15.1%** | 2,185 | 7.6% | 16,512 | 14.4% | 13,350 | 19.0% |
| ≥ 80% | **27,274** | **12.8%** | 1,949 | 8.3% | 13,767 | 11.9% | 11,558 | 16.5% |
| 8) GPS accuracy | | | | | | | | |
| DNN ≥ 50% & | | | | | | | | |
| GPS acc. ≤100 m | **4,871** | **2.9%** | 1,864 | 6.6% | 1,904 | 1.6% | 1,103 | 1.6% |
| GPS acc. ≤ 30 m | **4,035** | **1.9%** | 1,814 | 6.4% | 1,394 | 1.2% | 827 | 1.2% |
| GPS acc. ≤ 10 m | **2,932** | **1.4%** | 1,563 | 5.5% | 802 | 0.7% | 567 | 0.8% |
| DNN ≥ 80% & | **4,201** | **1.9%** | 1,660 | 5.9% | 1,586 | 1.4% | 955 | 1.4% |
| GPS acc. ≤100 m | **3,528** | **1.6%** | 1,616 | 5.7% | 1,183 | 1.0% | 729 | 1.0% |
| GPS acc. ≤ 30 m | **2,607** | **1.2%** | 1,400 | 5.0% | 696 | 0.6% | 511 | 0.7% |
| GPS acc. ≤ 10 m | | | | | | | | |

**Appendix 3: Detailed results of DNN accuracy level testing of major non-tree varieties submissions from Andhra Pradesh.**

Per range step a minimum of 30 images per crop were tested.

| Crop variety | Image count after filter step 6 | Range step: 50-60% | | Range step: 60-70% | | Range step: 70-80% | | Range step: 80-90% | | Range step: 90-100% | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Share of true DNN results | *Image share within range step* | Share of true DNN results | *Image share within range step* | Share of true DNN results | Image share within range step | Share of true DNN results | *Image share within range step* | Share of true DNN results | *Image share within range step* |
| **Bean** | 9383 | 94.59% | *14.86%* | 91.89% | *15.18%* | 94.87% | 17.96% | 97.50% | *22.59%* | 97.50% | *29.41%* |
| **Chickpea** | 3801 | 65.79% | *5.97%* | 55.26% | *4.87%* | 86.84% | 6.18% | 82.50% | *8.81%* | 100.00% | *74.16%* |
| **Cotton** | 67245 | 51.61% | *1.22%* | 74.19% | *1.32%* | 72.41% | 1.80% | 91.89% | *9.85%* | 100.00% | *85.81%* |
| **Cucumber** | 11982 | 79.49% | *9.11%* | 82.05% | *10.28%* | 87.50% | 14.59% | 95.00% | *29.19%* | 89.74% | *36.84%* |
| **Eggplant** | 18232 | 66.67% | *3.64%* | 52.78% | *3.88%* | 84.62% | 4.45% | 100.00% | *12.27%* | 87.50% | *75.76%* |
| **Gram** | 11019 | 83.87% | *7.72%* | 96.77% | *8.92%* | 91.89% | 9.71% | 100.00% | *20.61%* | 100.00% | *53.04%* |
| **Lentil** | 542 | 35.90% | *29.15%* | 46.15% | *18.27%* | 63.89% | 22.14% | 60.00% | *14.76%* | 66.67% | *15.68%* |
| **Maize** | 17128 | 89.74% | *4.12%* | 92.31% | *5.44%* | 100.00% | 7.36% | 97.50% | *21.23%* | 100.00% | *61.85%* |
| **Melon** | 8998 | 59.09% | *7.65%* | 36.36% | *7.32%* | 53.33% | 8.70% | 62.16% | *16.55%* | 91.89% | *59.78%* |
| **Millet** | 862 | 75.00% | *22.97%* | 80.00% | *18.68%* | 85.00% | 12.53% | 85.00% | *13.32%* | 100.00% | *32.60%* |
| **Onion** | 6491 | 27.03% | *3.96%* | 48.65% | *5.01%* | 68.42% | 4.64% | 92.11% | *10.54%* | 100.00% | *75.86%* |
| **Peanut** | 20492 | 96.43% | *2.21%* | 114.29% | *2.55%* | 82.35% | 3.40% | 100.00% | *7.72%* | 100.00% | *84.13%* |
| **Pepper** | 86334 | 48.00% | *3.17%* | 48.00% | *3.14%* | 62.96% | 3.90% | 87.10% | *13.19%* | 100.00% | *76.60%* |
| **Pigeonpea** | 4473 | 92.31% | *5.92%* | 123.08% | *6.89%* | 100.00% | 9.12% | 100.00% | *16.25%* | 100.00% | *61.82%* |
| **Potato** | 1589 | 58.82% | *14.08%* | 73.53% | *11.14%* | 69.70% | 9.64% | 67.86% | *12.27%* | 100.00% | *52.88%* |
| **Rice** | 97072 | 88.89% | *1.28%* | 107.41% | *1.69%* | 93.55% | 2.45% | 100.00% | *7.24%* | 100.00% | *87.34%* |
| **Sorghum** | 478 | 100.00% | *29.29%* | 105.41% | *22.80%* | 100.00% | 16.95% | 100.00% | *16.11%* | 100.00% | *14.85%* |
| **Soybean** | 1064 | 81.48% | *29.51%* | 100.00% | *21.05%* | 81.82% | 16.82% | 94.44% | *15.98%* | 100.00% | *16.64%* |
| **Sugarcane** | 2643 | 74.07% | *13.85%* | 92.59% | *11.20%* | 91.30% | 12.30% | 82.35% | *15.59%* | 100.00% | *47.07%* |
| **Tomato** | 44478 | 44.83% | *2.97%* | 58.62% | *3.00%* | 90.63% | 3.40% | 100.00% | *15.38%* | 100.00% | *75.24%* |
| **Wheat** | 1138 | 42.86% | *31.11%* | 52.38% | *24.69%* | 86.36% | 19.77% | 70.00% | *16.26%* | 89.29% | *8.17%* |

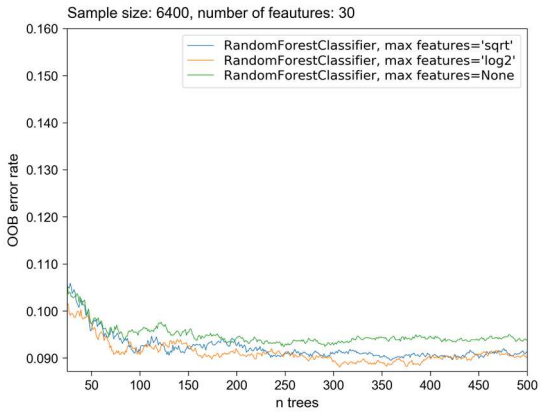
**Appendix 4: List of crops that were included in the training datasets with respective counts after filter step 6.**

The crops with a count above 1000 were included in the Andhra Pradesh top crops training data set**.**

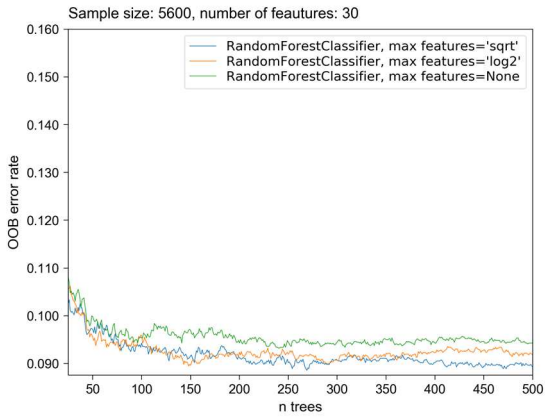
| Crop variety | Count after filterstep 6 | | |
|---|---|---|---|
| **Rice** | **8331** | Cabbage | 317 |
| **Pepper** | **6273** | Onion | 312 |
| **Tomato** | **4042** | Pumpkin | 292 |
| **Peanut** | **2795** | Wheat | 276 |
| **Maize** | **2520** | Millet | 181 |
| **Eggplant** | **1896** | Zucchini | 161 |
| **Bean** | **1695** | Lentil | 130 |
| **Gram** | **1488** | Okra | 104 |
| **Cucumber** | **1360** | Sorghum | 92 |
| **Cotton** | **1337** | Strawberry | 84 |
| Soybean | 807 | Pea | 73 |
| Melon | 653 | Sugarcane | 70 |
| Chickpea | 578 | Raspberry | 69 |
| Rose | 573 | Barley | 60 |
| Potato | 476 | Currant | 51 |
| Grape | 433 | Lettuce | 51 |
| Pigeonpea | 320 | | |

## Appendix 5: Out-of-Bag error plots

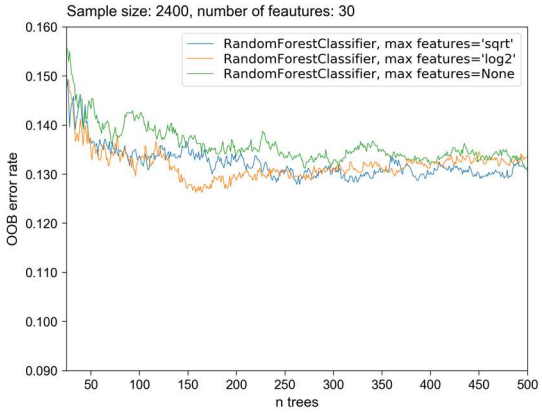
Feature subset (6)



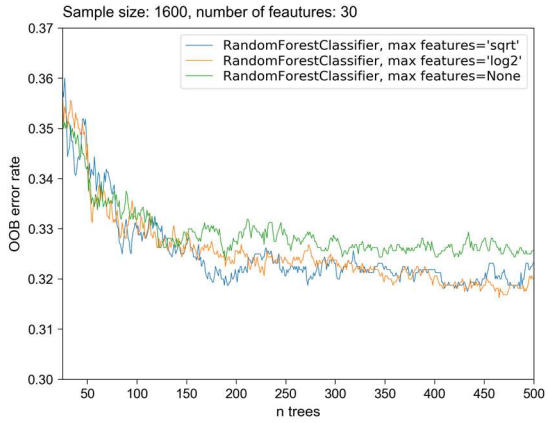

Feature subset (7)




Feature Subset (8)




Feature Subset (10)




## Appendix 6: Overall classification accuracies for all feature subsets and classified maps

| | 2017 | | 2018 | | 2019 | |
|---|---|---|---|---|---|---|
| **Feaure. subset** | **OA** | **area-adj. OA [95% CI]** | **OA** | **area-adj. OA [95% CI]** | **OA** | **area-adj. OA [95% CI]** |
| 2017 - 2019 submissions, DNN ≥ 50%, acc ≤ 100m, sample size 6400 | 81.61% | **86.87% [± 1.35]** | 86.06% | **92.25% [± 1.29]** | 86.06% | **88.76% [± 1.16]** |
| 2017 - 2019 submissions, DNN ≥ 50%, acc ≤ 30m, sample size 6400 | 81.87% | **86.54% [±1.39]** | 86.13% | **89.41% [± 1.20]** | 86.14% | **88.39% [± 1.17]** |
| 2017 - 2019 submissions, DNN ≥ 50%, acc ≤ 10m, sample size 6400 | 83.10% | **86.66% [± 1.41]** | 88.52% | **92.79% [± 1.26]** | 88.52% | **89.65% [± 1.12]** |
| 2017 - 2019 submissions, DNN ≥ 80%, acc ≤ 100m, sample size 6400 | 82.45% | **86.86% [± 1.37]** | 86.39% | **92.57% [± 1.23]** | 86.40% | **88.83% [± 1.11]** |
| 2017 - 2019 submissions, DNN ≥ 80%, acc ≤ 30m, sample size 6400 | 82.65% | **86.52% [± 1.41]** | 88.19% | **92.77% [± 1.24]** | 88.19% | **89.53% [± 1.11]** |
| 2017 - 2019 submissions, DNN ≥ 80%, acc ≤ 10m, sample size 6400 | 82.90% | **86.22% [± 1.43]** | 88.84% | **92.42% [± 1.32]** | 88.84% | **89.54% [± 1.15]** |
| 2017 - 2019 submissions, major crop varieties, DNN ≥ 80%, acc ≤ 10m, sample size 5600 | 83.68% | **86.02% [± 1.44]** | 89.03% | **91.22% [± 1.47]** | 89.03% | **88.78% [± 1.27]** |
| 2017 submissions, DNN ≥ 80%, acc ≤ 10m, sample size 2400 | 87.16% | **87.54% [± 1.39]** | 90.71% | **90.87% [± 1.41]** | 90.71% | **89.61% [± 1.23]** |
| 2019 only Plantix submissions, DNN ≥ 80%, acc ≤ 10m ,sample size 2400 | 77.94% | **84.75% [± 1.65]** | 84.32% | **92.04% [± 1.33]** | 84.45% | **88.39% [± 1.20]** |

## Appendix 7: Confusion matrices for the best and worst performing feature subset for the classification of the whole AOI

Confusion matrices for (a) 2018 classification of feature subset 8 (Best estimate location strategy: P & G submissions from 2017, all non-tree crop varieties DNN similarity ≥ 80%, location accuracy ≤ 10m, subset size 2400) and (b) feature subset 9 (Fused location provider: P sub-missions from 2019, all non-tree crop varieties DNN similarity ≥ 80%, location accuracy ≤ 10m, subset size 2400). [CRO = cropped, WAT = water, URB = urban, WOO = woody canopy, UNS/SP = unsown/bare land/sparse natural vegetation]

(1a)

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **467** | 0 | 9 | 7 | 16 | 500 |
| WAT | 0 | **200** | 0 | 0 | 0 | 200 |
| URB | 10 | 0 | **188** | 0 | 2 | 200 |
| WOO | 31 | 0 | 1 | **265** | 3 | 300 |
| UNS/SP | 32 | 5 | 25 | 2 | **285** | 350 |
| Sum | 541 | 205 | 223 | 274 | 307 | 1550 |

(1b)

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **481** | 0 | 5 | 8 | 4 | 500 |
| WAT | 1 | **198** | 0 | 1 | 0 | 200 |
| URB | 30 | 0 | **167** | 1 | 2 | 200 |
| WOO | 37 | 0 | 0 | **263** | 0 | 300 |
| UNS/SP | 137 | 3 | 11 | 3 | **198** | 350 |
| Sum | 686 | 201 | 183 | 276 | 20 | 1550 |

Cells populated with adjusted probabilities.

(2a)

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.5238** | 0.0000 | 0.0101 | 0.0079 | 0.0191 | **0.5608** |
| WAT | 0.0000 | **0.0188** | 0.0000 | 0.0000 | 0.0000 | **0.0188** |
| URB | 0.0032 | 0.0000 | **0.0609** | 0.0000 | 0.006 | **0.0648** |
| WOO | 0.0207 | 0.0000 | 0.0007 | **0.1766** | 0.0020 | **0.1999** |
| UNS/SP | 0.0147 | 0.0022 | 0.0111 | 0.0009 | **0.1267** | **0.1557** |

(2b)

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.6645** | 0.0000 | 0.0069 | 0.0111 | 0.0055 | **0.6880** |
| WAT | 0.0001 | **0.0175** | 0.000 | 0.0001 | 0.0000 | **0.0177** |
| URB | 0.0068 | 0.0000 | **0.0379** | 0.0002 | 0.0005 | **0.0454** |
| WOO | 0.0237 | 0.0000 | 0.0000 | **0.1686** | 0.0000 | **0.1923** |
| UNS/SP | 0.0220 | 0.0005 | 0.0018 | 0.0005 | **0.0318** | **0.0566** |

## Appendix 8: Confusion matrices for the classification of the four agro-climatic zones

Confusion matrices for the 2018 classification of feature subset 8 (Best estimate location strategy: P & G submissions from 2017, all non-tree crop varieties DNN similarity ≥ 80%, location accuracy ≤ 10m, subset size 2400). Cells populated with adjusted probabilities.

North Coastal Zone

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.4173** | 0.0000 | 0.0422 | 0.000 | 0.0092 | 0.4688 |
| WAT | 0.0000 | **0.0065** | 0.0001 | 0.0000 | 0.0000 | 0.0067 |
| URB | 0.0045 | 0.0000 | **0.1041** | 0.0000 | 0.0045 | 0.1131 |
| WOO | 0.0373 | 0.0000 | 0.0034 | **0.2947** | 0.0034 | 0.3387 |
| UNS/SP | 0.0044 | 0.0029 | 0.0102 | 0.000 | **0.0552** | 0.0727 |

Krishna-Godavari Zone

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.6095** | 0.0000 | 0.0031 | 0.0031 | 0.0000 | 0.6157 |
| WAT | 0.0000 | **0.0271** | 0.0000 | 0.0000 | 0.0000 | 0.0271 |
| URB | 0.0016 | 0.0000 | **0.0771** | 0.0000 | 0.0000 | 0.0787 |
| WOO | 0.0038 | 0.0000 | 0.0000 | **0.1854** | 0.0000 | 0.1892 |
| UNS/SP | 0.0045 | 0.0036 | 0.0072 | 0.0009 | **0.0733** | 0.0894 |

Scarce Rainfall Zone

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.5431** | 0.0000 | 0.0000 | 0.0000 | 0.0058 | 0.5489 |
| WAT | 0.0000 | **0.0107** | 0.0000 | 0.0000 | 0.0000 | 0.0107 |
| URB | 0.0016 | 0.0000 | **0.0368** | 0.0000 | 0.0016 | 0.0400 |
| WOO | 0.0015 | 0.0000 | 0.0000 | **0.0717** | 0.0000 | 0.0731 |
| UNS/SP | 0.0257 | 0.0032 | 0.0193 | 0.0000 | **0.2792** | 0.3273 |

Southern Zone

| Classification \ Reference | CRO | WAT | URB | WOO | UNS/SP | Sum |
|---|---|---|---|---|---|---|
| CRO | **0.4589** | 0.0000 | 0.0069 | 0.0000 | 0.0219 | 0.5463 |
| WAT | 0.0033 | **0.0000** | 0.0377 | 0.0000 | 0.0000 | 0.0410 |
| URB | 0.0033 | 0.0000 | **0.0377** | 0.0000 | 0.0000 | 0.0410 |
| WOO | 0.1380 | 0.0000 | 0.0049 | **0.0887** | 0.0148 | 0.2465 |
| UNS/SP | 0.0160 | 0.0000 | 0.0073 | 0.0015 | **0.1206** | 0.1453 |

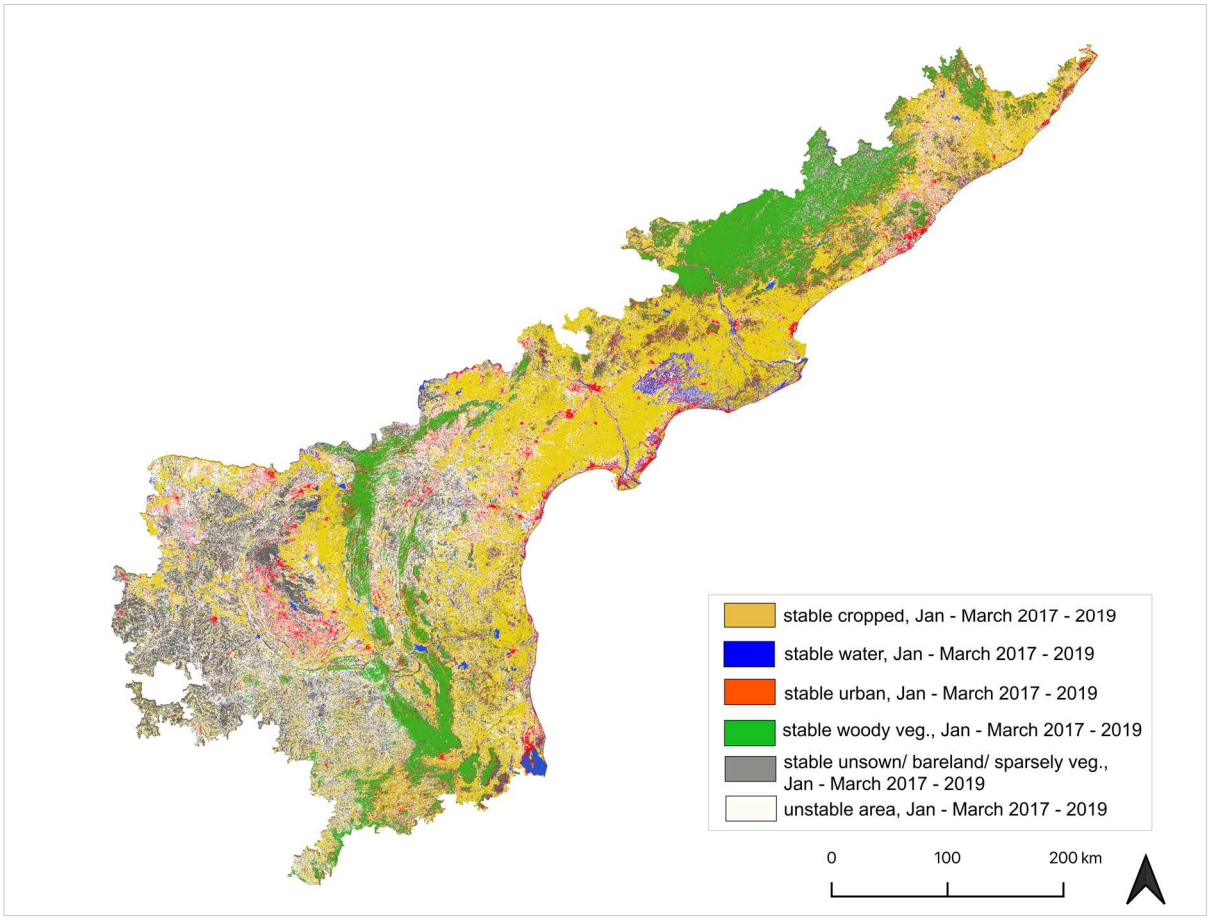
**Appendix 9: Cropped area map estimates for the timespan 1 Jan - 31 March of the entire AOI for the maps with the highest PA and UA accuracies.**

| Feature subset | 2017 cropped area map estimate in Mha [95% CI] | 2018 cropped area map estimate in Mha [95% CI] | 2019 cropped area map estimate in Mha [95% CI] |
|---|---|---|---|
| (6) DNN ≥ 80%, loc. acc. ≤ 10m | 10.33 [±1.60] | 10.99 [±1.36] | 9.72 [±1.44] |
| (8) best estimate loc. strategy, DNN ≥ 80%, loc. acc. ≤ 10m | 8.58 [±1.54] | 9.16 [±1.49] | 7.81 [±.1.44] |

**Appendix 10: Cropped area map estimates for the timespan 1 Jan – 31 March of the agro-climatic zones (feature subset 8 utilized)**

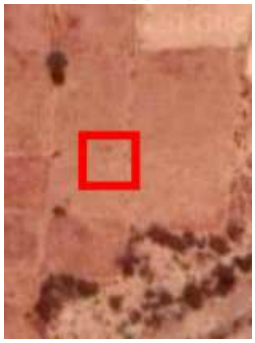
| Agro-climatic zone | Feature subset | 2017 cropped area map estimate in Mha [95% CI] | 2018 cropped area map estimate in Mha [95% CI] | 2019 cropped area map estimate in Mha [95% CI] |
|---|---|---|---|---|
| **North Coastal Zone** | (8) best estimate loc. strategy | 1.04 [±2.81] | 1.09 [±3.65] | 1.08 [±2.86] |
| **Krishna-Godavari Z.** | (8) best estimate loc. strategy | 3.58 [±2.23] | 3.66 [±1.11] | 3.36 [±1.14] |
| **Scare Rainfall Zone** | (8) best estimate loc. strategy | 1.40 [±2.72] | 2.1 [±2.09] | 1.20 [±3.11] |
| **Southern Zone** | (8) best estimate loc. strategy | 2.63 [±4.23] | 2.69 [±5.31] | 2.52 [±4.63] |

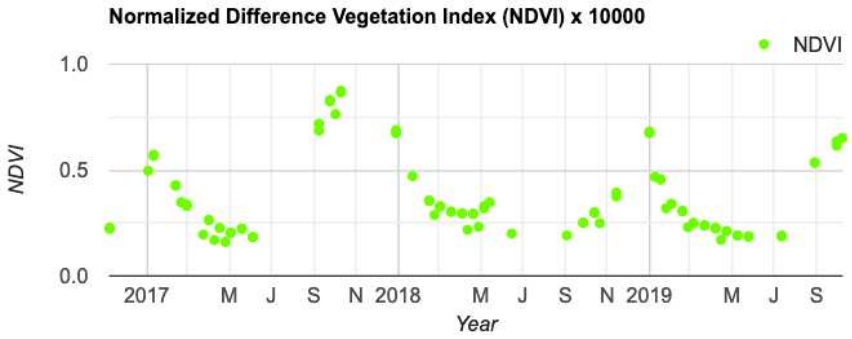
**Appendix 11: Aggregated map showing areas of stable map classes from 1 Jan - 31 March 2017 - 2019**



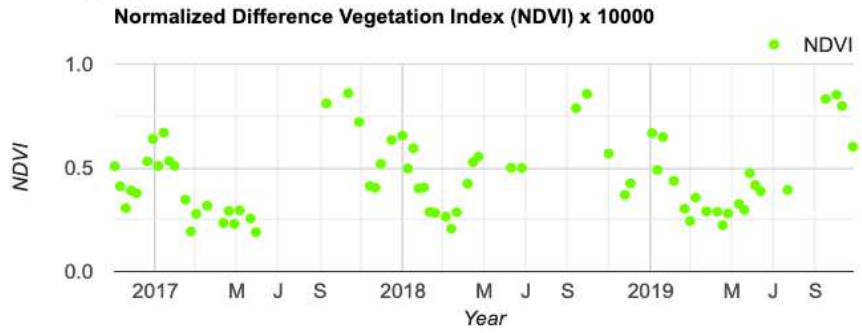
**Appendix 12: NDVI time series for selected example points showing the temporal differences in the cropping seasons in Andhra Pradesh**

8.1 Kharif cropped plot (late harvest)

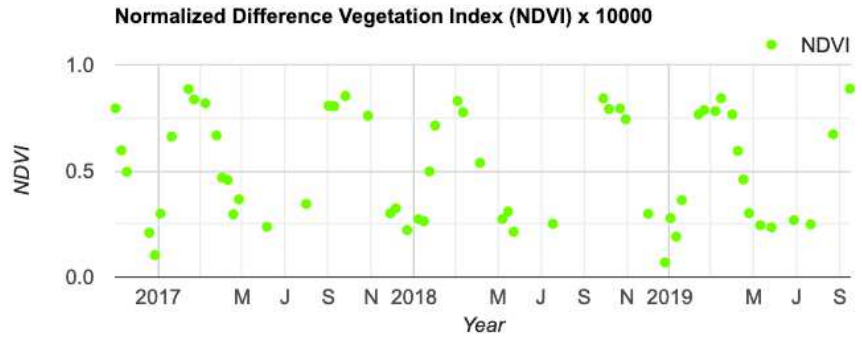




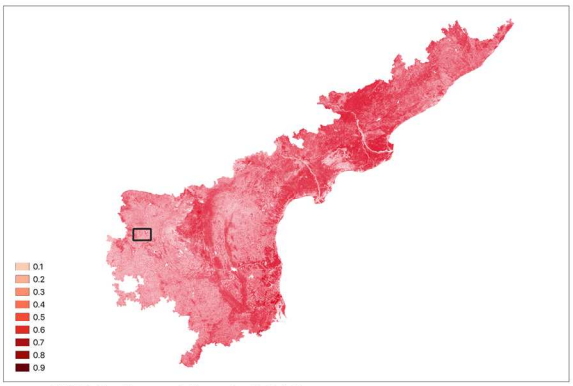

8.2 Kharif (early harvest) and Rabi cropped plot



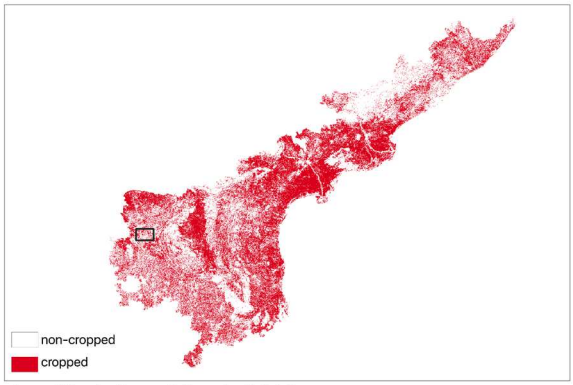

8.3 Plot cropped with rice irrigated, Kharif (early harvest) and Rabi cropped



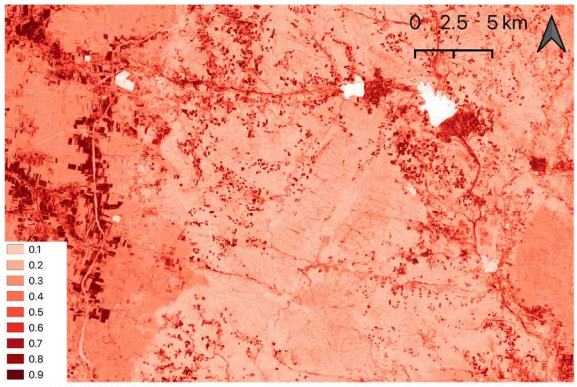

**Appendix 13: Maximum NDVI composites compared with classification result**



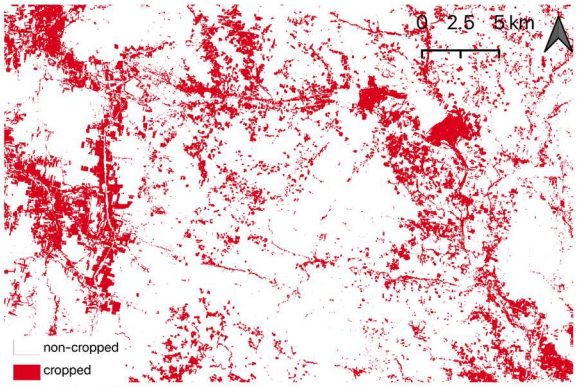

max NDVI Jan - March 2018



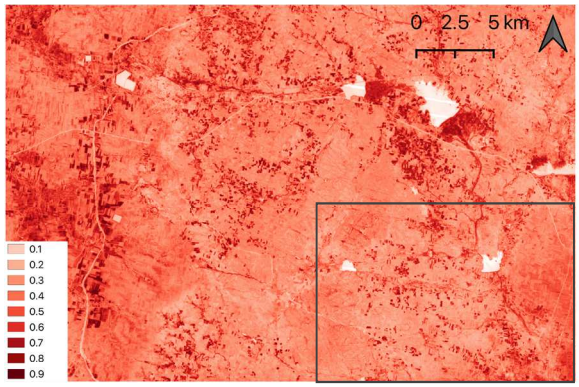

classified Jan - March 2018



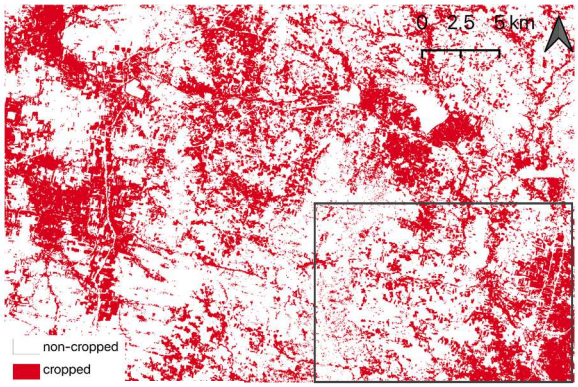

max NDVI2017



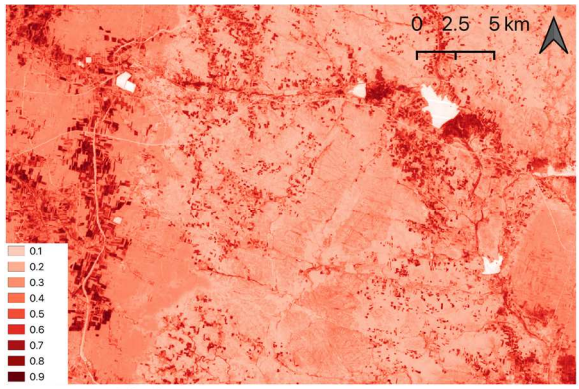

classified 2017



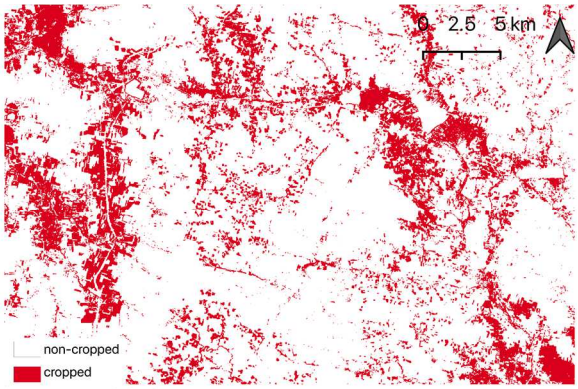

maxNDVI2018




classified 2018

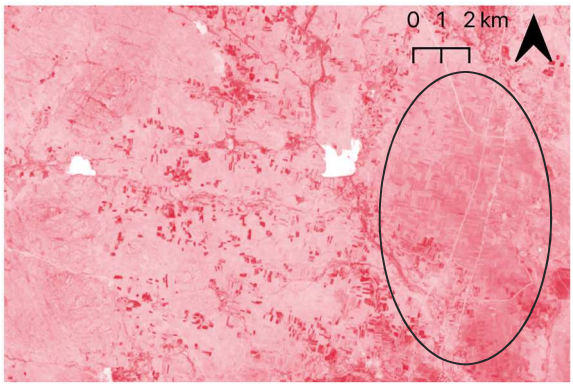


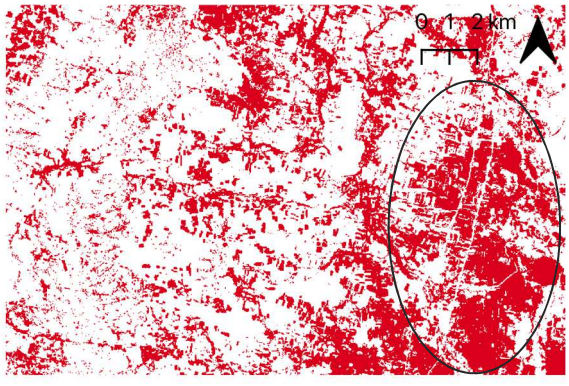

max NDVI2019



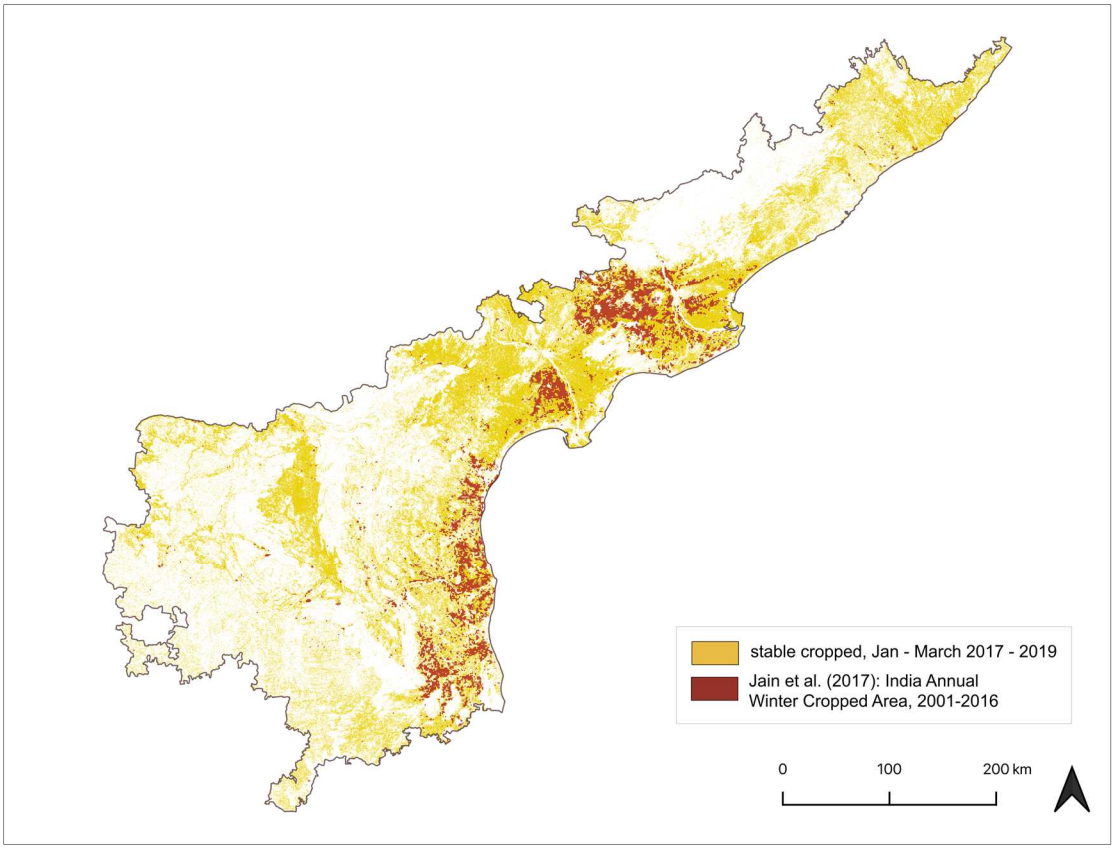

classified 2019

**Detail 2018**



**Appendix 14: Map overlay of Jain et al. (2017) with classified stable cropped area Jan - March 2017-2019**



ERKLÄRUNG:

Ich erkläre, dass ich die vorliegende Arbeit nicht für andere Prüfungen eingereicht, selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe. Sämtliche fremde Quellen inklusive Internetquellen, Grafiken, Tabellen und Bilder, die ich unverändert oder abgewandelt wiedergegeben habe, habe ich als solche kenntlich gemacht. Mir ist bekannt, dass Verstöße gegen diese Grundsätze als Täuschungsversuch bzw. Täuschung geahndet werden.

Berlin, den 28.02.2020

_____