Title
Monitoring agricultural behavior under climate change with cloud computing and satellite imagery

Author
Zhang, Minghui

Publication Date
2020

Peer reviewed|Thesis/dissertation

Monitoring agricultural behavior under climate change with cloud computing and satellite imagery

by

Minghui Zhang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Sally Thompson, Co-chair
Fotini K. Chow, Co-chair
Ashok Gadgil
Iryna Dronova

Summer 2020

Monitoring agricultural behavior under climate change with cloud computing and satellite imagery

Abstract

Monitoring agricultural behavior under climate change with cloud computing and satellite imagery

by

Minghui Zhang

Doctor of Philosophy in Engineering- Civil and Environmental Engineering

University of California, Berkeley

Sally Thompson, Co-chair

Fotini K. Chow, Co-chair

Understanding agricultural productivity under climate change is critical to helping global food systems tackle impending challenges in supply and demand. Crucially, this requires us to understand planting dates: by controlling the yield and cropping intensity of rainfed agriculture, planting dates could serve as a major adaptation strategy under climate pressure. Currently, the lack of spatiotemporally resolved crop timing information makes it difficult to produce insights into planting behavior, which is the result of complex human decisions made under varying socio-economic and climatic contexts. This data gap hinders our ability to understand how farmers will adapt to climate change by shifting planting dates, and may negatively impact the accuracy of yield predictions. This dissertation addresses this data gap by introducing a scalable method to estimate planting dates. I apply this method to generate insights into historical and future planting dates of soy (*Glycine max*) in the heavily agricultural state of Mato Grosso, Brazil (MT). My work begins with a remote sensing-based method to estimate field-scale (500 m) planting and harvest dates over large (100,000 km^2), and therefore computationally challenging areas, with sparse ground truth information. The method pairs (1) a timeseries analysis algorithm for MODIS imagery, implementable on the cloud computing platform Google Earth Engine (GEE), to extract 500 m phenological milestones and (2) proxy ground truth data based on Planet Labs imagery to relate phenological milestones to observed planting and harvest dates. Next, I build a statistical model of satellite-estimated planting dates as a function of the wet season onset. This model reveals several novel insights about agricultural behavior in Mato Grosso. First, traditional climatological definitions of wet season onset are less correlated to observed planting dates than alternative, easily observable definitions based on rainfall frequency, highlighting the need to explore farmer-relevant definitions of climate. Second, planting dates' sensitivity to wet season onset varies dramatically among fields and with cropping

intensity; this heterogeneous response produces a wide range of planting dates that are rarely included in crop models. Finally, a trend toward earlier planting dates, independent of wet season onset, reveals nonstationary behavior that cannot be captured with the historical survey data used by some yield prediction efforts. My findings suggest that under RCP 8.5 climate conditions, climatic windows will constrain planting dates relative to agronomically preferred times, and the feasibility of double cropping will be endangered for vulnerable portions of Mato Grosso. Both delayed planting dates and loss of double cropping suitability are problematic for an economy that largely depends on agribusiness and that is central to international soy supply. While Mato Grosso is the focus of this dissertation, the methods developed here lay the groundwork for similar studies globally. By introducing a scalable method to close the information gap on planting dates and generating new insights into the planting dates of tropical rainfed crops, this work provides a foundation for investigating planting date behavior and climate change adaptation in vulnerable, data-scarce agricultural regions worldwide.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Sally Thompson, for her steady and kind mentorship; my committee members, Tina Chow, Ashok Gadgil, Iryna Dronova, Avery Cohn and Paolo D'Odorico, for their encouragement and perspective; my co-authors, Gabriel Abrahao, Avery Cohn, and Jake Campolo for their expertise and significant contributions; my lab mates and classmates, Liya Weldegebriel, Gopal Penny, David Dralle, Octavia Crompton, Katya Rakhmatulina, Jeannie Wilkening, Morgan Levy, Karina Cucchi, Michaella Chung, Xue Feng, Gabrielle Boisrame, Dana Lapides, and George Greer for their camaraderie and support; and my parents, husband, and friends for giving me strength and community.

Finally, I would like to acknowledge the funding sources that have made this dissertation possible: NSF Graduate Research Fellowship Program, NSF EAR, and the Gordon and Betty Moore Foundation.

Chapter 1

Introduction

1.1 Agricultural productivity under climate change

1.1.1 Agriculture faces pressure under climate change

Population growth, dietary change, and increased biofuel use will require a doubling of agricultural production in this century [45], a challenge compounded by climate change. Generally, and especially in the tropics, climate change is expected to harm food production systems. A reduction in climatically suitable land and elevated likelihood of extreme weather events like heat waves, droughts, and flooding are expected worldwide [123]. Additionally, a hotter, drier climate is projected over tropical regions, where crops are heavily reliant on rainfall and already experience the upper limit of their temperature tolerance [25, 113]. In tropical climates, even moderate temperature increases of 1°C can produce declines in yield [25, 113]. Higher temperatures negatively impact crop growth by increasing photorespiration, reducing carbon assimilation per unit of water consumed, and decreasing nutrient absorption [69]. Consequently, the mean yield of major crops is expected to decline by 8% in the 2050s across Africa and South Asia [79], with maize yields declining sharply at temperatures above 30°C [122].

In addition to reducing the yield of individual crops, climate change may force shifts in agricultural practices that sharply decrease overall production. For example, one concern about agricultural adaptation to climate change is whether double cropping (the sequential planting of two crops in a single growing season) will decline in the face of changing temperature and precipitation seasonality [106, 124].

The damage caused by climate change may be mitigated with adaptation strategies [78, 111]. Management has such a strong influence on productivity that farmers may be able to minimize or even reverse the harmful effects of climate change through a broad range of strategies, including cultivar selection, increased cropping intensity, improved water management, nutrient and pest management, and agroforestry [4, 8, 19, 58, 63, 111, 137, 143]. The adaptation strategies implemented in agricultural systems will determine the health of economies, food systems, and societies globally. Successful adaptation could prevent global

food crises; unsuccessful adaptation may hasten them.

1.1.2 Planting dates are an adaptation strategy under climate change

As a primary control on the weather experienced by crops, the decision of when to plant is one of the most cost-effective adaptations to climate change. A knowledgeable farmer can select planting dates to optimize the exposure of crops to beneficial weather conditions and to avoid exposure to harmful conditions during sensitive phenological stages. Examples spanning production systems from sorghum in Italy [63]; to soybean in Austria [5], the US south east [14], and sub-Saharan Africa [147]; rice in Sri Lanka [34] and India [69]; wheat in Iran [101]; and corn in the US midwest [81] and Burkina Faso [149] demonstrate that adaptations in planting dates to avoid drought and heat stress could mitigate or reverse the negative effects of climate change on production.

However, planting dates of rainfed crops are limited by the timing of the wet season. The gains from changing planting dates may be reduced if planting dates are restricted to inopportune times by a delay in wet season onset (a date which I will refer to interchangeably as "onset"). A later onset necessitates delayed, and possibly suboptimal, planting of rainfed crops, in which crop growth stages can no longer be synchronized with optimum weather conditions. For example, a 1% decrease in wheat yield was observed for every day of delayed planting in northern India, a decline caused by heat stress during the grain filling period of crop growth [102]. In the US, soy yields decrease linearly with delayed planting due to reduced moisture supply [15]. Soy yields are also influenced by temperature. If the available planting dates force soy to experience hotter weather, the higher temperatures accelerate phenological development, allowing less time for biomass accumulation before maturity. This hastens the harvest date and reduces yield [127].

Much of the damage inflicted by climate change may appear indirectly as a decline in cropping intensity. Planting dates for rainfed crops rely heavily on the timing of water availability and may become ineffective as an adaptation strategy if the wet season is constricted. In an extreme scenario, the shortening of the rainy season may force growers to give up a double cropped system and settle for single cropping, effectively halving productivity. This issue is already present in Iran, where delayed planting of the first crop in a double cropped system reduces exposure to dry spells, but the postponement places the second crop at risk [101].

Planting dates therefore simultaneously create adaptation strategies to climate change, but also pose vulnerabilities to climate change. While shifting planting dates allows farmers to avoid dry spells and other unfavorable weather events, this adaptation is ineffective under extreme forms of climate change in which the number of possible crop rotations is constrained, or in which the possible planting dates are all suboptimal. Crucial to the understanding of how climate change will impact agricultural productivity is therefore quantification of planting dates and their sensitivity to non-stationary climatic variables such as the wet

season onset.

1.2 Existing planting data

1.2.1 Planting dates vary spatiotemporally because they depend on a variety of social, economic, and physical constraints

Planting dates are highly variable in space and time, because deciding when to plant crops involves inherent uncertainty [52, 103]: farmers must navigate the tradeoff between waiting for improved weather conditions [for example, greater water availability or lower frost risk, 82, 94] while providing enough time for phenological development and crop maturation. Double cropping places additional constraints on planting because the first crop must be planted early enough to enable the second crop to mature before the end of the wet season [106]. To navigate this uncertainty and decide to plant, farmers may draw on the weather experienced within a given season to date [94, 109, 134], seasonal weather forecasts [135], memory of planting dates in recent years and other subjective beliefs [75], and/or recommendations from agricultural extension services [23].

Even if the planting date can be optimized with respect to uncertainties in phenology and weather; economic, logistical, and social constraints may cause farmers to accelerate or delay planting, for example based on the availability of agricultural credit [1], crop prices [19], risk aversion [23, 44, 75], access to planting equipment and labor [36], use of irrigation [44], desired cropping intensity [106], or soil type (slow-draining soil hampers planting equipment) [23].

Because planting dates arise as uncertain decisions made by farmers [148], they are dynamic, variable, and difficult to predict [66]. Variations in planting and harvest dates can occur over spatial scales as small as individual fields [16], while interannual variations reflect volatility in climate, crop price, equipment availability, or technological development [66, 80, 81, 91]. This spatiotemporal variability means that planting dates must be understood at field-scales (500 m) and over individual years.

1.2.2 Existing global planting data are spatially aggregated, outdated, or based on outdated assumptions

Few datasets resolve planting and harvest dates at field scale for each year. The datasets that do are primarily ground surveys collected by researchers and agricultural organizations over small groups of agricultural fields, and are not easily scaled or generalized to other areas or years [20, 23, 88, 94, 116]. Globally, planting and harvest dates are often interpolated between sparse measurements and may represent outdated crop varieties and agricultural practices [150]. The Food and Agriculture Organization of the United Nations (FAO)'s AQUASTAT database provides information about planting and harvest dates worldwide at the country level [46]. At national levels, aggregated planting and harvest information can be found in

Europe through the Statistical Office of the European Communities (EUROSTAT) [39, 108]; in Japan, through the Ministry of Agriculture, Forestry and Fisheries (MAFF) [116]; and in China, through local agro-meteorological bureaus [104]. The most spatially resolved and up-to-date planting and harvest datasets available over large regions are from the US, where crop progress reports are available through the US Department of Agriculture (USDA) National Agricultural Statistics Service. However, even these reports are not spatially explicit at the field scale; rather, reporters assess the percent of crop planted or harvested over a county [141]. The Risk Management Agency of the USDA records planting dates in randomly sampled fields, but this information is not made public [140].

Datasets documenting the timing of crop planting and harvesting require improvement in most agricultural areas, but they are least adequate in areas that are especially vulnerable to global change. Large regions in Latin America, Africa, and the Asia-Pacific are vulnerable in terms of both physical response (how crops are biophysically impacted by climate change) and adaptive capacity of farmers and institutions [26, 84]. The effect of damaging extreme weather events and higher average temperatures is compounded where adaptive capacity is low [96]. The IPCC warns that the African continent "is most vulnerable to the impacts of projected changes because poverty limits adaptation capabilities" [26], due to constraints such as limited physical resources [57] and reliance on rainfed agriculture. In both Latin America and Africa, a predicted 10% decrease in maize yield by 2055 will be compounded by large spatial variability in yield: in some areas, extreme declines in yields will disrupt the livelihoods of rural families, while in others, increased yields may enable crop intensification and increased wealth [73]. Spatially-resolved information about crop planting and harvest dates could illuminate the limits of local adaptability, allowing targeted vulnerability assessments.

Extending planting and harvest date information globally, however, remains challenging. There are currently three approaches used to generate global datasets: crop progress reports, planting rules based on crop requirements, and estimation of optimal (yield-maximizing) planting and harvest dates. Examples of crop progress reports include SAGE [115] and MIRCA2000 [108], which are spatially interpolated 5 min-resolution global maps of planting and harvest dates based on national and subnational crop progress reports circa 2000. The second approach is illustrated by Iizumi et al (2019), who produced global, 0.5 degree maps of planting and harvest windows using a rule-based model of crops' physical requirements for water, heat and chill, along with field workability constraints derived from snow and soil moisture. In the third approach, crop models are used to estimate yield-maximizing planting and harvest dates [148]. All these methods can be critiqued. SAGE and MIRCA2000 are outdated and require spatial extrapolation in many locations [66, 108]. Rule-based calendars impose assumptions around the stationarity, homogeneity and accuracy of the rules used [55] – assumptions that produce large errors in the tropics where multiple cropping is common [115, 148] – and the requirement to force the calendar with daily weather data introduces uncertainty in sparsely gauged regions (including the neotropics) [66]. Finally, while optimal planting and harvest dates are informative of potential best practices, they are not a guide to actual farmer behavior [148]. Therefore, existing datasets outside of areas with regularly

updated, spatially-resolved agricultural census efforts provide limited insight into historical planting dates and farmer's responses to historical weather perturbations.

1.2.3 Planting date uncertainty introduces errors in crop yield predictions

A major source of uncertainty in crop yield predictions is a lack of knowledge about how planting date selections are affected by climate change. Currently, many efforts that predict crop yields under climate change introduce untested, but necessary, assumptions about planting dates and their relationship to climate variables (following the rule-based approach to estimating planting and harvest windows) [46, 150]. In the absence of high-quality planting date data, a common assumption is that planting date for tropical rainfed agriculture occurs either at wet season onset, or within a constant number of days after onset [148]. The idea that planting date occurs at wet season onset features heavily in efforts to model crop yield under climate change, which are usually performed under the implicit assumption that the best planting dates for tropical rainfed agriculture follow directly from precipitation patterns. A study of rainfed agriculture in sub-Saharan Africa predicts yields under the assumption that planting date occurs at the "optimal" time, defined as the onset of the rainy season [84]; in Sudan, the optimal planting for sorghum was exclusively based on dekadal (10 day) rainfall values in relation to the total mean annual rainfall [23]; in East Africa, automatic maize planting was triggered in the model when the soil profile was thoroughly wetted [138]; in southeast Asia, the average historical difference between monsoon onset and rice planting date is used as a constant for predicting future planting dates [95]; in precipitation-limited regions worldwide, it was assumed that crops were planted when 10-day precipitation totals reached a certain threshold [18].

The assumption that planting date occurs at wet season onset appears to be a reasonable approximation of real planting behavior: it relies on the idea that growers want to maximize the precipitation experienced by rainfed crops. However, it ignores the economic and technological context in which planting date decisions are made, and assumes that planting date sensitivity to climatic variables is uniform across all fields [115]. For example, Sacks et al (2010) attempted to explain the spatial pattern of SAGE planting data in temperature-limited regions (which are governed by freezing and thawing cycles) using temperature, and in precipitation-limited regions (which are not governed by temperature cycles) using the start of the rainy season [115]. They found the relationship between planting date and climate patterns is not always clear, especially in the low latitudes. In tropical regions with long rainy seasons and multiple cropping, the assumption that planting occurs at wet season onset can result in deviations between estimated and actual planting dates of more than five months [148]. These errors arise because climate is only one of many determinants of planting date. Projecting future planting date behaviors purely based on rainfall climatology may therefore significantly bias predictions of future agricultural behavior and yield.

Complicating the estimation of rule-based planting dates, and the resulting crop yields,

is the uncertainty in how the wet season onset should be defined. While climate projections themselves are a major source of error in crop yield models, even if the precipitation timeseries is perfectly known, there is still flexibility in defining the onset. Examples from literature span climatological definitions such as the anomalous accumulation or Stern definition [23, 86, 132]; metrics based exclusively on volume of rainfall, soil moisture, or relationship between precipitation and evapotranspiration [72, 121, 138]; and fuzzy logic algorithms that simultaneously consider rainfall depth, frequency and dry spell duration [37, 84]. Some definitions were chosen based on agreement with field-scale planting date observations, but many of the planting dates used in defining onset were based on aggregated country-level statistics or on expert knowledge [3, 121].

Ignorance about planting dates and how they respond to climate signals requires modelers to make assumptions about planting date rules, and has led to a diversity of definitions for the wet season onset. These unknowns, in turn, generate significant uncertainties in crop model results that are rarely quantified [151]. An improved understanding of planting dates' response to a variety of possible definitions for wet season onset can help clarify these assumptions. Observed and up-to-date planting information can be related to potential wet season onset definitions with a regression model, where the strength of correlation between variables represents the relevance of each onset definition.

1.2.4 Techniques to estimate spatiotemporally resolved planting dates are needed

The lack of high-quality global planting date data, and of adequate rule-based proxies to estimate planting dates from climate, motivates new techniques to observe crop timing. Ideally, these techniques will produce spatiotemporally resolved planting dates rapidly and without reliance on untested assumptions. Such planting date datasets would help us assess the impact of planting and harvest dates on productivity, quantify farmers' behavioral adaptations to historical climate variability, and improve predictions of agricultural yield in regions where significant climate change impacts are expected.

1.2.5 Satellite-based methods can derive planting and harvest dates, but are typically difficult to upscale

Methods based on remote sensing have the potential to produce the global, spatiotemporally resolved planting and harvest datasets that are needed to understand agricultural behavior under climate change. Remotely sensed imagery can be used to calculate vegetation indices of crop greenness, and timeseries analysis of these vegetation indices has been widely used to identify planting, harvest, and phenological stages including emergence, peak vegetative stage, and the yield-forming grain fill stage [20, 116, 156]. These methods require frequent, high quality measurement of vegetation indices, generating a smooth timeseries with high signal-to-noise ratio during the cropping cycle. A range of more-or-less complex timeseries

analysis algorithms is applied to the observed vegetation indices in order to identify peaks, inflection points and similar markers of crop growth. These timeseries indicators are then calibrated and validated with ground-truth planting and harvest dates. These approaches have been applied to agricultural locations across the US, Europe and Asia: in the midwest US, maize, wheat and soybean planting dates have been estimated using MODIS, achieving RMSE of less than 10 days compared to state-level Crop Progress Reports [112, 139]; in Italy, rice planting and harvest dates have been estimated at 250 m scale with mean average error of 10 days [20]; in Japan, planting and harvest dates at 1000 m scale were produced with RMSE of 12.1 days and 10.6 days, respectively [116]; and phenological information extracted for wheat at 1 km scale in Punjab, India and for wheat and corn at 30m in central China were both highly correlated to validation data collected at an aggregated scale [88, 104].

As promising as these results seem, the approaches described above cannot be easily upscaled globally, nor applied regionally in tropical and developing nations. Previous studies typically addressed areas on the order of 100 km^2, for which computationally intensive nonlinear timeseries analysis techniques are feasible. The scalability of such methods over regional to global scales, however, is questionable, and requires access to advanced geospatial computational infrastructure [112]. Accessible geospatial cloud computing tools such as Google Earth Engine support a limited range of analytical approaches, which would require new approaches to infer crop dates from vegetation indices that are compatible with available tools. Simplifying assumptions used in previous studies, including relatively rigid assumptions about crop phenology and timing, are not defensible on regional scales, particularly where single, double and triple cropping may co-occur.

Additionally, the quality and temporal resolution of optical data are not consistent. For the rainfed agriculture that comprises 80% of the planted global agricultural area [40], frequent cloud cover coincides with periods of rapid crop growth. Aerosol concentrations are often high in the tropics [77], adding uncertainty to measured vegetation indices. Some of these challenges could be addressed with newer sensors such as GOME-2's solar-induced fluorescence (SIF) and QuikSCAT's Ku-band backscatter, which could reduce noise from background reflectance and clouds, respectively [71, 89]. However, their relatively low radiometric accuracy and spatial resolution have prevented widespread use to date [139].

Finally, ground data availability is a major constraint in most parts of the world, where planting and harvest date observations are not made consistently or over broad areas. This creates difficulties for both calibration and validation of methods, forcing spatially resolved satellite-based planting and harvest dates to be validated over aggregated scales and years, using a limited set of surveys performed by individual researchers and their organizational partners, or using privately held datasets [20, 88, 104, 112, 116, 139]. Further, lack of field-scale crop cover maps generate difficulties in identifying where crops of a particular type are planted, making it difficult to produce crop-specific calendars [116, 139].

In light of these challenges, I aim to develop a globally relevant planting date estimation method that is scalable, cheap to employ, suitable for application in regions where agriculture is primarily rainfed, and practical under limited ground data.

1.3 Mato Grosso

In this work, I develop a planting and harvest date estimation method and apply it to rainfed soy in the state of Mato Grosso, Brazil from 2004 to 2014. The estimated planting dates illuminate agricultural response to historical variations in wet season onset, and enable predictions of planting behavior in the future.

1.3.1 Context

Located in central-west Brazil, Mato Grosso has been an agricultural hub for over 40 years (Figure 1.1). Its land area of 900,000 km^2 comprises three major biomes: Pantanal (tropical wetland, 62,000 km^2) in the south, Amazon (humid tropical forests, 481,000 km^2) in the north and Cerrado (tropical savannas, 360,000 km^2) in the center [22]. The state experiences a hot, semi-humid to humid climate, with nearly constant and homogeneous temperatures (22 to 26°C) year round [12], but a strong north-south gradient in rainfall. At the north of the state, annual precipitation exceeds 2000 mm, with a 3 month dry season; in the south, annual rainfall is 1000 mm with a 5 month (May-October) dry season.

Soy development in the Mato Grosso began in the 1970s following the introduction of cultivars adapted to the local climate and photoperiod, and has expanded to account for 27% of Brazil's soybean production over the past several decades [144, 1]. Continued technological advances in crop varieties have facilitated soybean-corn and soybean-cotton double cropping systems, with the proportion of double cropped systems rising from 6% in 2000 to 30% in 2006 [10]. By the mid-2010s, total row crop area covered nearly 100,000 km^2, mostly in the central Cerrado region [154], of which 70% was soybean [22] and 85% was double cropped [27]. A long rainy season and continued technological advances in crop varieties have facilitated rainfed soybean-corn and soybean-cotton double cropping systems [1, 30], making Mato Grosso a center of agricultural production for over 40 years [11]. The vast majority of the Mato Grosso's agriculture is rainfed; only 2.5% of row crop is irrigated [56].

Planting dates in Mato Grosso may respond to a range of climatic and socio-economic constraints, all of which obscure our understanding of soy planting decisions (see Figure 4.1) [1]. In addition to these considerations, planting dates are subject to an additional legal constraint to prevent pathogen outbreaks, in which planting is prohibited during a sanitary break from June 15 to September 15/30 [106].

1.3.2 Significance

Mato Grosso's continued productivity depends in part on whether its intensive cropping practices can be sustained in future climates, and is especially vulnerable to the shorter wet seasons that are expected with climate change [10, 11]. Because the majority of agriculture in this region is rainfed, planting dates and crop yields are expected to be sensitive to shifts in precipitation patterns [106, 48]. With planted area and cropping intensity both nearing capacity [130], a decrease in climatically suitable land or a shortening of the wet season would

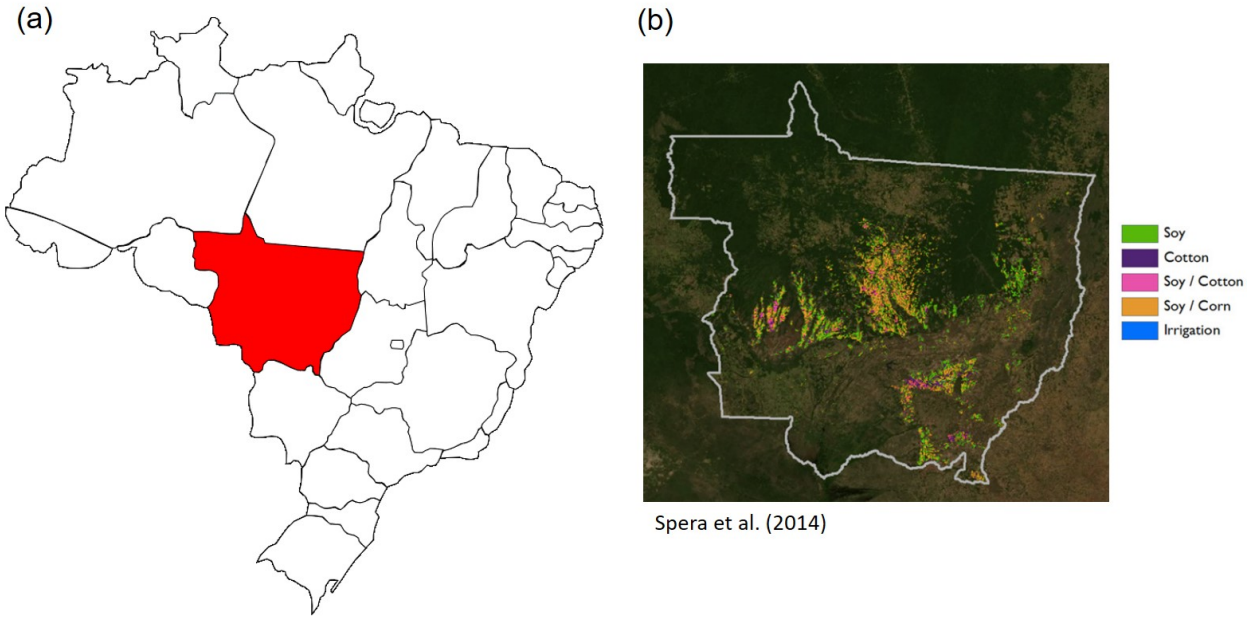


Figure 1.1: (a) Mato Grosso, Brazil and (b) its crop cover [130].

quickly force a change in agricultural practice [1]. Farmers who practice rainfed double cropping may prefer to plant as early as possible in the wet season [106], so a shortened wet season onset may render double cropping impossible, effectively halving the region's agricultural productivity [48].

Complicating the issue, Mato Grosso's agriculture is involved in a feedback loop in which agricultural expansion causes harmful, local-scale climate change. In Mato Grosso and other deforested tropical regions, the land use change associated with agricultural expansion cause local shifts in the water and energy balance. The lower albedo, lower surface roughness, and shallower roots resulting from conversion of natural vegetation into cropland decrease the ability of the land surface to maintain high evapotranspiration rates. In the tropics, lower evapotranspiration, especially during the transition between dry and wet seasons, leads to a reduction in the strength of tropical convection and subsequently a reduction in rainfall [118]. Simultaneously, a reduction in evapotranspiration causes a decline in latent cooling of the land surface, forcing a higher sensible heat flux and consequently a higher land surface temperature. In South America, locally induced climate change may contribute to a one-month reduction in the duration of the wet season by the end of the century, shortening the window over which rainfed cropping (and especially double cropping) is possible [31].

The risks that both global and local climate change will pose to agricultural productivity in Mato Grosso will be better understood with knowledge of how historical planting behavior responded to changes in the wet season onset, and extrapolating to behavior under delayed

onset conditions.

1.3.3 Technical challenges

In addition to its position as a vulnerable but important agricultural center, Mato Grosso presents many of the technical challenges that have prevented the estimation of high-resolution planting dates in much of the world. These challenges include sparse ground-truth data on planting dates, harvest dates and crop cover; a large (and therefore computationally intensive) study area; frequent cloud cover and aerosol interference that degrade satellite imagery; and a wide variety of farming practices that thwarts rules of thumb. Mato Grosso is therefore a challenging but important region for the estimation of planting and harvest dates, making it an ideal study site.

1.4 Research Approach

I introduce a rapid, affordable planting date estimation method that allows unprecedented insight into agricultural practice. In the following chapters, I develop a planting date estimation method that harnesses new remote sensing and cloud computing tools (Chapter 2), then apply it to rainfed soy in Mato Grosso, Brazil (Chapter 3). I then use these planting estimates to select the features of precipitation most relevant for planting decisions and to understand historical response to wet season onset (Chapter 4), and conclude with predictions of planting behavior under future climate scenarios (Chapter 5).

Chapter 2: Developing a scalable planting and harvest date estimation method

In Chapter 2, I develop and test a remote sensing-based method to estimate planting and harvest dates. The method extracts important crop development stages by applying a timeseries analysis algorithm to satellite imagery taken from MODIS, a 500 m satellite dataset which can track the progression of crop growth. The algorithm is designed to be implementable in the powerful cloud computing tool Google Earth Engine (GEE), a platform designed for large-scale satellite image analysis. In the absence of ground truth validation data, I leverage imagery from high resolution (3 - 5 m) satellites deployed by Planet Labs to generate ground-truth proxy planting and harvest dates. My method ensures scalability by avoiding region-specific assumptions, specialized timeseries analyses, and heavy reliance on ground-truth data. I show that the algorithm, simplified to accommodate GEE's limited analytic tools, is competitive with the complex timeseries algorithms found in literature.

Chapter 3: Estimation of planting and harvest dates in Mato Grosso, Brazil

In this chapter, I apply the planting estimation method to rainfed soy agriculture (*Glycine max*) in the state of Mato Grosso, Brazil (MT) from 2004 to 2014. To target my estimates to

the crop of interest, I train a crop cover classifier to extract the locations of double cropped and single cropped soy. The classifier, which is trained on a variety of phenological, spectral, and topographical features, is associated with a land cover classification error that must be combined with the error of the planting and harvest date estimates. These two error sources are combined with bootstrapping. The highly resolved planting and harvest date maps, despite uncertainty (bias and its confidence interval of 6.9 ± 16.5 days for planting and 1.8 ± 18.7 days for harvest), still provide insights into local agricultural practices. The estimates produced in this chapter are used to quantify the relationships between planting date, crop intensity, and wet season onset in Chapter 4.

Chapter 4: Quantifying planting date's sensitivity to wet season onset

Regression models can describe the sensitivity of soybean planting dates to wet season onset in Mato Grosso, Brazil from 2004 to 2014. This chapter, made possible by the 500 m-scale planting date estimates from Chapter 3, develops a statistical model of planting date as a function of wet season onset. Because it is unclear which features of precipitation are most relevant to planting decisions, the model is used to explore planting sensitivity to different definitions of wet season onset. Intuitively, I find that an easily observable definition based on the frequency of rainfall in a four week period is most highly correlated to planting dates; hard-to-observe climatological definitions are less correlated to planting behavior. Model results demonstrate that, within each 25 km region, fields planted in the 5th percentile are more sensitive to onset than fields planted in the 95th percentile. Similarly, double cropped fields are more sensitive to onset than single cropped fields. A trend toward earlier planting dates, independent of wet season onset, is also detected. The spatial variability in sensitivity to wet season onset, and the interannual trend, imply that spatially aggregated planting datasets, or those based on historical surveys, should not be used to predict future planting behavior.

Chapter 5: Predicting planting dates under climate change scenarios

Climate change (specifically, changes in the timing of the wet season) can impact agricultural yields by forcing planting to suboptimal dates or by destroying the possibility of double cropping. In this chapter, I assess these risks by predicting changes in planting dates and the feasibility of double cropping in Mato Grosso under RCP 8.5 conditions. In Mato Grosso, the wet season is generally expected to shorten through equal contributions from delayed onset and early demise, but the effect varies interannually and spatially. To account for interannual variability, I predict planting behavior under short, medium, and long wet season scenarios; for spatial variability, I split Mato Grosso into northeastern and northwestern regions to reflect broad spatial patterns in precipitation. Predictions are made under all combinations to delimit the set of possible outcomes for agriculture in Mato Grosso. Using the statistical model of planting dates developed in Chapter 4, I anticipate that by 2024, 45% of single cropped and 61% of double cropped soy in the vulnerable northeast must delay

planting (presumably to suboptimal days) if a short wet season is experienced. Under the same wet season scenario, an earlier demise will prevent 74% and 21% of double cropping in the northeast and northwest, respectively. These changes in planting behavior are concerning for not only the yield of individual soy crops, but also for the continuation of the lucrative double cropping practices on which this region depends.

Contributions of this work

1. Developed a scalable remote sensing-based method to estimate planting and harvest dates at high spatial resolution (500 m) over large areas (100,000 km^2) with sparse ground truth data.

2. Produced the first spatiotemporally resolved planting and harvest dataset in Mato Grosso, Brazil.

3. Quantified the sensitivity of planting dates to wet season onset based on the cropping intensity, the definition of wet season onset, and planting percentile (early vs late planting).

4. Discovered a trend to earlier planting dates, independent of wet season onset.

5. Predicted the consequences of changing wet season timing on the feasibility of preferred planting dates and double cropping in Mato Grosso.

Chapter 2

Developing a method to estimate field-scale planting and harvest dates

2.1 Introduction

2.1.1 A novel, scalable method to estimate planting and harvest dates

The selection of planting dates is a critical adaptation strategy to climate change, but our knowledge is inadequate for understanding planting dates' potential for supporting agricultural yields. The high cost of ground surveys have prevented planting date estimation at sufficient spatiotemporal resolution to isolate the various climatic, social, and economic controls on planting behavior, forcing efforts at predicting crop yields to rely on aggregated historical observations or on untested assumptions about planting response to climate.

Tracking the full spatiotemporal heterogeneity of planting dates requires a fast, field-scale estimation method. While satellite imagery can provide spatially resolved observations of crop development, existing methods to extract planting dates are complex and therefore limited to relatively small areas of around 100 km^2. In this chapter, I develop a satellite-based strategy to rapidly estimate field-scale planting and harvest dates of soy crop. To ensure a globally applicable method, I address common challenges associated with remote sensing-based estimation: large area of interest; frequent cloud and aerosol cover; and limited ground-truth data. Therefore, the method (i) leverages the power of geospatial cloud computing while working within the constraints imposed by available platforms; (ii) is robust to low-quality optical data; and (iii) augments or replaces sparsely available ground data on crop timing.

The estimation method is based on linearized harmonic timeseries analysis of MODIS imagery and was designed to be implementable on Google Earth Engine, a cloud geospatial platform. It provides a level of scalability that is difficult to achieve with more sophisticated approaches, while still maintaining estimation accuracy. The method addresses low-quality

optical data with smoothing techniques, and uses high-resolution imagery from Planet Labs microsatellites to provide validation data where survey information is unavailable.

With scalability and data scarcity in mind, my work was guided by the following research questions:

1. Can smoothing approaches can be used to compensate for cloud-induced data gaps in satellite imagery?

2. Can the simple, linear timeseries analysis tools available in Google Earth Engine be used to extract phenological parameters from MODIS images without significant loss of estimation accuracy, compared to complex or nonlinear methods?

3. Can ground-truth planting and harvest date data be supplemented or replaced with high resolution satellite imagery?

These research questions were answered using select rainfed soy fields in Mato Grosso.

2.1.2 Background on soy life cycle and planting practices

The timeseries analysis method and planting date estimation uncertainty can be more clearly understood in context of the soy life cycle and typical planting practices. In Brazil, soy is planted as dry seeds (not pre-sprouted seedlings), at a density of 300,000 plants per hectare (corresponding to roughly 20 seeds per meter, with rows 45 cm apart) [90, 42]. Seeds are planted between 2.5 cm and 4 cm below the soil surface, with optimal depth varying by soil type [131]. Once planted, the soy plant's life cycle begins with seed germination and emergence, a process in which a planted seed absorbs moisture from the soil (germination) and sprouts above the soil surface (emergence). The time between the planting date and emergence varies from four days to over two weeks, depending on soil moisture, temperature, seed planted depth, and soil type [9, 119, 152]. After emergence, the plant enters the vegetative growth phase, during which stems and leaves develop (six to eight weeks); followed by the reproductive period, during which flowering (one to two weeks) and soybean pod development (30 to 40 days) occur [152]. The final stages are senescence and maturity, during which photosynthesis slows, and the seeds and leaves turn yellow [152].

2.2 Methods

2.2.1 Datasets

I used a range of remotely sensed imagery and ground-based datasets.

MODIS imagery

The remotely sensed Enhanced Vegetation Index (EVI) tracks soy development over time and space. EVI is calculated from cloud-masked Moderate Resolution Imaging Spectroradiometer (MODIS) 8-day composite products (MYD09A1 and MOD09A1) at 500 m resolution from 2004 to 2014 [98, 99]. The 8-day composites are mosaics of images from 8-day periods chosen for the clearest atmosphere and best viewing angle.

Of the many remotely sensed sources available, MODIS offers both the spectral and spatial resolution required to calculate field-scale vegetation indices, and a history long enough to cover the study period of 2004 - 2014. Its higher temporal resolution (compared to Landsat [114]) makes it more appropriate for timeseries analysis in cloudy regions like Mato Grosso. The MODIS system comprises two satellites, Aqua and Terra, which cover every point on Earth in 1 - 2 days; in contrast, Landsat's revisit time is 16 days. MODIS' coarser spatial resolution (500 m vs 30 m) is not concerning because 500 m is still more resolved than the size of individual soy fields.

Crop cover

The locations of single and double cropped soy are available from a 9,000 point, 2003 - 2017 crop cover training dataset that was formed by intersecting a Landsat-based crop classification produced by Agrosatelite [2] with a roadside survey of Mato Grosso's agricultural areas conducted by Embrapa and the Kansas Biological Survey [76].

Planet Labs imagery

High resolution satellite images from Planet Labs [107] are used to generate field-scale planting date estimates in Mato Grosso. At 3 m and 5 m spatial resolution, Planet Labs imagery offer enough spatial resolution to clearly delineate soy fields, observe greenup soon after planting, and follow the progress of harvest equipment as bright, bare patches gradually replace mature, brown fields at the end of the growing season.

Images of soy fields in Mato Grosso were obtained from the PlanetScope (3 m) and RapidEye (5 m) satellites from August 1, 2016 to July 31, 2017 at three locations: 11 images of 32 km^2 at (-55.389, -11.868); 11 images of 55 km^2 at (-53.454, -15.396); and 16 images of 126 km^2 at (-57.731, -13.285). The three locations were chosen to (1) cover only soy fields and (2) represent all the potential sources of error in the estimation method. To ensure that the image locations contain soy, I chose combinations of fields and years that were reported as soy in the crop cover data. To ensure that the Planet Labs images are representative of all potential error sources, I included both single and double cropped fields and chose images located far apart. A large, persistent storm system would generate the same gap in the vegetation index timeseries over adjacent fields, potentially causing the images to miss a significant source of error that would have appeared if the storms occurred in a different pattern. Additionally, using images scattered over Mato Grosso ensures that the validation

dataset is not biased toward a single producer or toward practices that are local to a specific region.

MATOPIBA planting and harvest survey

Calibration data at high spatial resolution is required to link the phenological information derived from timeseries analysis to the planting and harvest dates.

While ground surveyed planting and harvest dates are not available over the study area of Mato Grosso, they are available for 90 soy properties from 2010 to 2017 in the MATOPIBA region of Northeast Brazil, comprising the states of Maranhão, Tocantins, Piauí and Bahia. Though I do not use this dataset to validate my estimates in Mato Grosso, I examine and compare this survey against Planet Labs imagery to quantify potential reporting errors.

Ground truth data like the MATOPIBA dataset are considered the gold standard for validation, but may be affected by recall bias: the MATOPIBA survey was conducted only during the 8th year of the dataset (2017). Additionally, many responders reported vague date ranges for planting and harvest (such as "early October"). To account for this, I introduced error bars to the original reported date. The widths of these error bars were chosen based on the wording of the reported date range. For example, a report of "2nd week of October" received a range of October 10 +/- 7 days; "early October" received an error bar of October 10 +/- 15 days; and "October" received an error bar of October 15 +/- 20 days.

2.2.2 Definitions

The agricultural year in Mato Grosso begins on August 1, in the middle of the dry season. To relate the calendar year to the agricultural year, I refer to "planting year" and "harvest year". For example, for the agricultural year commencing August 1, 2013, the planting year is 2013, and the harvest year is 2014. A planting date of 60 for harvest year of 2014 refers to September 30, 2013 (60 days after August 1, 2013).

Double cropping is an intensive cropping practice in which two crops are planted in succession during one agricultural year. I refer to these successive crops as the first and second crops [106]. Because soy in Mato Grosso can be single or double cropped, I refer separately to "single cropped (SC) soy" and "double cropped (DC) soy", using "soy" to denote both SC and DC soy.

2.2.3 Method Overview

The planting and harvest estimation method is a five-step process, as illustrated in Figure 2.1. First, I calculate a vegetation index that is quantitatively related to crop growth (Step 1). A timeseries analysis method retrieves phenological parameters from the timeseries of the vegetation index (Steps 2 and 3); finally, these phenological parameters are related to planting and harvest dates through an equation that was calibrated with proxy ground

truth data from Planet Labs (Steps 4 and 5). Each step is described in detail in the following sections, concluding with a sensitivity analysis of the algorithm choices from Steps 2 and 3.

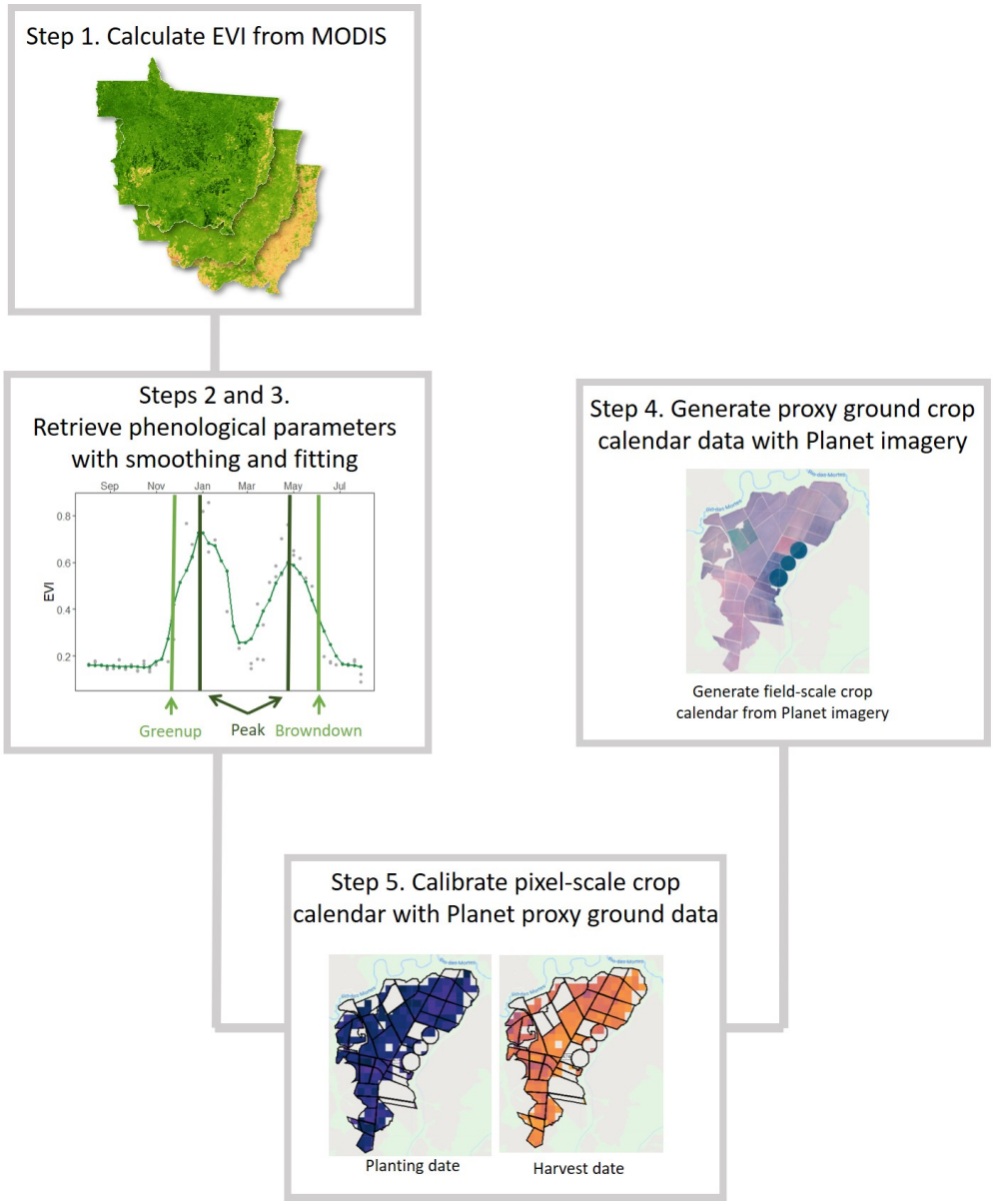


Figure 2.1: Method overview

2.2.4 Step 1: Calculate vegetation index from MODIS imagery

The enhanced vegetation index (EVI) is used to track soybean phenology. EVI is a proxy for greenness that does not saturate at high biomass and contains terms for aerosol correction and canopy background adjustment [64]. EVI is calculated from MODIS 8-day composites using Equation 2.1 [110]:

$$\text{EVI} = \frac{G * (NIR - red)}{NIR + C_1 * red + C_2 * blue + L} \tag{2.1}$$

where NIR, red, and $blue$ refer to the reflectances from the near infra-red, red, and blue bands, respectively. The calibration parameters for MODIS are given by: $G = 2.5$, $C_1 = 6$, $C_2 = 7.5$ and $L = 1$. I removed cloud-contaminated pixels from the EVI images using the "StateQA" band from the MOD09A1 and MYD09A1 data products, retaining only pixels labeled as "clear" (Figure 2.2a) [156]. The removal of the cloudy pixels produces EVI timeseries containing idiosyncratic and irregular gaps, impacting both EVI and its time derivative, $d\text{EVI}/dt$.

2.2.5 Step 2: Reduce noise in EVI profile through smoothing (RQ 1)

The resulting gap-containing timeseries provoked my first research question: can smoothing approaches can be used to compensate for MODIS data that are heavily affected by cloud- and aerosol-induced noise?

To address this question, I compare a range of procedures: (a) use of raw, gap-containing timeseries, (b) smoothing $d\text{EVI}/dt$ only, (c) smoothing EVI and $d\text{EVI}/dt$ once each, (d) smoothing EVI twice, and not smoothing $d\text{EVI}/dt$, and (e) smoothing EVI twice and $d\text{EVI}/dt$ once. In each smoothing step, the number and size of the moving average windows are selected through a set of sensitivity analyses, detailed below. In all cases, I compute $d\text{EVI}/dt$ using a forward difference method.

These procedures are compared in terms of: (i) the percent of soy area for which reasonable planting and harvest date estimates were obtained (as defined in Step 7, described in Chapter 3); (ii) accuracy of planting and harvest date estimates (as obtained by comparison to data described in Step 4); and (iii) robustness of estimates to the timing and number of missing points in the EVI timeseries. To address the third point, I take complete, cloud-filtered EVI timeseries soy pixels and randomly eliminate 1 to 10 EVI data points (out of the 25 points between August 1 of the planting year and April 1 of the harvest year, covering the timing of the first crop). Points are eliminated at random to make 50 timeseries. The 10×50 degraded EVI profiles are used to estimate phenological parameters under each of the five smoothing combinations.

The best of the trialed methods is incorporated in the main algorithm: this involved smoothing each pixel's annual timeseries with two successive 20-day moving average windows (Figure 2.2b), and smoothing $d\text{EVI}/dt$ once with a 40-day moving average window (Figure

2.2i). While these windows are large compared to the roughly 120-day growing cycle of soy, they are necessary to eliminate high noise caused by cloud gaps and aerosols. The width of these windows is justified because I am aiming to describe only broad features of the timeseries, such as the date of the peak and width of the soy curve, which are not degraded under the smoothing windows.

2.2.6 Step 3: Retrieve phenological parameters from EVI timeseries via linear fitting (RQ 2)

The "greenup" and "browndown" dates for crop i, $t^i_{greenup}$ and $t^i_{browndown}$, are defined as the dates of fastest increase and decrease in EVI. Here, i indicates the crop cycle, i.e. the first or second crop. The "peak" day for crop i is similarly defined as the date of maximum EVI, $t^i_{peak,num}$. These dates can be easily identified from the EVI timeseries with a high signal-to-noise ratio [62]. Peak and greenup dates are often used to estimate planting and harvest dates, which cannot be directly observed from EVI [116], and their high signal-to-noise ratio makes them relatively robust to data gaps [62, 104].

I fit the smoothed EVI timeseries to a 1st order harmonic function [28], similar to other functional forms used extensively in phenological studies [e.g. 53, 146]:

$$EVI = \beta_0 + \beta_1 \cdot t + A * cos(2\pi\omega \cdot t - \phi) \tag{2.2}$$

where the phase (ϕ) and amplitude (A) in Equation 2.2 can be calculated by linear regression of EVI on time (t), if frequency (ω) is known [125], and once the mean (β_0) and the linear trend (β_1) have been removed. I refer to this simplified case, in which frequency is known or predefined, as the "linearized 1st order harmonic" because the remaining parameters can be found by linear regression. I estimate ω using the "quarter period" (q), defined as the time difference between $t^i_{peak,num}$ and the preceding $t^i_{greenup}$ or succeeding $t^i_{browndown}$, where ω and q are then related as $q = \pi/2\omega$.

As shown in Table 2.1, peak, greenup and browndown dates are identified by searching for maximum EVI or maximum $\mid d\text{EVI}/dt \mid$ in four time windows corresponding to different rotations and phenological stages (see Figure 2.2, panels c - f, j - l).

The method avoids computing dates in the middle of the wet season due to high levels of cloud cover. Therefore, the quarter period, q, for the first crop is estimated as $t^1_{peak,num} - t^1_{greenup}$, and q for the second crop as $t^2_{browndown} - t^2_{peak,num}$ (Figure 2.2, panels k - m). Because the harmonic function applies to a single crop cycle, the method splits the timeseries into the first crop (from $t^1_{peak,num} - 2q$ to $t^1_{peak,num} + q$), and the second crop (from $t^2_{peak,num} - q$ to $t^2_{peak,num} + 2q$ see Figure 2.2(g-h)) prior to fitting. These windows improve method robustness in two ways. First, they maximize the number of EVI points available for fitting but avoid associating EVI observations with the wrong crop. Second, they avoid the final quarter of soy's crop cycle, leaving the method unaffected by sudden decreases in EVI due to harvesting that would invalidate the symmetric EVI shape on which the harmonic shape depends. These advantages rely on the assumption that soy harvesting (and by extension, planting of the

Description	Symbol	Date Window
Peak date of the first crop	$t^1_{peak,num}$	August 1 of the planting year to March 31 of the harvest year
Peak date of the second crop	$t^2_{peak,num}$	April 1 to July 31 of the harvest year
Greenup for first crop	$t^1_{greenup}$	August 1 of the planting year until $t^1_{peak,num}$
Browndown for the second crop	$t^2_{browndown}$	$t^2_{peak,num}$ until July 31 of the harvest year

Table 2.1: The date windows over which phenological stages were found.

second crop) occur after $t^1_{peak,num} + q$. This is reasonable because soy undergoes at least a month of grain filling after the end of the vegetative stage (roughly corresponding to $t^1_{peak,num}$) [152].

With ω for each crop cycle estimated from q, Equation 2.2 is fit to the EVI timeseries for each crop cycle, and A and ϕ are obtained from the fitted values. The phase is then used to calculate the "analytic" peak date for the first crop as $t^1_{peak,fitted} = \frac{\phi}{2\pi\omega}$. The analytic peak date is an estimate of peak greenness that is more robust to noise and data gaps than its numeric counterpart. The method does not calculate an analytic greenup or browndown date.

The linearized 1st order harmonic function is more prescriptive and simpler in form than other functions used to interpret phenological timeseries (e.g. double logistic functions [156], wavelets [116], cubic splines, asymmetric Gaussian and Savitzky-Golay filters [74]). This raised Research Question 2: does the linearized 1st order harmonic method extract phenological parameters from MODIS images without significant loss of estimation accuracy compared to nonlinear methods?

To answer this question, I compare my simpler fitting function, the linearized 1st order harmonic, against three more complex/nonlinear methods: (i) a Savitsky-Golay filter, (ii) a 3rd order harmonic curve with predefined frequency term (which I will refer to as "linearized 3rd order harmonic"), and (iii) a 1st order harmonic with nonlinear fitting of the frequency term (which I will refer to as "1st order harmonic"). Fitting methods (i) and (ii) offer

a comparison to the methods of the established timeseries analysis tool TIMESAT [74], while (iii) is a nonlinear version of my algorithm that fits the frequency parameter, avoiding numerical estimation of q. For each method I compare (1) the accuracy in predicted peak and greenup dates; (2) ability to handle the complex, rapidly changing EVI profiles associated with more than two vegetation cycles (usually occurring for triple cropped systems or fields with failed crops and significant weed growth prior to double cropping); and (3) scalability and ease of use. The Supporting Information for Chapter 2 provides details on each of these alternative fitting methods.

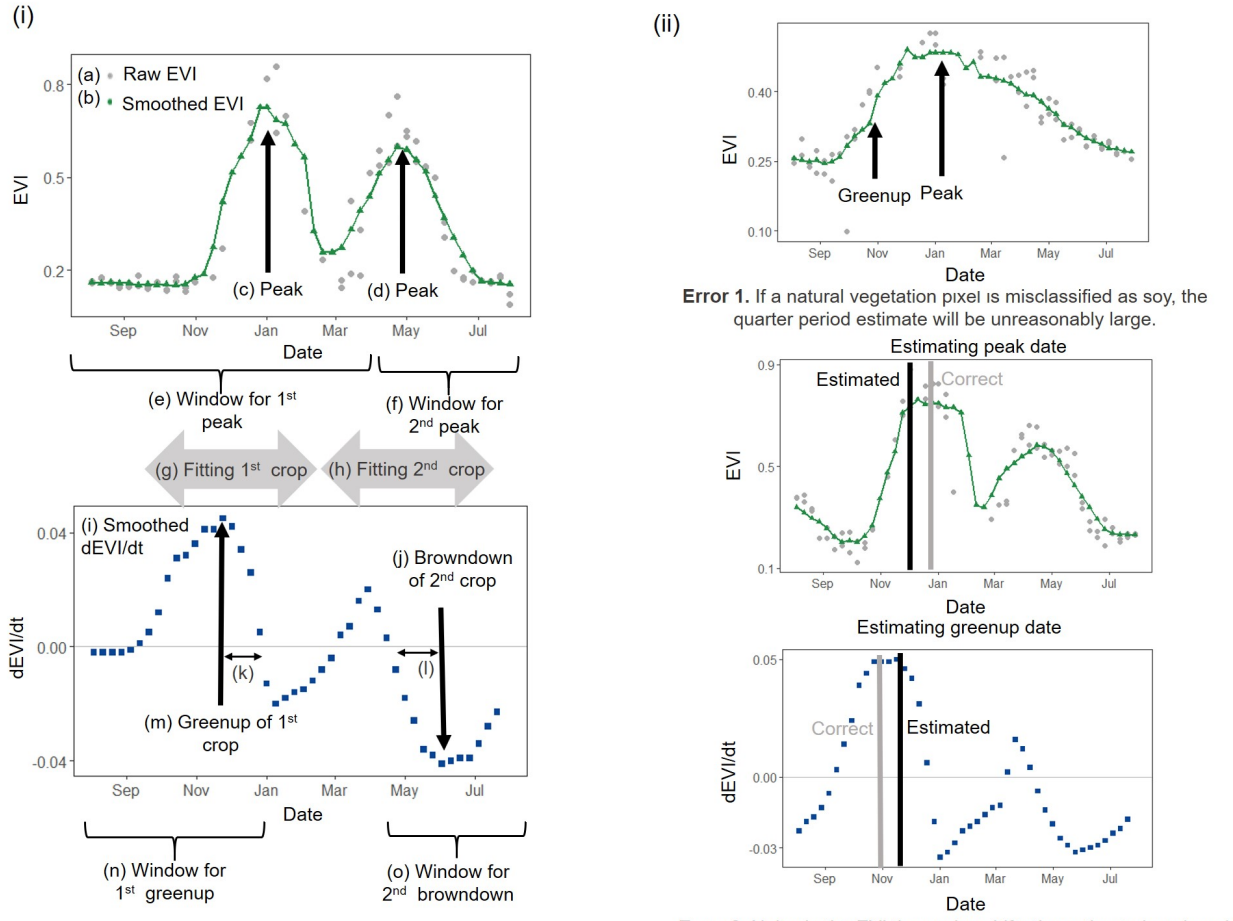
Estimates of the peak day, quarter period, planting and harvest dates from the three complex/nonlinear methods are compared to those from the linearized 1st order harmonic method for 15 randomly selected soy EVI timeseries across Mato Grosso. Additionally, I compare estimation errors of each fitting algorithm relative to the Planet Labs-derived validation dataset for 12 soy points. The tested points include triple cropped systems, allowing me to identify how each of the fitting curves handle these rapidly changing EVI profiles. To evaluate scalability and ease of use, I explore the sensitivity of estimated peak and quarter periods to the parameters of each method, on the basis that high sensitivity would make the method impractical for analysis over large areas.

Though I focus on the planting dates of the first crop, soy, this method can be applied to estimate the planting and harvest dates for both first and second crops. I use phenological estimates made for the second crop in crop cover classification (Step 6, described in Chapter 3) and quality control (Step 7, described in Chapter 3), but lack of crop cover data for the second crop prevents estimation of planting and harvest dates. A sensitivity analysis, detailed in Section 2.2.9, shows that estimates of the first crop's planting and harvest dates are independent of the presence or absence of a second crop, so the same methods are applied to single and double cropped pixels.

2.2.7 Step 4: Generate proxy ground-truth data from Planet Labs imagery (RQ 3)

Existing survey data from MATOPIBA are likely inadequate for ground-truthing soy in Brazil, due to excessive spatial aggregation and concerns about farmers' recall bias. Data in this survey are reported at the property level (on average 52 km^2); however, visual inspection of Planet Labs imagery at this scale shows that planting and harvest dates vary by as much as 3 months within a property. Additionally, quality assurance inspections of the survey data reveals multiple survey responses reporting identical dates across all years, suggesting farmer recall may have impacted the quality of the results. These concerns about the survey data lead to my third research question: Can ground-truth planting and harvest date information be supplemented or replaced with high resolution Planet Labs imagery?

To address this question, I explore the extent to which high-resolution Planet Labs imagery can be visually interpreted to identify planting and harvest activity, and the correspondence of their timing to the existing farm survey data in MATOPIBA (Figure 2.3). I



Error 1. If a natural vegetation pixel is misclassified as soy, the quarter period estimate will be unreasonably large.

Error 2. Noise in the EVI timeseries shifts the estimated peak and greenup. Here, underestimated peak date and overestimated greenup date created an unreasonably low quarter period estimate.

Figure 2.2: (i) Timeseries analysis method, (ii) the method breaks down when a natural vegetation pixel is misclassified as soy or when noise in the EVI timeseries causes the estimated peak dates to shift in unpredictable ways .

create a dataset of planting and harvest dates based on visual inspection of Planet Labs imagery, and refer to it as the proxy ground-truth dataset. The ground-truth dataset includes three distinct soy growing regions in Mato Grosso, from 1 August 2016 - 31 July 2017, each containing 40 - 80 soy fields. Additionally, I obtain images over two MATOPIBA properties to inspect the quality of survey responses. The locations of the Mato Grosso images and MATOPIBA surveyed properties are shown in Figure 2.3.

Farms in the images are manually delineated into fields based on the presence of clearly

visible paths. Visual classification of farming cycles (based on identifying bare soil, green vegetation, brown mature crops, and bare soil) is undertaken for each field. This allows a relatively precise estimation of harvest dates (illustrated in Figure A.1), based on the distinct geometric patterns caused by harvesting equipment.

Planting dates are more uncertain, and are reported by visual assessment over a 2 to 5 week date range preceding the first observed increase in greenness in the soy fields. This relatively large date range represents uncertainty introduced by (1) variable temporal resolution of cloud-free Planet Labs imagery, and (2) variability in the time between planting date and satellite-detected greenness. This second source of uncertainty arises from my attempt to estimate an on-the-ground action (planting date) using an event (detection of biomass above the sensor threshold) that occurs days or weeks later. Thus, a significant portion of the soy life cycle is traversed before greenness is observed by the satellite: (a) the germination and emergence stages (in which the soy sprout emerges from the soil surface), and (b) the early vegetative stage (in which the plant accumulates enough biomass to be detected by the satellite). Both stages can vary in duration, increasing uncertainty in the proxy ground truth data. First, the speed of seed germination and sprout growth depend on the depth of the planted seed, seed size, soil type, soil moisture, the presence of soil crusting, and temperature. These factors can introduce up to 2 weeks of variability [9, 119, 152]. Second, the number of days between *biological emergence* and *satellite-detected* "emergence" depends primarily on the planted density and presence of weeds, both of which influence whether greenness exceeds the sensor's detection limit. In Brazil, most farmers use a consistent plant density of 300,000 plants per hectare [90], but the presence of weeds and other natural vegetation may introduce more uncertainty in the proxy ground truth dataset. To account for this variability across Mato Grosso, my planting dates are reported with an uncertainty of 2 to 5 weeks. The final proxy ground-truth dataset consists of the earliest and latest possible planting and harvest dates.

2.2.8 Step 5: Calibrate an equation that relates phenological parameters to the planting and harvest dates

The peak and greenup dates are used to estimate the planting and harvest dates through equations calibrated to the proxy ground-truth data (described in Step 4):

$$\text{planting date} = t^1_{peak,fitted} - p * q \tag{2.3}$$

$$\text{harvest date} = t^1_{peak,fitted} + h * q \tag{2.4}$$

Where p and h represent the number of quarter periods (q) between the fitted peak date ($t^1_{peak,fitted}$) and the planting and harvest dates, respectively.

Given the uncertainties in the planting and harvest date information, I compute the RMSE as the difference (in days) between the estimate and the nearest date in the reported

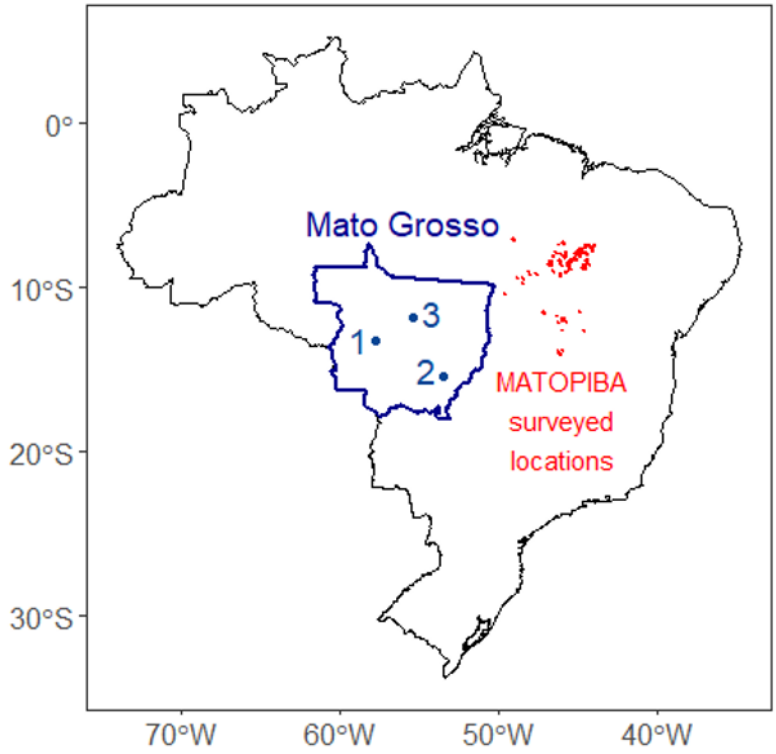


Figure 2.3: MATOPIBA survey points (red) and Planet Labs image locations within Mato Grosso (blue).

date range if the estimated date falls outside of the range, or zero when the estimate falls inside the range.

The best calibration RMSE of 2.5 and 1.6 days for planting and harvest date across all three Planet Labs imagery sites, respectively (Figure 2.3), is achieved by setting the planting date as $t^1_{peak,fitted} - 1.75q$, and harvest as $t^1_{peak,fitted} + 1.1q$. Take-one-out cross-validation, in which one Planet Labs imagery site is removed per calibration, estimates out-of-sample prediction RMSE as 2.92 and 1.61 days for planting and harvest date. I test the sensitivity of the estimated planting and harvest dates to variations in these (p, h) and other timeseries analysis parameters (moving window sizes from Step 2) that were settled by trial and error by comparing each version of the estimated calendar to the proxy ground-truth dataset. Section 2.2.9 details the selection of these parameters.

2.2.9 Sensitivity Analysis

In Steps 2 and 3 (described in Sections 2.2.5 and 2.2.6), I use smoothing windows to compensate for cloud-induced noise in the EVI timeseries, and calibrate parameters that relate two phenological indicators, peak date and quarter period, to planting and harvest dates. The size of the smoothing windows and value of the calibrated parameters are selected to minimize error as compared to the proxy ground-truth dataset obtained from Planet Labs imagery. As shown in Table 2.2, of the six parameters chosen through sensitivity analysis, five have a significant impact on the estimated planting date.

As expected, the "peak cutoff date" (Figure 2.2) has no effect on the estimate for the first crop, indicating that the method is independent of the cropping intensity. Here, the peak cutoff date refers to the date *after* which EVI values are ignored for first crop estimates, and the date *before* which EVI values are ignored for second crop estimates. In other words, EVI values after the peak cutoff date (second crop) do not make any impact on estimates based on EVI values before the peak cutoff date (first crop). EVI observations after the chosen peak cutoff date of April 1 are not considered for estimating the planting/harvest dates of the first crop, *regardless* of the cropping intensity. Because estimates of the first crop date are independent of the presence or absence of a second crop, the same methods can be applied to single and double cropped pixels.

I also test the effect of two successive smoothing windows on EVI. As both the peak and greenup are calculated numerically from the smoothed EVI, smoothing is essential. An overly wide smoothing window may merge two separate peaks into one, falsely increasing the estimated quarter period, while an overly narrow smoothing window would distort the location of the peak and greenup in unpredictable ways. The use of two successive smoothing windows balances the need for increased smoothing without blending a double cropped field into a single peak. Two moving windows for EVI decreases estimated error by 0.3 and 0.1 days for planting and harvest, respectively, compared to using only one window. Similarly, the sensitivity analysis reveals that the error-minimizing smoothing window size for $d\text{EVI}/dt$ is 40 days, with errors increasing rapidly (roughly 0.1 day of error per 1 day change in window size) if the window size shifts in either direction.

2.3 Results

2.3.1 RQ 1: Can smoothing approaches be used to compensate for cloud-induced data gaps in satellite imagery?

Moving average smoothing is often used to reduce the effects of noise, gaps, varying atmospheric composition and viewing geometry, and to therefore improve the robustness of phenological estimates made from satellite observations [62, 128]. Table 2.3 reports on the performance of five smoothing methods applied to the EVI and $d\text{EVI}/dt$ timeseries in terms of the quantity of pixels retained for analysis after smoothing, and the resulting errors in the

Timeseries analysis parameter	Effect on planting date error [days of error per unit change in parameter]	Effect on harvest date error [days of error per unit change in parameters]	Error-minimizing value
Peak cutoff [days]	0	0	April 1
First moving window for EVI smoothing [days]	0.01	0.1	20
Second moving window for EVI smoothing [days]	0.02	0.14	20
Smoothing window for dEVI [days]	0.1	0.1	40
Planting calibration [-]	21	-	1.75
Harvest calibration [-]	-	13	1.1

Table 2.2: The estimates are most sensitive to the planting and harvest calibration parameters, followed by dEVI/dt's smoothing window size.

planting and harvest dates, where the timeseries is fit using the linearized 1st order harmonic function (Section 2.2.6). As shown in Table 2.3, this method requires the use of smoothing to perform well across the three tested criteria, and performs best when the most aggressive smoothing method is used.

EVI smoothing	dEVI/dt smoothing	Fraction soy area remaining after quality masking (Step 7, Chapter 3)	Planting date error [days]	Harvest date error [days]
Double	Yes	0.85	2.5	1.6
No	No	0.47	9.1	9.1
No	Yes	0.63	4.3	3.6
Double	No	0.90	5.6	4.5
Single	Yes	0.79	2.8	1.7

Table 2.3: The effect of smoothing combinations on quality of phenological and planting and harvest date estimates for the linearized 1st order harmonic method. Errors are calculated relative to Planet Labs-derived validation data.

However, the importance of smoothing to obtain robust phenological estimates in the face of poor quality data varies with the function used to describe phenological variation. As shown in Figure 2.4, the experiments suggest that the linearized 1st order harmonic method is only robust if both EVI and dEVI/dt are smoothed, while the complex/nonlinear methods depend much less on smoothing. The results suggest that for complex/nonlinear curve fitting, the tendency for smoothing to lump separate peaks together may outweigh its slight advantage in providing stability. In contrast, for my proposed method, smoothing of both EVI and dEVI/dt is necessary to maximize the area over which planting and harvest date predictions can be made, to obtain accurate results, and to retain robustness as data quality deteriorates.

Although smoothing with moving average windows is only one of several available noise reduction methods, other noise removal techniques failed to improve the timeseries (e.g. the nadir bi-directional reflectance adjusted (NBAR) and atmospherically corrected MODIS data products [20, 156]), could not eliminate residual cloud and aerosol effects over Mato Grosso [61], or resulted in very sparse timeseries (the outcome of filtering images based on view angle and reflectance [50, 116]).

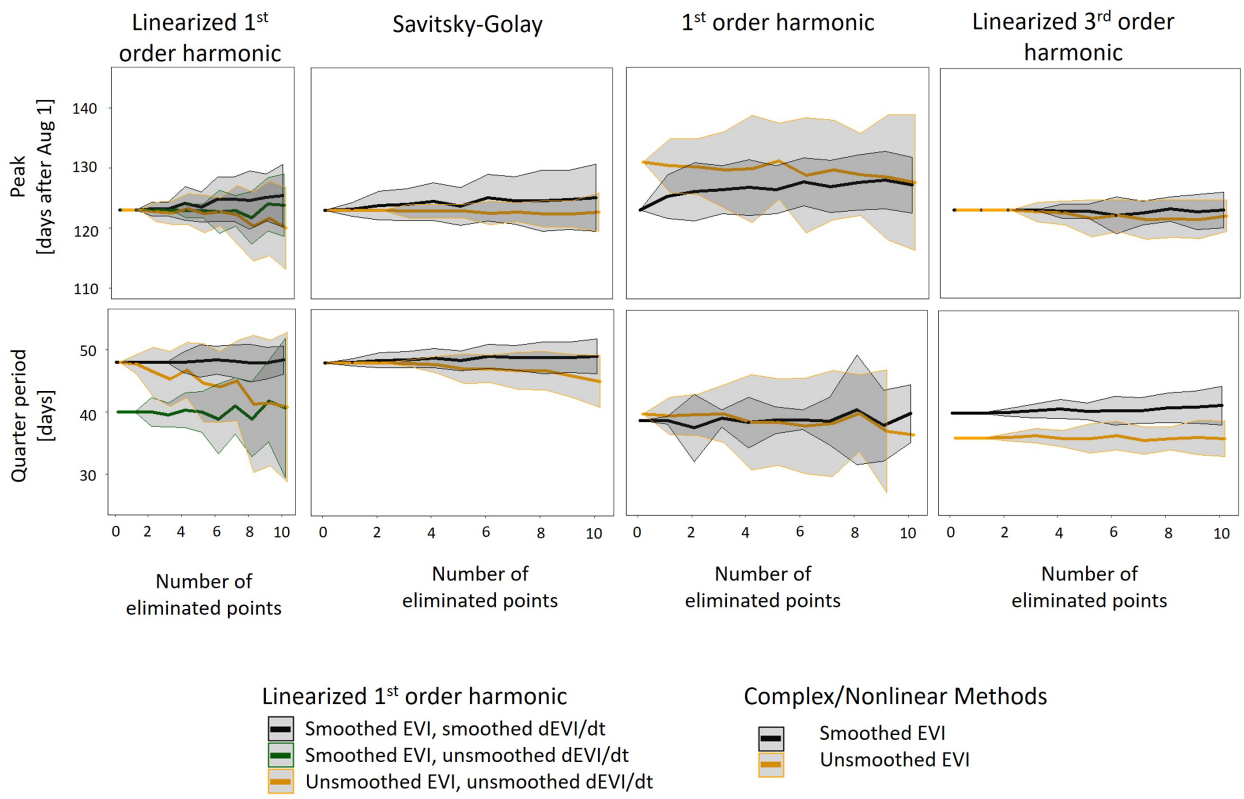


Figure 2.4: Predicted peak date and quarter period vary as input EVI data are progressively degraded for different fitting curves and under different smoothing regimes. The results suggest that the linearized 1st order harmonic function employed in this study requires that input data are smoothed to remain robust to loss of input data quality. For example, experiments using unsmoothed $d\text{EVI}/dt$ (green and yellow lines) generate different quarter periods from those that use smoothed $d\text{EVI}/dt$ (black line), while experiments that smooth both EVI and $d\text{EVI}/dt$ (black line) make more robust estimates of the peak under conditions of poor data quality. In contrast, smoothing is less important for the complex/nonlinear methods. These methods do not use $d\text{EVI}/dt$ to calculate any phenological parameters, so results show the effects of smoothing EVI only. Smoothing does not provide a clear benefit for stability in peak or quarter period under degrading data conditions, but may improve robustness to the location of missing data (reduced confidence interval). Differences in estimated quarter period between smoothed and unsmoothed experiments are approximately half (4 days) of the differences seen between smoothed and unsmoothed data for the linearized 1st order harmonic function.

2.3.2 RQ 2: Will the simple, linear timeseries analysis methods available in Google Earth Engine extract phenological parameters from MODIS images without significant loss of estimation accuracy compared to complex or nonlinear methods?

Table 2.4 shows the estimation difference between the linearized 1st order harmonic function and the three complex/nonlinear fitting functions for a range of phenological predictions. The linearized 1st order harmonic method estimates a peak that is slightly later (2 - 5 days) and a quarter period that is slightly larger (5 - 7 days) than those estimated by the more complex methods, presumably due to the tendency for the timeseries smoothing required by this approach (see Section 2.3.1) to lump closely spaced EVI peaks together. This lumping almost exclusively occurs on rare triple cropped areas. The longer quarter period translates into a planting date 7 - 10 days earlier and harvest date 9 - 10 days later, or a slightly longer estimated crop cycle length.

Amongst the other methods, there is close agreement between the nonlinear 1st order harmonic function and the Savitsky-Golay method, while the linearized 3rd harmonic function is inconsistent with the other methods by a large degree. These inconsistencies are partly attributable to a spurious local minimum that arises when fitting this functional form, and which tends to compress the estimated quarter period. While spurious minima can also occur when fitting the other functional forms, their occurrence is much less frequent (see Figure 2.5).

Additional methodological issues arise when fitting the different functions: for example, estimates of the date of minimum EVI (and thus the estimated quarter period) are sensitive to the time period used to fit each function for the 3rd order harmonic and Savitsky-Golay forms, reducing the scalability of the methods. Similarly, key data needed for fitting the nonlinear 1st order harmonic function are often unavailable in cloudy areas, reducing the area over which estimates could be produced.

Finally, the simple form of the linearized 1st order harmonic offers other advantages beyond robustness and the ability to maximize use of available data [142]. For instance, it is easily interpretable - the phase (ϕ) represents the date of peak EVI and measures soy's seasonality, the frequency (ω) measures its crop cycle length, and amplitude (A) measures the difference in EVI between bare soil and peak greenness, which is useful for land cover classification purposes.

Additionally, I also compare planting and harvest dates estimated from each of the timeseries analysis algorithms to Planet Labs validation data. The estimation errors in Table 2.5 show that while my linearized 1st order harmonic algorithm has a slightly higher planting date error, it performs better for harvest dates. The higher error in planting dates is mostly attributed to one triple cropped field, in which the smoothing algorithm combined two closely occurring peaks into one, overestimating the quarter period.

Estimation differences [days]	Nonlinear 1st order harmonic	Linearized 3rd order harmonic	Savitsky-Golay
Peak date RMS difference	4.8	5.1	4.9
Peak date difference	-2.3	-5.1	-2.5
Quarter period RMS difference	9.5	14.4	6.8
Quarter period difference	-6.7	-14.4	-5.7
Planting date RMS difference	9.4	20.0	7.5
Planting date difference	9.4	20.0	7.5
Harvest date RMS difference	13.8	21.0	12.2
Harvest date difference	-9.7	-21.0	-8.7

Table 2.4: Comparing the estimated planting and harvest dates and phenological parameters from my linearized 1st order harmonic algorithm to those from complex/nonlinear fitting curves. Estimation differences are calculated as complex/nonlinear estimate minus my estimate. Root mean squared (RMS) differences are also reported. These are calculated over 15 soy points in the point land cover dataset.

2.3.3 RQ 3: Can ground-truth planting and harvest date information be supplemented or replaced with high resolution satellite imagery?

MATOPIBA survey data and Planet Labs imagery collections coincide for two properties, enabling a direct comparison between the two datasets. Significant differences in the planting and harvest dates across fields within each property are visually evident in the Planet Labs

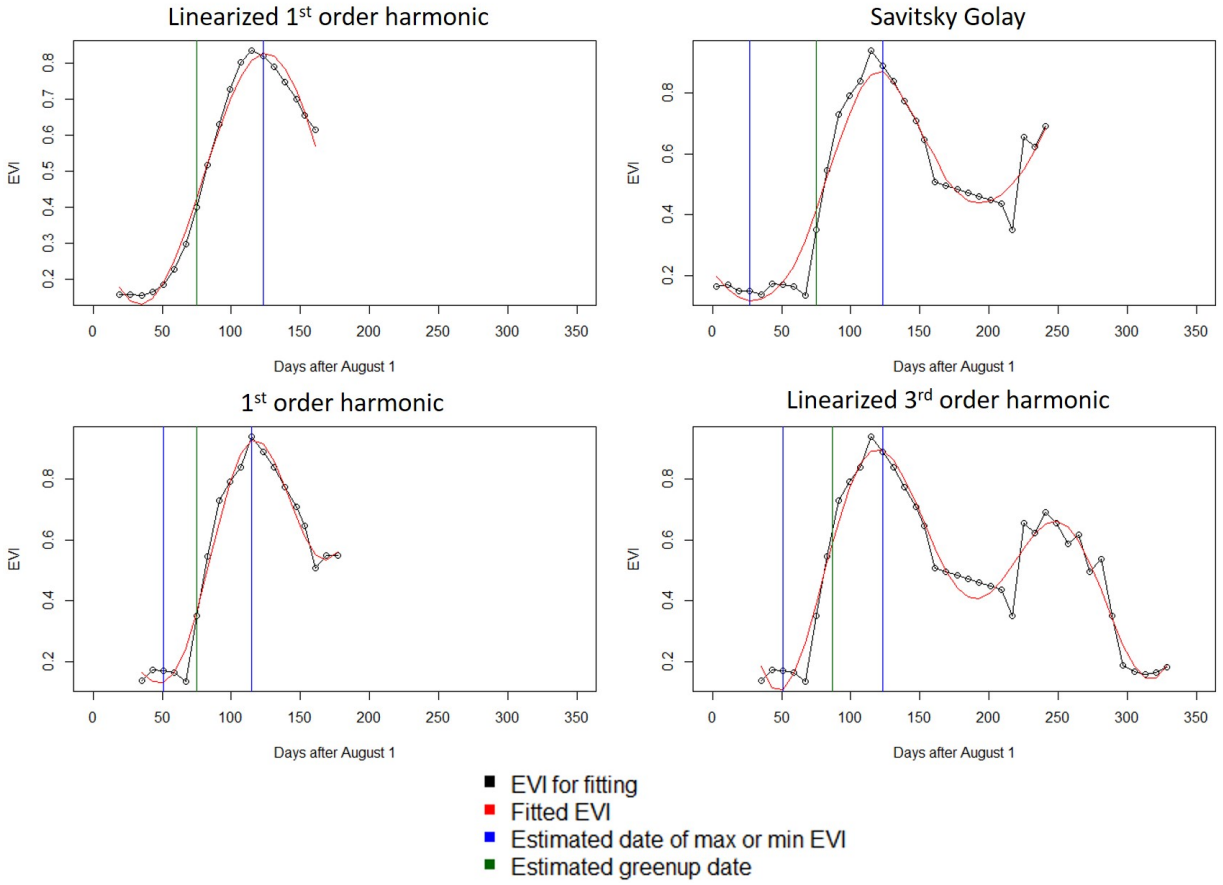


Figure 2.5: Estimated peak and greenup dates from my proposed linearized 1st order harmonic method and three complex/nonlinear methods. Though EVI data exists for the whole year, only the points used for fitting are displayed.

imagery. However, each property only reports one set of planting and harvest dates. For example, one property reported that their planting and harvest activity occurred over a period of 1 and 1.5 months, respectively. For the same property, Planet Labs imagery suggest that planting was spread over a 2 month period, and harvest activity over a 3 month period, an increase of 100%. This larger range indicates that at least some fields are not covered by the values reported in the MATOPIBA survey. For the other overlapping property, the reported planting and harvest dates are not consistent with greenup and browndown observed in Planet Labs imagery in 3 out of 4 of its fields, with differences of up to 1.5 months.

Selecting a preferred data source for calibrating and assessing error in planting and harvest dates entails a tradeoff between use of a proxy dataset based on visual interpretation of imagery, and the potential issues of spatial aggregation and human recall bias involved

Algorithm	Planting date error [days]	Harvest date error [days]
Linearized 1st order harmonic	10.2	0.6
Nonlinear 1st order harmonic	6.4	4.7
3rd order harmonic	1.75	5.3
Savitsky-Golay	1.9	5.7

Table 2.5: Comparing planting and harvest estimation errors from my linearized 1st order harmonic algorithm and from complex/nonlinear fitting curves for 12 soy points over the Planet Labs imagery locations. Errors are calculated as the estimate minus the closest date in the Planet Labs-derived range of plausible planting and harvest dates.

in survey datasets. Further limitations of the proxy dataset include: (i) dependence on the temporal resolution and quality of the satellite images, and (ii) large uncertainty that is propagated into the estimated planting and harvest dates. In this case, examination of year-to-year reported planting and harvest dates in the MATOPIBA dataset shows that some 5% of properties repeatedly report the same dates, indicating recall bias. Examination of two of these properties suggests that spatial aggregation could lead to large errors in reported crop timing. I conclude that proxy ground-truth datasets may be a worthwhile approach because it avoids the greater imperfections in some survey datasets.

2.4 Discussion

My results demonstrate that a simple, scalable estimator for planting and harvest dates over large areas can be constructed by a combination of smoothing EVI data, fitting a linearized 1st order harmonic function to represent crop phenological timeseries, and drawing on satellite-based proxies to supplement or replace sparse ground-truth data of varying quality. With a pixel-level bias of 6.9 and 1.8 days for planting and harvest respectively, my method provides comparable performance to more complex timeseries analysis techniques. It allows users take advantage of existing computational machinery in the Google Earth Engine (GEE) platform, enabling implementation of the algorithm across large scales. Thus, it allows us to benefit from the computational resources of GEE while retaining the ability

to handle complex timeseries.

The main limitations of the proposed methodology are:

(i) The dependence on smoothing could risk combining closely spaced phenological peaks, which might need to be distinguished in triple cropped systems, where first crops failed, or for fields with a significant weed greenness peak that precedes the first crop. This issue is not significant if peaks are sufficiently spaced in time (relative to the width of the smoothing window), do not occur frequently in the study region, and, if it does occur, leads to erroneous predictions that are removed during post-processing quality control of the planting and harvest date maps. Lumping peaks, however, could result in greater error in other regions, and should be critically evaluated based on known cropping practices. At present, these limitations are a necessary consequence of leveraging the power of geospatial cloud computing. As cloud platforms become more powerful and versatile, these constraints will presumably be lifted, diversifying the potential smoothing/fitting approaches.

(ii) The uniformity of the calibration parameters that relate soy's phenological development stages to the timing of planting and harvest is not well understood. In this study, I found that slightly different calibration parameters optimized the relationship between planting and harvesting when the MATOPIBA dataset versus the proxy Planet Labs-based dataset are used. It is not possible to determine if these differences arise from bias in one of the datasets, or from non-uniformity in the calibration parameters. Both are plausible causes: location significantly influences soy's rate of phenological development [6], while, as noted above, issues of recall bias could have impacted the MATOPIBA dataset. While the assumption that relationships between phenological milestones and planting/harvest dates are uniform has been used over regions as large as the US midwest [112] and is supported by soy's consistently symmetric and uni-modal EVI profile[117], the uniformity of calibration parameters will remain an important methodological issue to confront if upscaling these methods.

(iii) The proxy ground-truth data that I used was limited in its extent and resolution by currently available Planet Labs imagery, and limited in precision by the satellites' temporal resolution. Although it still enabled us to avoid errors due to uncertainty in farmer survey products, there remains further scope for rapidly advancing satellite technology to allow planting and harvest dates to be better constrained using high resolution imagery.

(iv) Crops like maize that have nonsymmetric EVI profiles would not be amenable to fitting with the linearized 1st order harmonic form; their major phenological parameters might be more accurately extracted by complex, nonlinear methods, which may not be readily incorporated into tools for geospatial cloud computing platforms at present.

In spite of these limitations, the methods developed here provide a scalable, reliable method to estimate soy planting and harvest dates across Mato Grosso's 0.9 million km^2 area.

2.5 Conclusion

This chapter demonstrates the power of emerging microsatellite and cloud computing resources to allow high-quality planting and harvest dates to be generated across large areas, especially in regions where the effort of collecting ground data would have been prohibitively high. I showed that the simple timeseries analysis methods currently available on cloud computing platforms, if paired with quality control methods like cloud filtering, smoothing, and post-estimation masking, are competitive with the more complex, nonlinear methods available in R and other popular timeseries analysis tools. While the use of a cloud geospatial tool does not improve the accuracy of planting and harvest date estimates at the pixel level compared to these alternative approaches, it provides an opportunity to calculate planting and harvest dates over every planted field, ensuring that the highly spatially variable responses of planting dates to climate change are recorded.

This chapter also demonstrated that imagery derived from Planet Labs can be used to extract proxy ground-truth data that is free of the spatial aggregation and human reporting errors that often degrade the quality of survey-generated ground data. Though the use of satellite imagery to generate ground-truth data introduces additional uncertainty in the data, this uncertainty will decline as the temporal resolution of microsatellites improves. As cloud computing platforms mature and as microsatellites proliferate, they will allow us to observe agriculture at unprecedented detail and scale. The high-resolution planting data made possible by these advances can be used to gain new insights into historical and future planting behavior over data-scarce regions.

Chapter 3

Calculation of planting and harvest dates of soybean in Mato Grosso, Brazil

3.1 Introduction

The lack of high-resolution planting and harvest dates has prevented accurate predictions of crop yields under future climate scenarios and has necessitated untested assumptions about the timing of planting dates. For example, many efforts to predict crop yield assume that planting occurs on the yield-maximizing date, and some efforts to define planting dates assume that planting occurs at wet season onset (a date that I will refer to interchangeably as "onset") [55, 66, 115, 148]. In South America, these assumptions could produce errors in planting date of up to five months [148]. And while planting date information is recorded in many national and sub-national reports, they may represent outdated agricultural practices and are typically highly spatially aggregated [150].

The scalable method for planting and harvest date estimation outlined in Chapter 2 allows me to create field-scale (500 m) planting datasets, updated each year. In this chapter, I create a field-scale soybean planting date dataset in Mato Grosso, Brazil from 2004 - 2014, based on timeseries analysis of MODIS imagery. I produce spatiotemporally resolved, observation-based maps of planting and harvest dates, with minimal reliance on untested assumptions and outdated or aggregated crop progress reports.

The production of soybean planting and harvest date maps requires knowing the location and cropping intensity of soy agriculture. The process of classifying pixels as single or double cropped soy introduces additional errors in the crop calendar at the regional scale. In addition to errors in planting and harvest date at a pixel known to be soy, uncertainties in the crop cover classification will introduce errors in regionally aggregated planting and harvest dates. For example, the planting date of double cropped soy within a 25 km region will have uncertainties associated with both the planting date error at individual pixels and the

misclassification of double cropped soy in the region. A third source of uncertainty originates from the proxy ground truth validation dataset derived from Planet Labs imagery. Due to cloud cover and two-week temporal resolution, the "true" values for each field are date ranges, instead of single dates, preventing the calculation of a single value for the estimation error. I combine these error sources to quantify the estimation uncertainty at pixel and aggregated scales.

The planting and harvest date maps generated in this chapter can be used as input data in crop models, or to extract insights about historical and future planting behavior. I focus on the second application. In this chapter, I examine spatial and interannual trends in planting dates, as well as their relationship to cropping intensity, the wet season onset, and the sanitary break; in Chapter 4, these maps form the basis for a statistical model of planting behavior as a function of climate and cropping intensity; in Chapter 5, this model is used to predict planting behavior under climate change.

3.2 Methods

3.2.1 Data

I use a range of existing agricultural maps, ground-based datasets and climate products.

Mapping products

The crop cover dataset of soy across Mato Grosso from 2004 to 2014 is classified using the point dataset of single and double cropped soy identical to the one used in Chapter 2, MODIS imagery, and NASA's Shuttle Radar Topography Mission (SRTM) containing 30 m topographic information [41].

Datasets identifying irrigated fields and agricultural regions are used to target rainfed agriculture. A center pivot irrigation map for 2014 from Brazil's National Water Agency (ANA) is used to mask out irrigated pixels [21]. Though I only have data for one year, center pivot is a permanent structure that, once installed, will not be dismantled for several years. This means that a field that was not irrigated in 2014 was likely not irrigated in past years. In the absence of center pivot data for all years in the study period, I create a conservative mask for center pivot by eliminating all pixels in previous years (2004 to 2013) that were irrigated in 2014.

Agricultural regions are identified with Mapbiomas v3, a 30 m resolution land cover map of Brazil from 1985 to 2017 produced by a group of universities, NGOs and technology companies [92]. It is the most reliable and comprehensive land cover map available for Brazil. Although it does not differentiate among individual crops, it does contain a class representing general row crop agriculture. Because the crop cover dataset does not contain training information for non-agricultural classes, I first classify all pixels in Brazil as one of three agricultural classes (single cropped soy (SC); double cropped soy (DC); and non-soy agriculture) and use Mapbiomas' agriculture class to mask out all non-agricultural pixels.

Climate data

Maps of wet season onset from 2004 to 2014 were calculated by Abrahao et al (2018), in which the anomalous accumulation (AA) method was applied to a gridded (0.25×0.25 deg) daily rainfall product produced through interpolation of 3625 rain gauges and 735 weather stations across Brazil [153].

In the anomalous accumulation (AA) method, the wet season onset date is defined based on the value of the anomalous accumulation [mm/day]:

$$AA(t) = \sum_{n=1}^{t} (R(n) - R_{\text{ref}}) \tag{3.1}$$

where $R(n)$ is the rainfall on day n and R_{ref} is a reference rainfall value, defined here as the agronomically significant threshold of 2.5 mm/day [12]. Here, $t = 1$ refers to July 1. The onset date is defined as the day at which the value of $AA(t)$ reaches its minimum [86].

3.2.2 Method overview

The steps below are numbered in continuation from Chapter 2; Figure 3.1 summarizes all steps from Chapters 2 and 3. Planting and harvest dates for soy are estimated within Mato Grosso (Step 6), then filtered with a quality control process (Step 7). Finally, uncertainty in the resulting planting and harvest dates is quantified (Step 8).

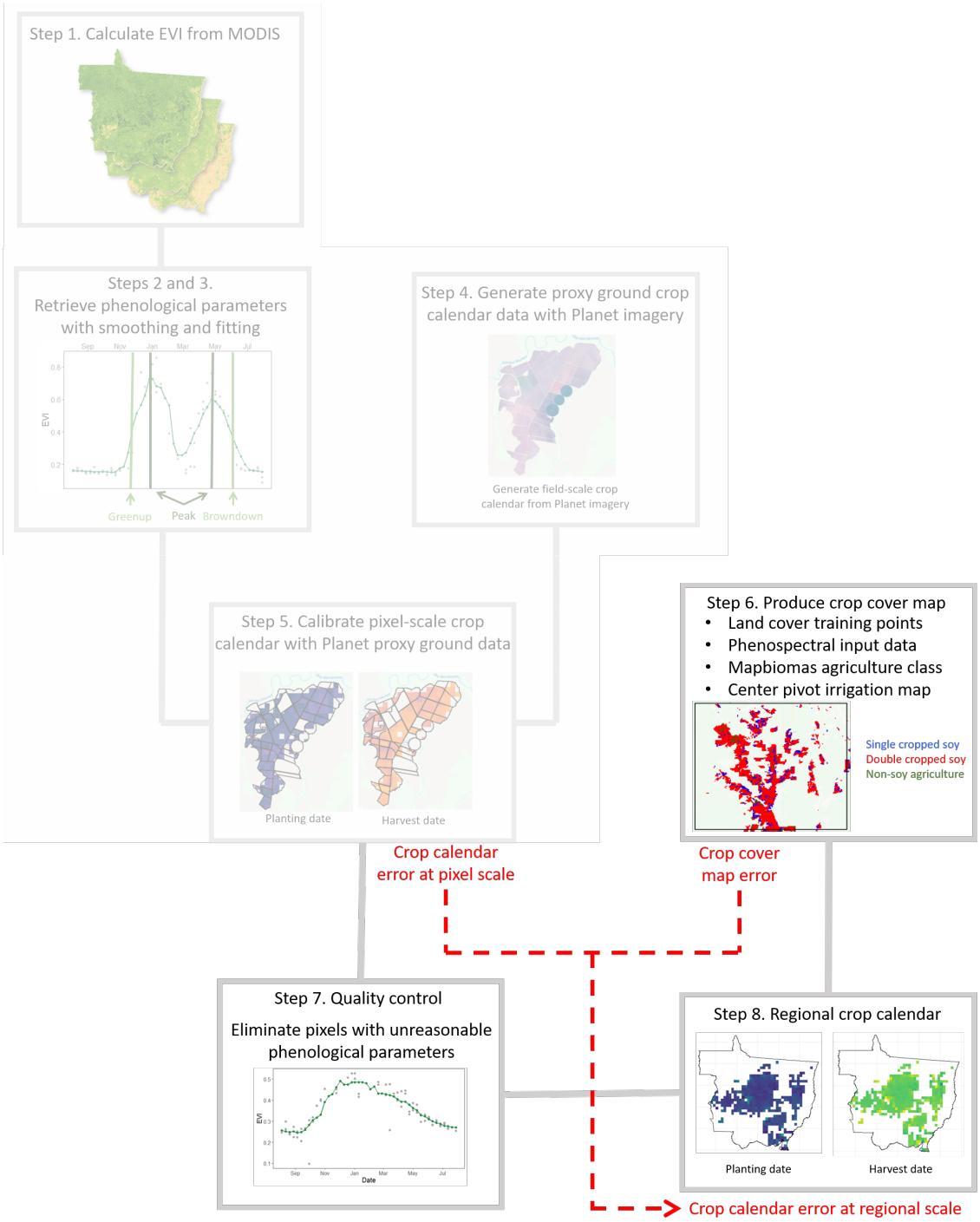


Figure 3.1: Method overview, continued from Chapter 2.

Step 6: Classify soy pixels across Mato Grosso to mask the estimated planting and harvest dates

Because a crop cover map of soy and its cropping intensities is not available over Brazil, I create a 500 m resolution map of soy trained on MODIS-derived phenospectral information over the crop cover dataset. I first use the crop cover dataset to create a classifier that categorizes all pixels in Mato Grosso as either soy agriculture or other agriculture, then eliminate non-agricultural pixels with the Mapbiomas dataset. Mapbiomas offers an agriculture class that lumps all crops, so while it cannot be used to target specific crops like soy, it can be used to highlight only the agricultural pixels after all pixels have been classified.

To map soy agriculture and its intensity, I adapted an existing crop classification technique tested for soy and corn in Parana State, Brazil [157]. In the Parana study, classifiers were trained on a set of phenological and spectral input data derived from MODIS. Similarly, in this study, I train a Cartesian classifier in GEE using topographic, phenological and spectral information derived from MODIS. All phenological and spectral input data (cloud-filtered and smoothed) are derived from the EVI timeseries calculated in Step 2 (described in Chapter 2), while topographic data (elevation, slope, aspect and hillshade, relevant because soy requires intensive equipment that functions best on flat, low-elevation land) are derived from NASA SRTM. The phenological and spectral information is shown in Figure 3.2. Because cloud filtering of the MODIS data in Step 2 (Chapter 2) introduces spatial gaps in the EVI images that, if not filled, would be propagated as gaps in the classifier's training data, the EVI images are gapfilled over space with a mean square kernel. This ensures that each point in the crop cover dataset contains a full set of input data.

To maximize the amount of relevant information included in the input data while minimizing the risk of overfitting, I explore (i) the spectral bands that most clearly separate single cropped and double cropped soy from other agriculture, and (ii) three sets of input data: phenological only, spectral only, and pheno-spectral. Studies of the spectral signature of each crop cover class reveal that the NIR and red reflectance (components of the vegetation index, EVI) during the wet season are best able to separate crop classes. Benchmarking the spectral information on phenological stages rather than on calendar dates allows the classifier to align input data across years and locations [157]. This alignment produces a classifier that is robust to interannual and inter-regional variations in the planting and harvest dates, sensitive to physiological and seasonality differences among crop types and crop intensity levels, and relevant in extrapolated contexts. The Supporting Information for Chapter 3 provides detail on the selection process for input data. To finalize the crop cover map, I combine the location and cropping intensity of soy, as classified using the method described above, with a map of agricultural areas (Mapbiomas).

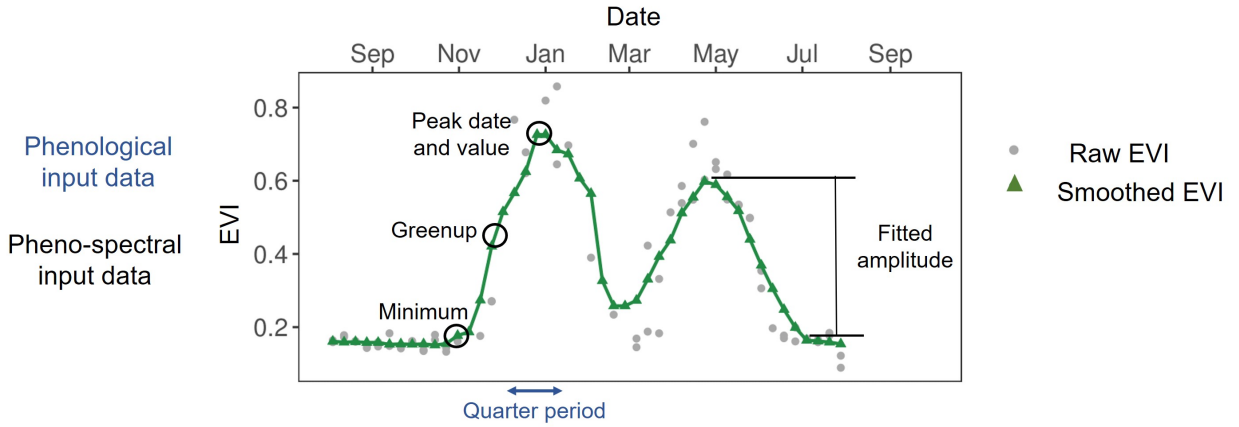


Figure 3.2: Input data for the crop classification.

Step 7: Quality control on planting and harvest dates

The soy cover map from Step 6 is used to eliminate non-soy pixels from the planting and harvest date maps produced in Steps 1 to 5 (Chapter 2). I also use a map of center pivot irrigation locations in 2014 to mask out potentially irrigated pixels from 2004 to 2014, under the assumption that non-irrigated locations in 2014 were also non-irrigated in previous years. Irrigated fields do not conform to the assumptions made in the timeseries analysis, and will correspond to fundamentally different planting dates and adaptation options than rainfed pixels.

In a final quality control step, I test the predicted planting and harvest dates against a series of rules intended to screen out implausible findings:

1. The peaks of the first and second crop must be more than 20 days apart.

2. Planting must occur between August 1 of the planting year and May 31 of the harvest year. That is, planting cannot occur within the dry season [1].

3. The crop cycle of soy (planting to harvest) must be between 60 and 150 days. This is based on observations that the average soy crop cycle is 120 days long, with short cycle varieties averaging a 90 day cycle [1] (See Figure 2.2 for an example of this error.)

4. A soy pixel must have a raw peak EVI of at least 0.8 during the growing season and a fitted EVI amplitude of at least 0.15. This filters out natural vegetation pixels that are misclassified as soy using EVI properties of soy that are established in literature. Soy pixels display a much higher seasonal change in EVI than forest and have a larger peak EVI value than savannah, so pixels can be filtered by peak EVI and the amplitude of the fitted EVI curve [50].

Step 8: Quantify errors for the planting and harvest dates

The planting and harvest dates generated from Steps 1 to 7 contain errors from two sources: (1) pixel-level errors associated with timeseries analysis and the calibrated equation, as well as uncertainties in the validation data itself; and (2) regional-level errors associated with the crop cover map. Pixel level errors can be quantified with the proxy ground-truth data. However, when pixel estimates are aggregated to regional scales, errors associated with the crop cover map must be included. While pixel level planting and harvest dates are valuable for understanding field-scale behavior, spatially aggregated planting and harvest dates highlight broader trends across time and space. Therefore it is worthwhile to aggregate estimates to regional scales, even though additional errors due to misclassified soy cover will appear.

First, I describe pixel-level errors using a probability distribution. Since the ground-truthing data are reported as a range of plausible planting/harvest dates, the calibration data contain uncertainty. I use the law of total probability, represented in Equation 3.2, to aggregate the error and its uncertainty at individual pixel scales into an error distribution that describes all pixels. Uncertainty in pixel-level error is modeled as a uniform distribution spanning from the difference between the estimate and the lower bound, and the difference between the estimate and the upper bound of the reported range. Each pixel in the dataset contains its own unique pixel-level errors, which are then aggregated into a single pixel-level error distribution, p(x), describing all pixels, using Equation 3.2. The distributions of the error bounds, a and b, are found by fitting a normal distribution to a and b values found at individual pixels. Equation 3.2 is solved numerically in R, with planting and harvest error distributions treated separately.

$$p(x) = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \frac{H(x-a) - H(x-y-a)}{y} * p(a) * p(y)\, da\, dy \tag{3.2}$$

Where:

a = estimate - upper bound of reported range

b = estimate - lower bound of reported range

y = b - a

H = Heaviside function

p(a) = probability density distribution of a

p(y) = probability density distribution of b

After pixel-level errors are defined, I introduce error due to land cover misclassification. Because the planting and harvest date estimation method is independent of land cover classification, the misclassification in land cover contributes to planting and harvest date as a

mislabeling of pixels when aggregating. I simulate the error introduced by misclassification through bootstrapping: I generate a "true" land cover map for a 25 km region containing the average proportion of single cropped soy, double cropped soy and non-soy agriculture found in Mato Grosso. From this, I generate many "erroneous" land cover maps in agreement with the confusion matrix. For each erroneous land cover map, I calculate the median planting and harvest dates for single cropped soy, double cropped soy, and all soy pixels. The difference between this median and the corresponding median for the "true" land cover map represents the error introduced to the aggregated planting and harvest dates by erroneous land cover classification. The total error at the 25 km aggregated scale is a simple sum of the pixel level estimate error and the error introduced by the land cover classification.

Application and evaluation

Steps 6 - 8 generate maps detailing planting dates, harvest dates, and cropping intensities (and their respective errors) for soy agriculture in Mato Grosso. In the following section, aggregated planting and harvest dates for single and double cropped soy are separately visualized as maps and histograms to detect the relationships between crop timing and cropping intensity, year, and location. Results reported for the year 2014 refer to planting year of 2013 and harvest year of 2014. I also calculate the delay between estimated planting dates and wet season onset, and observe its change over time.

Finally, to benchmark the performance of this method in Mato Grosso, I compare my aggregated planting and harvest dates to those reported in (1) the SAGE dataset, which compiles national and subnational planting and harvest date statistics circa 2000, and (2) weekly crop progress reports from Brazil's Instituto Mato-Grossense de Economia Agropecuaria (IMEA) agency.

3.3 Results

3.3.1 Crop cover classification

Figure 3.3 shows the soy cover map for a sub-region of Mato Grosso and the change in total area of each land cover class from 2004 to 2014. The vast majority of the agricultural area in Mato Grosso throughout this decade is double cropped soy. The timeline of crop area (Figure 3.3b) shows a steep increase in soy cropped area over the analysis period, almost all of which is double cropped soy.

The zoomed-in map of crop cover (Figure 3.3a) reveals that most of the soy is arranged in a dendritic pattern, clustered around major roads. The most established (oldest) soy areas are nearest to the roads, and newer soy areas mainly extend from existing soy areas, suggesting a close relationship between transportation networks and soy agriculture [105].

The soy cover maps are consistent with reported patterns of land use in Mato Grosso [27]. The soy map has an overall accuracy of 82.5%; the confusion matrix, producer's accuracy

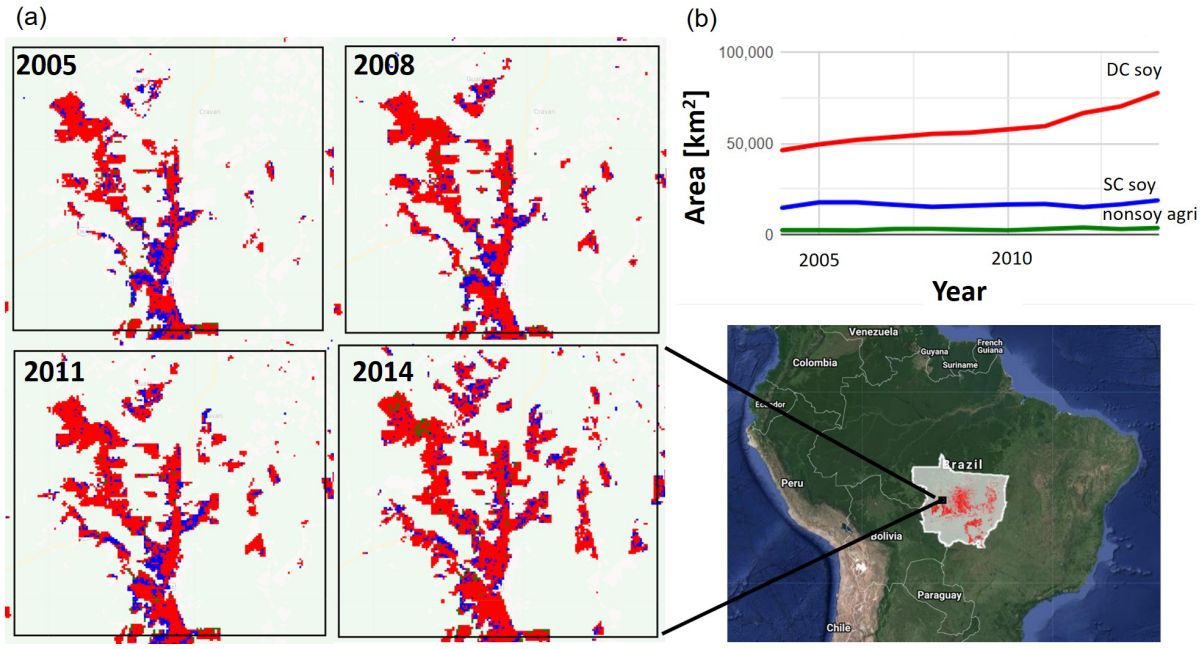


Figure 3.3: (a) Land cover over the selected years in the study period, revealing the expansion of double cropped soy. (b) The dominance of double cropped soy can be seen in both the timeseries and crop cover map over Mato Grosso. Here, the year corresponding to the land cover map is the harvest year: a double cropped pixel in the 2014 land cover map was double cropped from August 1, 2013 to July 31, 2014.

and consumer's accuracy are shown in Table 3.1. This is comparable to the 87% accuracy reported for a similar study in Parana [157].

3.3.2 Spatial pattern and variation in planting and harvest dates

Figure 3.4 shows quality-controlled, pixel-scale estimates of the planting and harvest dates for two out of three Planet Labs imagery locations in Mato Grosso. At pixel scale, my data reveal large differences in the timing of soy agriculture across adjacent fields, showing that neighboring fields of 1 - 2 km in size can have planting dates that differ by more than one month, and harvest dates that differ by more than two months.

Figure 3.5 shows median planting and harvest dates over 25 km cells for single and double cropped soy for selected years between 2004 and 2014. The maps display interannual and regional variation in the planting and harvest dates.

The long-term spatial pattern of planting and onset dates, averaged over 2004 to 2014, is shown in Figure 3.6. For both cropping intensities, central Mato Grosso is planted earlier,

	Single cropped soy	Double cropped soy	Non soy agriculture	Producer's accuracy
Single cropped soy	440	548	3	44%
Double cropped soy	301	4367	53	92%
Non soy agriculture	9	128	107	43%
Consumer's accuracy	59%	86%	66%	82.5%

Table 3.1: Soy map accuracy and confusion matrix.

while areas closer to the border are planted later. Onset likely plays a large role in determining these the regional differences in planting dates, as the spatial patterns in planting roughly follow patterns of onset. However, onset cannot explain all of the regional variability: time-averaged onset is more spatially homogeneous compared to time-averaged planting dates. The difference in spatial pattern of planting dates (and of planting dates after the influence of onset is removed, Figure 3.7) between the two cropping intensities also hints at non-climatic controls on planting decisions. Single-cropped soy appears to have a stronger spatial pattern in planting dates than double cropped soy, though they experience the same onset. Additionally, the similarity in planting dates between the two cropping intensities changes in space: in the central part of Mato Grosso, single and double cropped soy are planted at very similar times, but the difference widens on the edges of the state.

3.3.3 Sanitary break is not a hard limit for early planting, but wet season onset may be influential

Figure 3.8a shows histograms of planting and harvest dates for single and double cropped soy from 2004 to 2014, overlaid on Mato Grosso's median wet season onset for the corresponding year in blue; Figure 3.8b shows the delay between planting date and onset for single and

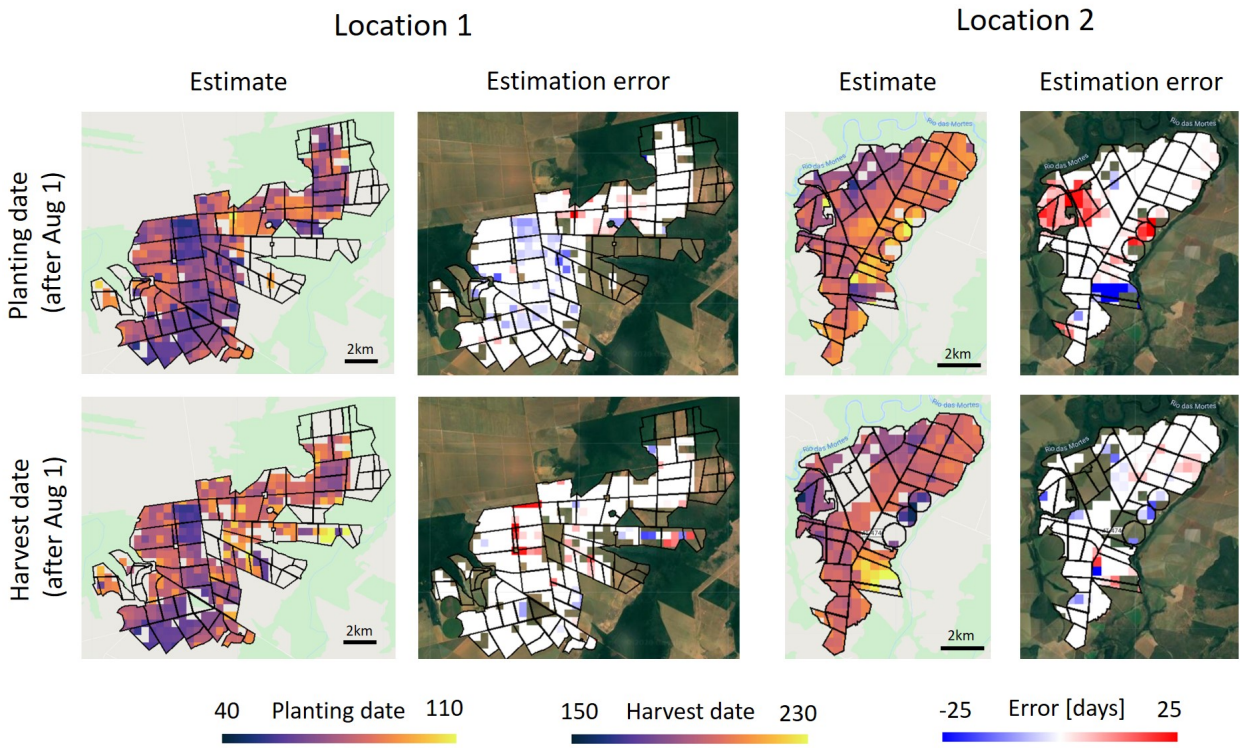


Figure 3.4: Estimated pixel-scale planting and harvest dates and their estimation errors for Planet Labs data locations (as labeled in Figure 2.3). The pixels in these maps were quality masked as described in Step 7. Some fields did not contain reported data because there were not enough Planet Labs images to construct a range of possible planting/harvest dates less than 1.5 months long. The error shown in this figure is defined as the distance between estimate and the nearest date in the reported range, and is calculated for each individual pixel. This error is in contrast to the "averaged" pixel-scale errors generated in Step 8, which is a single error distribution applied across all pixels.

double cropped soy from 2004 to 2014. These data reveal that the delay consistently (with the exception of 2010) decreased from 2004 to 2014 for both single and double cropped soy. During 2010, there was an anomalously early onset of the wet season, but planting dates did not shift to a correspondingly early time. By 2014, the delay for both double and single cropped soy were both at an all-time low of 19 and 30 days, respectively.

For most soy in Mato Grosso, planting dates occurred much later than both the wet season onset and the sanitary break, indicating that neither of these constraints becomes a hard limit for early planters. These results indicate that the mean planting date for double cropped fields is 89 days after August 1 (late October), while the mean planting date for

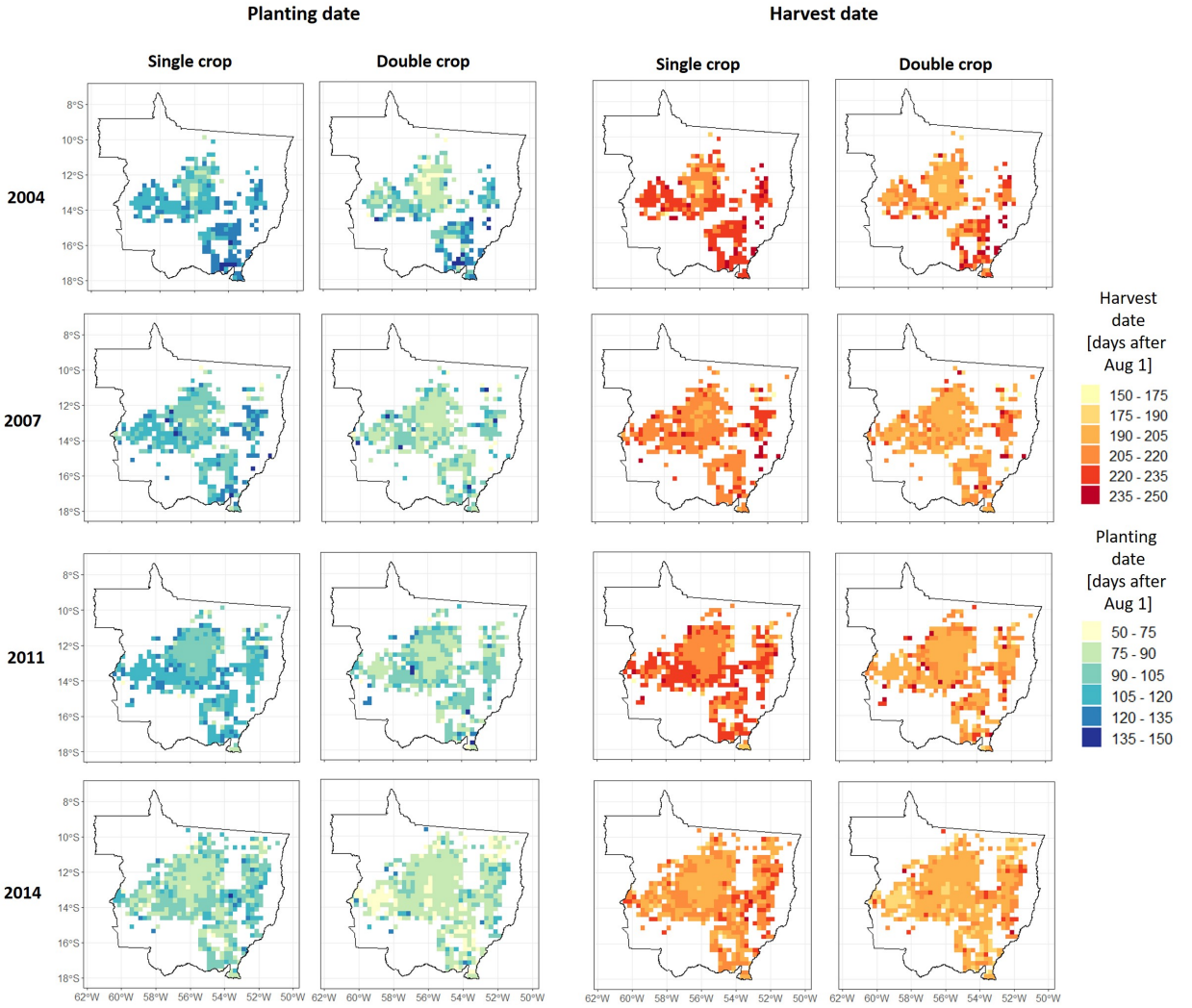


Figure 3.5: Estimated median planting and harvest dates, in days after August 1 of the planting year, over 25 km cells.

single cropped fields is 98 days after August 1 (early November) (Figure 3.8a). Thus, the average planting date for double cropped soy is over a month after the end of the sanitary break. Similarly, Figure 3.8a reveals a delay between planting and wet season onset of at least 19 days, but up to three months for late-planted single cropped soy and up to two months for late-planted double cropped soy. Such a delay in planting may raise the suspicion that the onset date itself is estimated too early. However, as the onset is defined based on the agronomic requirements of soy seedlings in Mato Grosso, it is deemed the most appropriate onset definition for this application [1, 12].

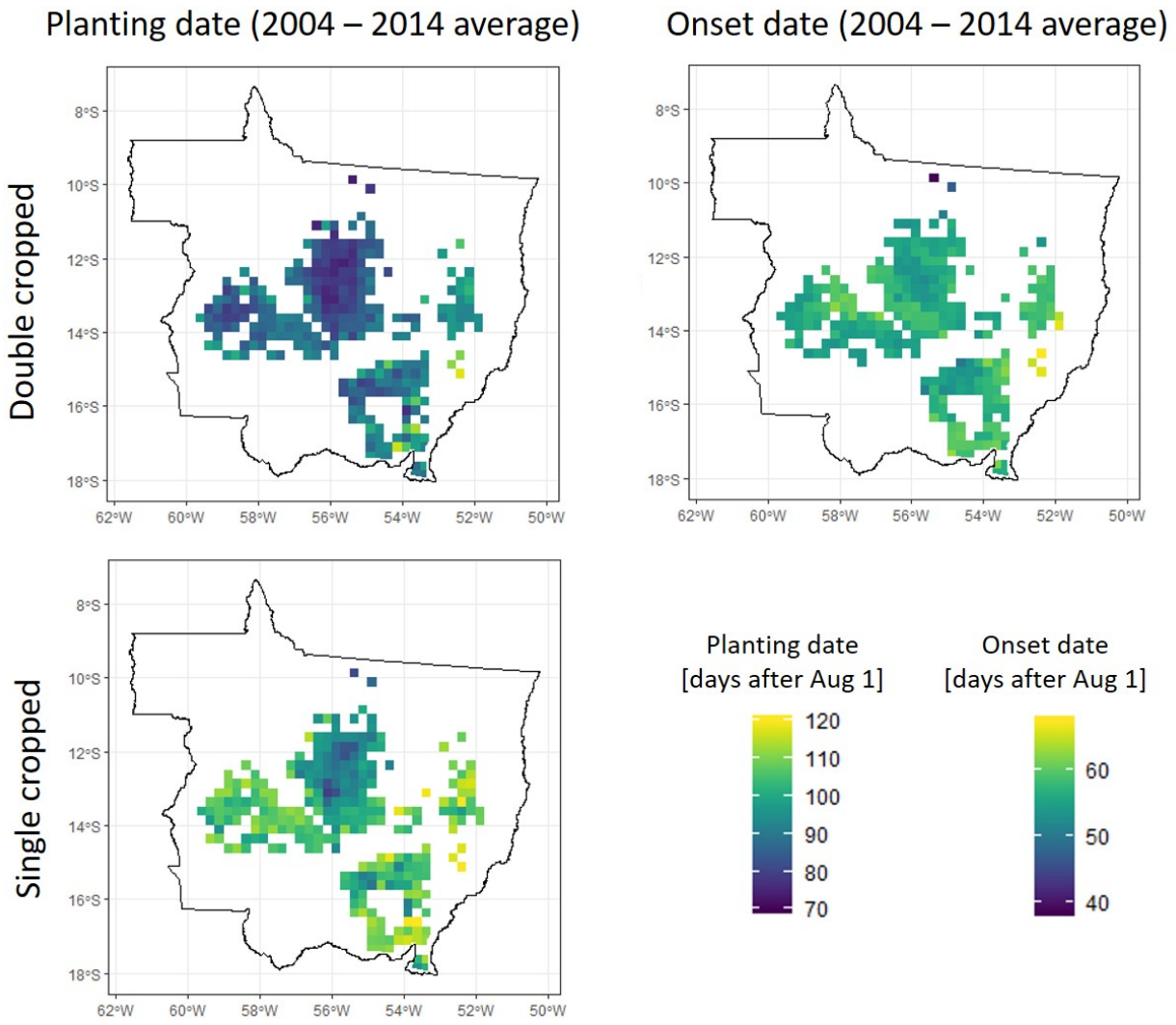


Figure 3.6: Estimated mean planting and onset dates, averaged from 2004 to 2014, in days after August 1 of the planting year. The areas shown represent only soy planted in all years of the study period.

However, it is clear that the onset and sanitary break do have some impact on crop timing: planting dates almost never occur before these two constraints (Figure 3.8a). Among the two limits, the wet season onset likely exerts the stronger pull on behavior. Except for the anomalously early-onset year of 2010, onset occurred after the end of the sanitary break, suggesting that the sanitary break rarely becomes relevant for planting behavior compared to onset (Figure 3.8a). Additionally, as planting dates moved closer to the onset, the probability

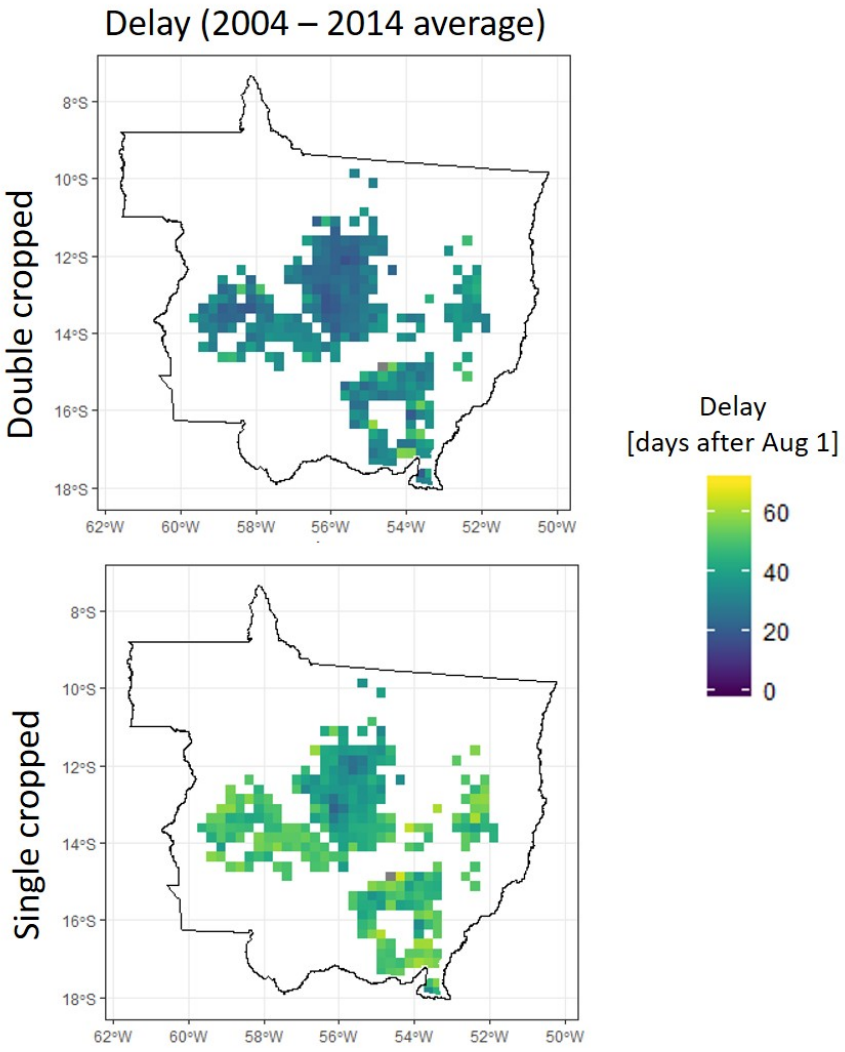


Figure 3.7: Estimated delay between median planting dates and onset, averaged from 2004 to 2014, in days after August 1 of the planting year. The areas shown represent only soy planted in all years of the study period.

density distribution of planting dates became more concentrated. This "piling up" effect is most obvious in 2014 and suggests that farmers collectively pay closer attention to onset as they attempt to plant earlier each year (Figure 3.8b).

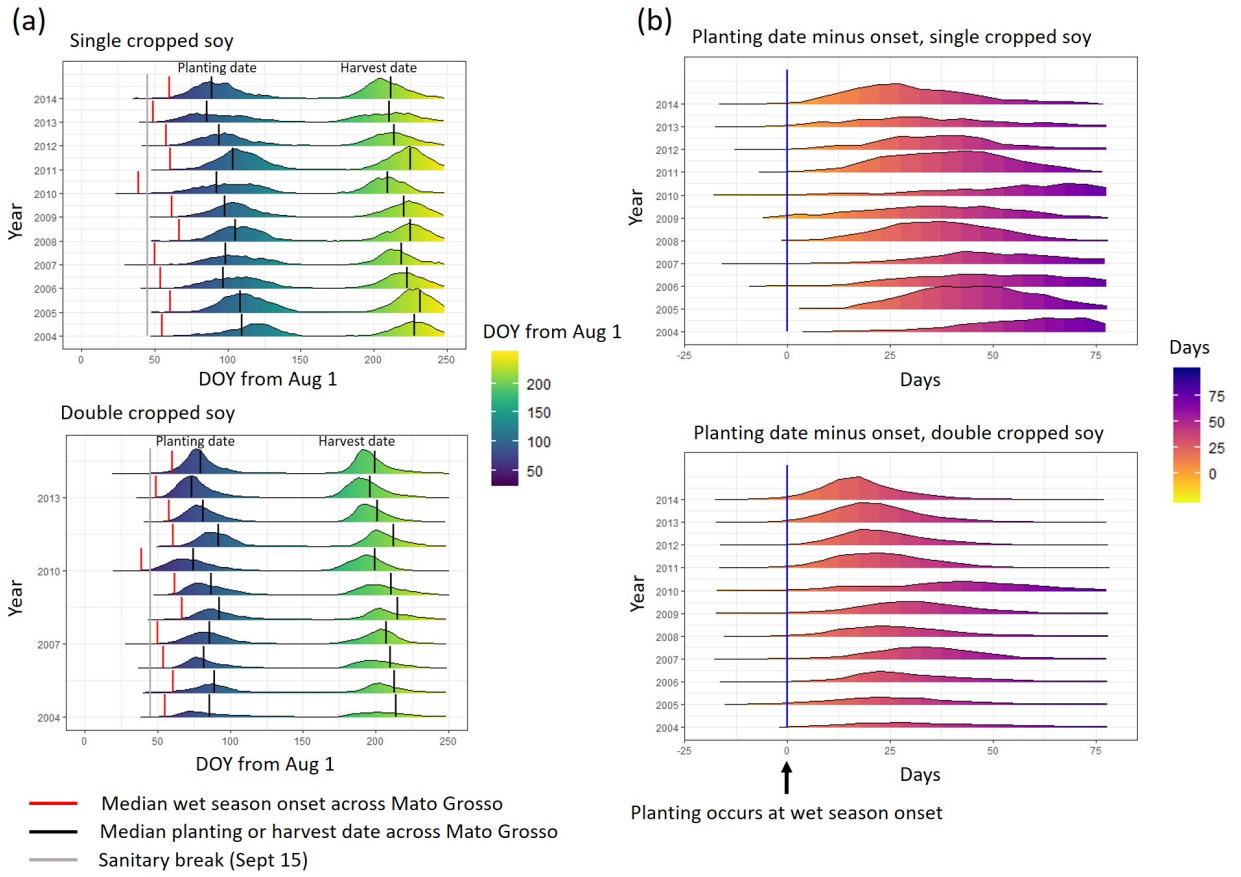


Figure 3.8: (a) Histogram of estimated planting and harvest dates for single and double cropped soy across Mato Grosso from 2004 to 2014. The median wet season onset across the state is shown in red vertical lines, median planting and harvest dates in black vertical lines, and the sanitary break in a gray vertical line. (b) The delay between planting and onset dates across Mato Grosso. Positive value indicates that planting occurs after onset. A delay of zero is represented by the blue vertical line.

3.3.4 Double cropped soy is planted earlier than single cropped soy

Double cropped soy was consistently planted earlier than single cropped soy, although this gap shrank over time - the difference between median planting date for double and single cropped soy across Mato Grosso ranged from 24 days in 2004 to 10 days in 2014 (Figure 3.8). This is mostly associated with earlier planting of single cropped soy, rather than later planting of double cropped soy. Double cropped soy planting dates also appear more sensitive

to wet season onset than single cropped soy. For example, in 2010, double cropped soy was planted earlier to match the unusually earlier onset, while single cropped soy did not adjust as strongly to the earlier onset. The left side of double cropped planting date histogram in Figure 3.8 pushes against the wet season onset, while the right side of the histogram tapers off around 110 days after August 1 - a cutoff consistent with the need to harvest soy in time for the second crop to be planted. Single cropped soy, in contrast, was less constricted on both ends and its probability density distributions are much wider.

3.3.5 Validation with other existing crop calendars

Though my estimates cannot be validated with a spatially resolved dataset, I ensure that my estimates agree with spatially aggregated statistics. First, my estimated planting and harvest dates indicate a mean crop cycle length of 112 days, consistent with reported values [1]. I also compare my estimates to two existing crop calendars. The SAGE dataset estimates that in Brazil, soy is planted in late November and harvested in late March, equivalent to 120 and 240 days after August 1, respectively [115]. This estimate is closest to my planting and harvest date estimates for single cropped soy in 2004 - close to the year during which the SAGE data were collected and in a period of time when single cropped soy was the dominant planting intensity. While SAGE data represent both irrigated and rainfed cropland at the global level, only 2.5% of Mato Grosso's row crop was irrigated as recently as 2017 [56], so almost all soy in Mato Grosso was rainfed during the SAGE period. Second, a weekly crop progress report for Mato Grosso's soy is available from the Instituto Mato-Grossense de Economia Agropecuaria (IMEA) agency [67]. The date at which 50% of Mato Grosso is planted is comparable to the reported values, as shown in Table 3.2.

3.3.6 Estimation error for planting and harvest dates

The trends and relationships described above must be placed in context of the planting and harvest date error. To quantify the pixel-scale error in the planting and harvest date estimate for known soy fields, I compare these estimates to the planting and harvest dates observed directly from Planet Labs imagery. The agreement between the estimates obtained from analyzing EVI timeseries and those made directly from Planet Labs imagery provides an estimate of the accuracy of the timeseries analysis algorithm, including internal calibrated parameters. These pixel-level errors quantify the accuracy of the planting and harvest date estimation method itself, independently of errors in the land cover map. Following Equation 3.2, the pixel level bias and its confidence interval is 6.9 ± 16.5 days for planting and 1.8 ± 18.7 days for harvest. This error includes the differences between the estimated and reported planting and harvest dates, and the uncertainties in the reported planting and harvest dates themselves.

Planting and harvest date errors at the 25 km aggregated scale are shown in Table 3.3. The values shown here combine errors in land cover classification and the uncertainties in pixel-scale estimates.

Year	IMEA-reported date of 50% planted	Estimated date of 50% planted (SC, DC)	IMEA-reported date of 50% harvested	Estimated date of 50% harvested (SC, DC)
2013	October 25	October 25, October 14	February 25	February 12, February 27
2014	October 24	October 29, October 20	February 20	February 16, February 28

Table 3.2: Comparing the estimated planting and harvest dates to IMEA's weekly crop progress reports. Unfortunately, IMEA does not report crop progress separately for single and double cropped soy.

Total error [days]	Single cropped soy	Double cropped soy	All soy (single + double cropped)
Planting	6.9 ± 18.7	6.9 ± 17.5	6.9 ± 17.4
Harvest	1.9 ± 21.3	1.8 ± 19.9	1.8 ± 19.8

Table 3.3: Planting and harvest date error at aggregated scales. This combines pixel level errors in planting and harvest date estimates and land cover classification errors.

Pixel scale errors (RMSE of 6.9 ± 16.5 days for planting and 1.8 ± 18.7 days for harvest) are comparable to those in other satellite-based studies of soy agriculture (mostly in the US and validated on USDA NASS Crop Progress Reports at the county level). For example, Urban et al (2018) achieved an RMSE of about 5 days for soy planting dates in the central US; RMSEs of 3.2 to 6.9 days were associated with different soy phenological stage estimates in a study across 3 regions in Eastern Nebraska [155]; and an RMSE of 5.3 days was achieved for estimates of the start of the soy season in the Midwestern US [112]. While the mean error in this study is comparable to these past efforts, the precision of the error is lower due

to the uncertain nature of planting and harvest dates derived from Planet Labs imagery.

Important features of planting behavior can be detected despite the error and imprecision. Differences between median planting dates of single and double cropped soy (10 to 24 days, depending on the year), the trend toward earlier planting (13 to 20 days over the study period, for double and single cropped soy), and the delay between planting date and onset (29 to 41 days) all appear clearly against the magnitude of RMSE. Despite the errors introduced from the simple timeseries analysis method, the estimates reveal important patterns in the planting and harvest dates' changing relationship to cropping intensity and wet season onset. While imprecision at *pixel* scale (17 to 21 days) is still sizeable compared to the detected differences, the imprecision of these *aggregated* differences will be much lower.

3.4 Discussion

3.4.1 Soy Agriculture and its timing in Mato Grosso

My results suggest that the common assumption that the onset controls the planting date of tropical rainfed crops is too simplistic, and often incorrect. Instead, Figure 3.8 reveals a delay between planting and wet season onset of up to 3 months for single cropped soy and up to 2 months for double cropped soy, a window first posited by Abrahao et al (2018) based on soy's photoperiod and climatological requirements. Averaged across all years, the delay between onset and planting date was 41 days for single cropped soy and 29 days for double cropped soy. The smaller delay over double cropped fields is important because much of Mato Grosso's agricultural revenue depends on the feasibility of double cropping [11]. If onset is delayed sufficiently to make the earlier planting dates of double cropping impossible, the state may suffer a loss of profit. Further, the magnitude of delay changed over time for all soy: the delay was larger in earlier years, suggesting factors that allow farmers to plant closer to the onset each year. The presence of a delay, and its change over time, means that planting and harvest date estimates in this region should not be based upon precipitation data alone. This also means that global datasets depicting planting and harvest dates and which rely on the assumption that planting date occurs at the onset of the wet season may be up to 3 months in error.

The delay between planting dates and the wet season (and occasionally the sanitary break, if it occurs later than the onset) could be attributed to a variety of non-climatic factors: (1) the time (up to 4 weeks) needed to complete planting operations [106], (2) the result of logistical and economic constraints that delay farmers from planting at a desired time [148], or (3) a deliberate choice to improve soy yields: simulations suggest that moving planting date from Sept 25 to Oct 5 increases soy yields by increasing the precipitation received by crops [106]. The reasons leading to the rift between the onset and the observed planting dates have significant implications for the future of double cropping in Mato Grosso, and should be examined further.

More evidence of the non-climatic constraints on planting can be found in the spatial

patterns of planting dates, emphasizing the need for high-quality planting data that can isolate these effects. At pixel scale, my maps of soy planting and harvest dates reveal that differences in crop timing between nearby fields are comparable to or greater than interannual and regional differences in the planting and harvest dates. This suggests that field-scale knowledge is essential to characterizing planting and harvest dates across a property, an insight that should inform future survey design. At regional scales, a distinct spatial pattern of planting dates emerges: the center of the state is planted earlier than most other areas. This could be a result of the uniquely favorable climate, soil and topography surrounding the BR163 highway running north and south across the state [105]. While the spatial patterns of planting for both double and single cropped soy roughly match the spatial pattern of onset, it is clear from the spatial patterns remaining in the delay (in which the effect of onset is taken out) that planting dates follow a spatial pattern that is independent of onset (Figure 3.7). Importantly, this pattern differs by cropping intensity: planting dates and delays for double cropped soy appear more spatially homogeneous. This discrepancy may stem from differences in resource access or optimal planting time. Single cropping practitioners typically lack the resources to complete the more expensive (but more profitable) double cropping operations, and farmers located away from the central highway network may be unable to plant as early as they desire due to logistic constraints [105]. On the other hand, the spatial homogeneity in double cropping may result from the necessity of planting as early as possible to allow the second crop to mature before the dry season begins [1, 106]. Single cropping practitioners have more flexibility in choosing a yield-optimizing time, which may contribute to the spatial heterogeneity. In the future, the state's growing transportation network (such as the expansion north into the Amazon) and shifting cropping practices may create a changing spatial pattern in planting dates, highlighting the importance of spatiotemporally resolved planting information [29, 105].

3.4.2 Implications for planting date and crop yield predictions under climate change

These insights could aid efforts to assess the impact of climate change on Brazilian agriculture in a way that reflects the diverse non-climatic factors that go into planting decisions [60, 111]. Globally, crop modeling efforts often resort to approximations in which planting is triggered based on precipitation, temperature, or soil moisture [56, 72, 133, 38, 121]. While these rules were developed in tandem with crop requirements and, in certain cases, historical observations, there is no guarantee that socio-economic constraints will allow farmers to plant in strict accordance with crop requirements, or that rules derived from historical observations will continue to apply under new crop varieties and agricultural practices. Indeed, these rules were found to perform well on a global or continental scale, but are less appropriate on smaller scales [35]. It would be beneficial to incorporate what we now know about the sizeable and shifting delay between onset and planting dates into crop yield predictions for Mato Grosso.

In addition to improving predictions of crop yield, my spatially resolved, updated plant-

ing dates are key to understanding adaptation choices at the individual level. Adaptive behaviors, such as shifts in planting dates and cropping intensity, are the result of heterogeneous information access, resources, and beliefs [68]. The nuances of adaptation, and their implications on crop yield, profitability, and food security, are nearly impossible to predict when disparate individuals are lumped together [60, 111]. My planting date maps are a necessary step towards a more robust understanding of current planting decisions and how they can be optimized.

3.4.3 Planting dates extend phenological information

As discussed in Chapter 2, the estimation of planting dates from remotely sensed data is limited by the spatial and temporal resolution of the sensor, and by the potentially high variability in duration between planting and satellite-detected emergence. This increases uncertainty in planting date estimates, and may mean that planting dates are not necessarily the most suitable input for predicting crop yields. Indeed, the reproductive stage of soy, during which seeds are developing, is the most critical to seed growth and therefore the most highly predictive of crop yield [7, 17]. The reproductive stage is more easily estimable from satellite imagery because of its higher above-ground biomass, and may be a better option for studies in which greenness at this stage is already known. Examples include yield forecasting, which predict crop yields for actively growing crops before they are harvested [17], or predicting yields in cases where the future phenological trajectory is known [136]. However, because phenology is a complex function of weather and adaptive strategies [23, 44, 75], it is unlikely that the reproductive stage can be forecast accurately without first understanding the preceding decisions, such as planting date. Lack of knowledge about planting dates would be a major shortcoming in efforts to predict yields under climate change. Planting and harvest dates reveal adaptations (such as switching between cropping intensities and crop varieties) that propagate into subsequent phenological development and influence yields in unpredictable ways. Understanding soy development from the beginning of the life cycle can also help to capture the full effect of climate change on yield. For example, the expected delay in wet season onset across Mato Grosso may disproportionately affect development in the early growth stages. Thus, knowledge of planting dates extends our understanding of crop growth and yield by more explicitly including the management decisions and weather variability that occur at the start of the growing season.

3.5 Conclusions

A realistic understanding of planting behavior, and consequently an accurate depiction of future yields, is only possible with updated, highly resolved planting data. My field-scale planting dates for Mato Grosso may aid in developing a mechanistic framework of how farmers respond to weather variability. This mechanistic understanding is indispensable because planting dates and other adaptation decisions are ultimately made on the individual

level, in response to a variety of internal and external factors [68]. In dynamic areas like Mato Grosso, where climatic, technological, and economic trends cause the planting and harvest dates to shift on the time scale of years rather than decades, quickly updated planting date information is essential to forming a timely response to climate change.

The planting dataset produced here provides rich information about individual farmers' behavior - knowledge that will allow extrapolation to behavior and productivity under future scenarios. Planting date variability across fields can be equal to variability across the state, indicating the importance of understanding planting decisions at the farm scale. At the regional (25 km) scale, a consistent spatial pattern emerges: central Mato Grosso tends to be planted earlier than other areas, indicative of differences in climate or availability of planting equipment. In addition to spatial variability, there is an interannual trend towards earlier planting, independent of onset, which may reflect improvements in planting technology, soy variety, or the transportation network. This trend was possible because average planting at the start of the study period (2004) occurred at least a month after the end of the sanitary break and the median wet season onset, suggesting either deliberate delays to maximize yields, or involuntary delays caused by logistic limitations. The assumptions that planting occurs at wet season onset or immediately after the sanitary break are therefore untrue for most of Mato Grosso. Finally, the results confirmed that double cropped soy is planted earlier than single cropped soy, but the difference between their planting dates shrank over the study period. The rapid approach of double cropped planting dates toward the wet season onset between 2004 and 2014 suggests that a delay in wet season onset due to climate change could render double cropping impossible, or force planting to suboptimal dates. The estimation strategies used for Mato Grosso may be useful for risk assessment in regions such as southern Asia and southern Africa, which face not only the most severe consequences of warming, but also data scarcity and limited adaptive capacity [54, 84, 87].

Chapter 4

Modeling the sensitivity of planting date selection to wet season onset

4.1 Introduction

Planting date shifts are a primary adaptation strategy under climate change, and in some agricultural regions, they are projected to more than compensate for its harmful effects. However, planting dates for rainfed crops rely heavily on the timing of water availability and may become ineffective as an adaptation strategy if the wet season is constricted. A crucial piece of understanding how climate change will impact agricultural productivity will, therefore, require quantifying planting dates and their sensitivity to climatic variables such as wet season onset. This chapter describes the sensitivity of soybean planting dates to wet season onset in Mato Grosso, Brazil from 2004 to 2014.

Before planting dates can be related to the wet season onset, the onset itself must be defined. The selection of the definition for wet season onset, and of the precipitation data from which it is calculated, both introduce uncertainty in the definition of wet season metrics. Precipitation data sources derived from satellite data products or interpolated from rain gauge observations may introduce biases that propagate into measures of precipitation seasonality. A study of gridded precipitation datasets over Brazil's Cerrado (the rainforest-savannah transition region which includes Mato Grosso) proposed a method to select precipitation datasets based on agreement with streamflow measurements. Their method exploits the natural correlation between rainfall and streamflow: because these variables are measured independently, streamflow measurements can inform the selection of a hydrologically relevant precipitation dataset [85].

Similarly, the most suitable definition of wet season onset in the context of planting decisions will have the highest correlation to observed planting dates, and will presumably highlight the attributes of precipitation that are most relevant in decision-making. This idea has been applied in Mali: four onset definitions were tested, and the definition that produced the closest agreement between onset date and planting windows was chosen as the

start of the growing season. However, two drawbacks affected the quality of the findings. Because historical planting dates were unknown, the study relied on expert knowledge to define planting windows. Further, all the definitions highlighted related features of precipitation: they were based on total rainfall accumulation over a certain number of days and no subsequent dry spell, with variations on the thresholds used [3]. This effort can be extended by using observed and spatially resolved planting dates, and by testing a broader range of onset definitions. In this chapter, I develop regression models to extract the sensitivity of soy planting dates to various definitions of onset. The definition with highest correlation to planting dates is selected for further analysis.

A statistical model that defines the effect of climate on planting dates must control for the impact of non-climatic factors on planting date. Figure 4.1 displays a conceptual diagram of planting date decisions. Delayed agricultural credit to purchase seeds [1]; slow transportation of heavy planting equipment [36]; the physical time required to plant a field; perception of climate risk [23, 44, 75]; and the desire to avoid harmful climate during sensitive phenological stages [59, 70, 31] may all contribute to a delayed planting date. In contrast, high crop prices [19], use of irrigation [44], and multiple cropping [106] may advance the planting date.

Additionally, Mato Grosso's unsteady institutional and economic background cause it to experience more volatile crop timing than regions with more established agricultural practice [29]. The continued development of new soybean varieties, improving transportation network, and variations in planting equipment availability and timing of agricultural credit disbursement continue to influence cropping intensities and planting dates [1, 43, 51].

These socio-economic controls mean that farmers may not respond perfectly to climate, and that sensitivity to climate may vary widely between individuals. The presence of confounding factors and possibility of nonstationary behavior inform the development of the model, which aims to describe the heterogeneous planting response to wet season signals across Mato Grosso. My effort to define planting date's sensitivity to precipitation regimes can be summarized with two related research questions:

1. What features of precipitation are most relevant to planting decisions in Mato Grosso?

2. What is the sensitivity of planting date to wet season onset in Mato Grosso?

Though Mato Grosso is the focus of this study, the insights and techniques introduced here can be applied to rainfed agriculture worldwide. Regression of various onset definitions with observed planting dates can clarify the features of precipitation that are most relevant to decision-makers at the farm level. In turn, these regressions may reveal a large difference in onset sensitivity for different cropping intensities and fields. Because adaptation strategies are implemented by individuals, quantifying the diversity of responses to onset is crucial to accurately predicting agricultural adaptation and yields under changing precipitation regimes.

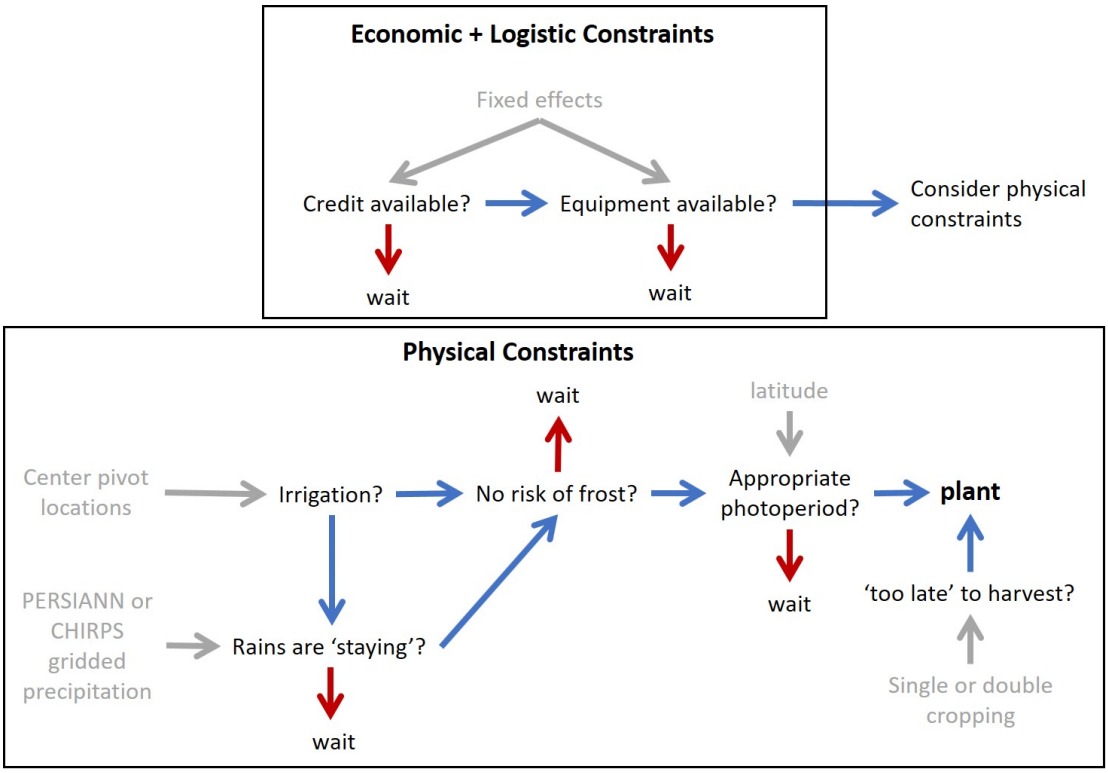


Figure 4.1: Economic, logistic, and physical (climate) constraints all exert a pull on planting date. They must be considered either explicitly through the inclusion of explanatory variables in the regressions, or accounted for with fixed effects. Gray arrows indicate data sources; blue arrows indicate that the constraint is met; and red arrows indicate that the constraint is not met (and therefore delays planting date).

4.2 Methods

4.2.1 Data

Precipitation

Two daily, gridded precipitation datasets are separately used to calculate wet season onset: (1) PERSIANN, available globally from 1983 to the present at 0.25 degree scale, estimates precipitation using infrared satellite data and is adjusted with the Global precipitation Climatology Project (GPCP) monthly product [13]; and (2) CHIRPS, available globally from 1981 to the present at 0.05 degree scale, estimates precipitation based on rain gauge and satellite-based cold cloud measurements [49]. Each of these datasets was used, individually, to calculate the wet season onset date.

The choice of rainfall dataset introduces data selection uncertainty that will impact the

calculation of precipitation statistics such as the wet season onset. Precipitation over areas with sparse rainfall gauge density are generally better described with remotely sensed products. Though remotely sensed precipitation may suffer from bias that worsens with complex topography, the flat terrain of Mato Grosso and low rain gauge density (fewer than 15 per 10^4 km^2) make it the preferred option. A study of rainfall datasets available over the Cerrado region found that PERSIANN was the best performing gridded dataset, with high correlation to streamflow observations and to interpolated in-situ rainfall measurements. Other gridded products tested were the Global Precipitation Climatology Project (GPCP, 1 degree), Climate Prediction Center Unified Gauge-Based Analysis of Global Daily Precipitation (CPC, 0.5 degrees), and the Tropical Rainfall Measuring Mission (TRMM, 0.25 degrees) [85]. Gridded CHIRPS data was not examined, but I use it to exploit its high spatial resolution (0.05 degrees) and gain a fine-grained understanding of sensitivity to wet season onset.

Planting dates

Planting dates of rainfed single (SC) and double (DC) cropped soybean from 2004 to 2014 in Mato Grosso were created in Chapter 3 and used here. Though the planting dates were estimated at MODIS (500 m) scale, they are aggregated to the spatial resolution of the onset data to which they were compared (25 km for PERSIANN-derived onset, and 5 km for CHIRPS-derived onset). For each 25 $\times$ 25 km or 5 $\times$ 5 km grid cell in the onset map, I calculate the 5th, 25th, 50th, 75th, and 95th percentile of the 500 m pixel-scale estimates within that cell. Single and double cropped pixels are aggregated separately to produce a total of ten aggregated planting date maps for each year, representing a combination of (single cropped soy, double cropped soy) $\times$ (5th, 25th, 50th, 75th, and 95th percentile). As in previous chapters, I refer to planting date in days after August 1. To relate the calendar year to the agricultural year, I refer to "planting year" and "harvest year". Results reported for the year 2014 refer to planting year of 2013 and harvest year of 2014.

Political boundaries

Individual farm property boundaries circa 2010 are reported by the National Environmental Registry of Rural Properties (CAR) [126].

4.2.2 Regression model

A regression model is built to simultaneously (1) select an agriculturally relevant onset definition, and (2) to quantify the sensitivity of planting date to onset (and how it varies for different cropping intensities and percentiles of planting within an area). To achieve this, I rely on a statistical model that allows an intuitive interpretation of the planting-onset relationship. In this section, I describe the model specification; the Supporting Information for Chapter 4 details the range of model specifications that were tested and the evaluation metrics that guided my choice.

I selected an ordinary least squares (OLS) regression with fixed effects (FE), shown in Equation 4.1, in which planting date is the response variable, each onset grid cell i is assigned a fixed effect α_i, with onset date and year included as additional predictors. OLS regression relates the response in a dependent variable (planting date) to its explanatory variables by fitting a line that minimizes the sum of the squares of residuals between the fitted equation and the data. In Equation 4.1, β_{onset} and β_{year} are the coefficient of the wet season onset and year, interpreted as the number of days that planting shifts for unit change in onset and year, respectively, with all other variables held constant. They are the estimated quantities of interest, and will be referred to as the "onset coefficient" and "year coefficient". In this generalized equation, *onset* can refer to any pairing of onset definition and precipitation dataset, and *plant* is aggregated to the scale of the onset dataset and may refer to any combination of cropping intensity and aggregation percentile: (single crop, double crop) $\times$ (5th, 25th, 50th, 75th, and 95th percentile). The uncertainty associated with the onset and year coefficients is a combination of the standard error of the individual coefficient estimates and the error in planting date estimates. This uncertainty is evaluated through bootstrapping.

$$plant = \alpha_i + \beta_{onset} * onset + \beta_{year} * year + \eta_i \tag{4.1}$$

Fixed effects background

I chose to include fixed effects because there are few datasets available to control for the economic and logistic variables that affect planting dates (see Figure 4.1). Fixed effects regression extends linear regression to control for unobserved explanatory variables that are constant in time but variable over space. This is accomplished by fitting separate lines to observations in pre-defined "groups" of space, rather than pooling all observations to fit a single line (as in OLS). The fixed effect term, α_i, represents the separate intercepts for each fixed effect unit, i. The inclusion of fixed effects allows the model to separate the group-specific effects on planting date from the effects associated with changes in the explanatory variables. By fitting a unique intercept to each group, fixed effects absorb differences in baseline planting behavior produced by spatially varying constraints. Ignoring fixed effects may result in omitted variable bias of the onset sensitivity. For example, if an inverse relationship exists between wet season onset and speed of access to planting equipment, a pooled OLS that ignores the access variable may confound the effect of a delayed onset with faster access and produce a lower estimate of the onset coefficient. While the fixed effect terms cannot control for spatiotemporally varying missing predictors, it can compensate for missing time-invariant factors like soil type, access to transportation networks, and access to credit and equipment. These are unobserved but may play a role in planting date decisions. I will refer to pooled OLS models as $\text{OLS}_{\text{pooled}}$, and OLS models with fixed effects as OLS_{FE}.

Model selection

The OLS_{FE} model specification was selected through a series of exploratory regressions, in which a variety of model types, observation scales, sampling grid sizes, and predictors were tested. This specification was selected from a set of possibilities based on model diagnostics, predictive accuracy, interpretability, and robustness under the condition of missing and/or spatiotemporally autocorrelated predictors. The Supporting Information for Chapter 4 provides the set of specifications explored and justifies the final choice through a series of selection criteria.

Model evaluation metrics

For both OLS_{pooled} and OLS_{FE} models, unbiased coefficient estimates are only guaranteed if several OLS assumptions are met. I tested the chosen OLS_{FE} model specification for adherence to the following assumptions:

- Residuals have zero mean.
- Residuals are normally distributed. This is tested with a QQ plot.
- Residuals are exogenous. This is tested by calculating the correlation coefficient between the residual and each of the predictors.
- Residuals are independent and not autocorrelated. This is tested with Durbin-Watson for temporal autocorrelation and Moran's I for spatial autocorrelation.
- Predictors are not multicollinear. This is tested by calculating the correlation matrix for the selected predictors.
- Residuals are homoscedastic. This is tested by looking for constant variance in residuals in a residual-fitted value plot.

If these assumptions are met, the residual contains no useful information about planting date.

These model selection and evaluation steps were performed once for each onset definition tested. This ensures that the most appropriate model is specified for each onset definition before their modeled coefficients are compared.

4.2.3 Selecting an agriculturally relevant onset definition

Onset definitions

Several definitions of the wet season onset exist in the literature, but it remains unclear which features of precipitation are most often perceived as the "start of the wet season" by farmers. I explore six definitions of onset that have been applied in South America,

Africa, and India, and quantify their correlation to the planting date decisions made in Mato Grosso based on the magnitude of the fitted onset coefficient. An onset definition with a higher onset coefficient indicates that observed planting dates are more correlated to the features of precipitation described in that definition.

The six onset definitions are:

1. The anomalous accumulation (AA) method. The AA method is a standard climatological definition of wet season onset. In AA method, the wet season onset date is defined based on the value of the anomalous accumulation [mm/day]:

$$AA(t) = \sum_{n=1}^{t} (R(n) - R_{\text{ref}}) \tag{4.2}$$

   where $R(n)$ is the rainfall on day n and R_{ref} is a reference rainfall value, defined here as the agronomically significant threshold of 2.5 mm/day [12]. Here, $t = 1$ refers to July 1, the middle of the dry season. The onset date is defined as the day at which the value of $AA(t)$ reaches its minimum [86].

2. Depth method. This defines onset as the first day after August 1 with rainfall depth over some threshold. This has been used to define onset in Rondonia, Brazil [24].

3. Volume method. The volume method defines onset as the first time total rainfall over 10 days exceeds a certain depth, where the subsequent 15 days experiences total rainfall over a certain depth. The onset occurs at the end of the 10 day period. This has been used to define onset in West Africa [83].

4. Frequency method. The frequency method defines onset as the end of a 4 week period in which the number of rainy days (depth $\geq$ 1 mm) exceeds a certain number. This has been used to define onset in the Brazilian Cerrado, a region south of the Amazon which includes the state of Mato Grosso [129].

5. Pentad method. The pentad method defines onset as the start of the earliest pentad with total rainfall greater than a certain depth, with the previous pentad has less total rain and the following pentad with more total rainfall. It has been used to define onset in the Amazon [93].

6. Monsoon method. The monsoon method defines onset as the first wet day (rainfall $\geq$ 1 mm) that is not followed by a 10 day dry spell (total rain less than a given depth). It has been used to characterize onset in India [68].

Because each onset definition involves some threshold(s) of depth or frequency, I test a variety of threshold values. For each onset definition, a "sensible" set of threshold values for Mato Grosso was found by testing a large range of potential threshold values, and choosing

the values that produce "reasonable" onset estimates (in which all locations in the state have onset dates within the agricultural year). I calculate onsets with each definition and threshold combination, listed in Table 4.1, using both PERSIANN and CHIRPS data sources. PERSIANN produces onset estimates at 25 km resolution; CHIRPS produces onset estimates at both its native resolution of 5 km or aggregated to 25 km. This set of onset variations allows me to simultaneously explore the features of precipitation most closely followed by farmers, but also explore the effect of spatial resolution and precipitation data source on the modeled relationship between planting date and onset. Computations are performed in Google Earth Engine (GEE), a cloud computing platform offering easy access to PERSIANN and CHIRPS datasets.

Onset definition	Threshold(s)	Threshold values tested
Anomalous accumulation	Agronomically significant value	2, 2.5, 3 mm
Frequency	Minimum wet days in 4 week period	5, 8, 10, 12, 14 days
Depth	Minimum depth of rainfall in one day	10 mm
Monsoon	Minimum total rainfall in 10 days	30, 40 mm
Pentad	Minimum total rainfall in 5 days	8, 10, 15 mm
Volume	(Minimum total rainfall in 10 days preceding, minimum total rainfall in 15 days following)	(15, 30); (15, 40); (20, 20); (20, 30); (30, 40)

Table 4.1: Thresholds values tested for each onset definition. These options were tested for each precipitation dataset; for CHIRPS-derived onset, both 5 km and 25 km scales were tested.

Selection criteria

The onset definition (and accompanying threshold value) with the highest estimated onset coefficient in the chosen OLS_{FE} specification is selected as the one to which farmers are most sensitive. However, each onset definition and threshold comes with many values for the onset coefficient: corresponding not only to variations in the precipitation data source, but also to the different cropping intensities and percentiles of planting date that are modeled.

I choose the "best" onset definition as the one that generates the highest average onset coefficient for all percentiles of double cropped soy because the dynamics of double cropped planting dates are expected to be more sensitive to wet season onset and more vulnerable to climate change.

Examining spatial and temporal patterns in wet season onset definitions

To illuminate why planting dates appear more responsive to certain definitions, I compare the spatial patterns and temporal trends associated with each onset definition. I test three mechanisms that may allow certain definitions to perform better than others: (1) a higher spatial variability in estimated onset would increase the variation of planting dates that are attributed to onset, in the absence of any systematic spatial or temporal differences among onset definitions; (2) consistent, long-term spatial biases among definitions allow some to capture spatial planting patterns better than others; and (3) differences in temporal patterns among definitions allow some to capture interannual variation in planting better than others.

4.2.4 Sensitivity and robustness tests

Even with model specification and onset definition selected, the model results are still susceptible to missing predictors, planting date estimation error, and, for OLS_{pooled}, to shifts in the sampling grid position (see Supporting Information for Chapter 4). Because my aim is to quantify the difference in sensitivity to onset for subsets of soy agriculture (depending on cropping intensity and percentile), I perform robustness tests to ensure that these differences can be detected in spite of the uncertainties.

4.3 Results

Though the selection of model specification and wet season onset definition are described as separate tasks, they are choices made in tandem. The full set of combinations between model specification and onset definition is tested to ensure that the model specification does not impact the choice of onset definition, and vice versa. However, in the interest of space, not all combinations are reported: here, I report model selection criteria based on the selected onset definition; and onset definition selection criteria based on the selected OLS_{FE} specification. Model results (estimated coefficients and uncertainties) are reported only for the selected

onset definition here; results for one alternative definition are described in the Supporting Information for Chapter 4.

4.3.1 Onset definitions

I choose the "best" onset definitions as the definitions and their thresholds with the largest onset coefficient for double cropped soy under the chosen OLS_{FE} specification across all percentiles, precipitation data sources and aggregation scales.

Figure 4.2 displays the onset coefficients for the 5th, 25th, 50th, 75th, and 95th planting date percentiles. The frequency and volume definitions have the highest onset coefficient, a trend that is robust to not only different cropping intensities and percentiles, but also to different precipitation data sources and aggregation scales (5 km or 25 km). The ideal thresholds for the frequency definition are 10 and 8 days, and the ideal thresholds for the volume definition are (15mm, 40mm) and (20mm, 30mm). While both definitions are best regardless of the precipitation data used, onsets calculated with PERSIANN tend to have higher onset coefficients for double cropped soy, and those calculated with CHIRPS have higher onset coefficients for single cropped soy. The top onset definition for double cropped soy is $\text{frequency}_{10,\,\text{PERSIANN}}$, while the top onset definition for single cropped soy is $\text{frequency}_{8,\,\text{CHIRPS}}$.

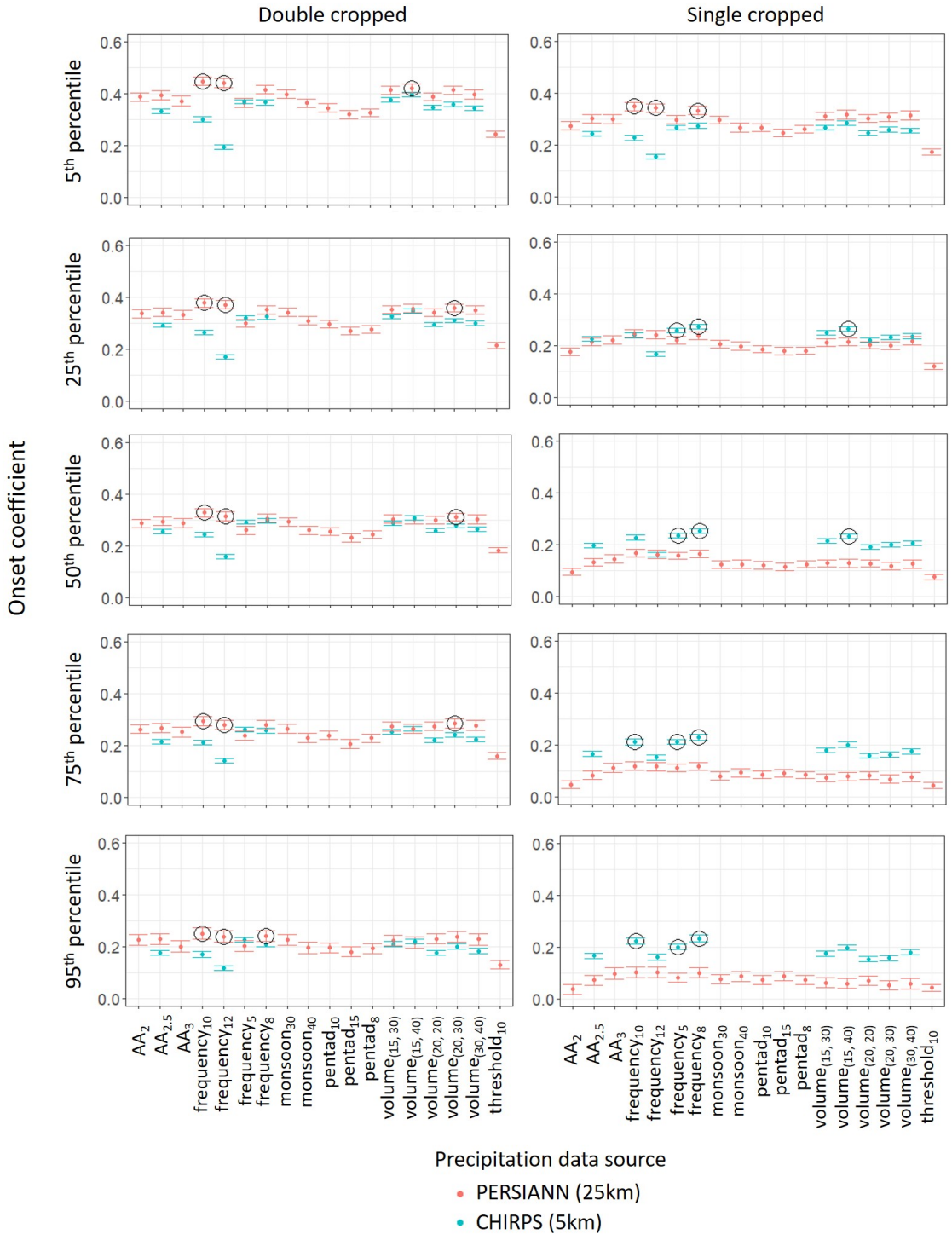


Figure 4.2: Onset coefficients calculated with each onset definition and planting date percentile. Circles indicate the top three onset coefficients, and error bars denote standard error.

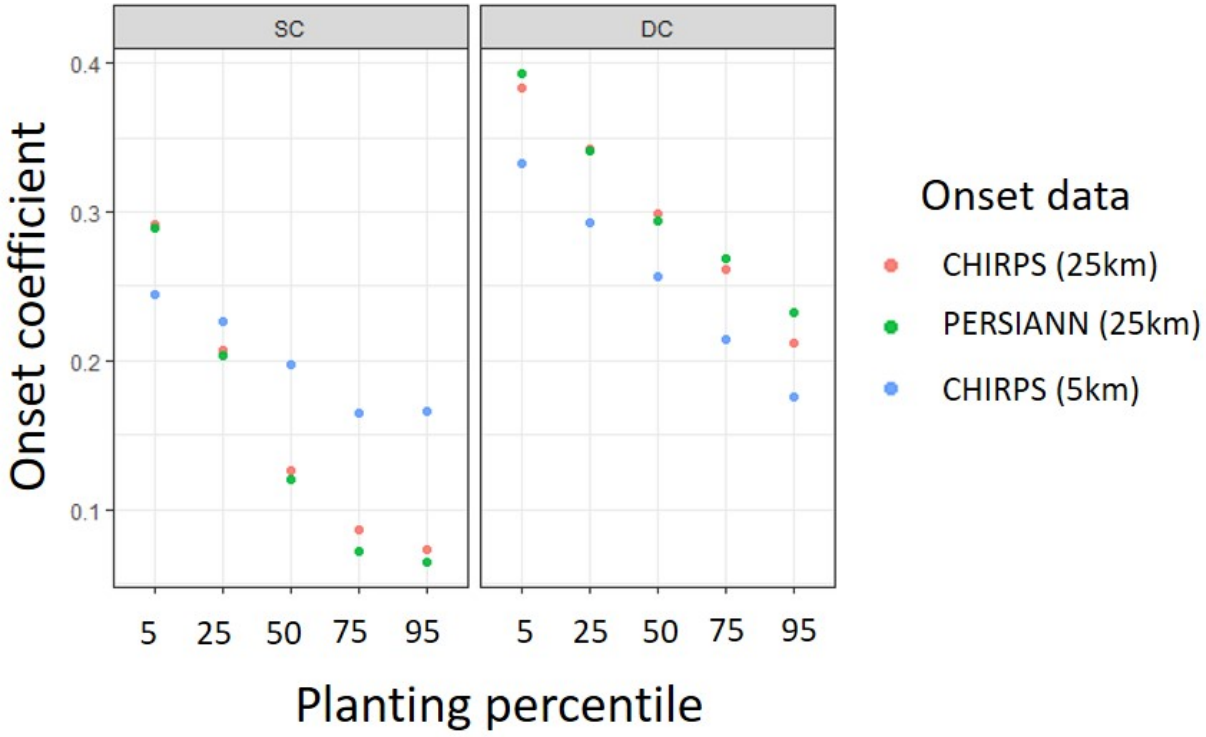


Figure 4.3: The scale makes a bigger difference in onset coefficient than the precipitation data.

Two factors may explain the differences between onset calculated for PERSIANN versus CHIRPS: (1) the difference in spatial resolution and (2) differences in the depth of precipitation detected. I investigate the first possibility by aggregating CHIRPS to 25 km scale, eliminating the mismatch in scale. Figure 4.3 shows that the difference between CHIRPS- and PERSIANN-derived onset coefficients stem primarily from the scale mismatch. When CHIRPS data are aggregated to PERSIANN scale, the estimated onset coefficients become almost identical.

I choose the $\text{frequency}_{10,\ \text{PERSIANN}}$ definition of onset for analysis because it creates the highest onset definition for double cropped soy, the category of higher interest. Planting dates for single cropped soy align most closely with the $\text{frequency}_{8,\ \text{CHIRPS}}$ definition. I report results calculated under this alternative onset definition in the Supporting Information and confirm that the same insights and patterns can be found if $\text{frequency}_{8,\ \text{CHIRPS}}$ were used.

4.3.2 Comparison of onset definitions

I explore the varying agreement between planting and onset based on the hypothesis that each onset definition produces differing levels of spatial variability or systematic biases in spatiotemporal patterns. I focus on the differences between the best onset definitions (which are based on frequency and volume) and a baseline definition, $AA_{2.5,\,\text{PERSIANN}}$, a common climatological definition that has been adapted and used for soy agriculture in Brazil [1, 56].

First, I test the possibility that the higher onset coefficients associated with the frequency and volume definitions are due to higher spatial variability in the onset estimate (in the absence of systematic spatial or temporal differences). Higher spatial variability could be the result of a definition that is more sensitive to small variations in precipitation, or the result of noise in the onset estimate. Both may allow the OLS_{FE} model to attribute more of the planting date variability to onset, resulting in a possibly spurious high onset coefficient. However, Figure 4.4 indicates that the definitions with higher onset coefficients do not have higher spatial variability within each year. Therefore, the higher onset coefficients cannot be attributed to noise or higher sensitivity to the precipitation signal.

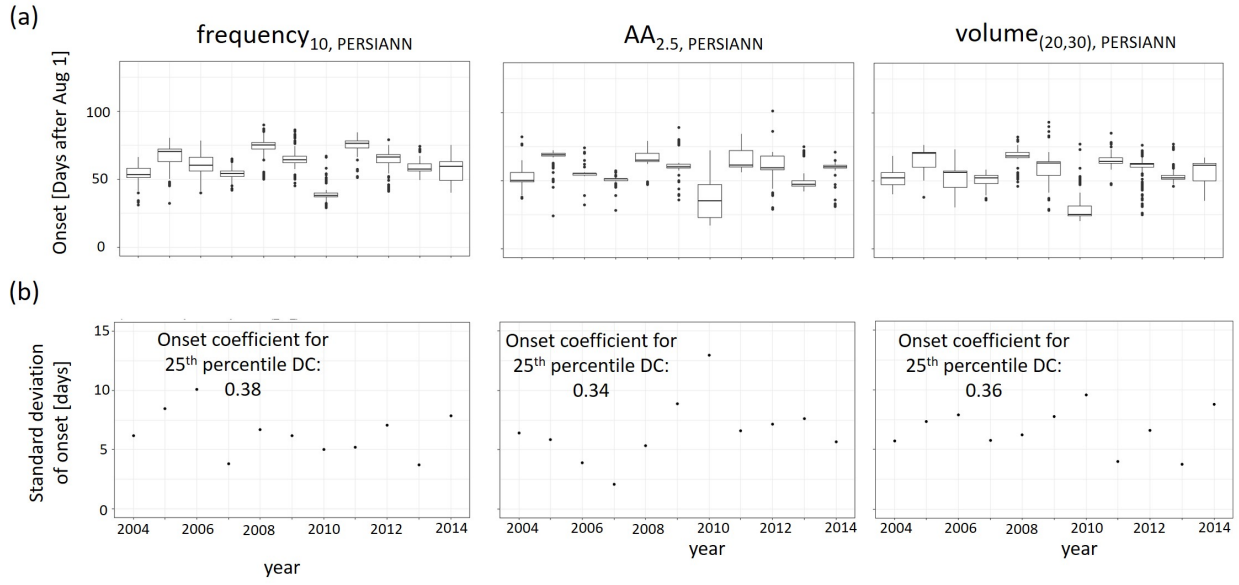


Figure 4.4: (a) Boxplot of estimated onset within each year, (b) standard deviation of estimated onset within each year. Higher within-year spatial variability is not associated with higher onset coefficients.

Second, I test the hypothesis that consistent, long-term spatial biases among definitions cause differing agreement to planting observations. I isolate the long-term spatial pattern of

onset from the interannual variability by transforming onset values into within-year quantiles, and taking the average of the quantiles across the study period. These maps, shown in Figure 4.5, indicate systematic differences in the spatial pattern of onset, especially in the eastern region. The frequency definition tends to estimate relatively late onset values in the east, a pattern that does not exist for the AA definition. This systematic difference in spatial pattern may help explain why the frequency definition is more closely associated to observed planting data.

In addition to spatial differences in quantiles of onset, there is also a systematic long-term difference in the value of onset among the definitions. As shown in Figure 4.6, the AA definition estimates onsets that are later on the edges of MT and earlier in the center compared to the frequency- and volume-based definitions. The frequency and volume definitions produce values that are more similar to each other than to the AA definition, but still have systematic spatial differences. The frequency-based definition produces onset estimates that are later on the northeast quadrant of Mato Grosso, and earlier onsets on the western side of MT, compared to volume-based definition. The systematic spatial biases among the onset definitions means that different definitions might pick up different aspects of the spatial pattern in planting dates, shown in Figure 3.6.

Finally, I test the hypothesis that differences in temporal patterns among the definitions cause differing agreement with planting observations. I isolate temporal variability of onset from spatial variability by taking the mean and standard deviation of onset over space for each year. Figure 4.7 shows that interannual patterns in onset are similar for the frequency, volume and AA definitions.

The differences in onset coefficients therefore most likely arise from systematic spatial differences in onset produced by each definition. The long-term spatial pattern in planting dates matches better with frequency definitions, suggesting that the frequency definition is better able to capture the spatial variability in the most important features of precipitation.

4.3.3 Model evaluation metrics

Prior to reporting estimated coefficients, I confirm that the ten fitted models (one for each cropping intensity and percentile) satisfy the linear regression assumptions listed in the Methods section. Residual plots confirm that the residuals have zero mean, are uncorrelated with the fitted value, and are homoscedastic; the QQ plot confirms that the residuals are normally distributed (Figure 4.8); Durbin-Watson and Moran's I, reported in the Supporting Information for Chapter 4, show that the residuals are not temporally or spatially autocorrelated; and a correlation matrix shows that the predictors are not multicollinear and that residuals are exogeneous (Table 4.2).

The model cannot be directly used for accurate predictions of planting date under climate change. The prediction error is about 10 days, higher than the expected magnitude of planting date change: if onset is delayed by a month, planting date would only change by 3 - 5 days. Instead, I use the onset and year coefficients to discuss the sensitivity of planting date to a unit change in onset. In the following section, I use robustness tests to show that

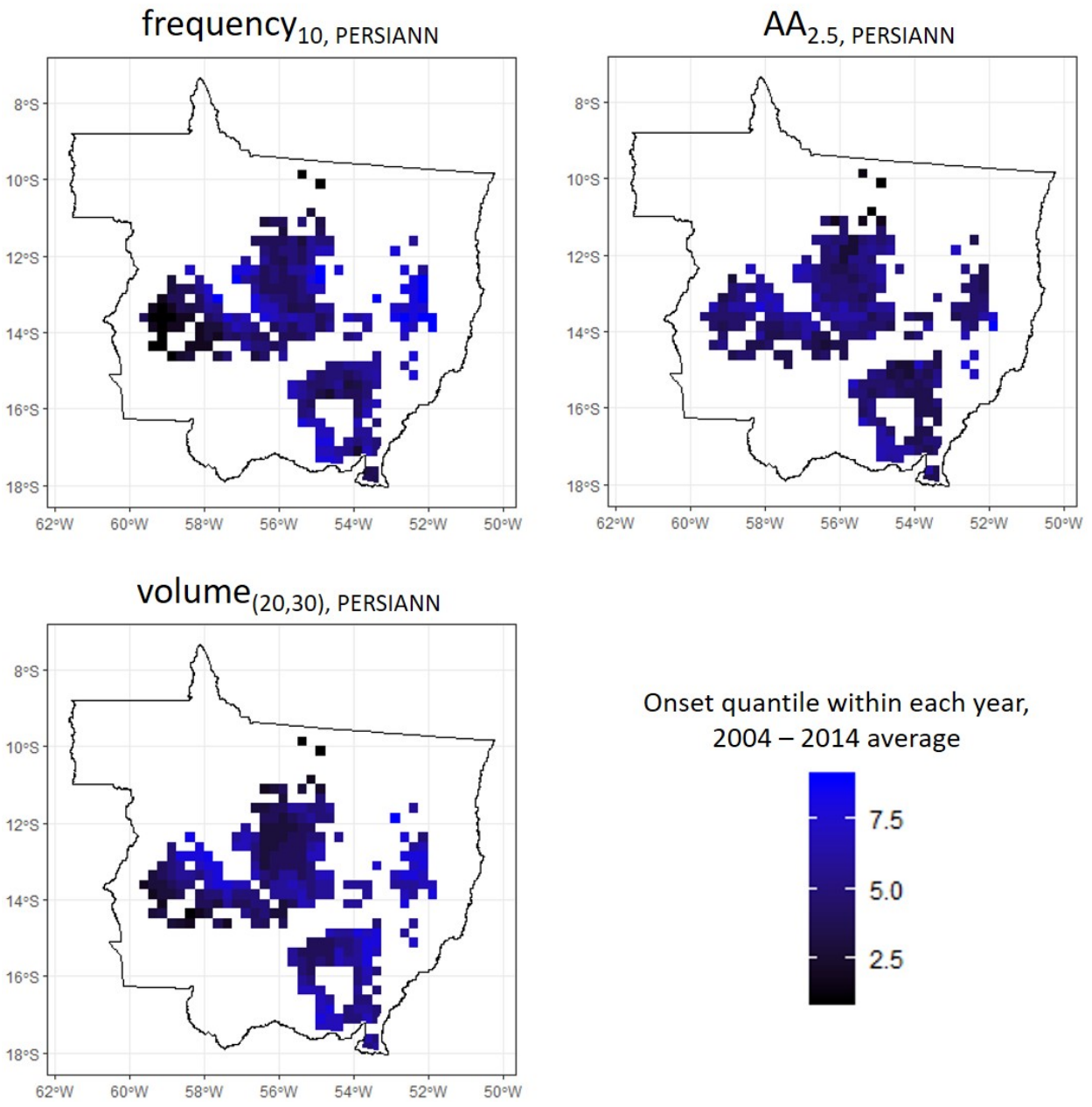


Figure 4.5: Quantiles of onset within each year were averaged from 2004 to 2014 to produce a map of spatial variability in onset.

the onset coefficient estimates, and the differences in the onset coefficient among cropping intensities and planting percentiles, are significant under sources of uncertainty.

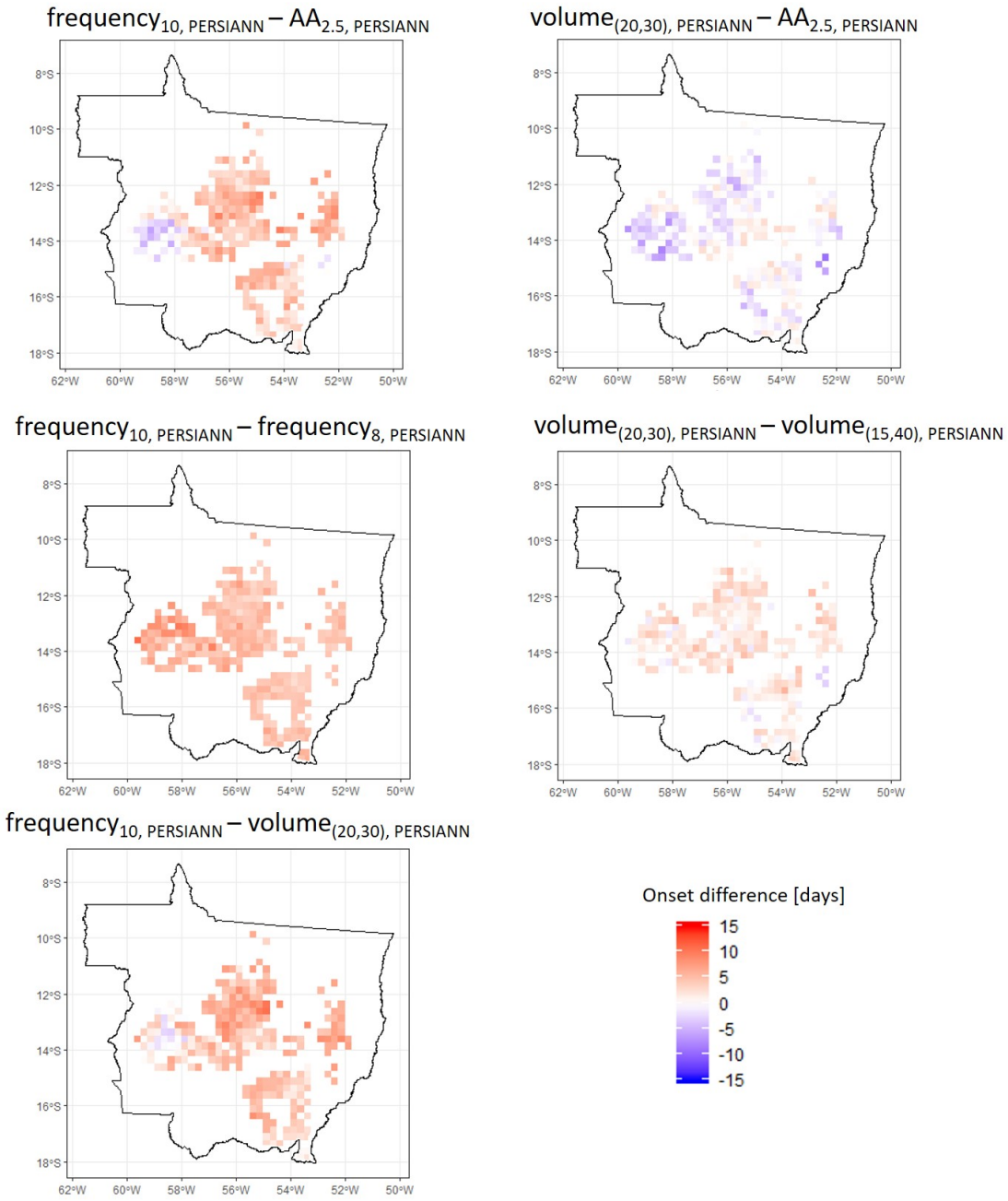


Figure 4.6: Long-term spatial patterns in the difference between onset estimates.

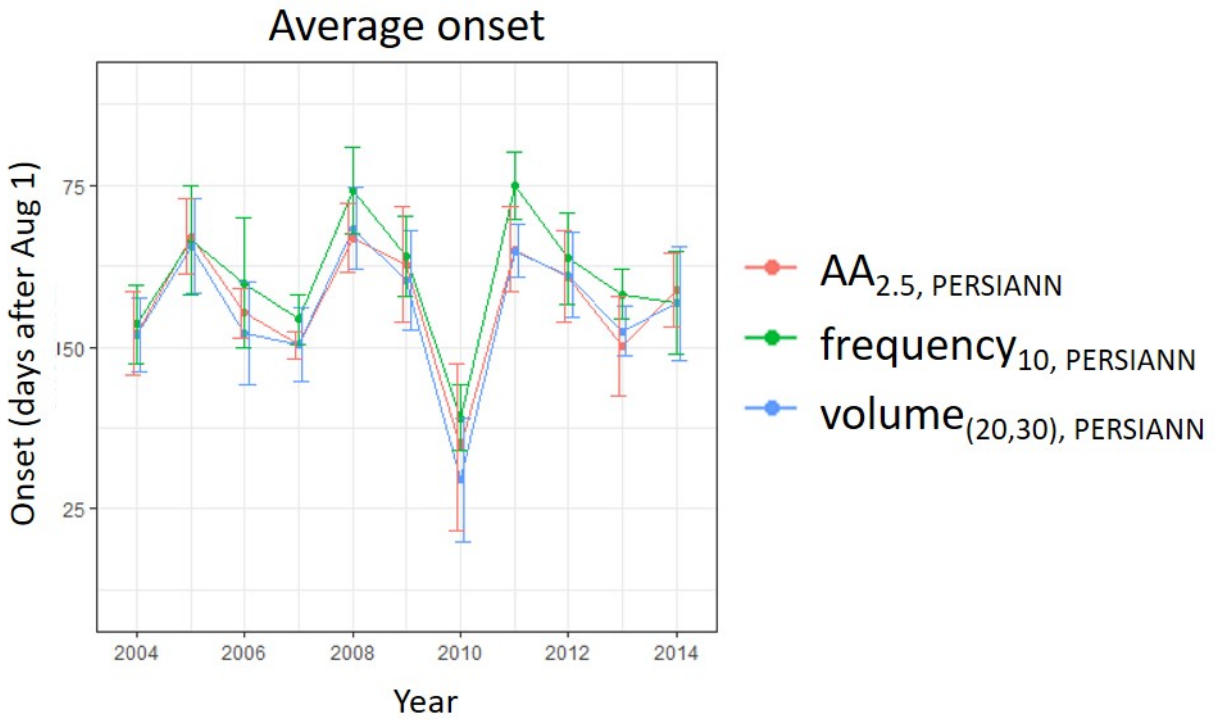


Figure 4.7: Temporal pattern, averaged over space, for each onset definition. Error bars represent standard deviation of onset within each year (i.e. the spatial variation).

	Onset	Year	Residual
Onset	1	-0.024	-0.014
Year	-0.024	1	4.9×10^{-4}
Residual	-0.014	4.9×10^{-4}	1

Table 4.2: Correlations show that predictors are not multicollinear and that residuals are exogenous. These are correlations for DC, percentile25, but correlations are similar for other intensities and percentiles.

4.3.4 Sensitivity and robustness tests

In this section, I confirm that the onset coefficients, and the differences in onset sensitivity among cropping intensities and planting percentiles, are robust to uncertainties caused by missing predictors and errors in planting date estimates. These robustness tests show that

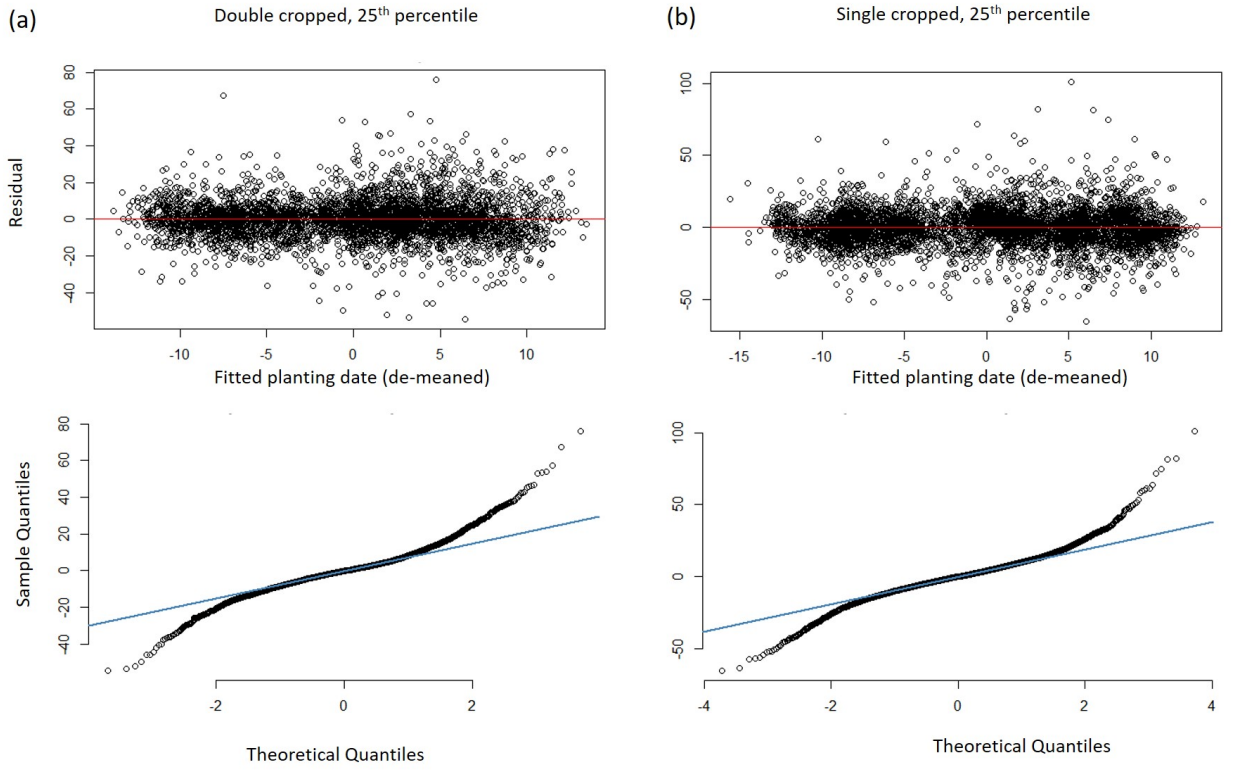


Figure 4.8: Residual and quantile-quantile plots confirm homoscedastic, nearly normal residuals for (a) double and (b) single cropped soy planted in the 25th percentile. Similar results are observed for other percentiles.

despite uncertainty, trends and differences in behavior among soy intensities and planting percentiles are still observable.

I test the robustness of the onset coefficient to missing predictors by eliminating known, important predictors such as year and location from the model and observing the change in the onset coefficient. I perform these tests in both the $\text{OLS}_{\text{pooled}}$ and OLS_{FE} specifications. Though the final model specification is OLS_{FE}, the $\text{OLS}_{\text{pooled}}$ specification enables me to eliminate location-related predictors; these are automatically included in the fixed effects terms at the heart of the OLS_{FE} specification. Table 4.3 shows that the onset coefficient is robust even when I have eliminated all predictors except onset. Even with all predictors but onset eliminated, the onset coefficient changes by a maximum of 0.04 compared to a model with the full set of known predictors. Similarly, Figure 4.9 shows that the OLS_{FE} specification is also robust when year is eliminated, and that both OLS_{FE} and $\text{OLS}_{\text{pooled}}$ are robust to eliminated location and year predictors across all intensities and percentiles. While the effect of missing predictors for which there is no data can never be quantified, the

stability of the onset coefficient when known predictors are missing is encouraging.

Additionally, as shown in Table 4.2, the residuals are uncorrelated to onset. This indicates that the onset coefficient is unbiased, even if the residuals could have been modeled with more predictors. Together, these tests indicate that the estimate of the onset coefficient is unbiased.

	SC onset coefficient	DC onset coefficient
OLS_{FE}, all predictors	0.25	0.38
OLS_{FE}, only onset	0.26	0.39
OLS_{pooled}, all predictors	0.26	0.37
OLS_{pooled}, only onset	0.29	0.40

Table 4.3: Onset coefficients estimated by OLS_{FE} and OLS_{pooled}. All predictors means that onset, year, latitude, longitude, and region were used. I report model results for the 25th percentile of planting here.

4.3.5 Planting date sensitivity to wet season onset

The onset coefficients for the chosen OLS_{FE} specification and onset definitions are summarized in Figure 4.10. The error bars represent bootstrapped uncertainties in planting date estimates. As expected, the onset coefficient changes with the cropping intensity and planting date percentile. The onset coefficient is higher for soy that is planted early (double cropped soy and soy in the 5th percentile) compared to soy that is planted later (single cropped soy and soy in the 95th percentile). For double cropped fields, the onset coefficient ranges from 0.5 at the 5th percentile to 0.25 at the 95th percentile; for single cropped fields, the onset coefficient ranges from 0.35 at the 5th percentile to 0.1 at the 95th percentile. These coefficients are statistically significant at the 5% level and follow naturally from the fact that growers who plant early more likely to be affected by changes in the onset of the wet season than those who plant late.

Additionally, my results indicate that planting date became earlier with each successive year, independently of the onset. The year coefficients are shown in Figure 4.10 and are statistically significant at all cropping intensities and planting percentiles, indicating that the trend to earlier planting dates affects all soy growers. However, the trend is stronger for single cropped fields than for double cropped fields, possibly because single cropped planting dates are generally later and have more flexibility to advance. The 95th percentile has consistently the smallest trend, representing the fields that are planted late. These quantitative results agree with the qualitative findings from Chapter 3.

Figure 4.10 also shows that despite uncertainty from errors in planting date data, differences in the onset coefficient for the various cropping intensities and percentiles are observable. To account for the effect of planting estimate error, I bootstrap 1,000 simulated

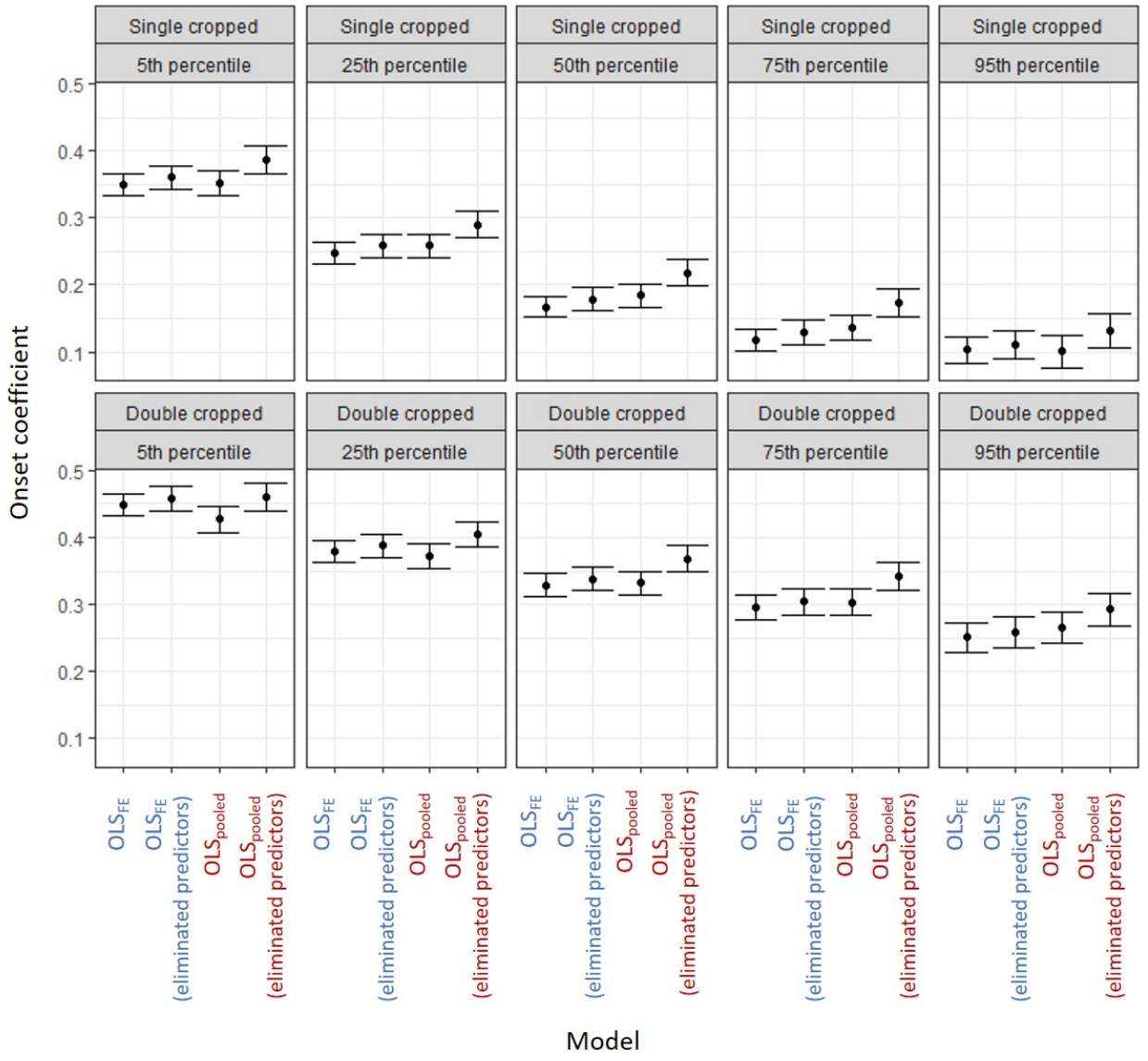


Figure 4.9: The onset coefficient is robust to eliminated predictors in both the OLS_{FE} and OLS_{pooled} specifications. Error bars represent standard error.

datasets of the same size as the original dataset. The bootstrapping method replaces each planting observation in the dataset assuming a normal distribution with mean equal to the observed value and error of 6.9 days (calculated in Chapter 3). Each simulated dataset is then fit to the OLS_{FE} model to create 1,000 bootstrapped onset coefficients for each percentile $\times$ intensity. Next, I use unpaired, two-sided t-tests to confirm their statistical significance. For

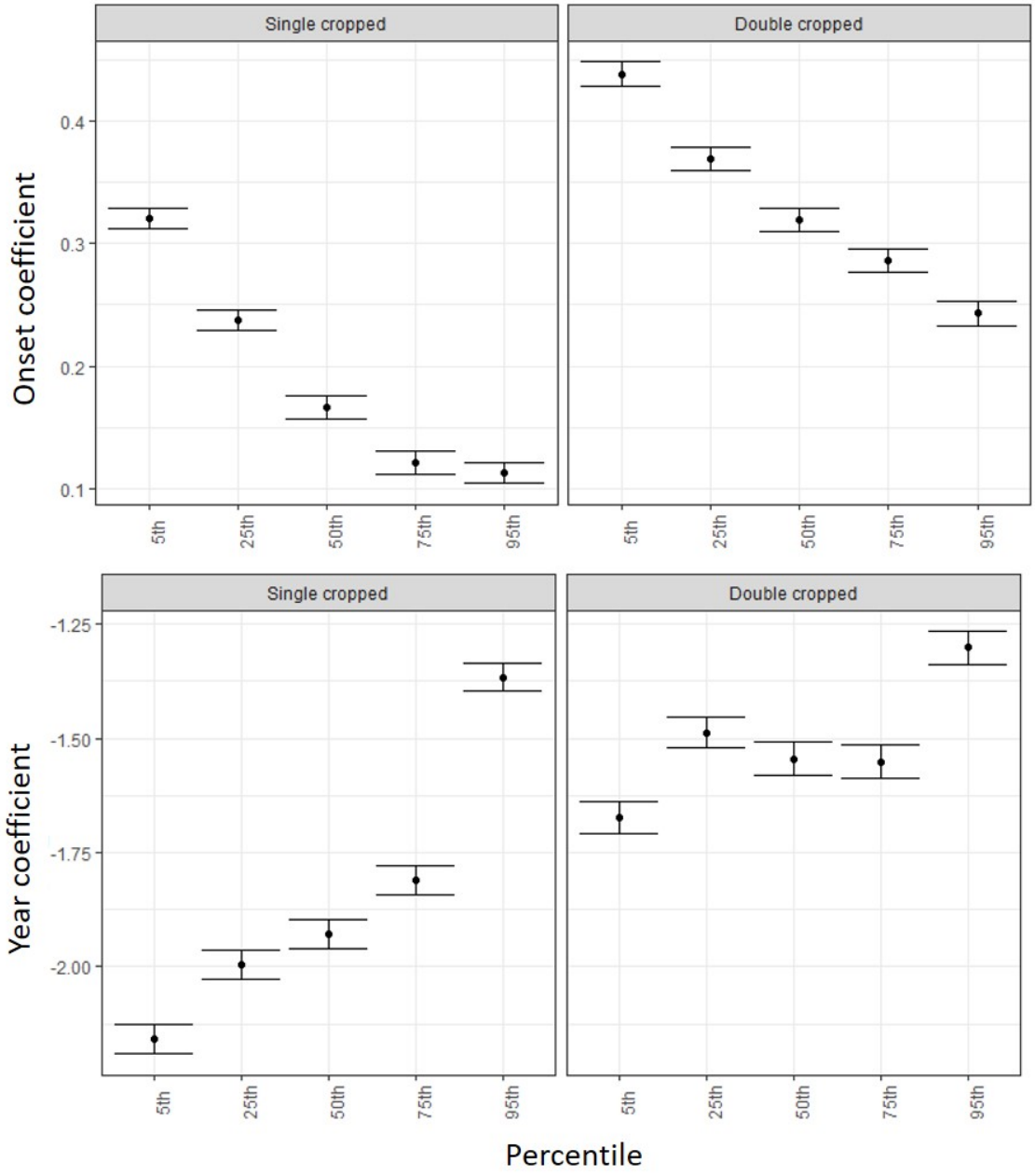


Figure 4.10: Onset coefficients appear statistically different among cropping intensities and percentiles, despite uncertainty. Error bars represent the standard deviation of 1,000 bootstrapped coefficients, reflecting planting date estimation error.

each percentile, I perform a two-sided t-test for the onset coefficient of single versus double cropped soy; and for each cropping intensity, I perform two-sided t-test for the onset coeffi-

cient of adjacent percentiles (5th vs 25th; 25th vs 50th, etc). The p-values of all t-tests are below the threshold of 10^{-15}, indicating that the different cropping intensities and planting percentiles do have statistically different sensitivities to onset. Thus, the signals of interest exceed the noise, allowing insight into planting date sensitivity to onset despite uncertainty.

4.4 Discussion

4.4.1 Research Question 1: Planting dates' sensitivity to onset

While model results shed light on the heterogeneous and nonstationary nature of planting behavior in Mato Grosso, the model itself cannot describe the full range of planting behavior. The OLS_{pooled} models only explain 16% - 22% of total planting date variability; OLS_{FE} models explain 42 - 57%, an increase that may be attributed to the capture of time-invariant factors such as the transportation network and soil type [36]. The remaining unexplained variability may come from spatiotemporally varying factors such as access to agricultural credit and perception of agricultural risk, for which there is not yet data [1, 23, 44, 75].

While much of the variability in planting dates is not captured in the models, the stability of the onset coefficient to the elimination of known predictors is encouraging, and indicates that the coefficients represent an unbiased estimate of planting date's sensitivity to onset. The low explanatory power does not necessarily result from a bad model; rather, it may result from noisy ground level behavior (for which data are not available) and random error in the planting date estimates. While the possibility of missing predictors prohibits the use of these models for prediction of absolute planting dates, the coefficients that are fitted provide new insights into planting decisions.

For example, results show that planting date, and its sensitivity to onset, has rich local variability: within individual 25 km cells, a field planted in the 5th percentile will be more sensitive to onset than a field planted in the 95th percentile; a double cropped field will be more sensitive than a single cropped field. These spatial nuances are not captured under the assumption that planting date occurs at wet season onset, and are commonly absent in global aggregated datasets for planting dates and in efforts to predict crop yield under climate change. This aggregation is problematic: because adaptation strategies are implemented by individuals, quantifying the diversity of responses to onset is crucial to accurately predicting agricultural adaptation and yields under changing precipitation regimes.

Finally, the presence of a trend towards earlier planting dates, independent of onset, indicates technological or logistical progress enabling growers to plant closer to the wet season onset in at the end of the study period, 2014, than at the beginning, 2004. If this trend continues, planting dates may become increasingly sensitive to wet season onset. This trend also means that planting date datasets based on older survey information, such as MIRCA2000, may be outdated in regions with developing agricultural practice.

4.4.2 Research Question 2: Best onset definition

While farmers are sensitive to wet season onset, they may refer to different features of precipitation than those used by climatological definitions. The worst-performing onset definition is the depth definition, with an onset coefficient of only 0.25 for the most sensitive group, double cropped soy at the 5th percentile. It's possible that this is least favored by farmers as an indication of onset because one large rainfall event contains less information about the arrival of the wet season compared to the arrival of many rainfall events over a longer period of time. Additionally, its reliance on comparing the rainfall on a single day to a predetermined threshold makes the onset date (and therefore onset coefficient) highly sensitive to small variations in the threshold value and in the precipitation data.

The climatological definition of onset based on anomalous accumulation (AA) has average performance, with onset coefficients ranging from 0.1 to 0.4 for SC-95th percentile and DC-5th percentile, respectively. AA's moderate performance indicates that other features of precipitation are more relevant in this context. The AA definition requires farmers to predict future precipitation during the agricultural year, making it impossible to observe in real time. In contrast, definitions based on the frequency of rainy days or the total volume of rain in a month are easy to observe and calculate.

While the pentad and monsoon definitions are also based on easily identifiable cumulative volume metrics, they have lower performance than the frequency and volume definitions. This could be because they observe rainfall over shorter periods of time. Rainfall signals over 11 days (monsoon definition) or 15 days (pentad definition) may not carry as much information for risk-averse farmers as definitions based on information from 28 days (frequency definition) or 25 days (volume definition).

Farmers' perception of the start of the wet season is therefore best captured by the frequency definition, one of the more easily observable precipitation-based measures of wet season onset. The frequency definition in Brazil is also used for cropping intensity decisions. A study in the Cerrado found that cropping intensity was more closely related to a frequency definition than to the AA definition [129]. The importance of the frequency definition for both planting behavior and cropping intensity should encourage wet season projections to favor easily observable definitions over climatological ones.

However, precipitation may not be the only indicator of onset, opening up possibilities for future investigation. In Niger and Kenya, farmers are likely to use soil moisture at a certain depth [94, 97] as indication that the wet season has begun. In some cases, decisions are not based on water availability at all: farmers in southeastern Kenya use indicators such as budding of trees, animal behavior, and wind direction to detect the rainy season [115]. These onset definitions are difficult to observe remotely, but could be approximated by merging temperature and soil moisture datasets with precipitation signals [145]. It may be worthwhile to examine these non-precipitation climate variables in defining the wet season.

4.5 Conclusions

The results reveal two important insights: (1) considerable heterogeneity exists in planting date's relationship to onset, and (2) the way in which onset is defined causes a significant change in planting dates' sensitivity to onset.

The sensitivity of planting dates to wet season onset is highly variable across cropping intensities and fields, suggesting that different growers will respond to onset depending on their unique constraints. In Mato Grosso, onset sensitivity changes between cropping intensities (double cropped soy is more sensitive than single cropped soy) and across fields (early-planted fields are more sensitive than late-planted fields). Further, planting dates can shift over time, independently of climate, suggesting technological advances in crop variety or the expansion of the transportation network. This means that global planting datasets that rely on climate-based assumptions and old survey data may be inappropriate. Similarly, areas with developing agriculture or areas affected by a changing climatic, economic, or logistical context may also be badly estimated by existing global datasets. The heterogeneous and changing planting behavior over Mato Grosso, if ignored, could generate grossly incorrect agricultural yield projections.

The definition of the wet season onset also impacts onset sensitivity: climatological definitions, such as those based on deviations from annual mean precipitation, are less correlated to planting behavior than definitions based on easily observable features of precipitation, such as the frequency of rainfall events. Agricultural studies that use climatological definitions of onset should acknowledge that their definition may not be the most reflective of farmers' behavior.

These findings reveal that planting dates have a more complex relationship to wet season onset than previously believed, signifying the need for a closer look at common global planting date maps and yield prediction efforts. Though Mato Grosso is the focus of this study, this work provides insight into planting behavior in agricultural systems worldwide. For example, the assumption that planting date uniformly occurs at climatological onset is likely untrue elsewhere. A better understanding of farmers' response to precipitation signals will impact predictions of how planting dates, and therefore agricultural productivity, will respond to climate change.

Chapter 5

Predicting planting dates under climate change scenarios

5.1 Introduction

Climate change has contributed to a historical shortening of the wet season in Brazil, and projections suggest this trend will continue. In the state of Rondonia in southwestern Amazonia, deforestation since the 1970s has delayed wet season onset by an average of 11 days in the past three decades [24]. Similarly, a study of onset over South America found an average onset (dry season end) delay of 11.5 +/- 2 days per decade, and attributed this delay primarily to global biogeochemical climate change [48]. Climate simulations also predict that deforestation in the Cerrado region of Brazil will cause a one-month reduction in the length of the wet season (from 6 months to 5 months), contributed by both delayed onset and earlier demise [31]. These observations are concerning for rainfed agriculture in Mato Grosso, whose significant area of double cropped soy relies on a long rainy seasons to support two sequential crops.

The choice of planting date is both a major adaptation strategy under climate change and vulnerable to its effects. While planting later helps crops avoid drought when the onset is delayed, it may reduce the likelihood of a successful second crop. The feasibility of double cropping would be further diminished under an accelerated wet season demise. Understanding the degree to which planting dates will adapt to changing wet season onset, and what delayed planting will mean for crop yields and intensive cropping practices, is crucial for predicting agricultural productivity under climate change.

While many studies have highlighted planting dates' potential as an adaptation strategy, there is still a need for a realistic understanding of planting behavior and its future trajectory. Studies that simulate optimal, yield-maximizing planting dates under climate change scenarios may recommend unrealistic dates, or dates that prevent intensive cropping systems. In Cameroon, optimized planting dates were three months later than traditionally observed planting dates [84]; in Sudan, the recommended planting date was two to four weeks earlier

than actual practice [23]. The large gap between actual and optimal dates suggests that the optimized scenario may not be implemented without systematic (and possibly unlikely) changes in agricultural practice. Additionally, it is implausible that all farmers, in their various socio-economic contexts, will adapt to an equal degree.

As an alternative to potentially unrealistic yield-optimizing planting dates, historical models of planting behavior can form the basis of yield predictions. Regression models from Chapter 4 have shown that the wet season onset correlates to planting dates, but this association (1) does not perfectly follow onset (i.e. the planting date changes by less than 1 day for every 1 day delay in onset), and (2) changes between cropping intensities (double cropped soy is more sensitive than single cropped soy) and across fields (early-planted fields are more sensitive than late-planted fields). The heterogeneity in onset sensitivity indicates that while all farmers will adjust planting behavior under climate change, this adaptation will vary spatially. Additionally, these models demonstrate that planting dates have trended earlier each year, independently of onset, suggesting technological factors may be influencing planting dates. The uneven impact of changing wet season timing and the trend towards earlier planting should be embedded in predictions of agricultural yield.

In this chapter, I predict planting behavior based on the regression model developed in Chapter 4. These predictions reflect a realistic, short-term (10 year) change in planting behavior in Mato Grosso as a response to expected perturbations in wet season timing. Given that both delayed wet season onset and earlier wet season demise are expected over Mato Grosso, I answer the following questions about planting and cropping behavior:

1. How will delayed wet season onset impact planting dates in Mato Grosso?
2. How will earlier wet season demise impact feasibility of double cropping in Mato Grosso?

5.2 Methods

5.2.1 Projected wet season

Projected wet season onset and demise dates were simulated by Costa et al (2019). First, daily precipitation timeseries were simulated using the HadGEM2-ES CMIP5 model; wet season metrics (onset and demise) were then extracted using the anomalous accumulation method with threshold of 2.5 mm. Precipitation timeseries were generated based on historical climate for the period 1970 - 2003, and based on the RCP8.5 scenario for the period 2006 - 2049. Figure 5.1 depicts simulated wet season metrics for northwest and northeast Mato Grosso.

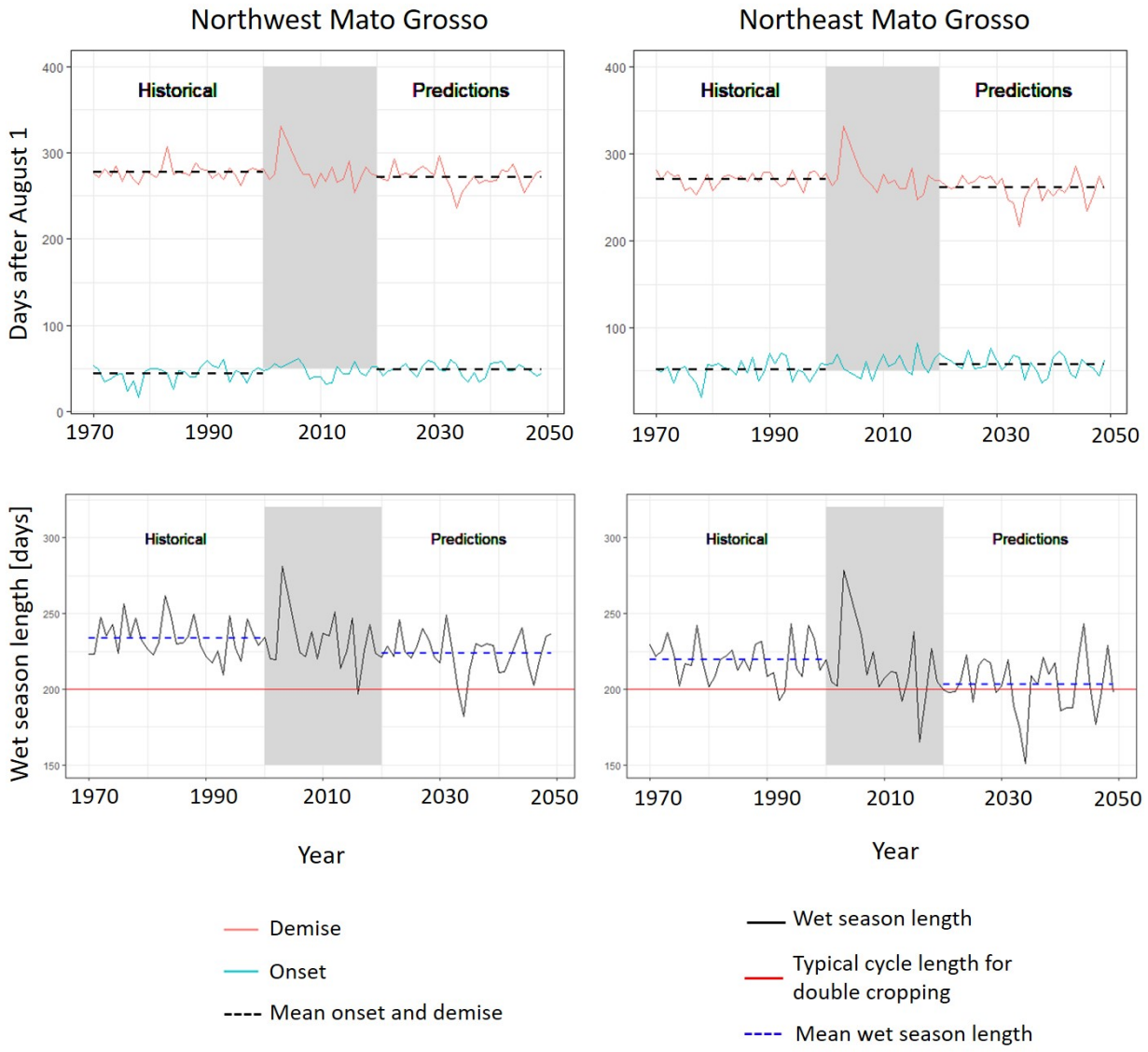


Figure 5.1: Costa et al (2019)'s historical and predicted wet season onset, demise, and length. Historical and predicted years are separated with a 20-year gap, 2000 - 2020. A red line is drawn at 200 days, the expected crop cycle length for double cropped soy.

5.2.2 Prediction scenarios

The magnitude of delayed onset and earlier demise in Mato Grosso are projected to vary interannually and spatially, with northeastern Mato Grosso more vulnerable to shorter wet seasons than northwestern Mato Grosso [32]. The simulated onset demise dates are therefore

aggregated separately to northwestern and northeastern Mato Grosso. Northwestern and northeastern Mato Grosso are defined as the area north of 15°S and west or east of 54°W, respectively, a line where precipitation patterns experience an abrupt shift. No simulations were made for areas south of 15°S [32]. Within each region, I account for interannual variability by making predictions under the prototypical scenarios of early, medium, and late onset/demise (corresponding to the 10th, 50th, and 90th percentiles).

Due to a likely incomplete set of predictors, the model provides a *relative* change in planting date in response to a change in onset (i.e. the number of days planting is delayed for every one day delay in onset) rather than an absolute planting date in response to an *absolute* onset date. To this end, for each of the two regions, I calculate changes in wet season onset demise using the average simulated values during 1970 - 2000 as a baseline. Future wet season metrics for the prototypical scenarios are calculated by taking the 10th, 50th, and 90th percentiles of simulated wet season onset and demise during 2020 - 2049. The differences between these percentiles and the historical baseline become the onset delay or demise acceleration for prototypical early, medium, and late onset/demise years. I exclude simulations from the transition period of 2001 - 2019 to ensure a clear switch from historical to future regimes.

Though predictions of wet season metrics are available until 2049, I only make predictions for ten years (2024) following the modeled period (2004 - 2014). I limit predictions to 10 years in the future because non-stationarities may exist in the onset sensitivity, in the nature of the unobservable variables characterized by fixed effects, and in the interannual trend towards earlier planting (as observed in Chapter 3). The statistical model is unable to capture these long-term changes in behavior and may bias predictions if applied to the far future. Assuming that the distribution of wet season metrics remains stationary over the period 2020 - 2049, I use simulated onset and demise during these 25 years to define a set of plausible wet season scenarios in 2024: this allows me to account for the high interannual variability in wet season timing expected in Mato Grosso. Figure 5.1 suggests that the stationarity assumption is valid.

Predictions are made under three wet season cases:

1. Early (10th percentile) onset and late (90th percentile) demise. This represents the best case scenario, in which the wet season length is longest.

2. Median (50th percentile) onset and median demise. This represents a moderate wet season scenario.

3. Late onset and early demise. This represents the worst case scenario, in which the wet season length is shortest.

These wet season cases are chosen to represent a moderate scenario and best- and worst-case bounds. The three scenarios are equally likely under the assumption that onset and demise are uncorrelated, in which a late onset is not more likely to be offset by a late demise or exacerbated by early demise. Figure 5.2 shows that the independence assumption

is warranted: there is no statistically significant correlation between onset and demise in the 1970 - 2049 study period. The possibility of random combinations of onset and demise dates is confirmed in Figure 5.1, which displays a 125-day range for the wet season length, contributed by independent variations in onset and demise.

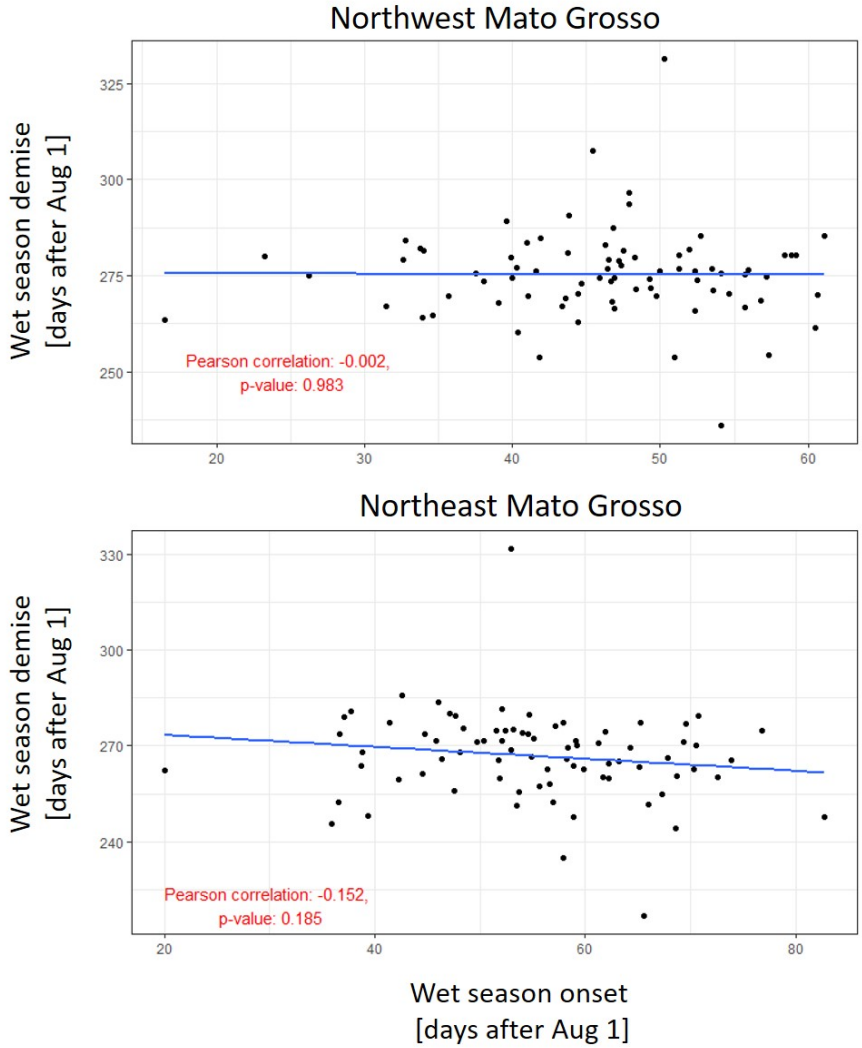


Figure 5.2: Correlation between simulated wet season onset and demise.

Table 5.1 shows the 10th, 50th, and 90th percentiles of onset delay and demise acceleration expected in 2020 - 2049 relative to the historical period of 1970 - 2000 [32]. These percentiles account for the large interannual variability in wet season timing: the vulnerable northeast region of Mato Grosso may experience an onset up to 21 days later and demise up to 26 days earlier [32].

Region	Percentile	Onset change [days]	Demise change [days]
Northeast	10	-9.16	-26.42
Northeast	50	6.99	-7.70
Northeast	90	20.92	4.06
Northwest	10	-4.48	-22.83
Northwest	50	3.91	-3.69
Northwest	90	13.64	9.93

Table 5.1: Projected changes in wet season timing in years 2020 - 2049, subtracted by the average onset or demise from 1970 - 2000 [32].

5.2.3 Model for planting predictions

For each of the three wet season cases, I predict the cumulative distribution function (CDF) of 2024 planting dates, using 2014 observations as the baseline. The 2014 CDFs represent planting dates observed in an average 25 km cell within either northeast or northwest Mato Grosso, and the future planting CDF is predicted using the percentile- and cropping intensity-specific coefficients found in Chapter 4. Though I selected a frequency-based definition of onset as the most relevant for Mato Grosso soy in Chapter 4, predictions of wet season metrics for Mato Grosso are only available under the climatological anomalous accumulation definition. To maintain consistency with available climate predictions, I use onset and year coefficients that were fitted for the $AA_{2.5,\ \text{PERSIANN}}$ onset definition. The coefficients used for prediction are shown in Table 5.2.

5.2.4 Predicted planting metrics

While predictions can account for both the onset and year trend effects, I additionally report planting dates under only the onset effect. This eliminates a major uncertainty in predictions, as it is unclear whether the trend toward earlier planting dates will continue to 2024. It also isolates the impact of delayed onset from other, potentially opposing, directions. Together, predictions under onset alone and under onset+year effects are used to calculate five metrics:

1. The average planting delay in a 25 km region due to delayed onset. The average planting delay is calculated as the area between the cumulative distribution functions (CDFs) of planting dates observed in 2014 and planting dates predicted under the future onset scenario. This delay only occurs if the trend toward earlier planting does not continue; otherwise, the trend towards earlier planting overpowers the onset-controlled delay in planting. In this case, delay is reported as a negative number.

Cropping intensity	Planting percentile	Onset coefficient	Year coefficient
SC	5	0.27 +/- 0.02	-2.13 +/- 0.06
SC	25	0.19 +/- 0.02	-1.98 +/- 0.06
SC	50	0.12 +/- 0.02	-1.92 +/- 0.06
SC	75	0.08 +/- 0.02	-1.80 +/- 0.06
SC	95	0.073 +/- 0.02	-1.36 +/- 0.07
DC	5	0.39 +/- 0.02	-1.64 +/- 0.06
DC	25	0.33 +/- 0.02	-1.46 +/- 0.06
DC	50	0.29 +/- 0.02	-1.51 +/- 0.06
DC	75	0.26 +/- 0.02	-1.52 +/- 0.07
DC	95	0.22 +/- 0.02	-1.29 +/- 0.08

Table 5.2: Onset and year coefficients estimated by OLS_{FE}, using $AA_{2.5,\,PERSIANN}$ onset definition. Error bars are bootstrapped standard deviations representing planting estimation error.

2. The percent of soy fields whose "preferred" planting date will no longer be feasible due to delayed onset, assuming the interannual trend in planting continues and pushes the "preferred" planting date earlier each year. This is calculated as the percent of the predicted planting dates that fall before the predicted onset. Graphically, it is the point at which the predicted planting CDF intersects with the predicted onset.

3. The percent of historically double cropped soy that would no longer continue double cropping, due to a combination of delayed onset and earlier demise. This represents the percent of soy that is planted too late to allow the second crop to mature before the end of the wet season. The "last day to harvest" for double cropped fields is 30 days after the wet season demise, representing 20 days of soil moisture use and 10 days of grain drying. The "last day to plant" is, in turn, 200 days before the "last day to harvest" (90 days for the first crop and 110 days for the second crop) [1]. Graphically, it is the point at which the predicted "last day to plant" intersects with the predicted planting CDF.

4. The critical threshold of onset change that would cause the earliest (5th percentile) of predicted planting dates to experience wet season onset as a hard limit. This represents the leeway available for onset delays before historical sensitivities to onset must change for early-planted fields. Graphically, it is the change in onset at which onset intersects with the 5th percentile point of the predicted planting CDF. I also report the likelihood that this critical threshold of onset change will be reached, based on climate projections for 2020 - 2049.

5. The available planting window, in days. This is delimited by the "too early to plant"

date set by the onset, and the "too late to plant" date set by the demise and cropping intensity. If the "too late to plant" date occurs after the "too early to plant" date, the available planting window is zero.

These metrics are calculated under each wet season case and region (northwest and northeast Mato Grosso).

5.3 Results

5.3.1 Planting predictions under bounding scenarios

The planting responses to the wet season scenarios listed in Table 5.1 are predicted for each percentile and cropping intensity using Table 5.2. The predicted planting CDFs are shown in Figures 5.3, 5.4, and 5.5.

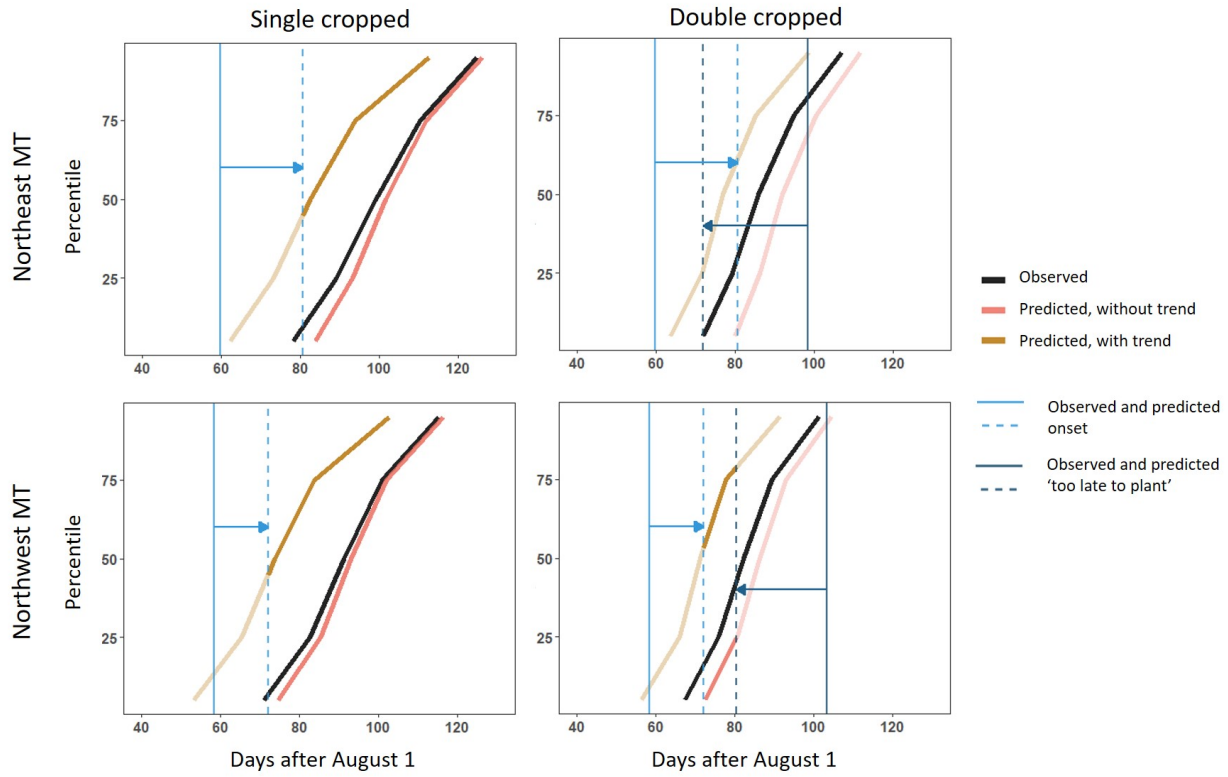


Figure 5.3: Observed and predicted CDF for planting dates within a 25 km cell that experiences the worst case scenario of late onset and early demise (worst case scenario). The "too late to plant" dates for single cropped soy are 182 and 190 days after August 1 for northeast and northwest Mato Grosso, respectively, and do not appear on the plots.

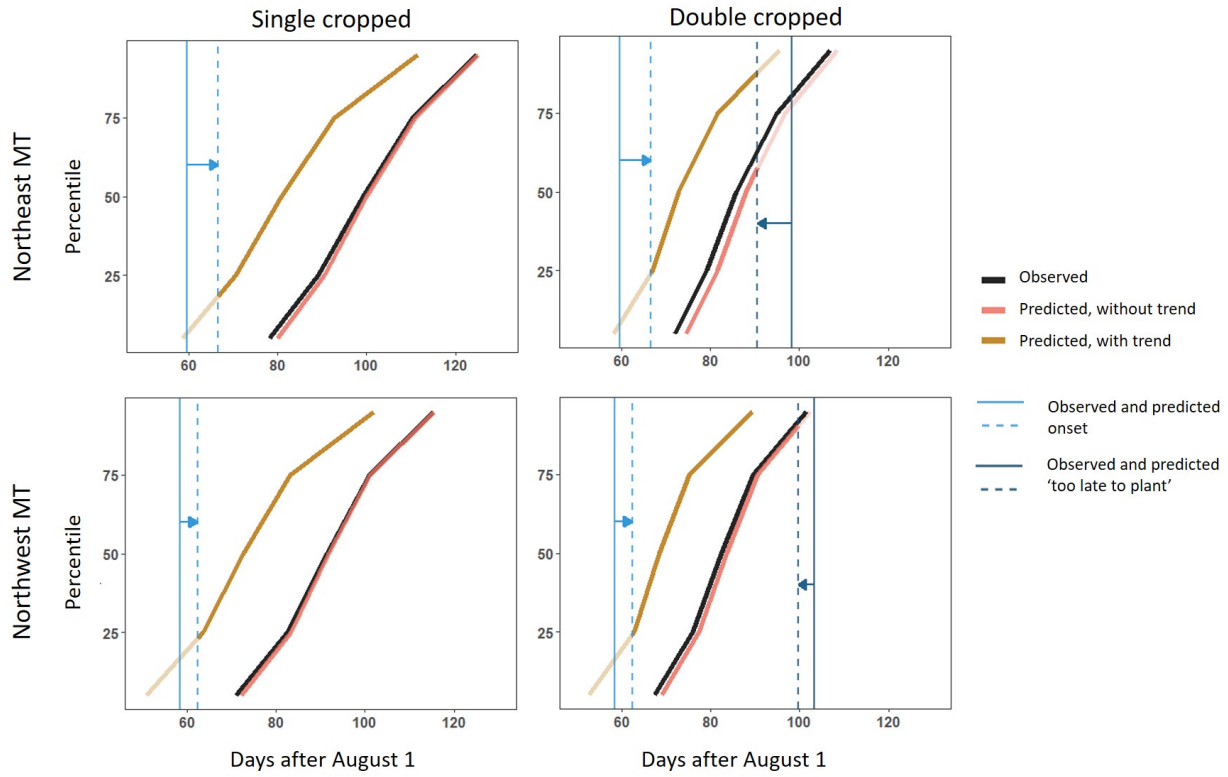


Figure 5.4: Observed and predicted CDF for planting dates within a 25 km cell that experiences medium onset and medium demise (moderate scenario). The "too late to plant" dates for single cropped soy are 200 and 210 days after August 1 for northeast and northwest Mato Grosso, respectively, and do not appear on the plots.

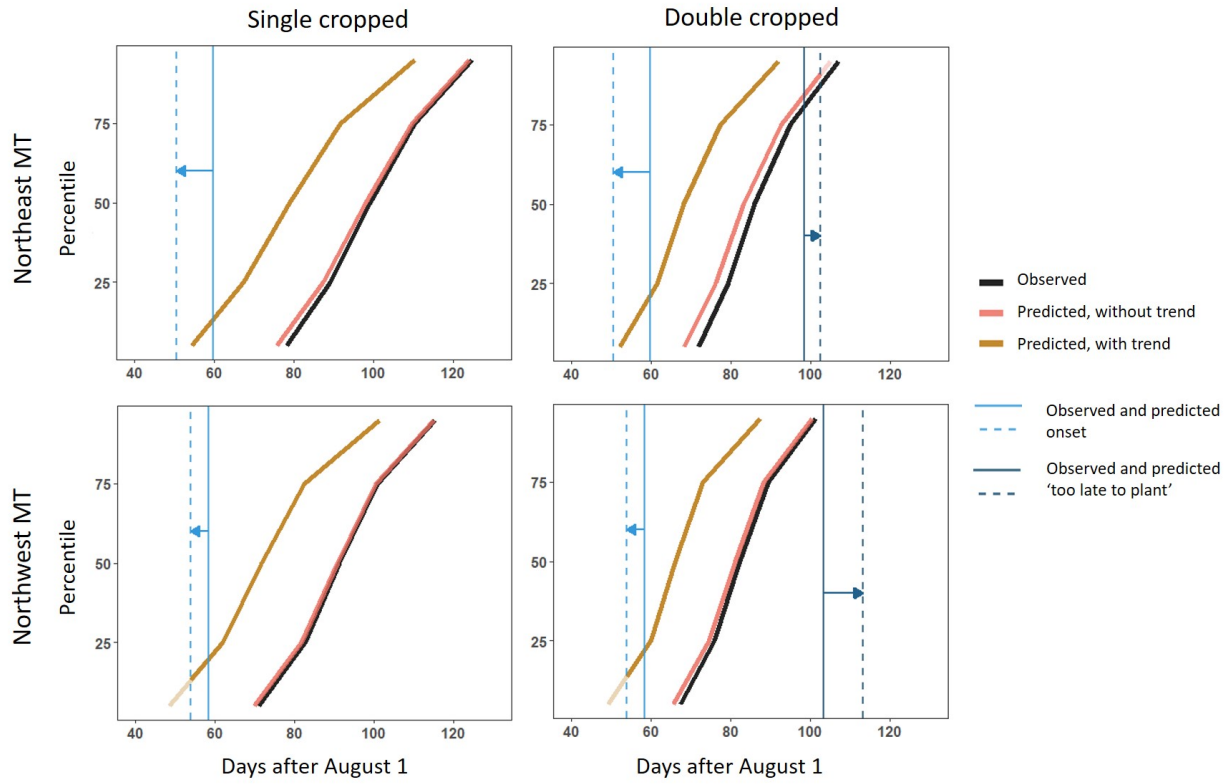


Figure 5.5: Observed and predicted CDF for planting dates within a 25 km cell that experiences early onset and late demise (best case scenario). The "too late to plant" dates for single cropped soy are 212 and 223 days after August 1 for northeast and northwest Mato Grosso, respectively, and do not appear on the plots.

In these figures, the observed CDF represents the mean value of the 25 km cell-scale planting dates observed in 2014. For example, the observed 5th percentile for the northeast region is the mean of the 5th percentile in planting dates observed in each individual 25 km cell located in northeastern Mato Grosso in 2014. The predicted CDFs represent expected planting dates in 2024 under the influence of onset or onset+year trend, *assuming that the onset and demise dates do not become the primary constraint on planting dates and cropping intensity*. However, it is expected that delayed onset and earlier demise *will* eventually become the dominant constraint for planting and cropping intensity for certain areas, impacting the shape and extent of future planting CDFs. While I am unable to predict the exact shape of these future CDFs, I account for these hard constraints by calculating the extent to which planting dates must change.

5.3.2 Predicted planting metrics

The true planting CDF will be the combination of (1) shifts of the planting CDF due to changes in onset and in response to interannual trends, and (2) hard limits on planting date and cropping intensity placed by delayed onset and earlier demise. Five metrics, derived from the predicted CDFs, serve as indicators of the many ways in which wet season timing can impact planting behavior. They are reported in Figures 5.6, 5.7, 5.8, 5.9, and 5.10, respectively.

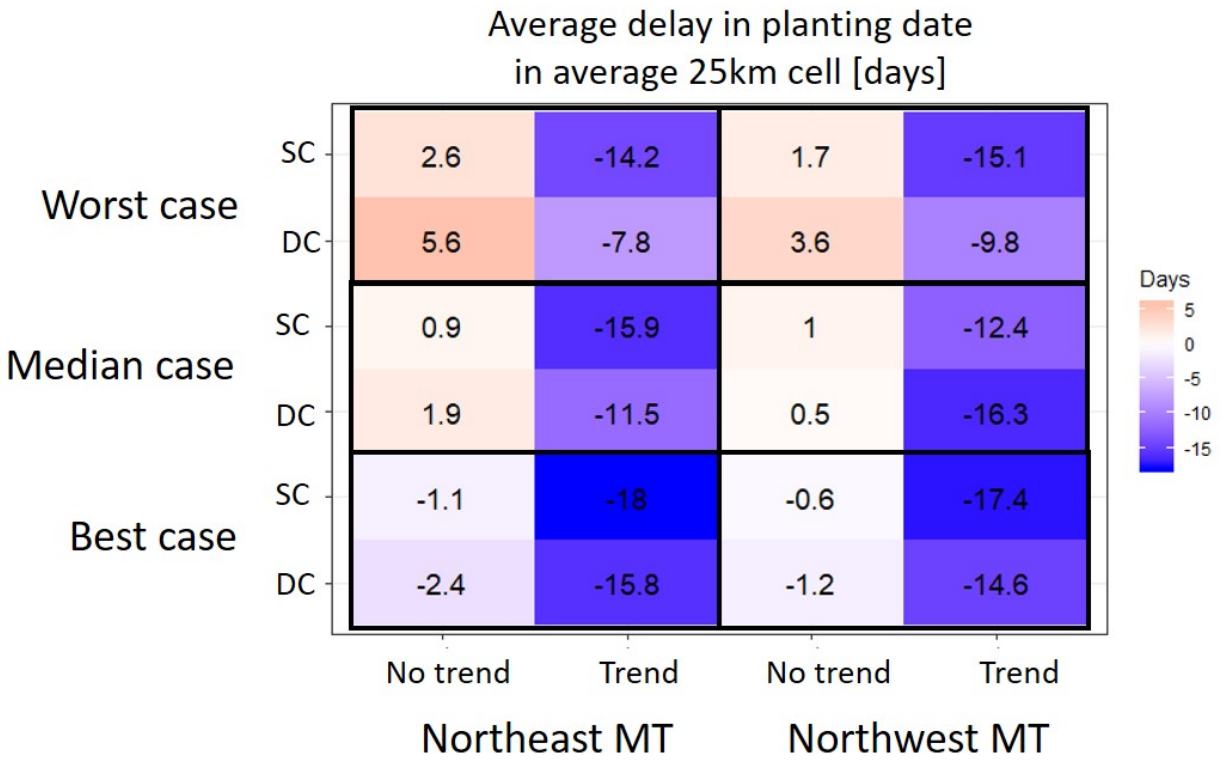


Figure 5.6: Projected changes in average planting date within a 25 km cell. Negative values indicate future planting dates that are earlier than 2014 values.

First, the area between the observed planting date CDF and predicted CDF represents the average change in planting date experienced by a 25 km cell, tabulated across all percentiles (Figure 5.6). This is done because the magnitude of delay will vary by field: farmers who plant in the 5th percentile will be more heavily affected than those who plant in the 95th percentile. These variations were summarized as an average delay in planting across all percentiles. It is important to note that these planting date changes are calculated solely

based on estimated onset and year coefficients, and are not bounded by the "too early to plant" and "too late to plant" limitations. Much of the advancement in planting dates predicted under the trend will be impossible due to delayed onset, and some of the delayed planting dates predicted without the trend will be impossible due to earlier demise. The average changes in planting dates, therefore, serve as bounding values for the expected change in behavior. If the trend does not continue, planting delays of 3.6 and 5.6 days are expected for double cropped soy in northwest and northeast Mato Grosso, respectively, under the worst case wet season scenario. This delay may impact agricultural yields: wheat in northern India experienced a 1% decrease for every day of delayed planting, caused by heat stress during the grain filling period of crop growth [102]. A similar decline may occur in Mato Grosso. However, under the worst wet season scenario, many of these delayed planting dates will not produce a successful second crop (in other words, they would be after the "too late to plant" limit). Thus the average change in planting, while indicative of the degree to which planting dates will shift, is insufficient; the "too early to plant" and "too late to plant" limitations should be included.

To this end, I calculate the second and third metrics: the percent soy area whose "preferred" planting dates will not be feasible under the projected onset delay and the percent double cropped area that not be feasible under the projected onset and demise scenarios (Figures 5.7 and 5.8).

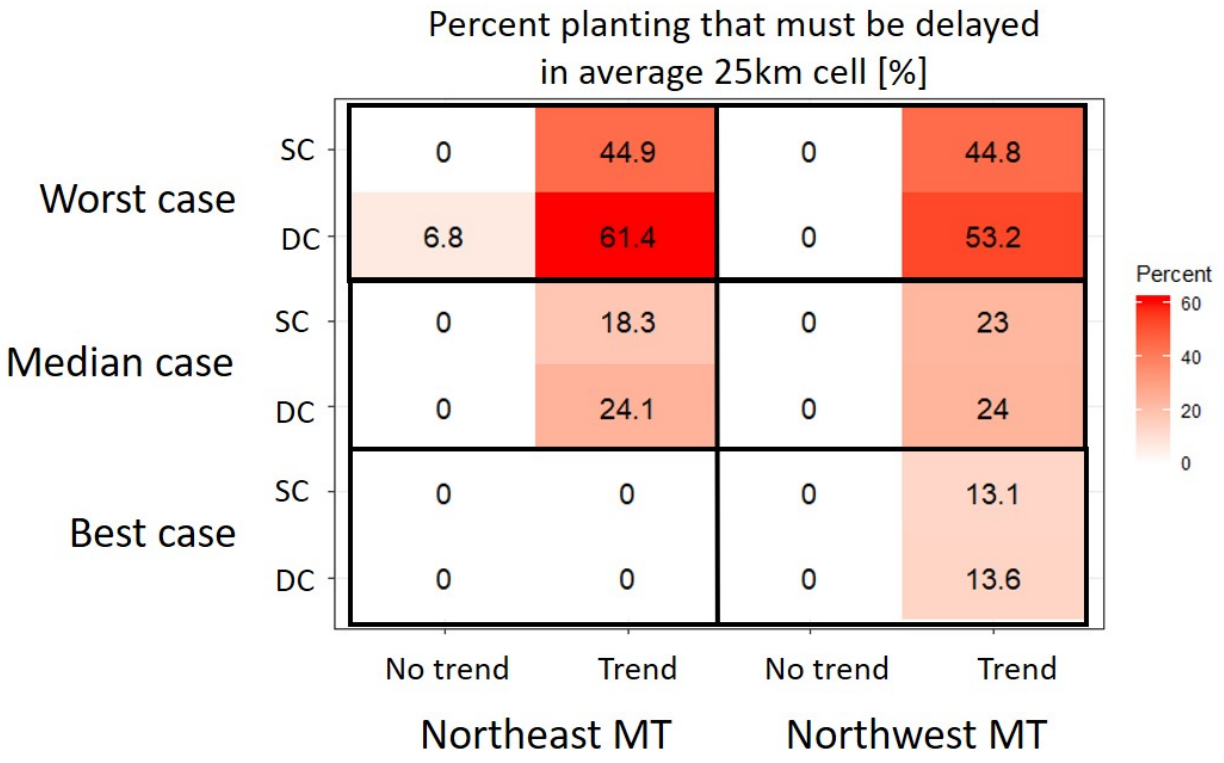


Figure 5.7: Projected percent of soy area whose planting dates will be affected by onset delay.

The percent of soy that will be affected by delayed onset depends on whether the current trend of earlier planting dates continues. Technological progress may continue to push planting dates earlier and earlier until onset becomes a hard limit. Under the worst case onset delay, the trend toward earlier planting dates cannot continue for 61% of double cropped soy and 45% percent of single cropped soy in the vulnerable eastern part of Mato Grosso. A median onset year will force 18% of single cropped and 24% of double cropped soy to delay planting in the northeast. Thus, if the interannual trend towards earlier planting persists, the difference between the "desired" planting date and "feasible" planting date will grow, and onset will become the primary constraint for some planting dates. If, however, the trend toward earlier planting dates does not continue, the delayed onset does not "catch up" with the predicted planting CDF and delayed onset will not become the primary constraint for planting.

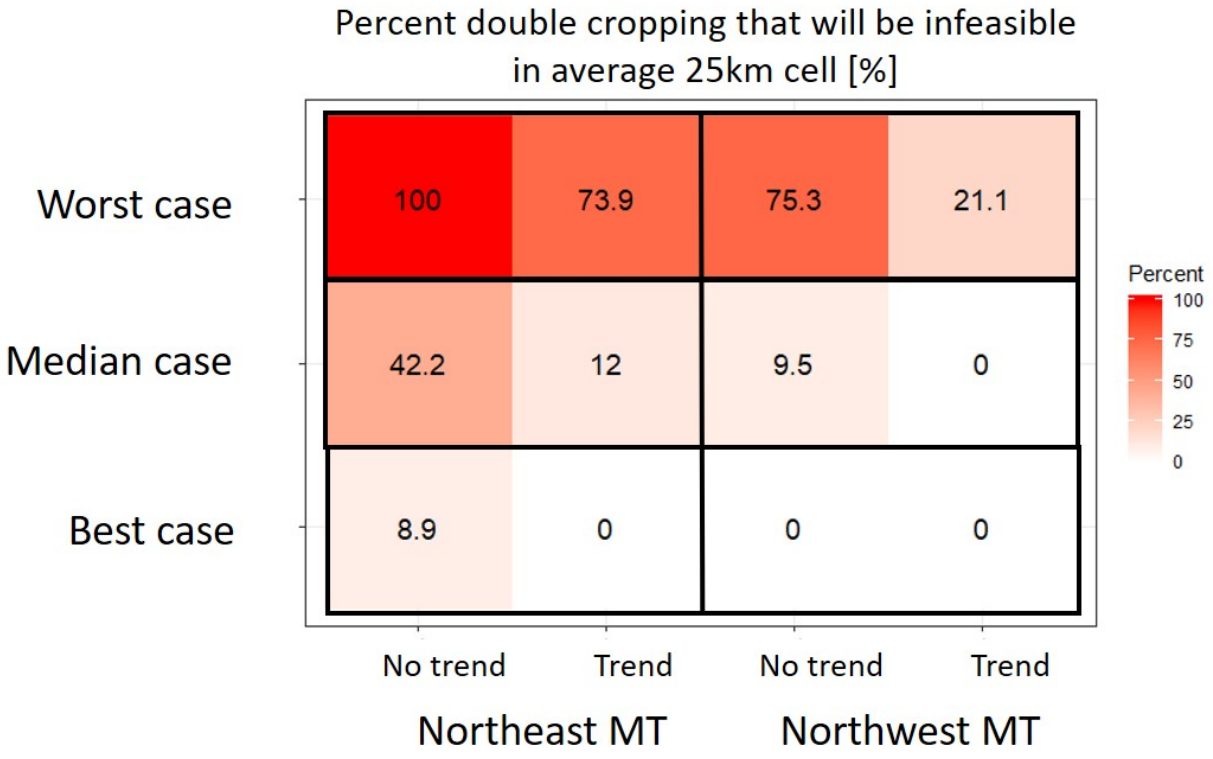


Figure 5.8: Projected percent of currently double cropped soy that will need to give up double cropping.

Even if they aren't directly affected by delayed onset, farmers may contend with an accelerated demise. While an earlier demise does not affect the planting date decision itself, it may reduce the land area suitable for double cropping. Conservative estimates put the crop cycle lengths of single and double cropped soy at 90 and 200 days, respectively [1]. With historical onset from 1970 - 2014 averaging 219 days for northeast and 233 days for northwest Mato Grosso, much of the state is already nearing the limit for double cropping suitability. A constriction of the wet season, expected to average 17 and 8 days in the northeast and northwest respectively, could make double cropping impossible for many of the currently double cropped fields. A medium demise date in the northeast would destroy 42% of double cropped area if the trend toward earlier planting does not continue; even the trend persists, 12% of double cropped area will be destroyed. An early demise would be catastrophic for double cropping: 74% and 21% of double cropped soy will be affected in northeast and northwest Mato Grosso, respectively, if technology allows planting dates to shift about 15 days earlier by 2024; 100% and 75% will be affected if this trend is absent. However, these percentages should be viewed two caveats. First, these metrics were calculated assuming

that planting occurs exactly as predicted. It is possible that farmers will collectively plant in a smaller time frame, creating a steeper planting CDF and allowing a larger percentage of double cropping to succeed. Second, in 2014, about 20% of double cropped soy in northeast Mato Grosso was planted after the "too late to plant" cutoff, suggesting flaws such as insufficiency of climatological definition of wet season timing used, an error in the crop cover map, and/or error in the planting date estimate. Fortunately, this discrepancy translates to less than 10 days and the relative impacts of each onset/demise scenario still hold.

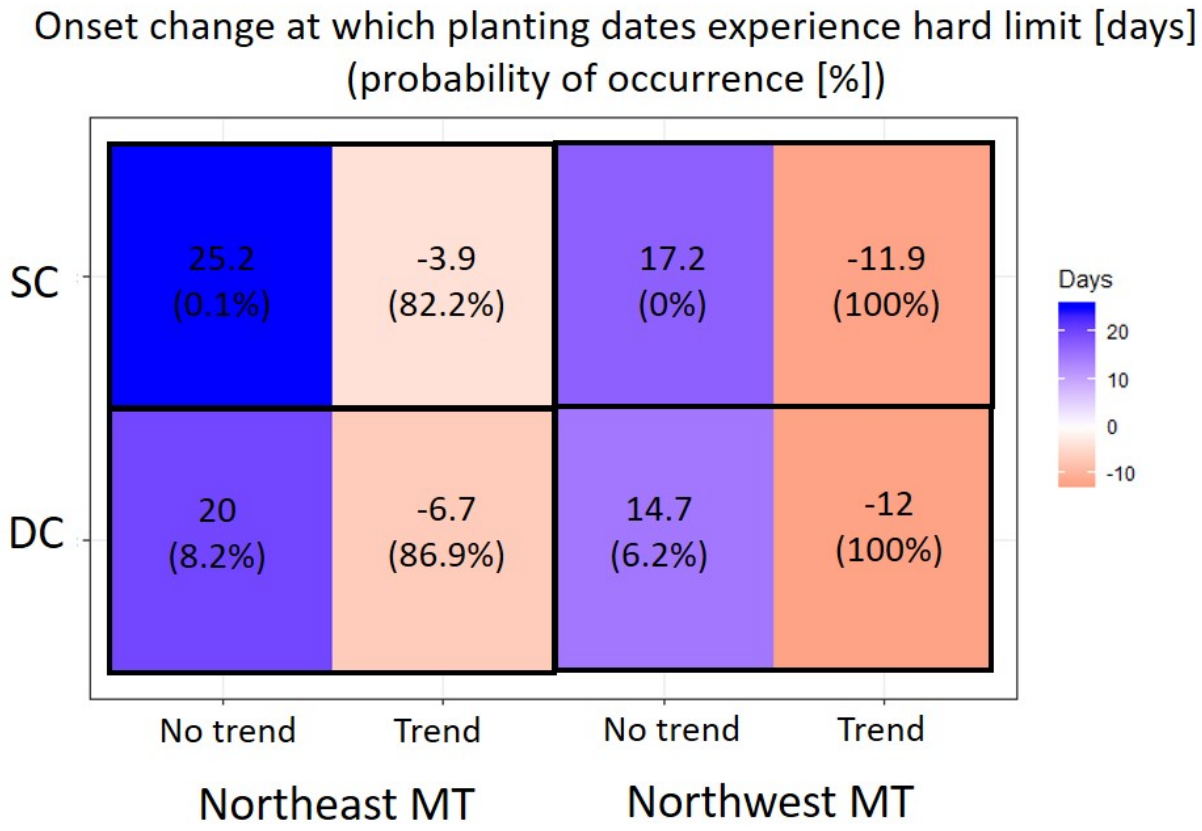


Figure 5.9: The change in onset, relative to 2014 values, that would cause at least the 5th percentile of predicted planting dates to experience wet season onset as the primary constraint to planting. Changes are calculated as future onset minus 2014 observed onset. The likelihood of these changes is also reported.

In addition to exploring limits on planting dates imposed by the three wet season scenarios, I also examine the threshold of onset change at which farmers would begin to experience wet season onset as a hard limit to planting. Historically, onset sensitivity was below unity:

a one-day delay in onset resulted in less than one day of delay in planting. I define the critical threshold of onset change as the point at which the delayed onset would overtake planting if historical sensitivity persists. At this point, farmers will presumably become perfectly sensitive to onset and experience the onset date as a hard limit to planting. Figure 5.9 shows the change in onset (relative to 2014 values) that would cause the 5th percentile of predicted planting dates to experience wet season onset as the primary constraint, and the likelihood of this critical change based on 2020 - 2049 climate projections. Though the northeast is more vulnerable to extreme onset delays, it is slightly less likely to experience onset as the primary constraint because historical planting dates were farther from onset. Negative onset changes in Figure 5.9 mean that onset will need to occur even earlier than 2014 values to allow the 5th percentile of planting to continue the trend toward earlier planting, a highly unlikely scenario. The trend to earlier planting for the 5th percentile, historically stronger than the trend of higher percentiles, is likely impossible for both regions. Without the trend, only a small minority of double cropped soy will experience wet season onset as the primary constraint, and current planting sensitivities will be more secure. Unfortunately, a cessation of the trend would put double cropping practices at risk as demise becomes earlier. Planting practices are therefore limited on both ends of the wet season.

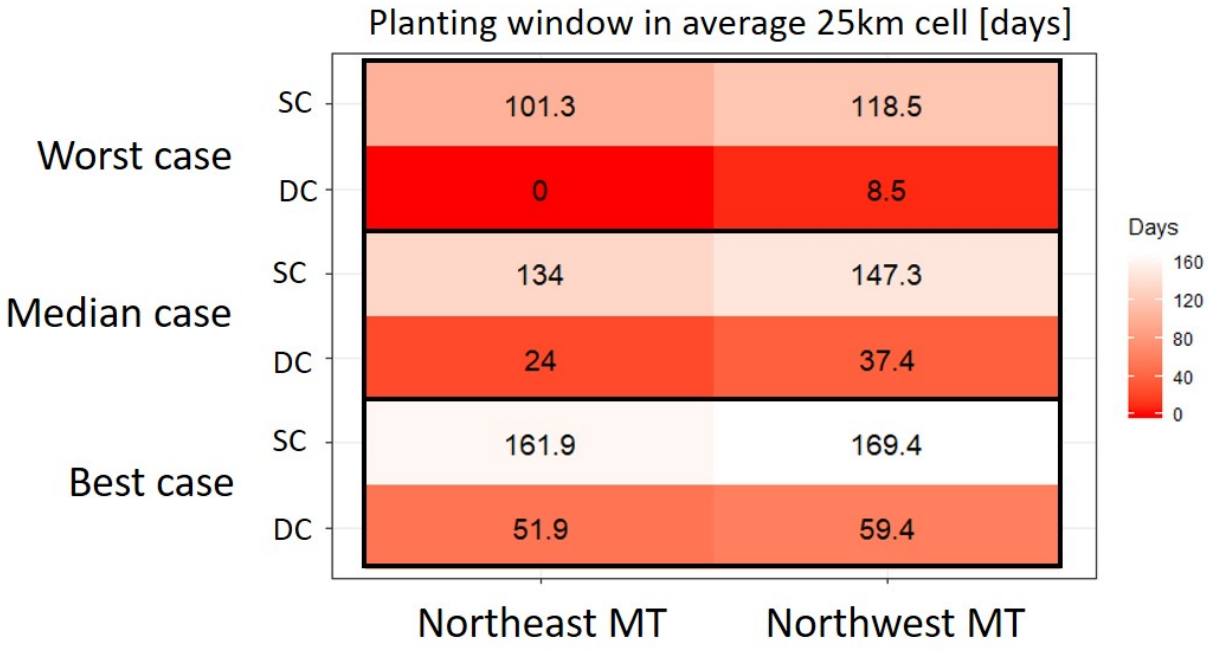


Figure 5.10: Projected number of days available for farmers to plant for a given cropping intensity.

The planting window, or the number of days available for planting, summarizes the extent of farmers' planting options (Figure 5.10). In agreement with projections from Pires et al (2016), severe impacts on double cropping are expected under the worst case scenario: in northeast Mato Grosso, the double cropping window shrinks to 8 days; in northeast Mato Grosso, double cropping becomes impossible (window of zero days). Even nonzero planting windows, when too small, may make it impossible for planting equipment to cover the whole property [106]. The planting windows observed in 2014 were 35 days for double cropped and 45 days for single cropped soy; therefore, the 24-day planting window for double cropped soy under a moderate scenario may still be disruptive.

5.4 Discussion

5.4.1 Predicted planting behavior in Mato Grosso

Mato Grosso is projected to experience a shorter wet season, an effect equally attributed to delayed onset and earlier demise. My predictions of planting dates' response to a delayed wet season onset are made under two behavioral scenarios: (1) planting dates are only affected by delayed onset and there is *no* trend to earlier planting after 2014, and (2) planting dates are affected by delayed onset and the trend to earlier planting *continues*. Planting dates under these behavioral scenarios will experience the negative impacts of wet season change differently. In general, planting dates that trend earlier each year will experience severe constraints related to delayed onset, while planting dates that do not trend earlier will primarily experience cropping intensity constraints due to earlier demise. Because it is impossible to anticipate whether and how much planting dates will trend earlier in the future, I discuss the two behavioral scenarios as bounding cases.

In the first bounding case, the trend toward earlier planting continues and planting dates will become about 15 days earlier in 2024 compared to 2014 values. Here, planting dates will come under two opposing forces: (1) a trend toward earlier planting, which pushes planting dates about 15 days earlier compared to 2014 values; and (2) the relatively smaller delay in planting due to delayed onset. The trend towards earlier planting is presumably made possible by improved technology, crop varieties, and transport networks that allow farmers to plant closer to the start of the wet season. While this trend cannot continue indefinitely, historically observed delays of around 30 days between onset and planting indicate that the trend is technically sustainable for at least another decade. The sanitary break in Mato Grosso, which ends on Sept 15/30, may also discontinue the trend [106], but average historical planting dates are a month later than the sanitary break. Therefore it's likely that a delayed onset, rather than external constraints like the sanitary break, would become the primary limitation for this trend by 2024. Assuming that the predicted planting distributions under the influence of delayed onset and interannual trend (Figures 5.3, 5.4 and 5.5) represent the "desired" planting date given technological and wet season onset sensitivities, I calculate the percent of soy area that must push planting later than the desired date due to a delay in

onset. In 2024, a median level of onset delay (7 days) in the vulnerable northeast region will delay the desired planting dates of only a minority of double cropped soy (up to 24%), but a late onset year (21 days) would coerce a majority (61%) of double cropped soy to delay planting. However, because the magnitude of the year coefficient may change (and likely decrease as technological advances push against climatic limits), these predictions are uncertain. This is an important caveat because the trend toward earlier planting comprises the majority of the difference between observed and predicted planting date distributions.

In the second bounding case, the trend toward earlier planting does not continue and planting dates are most likely delayed relative to 2014 values. This is the result of a soft limit imposed by delayed onset, in which onset does not directly touch planting dates, but does cause a delay in planting. During a late onset year, single cropped fields in the vulnerable northeast region will delay planting by an average of 3.6 days, while double cropped fields in the northeast region will delay planting by an average of 5.6 days. The delayed planting dates may force important phenological stages of the crop to be exposed to extreme high heat during the middle of austral summer. Additionally, when combined with earlier demise, the delay in planting will render double cropping impossible for large parts of the northeast: in the worst wet season case, all double cropping will be destroyed. While changes in the wet season during most years will generate milder consequences on double cropping, the chance of a severely disruptive year is still cause for concern. The decrease in cropping intensity, even occasionally, may have disastrous consequences for a state whose agriculture is 85% double cropped [27].

These bounding cases show that the necessary changes in planting behavior under delayed onset and/or earlier demise may be problematic for agricultural productivity for several reasons. First, the yield of individual crops may be impacted: planting dates that shift in response to either a hard or soft limit imposed by wet season onset may become suboptimal. Second, productivity may experience a sharp decline if double cropping becomes infeasible for some or all of currently double cropped areas. Third, the smaller planting windows imposed by delayed onset and earlier demise will require faster planting, which may be physically impossible or logistically difficult. Productivity may be affected by some or all of these avenues, depending on the presence of the trend, on the region, and on the year-specific wet season timing.

5.4.2 Caveats for predicted results

While the predicted results indicate that changes in wet season timing will disrupt planting and cropping practices in Mato Grosso, these predictions should be used cautiously. Because these results are based on a statistical model of historical (2004 - 2014) cropping practices, they lack a process-based understanding of variables that affect planting date and cropping intensity. It is unknown whether the fixed effects found in the regression models will persist in the future, whether new variables that influence planting dates will emerge, whether improved and short-cycle soy varieties will allow double cropping to survive under wet seasons shorter than 200 days, or whether historical sensitivities to onset will remain the same.

Additionally, I have shown in Chapter 4 that the climatological anomalous accumulation (AA) definitions of onset and demise are not the most relevant for planting decisions; wet season predictions based on the AA definition are therefore imperfect representations of behavior.

Other adaptive behaviors may used to alleviate the pressure on Mato Grosso's soy agriculture - these are not captured in the regression models. Farmers may respond to these stressors by expanding soy planted area to beyond what was observed in 2014. Because fixed effects cannot be calculated in new locations, it is impossible to use the OLS_{FE} models to predict planting date behavior in "new soy" areas located beyond the spatial extent of the training dataset. Therefore, while soy extensification may play a role in the future agricultural yields of Mato Grosso, I only predict planting date for historically planted soy regions. Likewise, although the onset scenarios and regions are representative of the broader spatial and interannual variations, it is possible that more extreme wet season timings may occur within each region and year. These extremes are overlooked in the predictions. Dramatic changes such as the replacement of small properties with large, mechanized agribusinesses may cause sudden shifts in planting behavior and yield [56]. Irrigation may also be used to avoid delaying the planting date. Although only 247 thousand ha of land is currently irrigated in Mato Grosso, improved power infrastructure may boost irrigated area in the future [56]. These adaptations could disrupt planting dates, but are ignored in the statistical model.

5.4.3 Understanding adaptation capacity

Adaptations such as planting date are central to agriculture's response to climate change, and planting dates are the most frequently tested adaptation strategy in crop models [151]. However, the lack of better information on planting behavior forces many crop yield projections to rely on planting scenarios or assume that farmers perfectly follow precipitation, soil moisture, or temperature patterns. The studies usually exclude other practices that impact planting dates, such as crop rotations, tillage practices, poverty, and lack of information [33, 47, 151]. My observation-based predictions of planting dates provide a more realistic idea of the speed and extent to which planting dates actually respond to climate variability, and implicitly account for imperfect responses to climatic variables and for the spatially heterogeneous response among individual farmers. As such, my predicted planting scenarios can generate more accurate predictions of crop yield under climate change.

5.5 Conclusions

In Mato Grosso, delayed onset and earlier demise under RCP 8.5 conditions will force planting dates to later (possibly suboptimal) times while decreasing the likelihood of a successful second crop. Under the first bounding case, in which the trend toward earlier planting dates continues into the future, planting dates will be delayed (and likely suboptimal) for 61% of

double cropped soy during the worst case wet season scenario in the vulnerable northeastern Mato Grosso (onset delayed by 21 days). More than 80% of years between 2020 and 2049 will experience a wet season onset that is delayed enough to become a hard limit to planting for 5th percentile double cropped soy. Under the second bounding case, in which the trend toward earlier planting does not continue, the demise of the wet season plays a larger role. In the worst wet season case, a wet season demise that arrives 26 days earlier will make 100% of double cropping infeasible in northeastern Mato Grosso (unless new short-cycle varieties are introduced). While these predictions come with significant uncertainties and should be taken with caution, they are based on the most spatiotemporally resolved planting observations available and are indicative of the fragility of Mato Grosso's agricultural practices. Future work can improve planting predictions by calculating projected wet season metrics based on features of rainfall that are most relevant for decision-making, quantifying the likelihood of alternative adaptations that make planting dates less vulnerable to the wet season (such as irrigation, new crop varieties, and agricultural extensification), and accounting for non-stationarities in planting behavior.

Similar predictive efforts would improve our understanding of agricultural yields worldwide. The magnitude of change in the wet season is expected to be concerning in many rainfed regions: in Malawi, the RCP8.5 climate scenario will shorten the growing season by 20 - 55 days by midcentury [145]. El Nino events in Indonesia are expected to increase in the probability of a highly disruptive 30-day delay in monsoon onset from 9 - 18% in 2007 to 30 - 40% in 2050 [100]. In Burkina Faso, the rainy season onset will be delayed by an average of one week in 2021 - 2050 compared to the 1971 - 2000 baseline under the A1B scenario, [65] and in West Africa the combined effect of delayed onset and earlier demise will cause a 20% reduction in the length of the growing season by 2050 [120]. Because my planting date estimation method and statistical analysis are scalable, they can be applied over agricultural areas like these to evaluate the sensitivity of planting date to wet season onset. These analyses can help to quantify risk to agricultural productivity, especially in vulnerable tropical and developing regions.

Chapter 6

Conclusions

6.1 Summary of Findings

The lack of updated, resolved planting date information impedes our understanding of how planting dates have responded to weather variability in the past, and consequently how they will evolve under climate change. This is a major deficiency in attempts to improve and predict future crop yields. Because planting dates are a major control on agricultural production (both directly by controlling the weather experienced by crops and indirectly by influencing cropping practices) and are a result of climatic, economic and social factors, we need to understand how they have behaved in the past before making extrapolations to the future.

In this work, I introduced a new, remote sensing-based estimation method for planting and harvest dates (Chapter 2). This method is uniquely scalable and addresses challenges that have previously prevented the creation of high resolution planting and harvest maps: lack of ground data and low computational resources. It avoids untested assumptions, sidesteps the requirement for ground truth calibration and validation data with microsatellite imagery from Planet Labs, and relies only on simple algorithms implementable in the cloud computing platform, Google Earth Engine. Additionally, I designed the method to be appropriate for locations that experience high cloud cover and aerosol interference during the growing season. Numerical experiments indicate that the smoothing methods implemented to address atmospheric noise produce estimates that are competitive with methods that reduce noise through complex, nonlinear timeseries analysis.

With this method, I produced updated planting and harvest date maps for soy in Mato Grosso, Brazil at unprecedented detail and scale (Chapter 3). The finer-resolution understanding of planting date behavior made possible by these maps is crucial for a region that is both vulnerable to climate change and subject to a shifting technological and economic context.

Regressions of planting date against the date of wet season onset over Mato Grosso reveal statistically significant differences in farmers' sensitivity to onset over small areas (Chapter

4). Fields planted relatively early (in the 5th percentile) within a 25 km region are up to three times more sensitive to the onset than fields planted late (in the 95th percentile). Cropping intensity also exerts strong control on the onset sensitivity: double cropped fields are almost twice as sensitive to onset as single cropped fields. Additionally, I discovered that farmers in Mato Grosso planted earlier each year from 2004 to 2014, a trend that occurred independently of wet season onset. Technological advances in crop variety or the expansion of the transportation network may have contributed to this trend. The heterogeneous and changing planting behavior over Mato Grosso, if ignored, could generate grossly incorrect agricultural yield projections. These regressions also revealed additional uncertainty surrounding the definition of the wet season: it is unclear which features of precipitation are most relevant for decision-makers. I found that climatological definitions such as the popular anomalous accumulation method are not as correlated to observed planting dates as definitions based on easily observable metrics, such as frequency of rainfall. Future studies of planting dates' sensitivity to weather may benefit from carefully chosen definitions of climatic variables.

Finally, predictions of planting behavior under future climate imply that delayed onset and earlier demise may become a primary constraint for agricultural productivity in Mato Grosso (Chapter 5). While high interannual variability in wet season timing and nonstationarity in planting behavior prevent the prediction of a precise impact, the bounding cases suggest that delayed onset and earlier demise may delay planting to suboptimal times, inhibit lucrative double cropping practices, and/or constrain the available planting window to a fraction of the historical range. These effects are concerning for a region whose economy depends on agribusiness.

These insights into historical and future planting behavior are only possible with updated, highly resolved planting data. By introducing an estimation method that closes the information gap on planting dates without the need for expensive ground survey data, this work helps to reduce uncertainty and error in crop yield models and propels understanding of how agriculture will adapt to future challenges. This information will be valuable for vulnerable agricultural regions such as southern Asia and southern Africa, which face not only the most severe consequences of warming, but also data scarcity and limited adaptive capacity [84, 87].

6.2 Future Work

Future work could improve the robustness of my estimation method to heterogeneous land cover and cloudiness, reduce reliance on irrigation maps, and deepen our understanding of the social and economic drivers of planting dates.

Fusion of satellite data sources

While my planting date estimation method was designed to be applicable to agricultural regions worldwide, it is expected to have varying degrees of success. Extensions of the

method to regions with less cloud cover, better air quality, less intensive cropping and to crops with longer crop cycles would be highly feasible: these features of the cropping system and environment would increase data availability, slow down the pace of phenological (crop growth) change, and allow less aggressive smoothing. Heavily irrigated, arid areas dominated by single cropping, such the central US and California, would be ideal candidates for this method. However, in regions with complex topography, fragmented cropland, where crop phenological change happens rapidly relative to the 8-day MODIS resolution, or where critical phenological periods (such as the date of maximum EVI change) occur during times of maximum cloud cover, the method may be of less value. Because smoothing risks merging separate crop peaks, areas dominated by triple cropping, frequent failed first crops that partially green up, or significant weed growth will overestimate the length of the crop cycle. The fusion of multiple satellite sources with varying spectral range and spatiotemporal resolution would improve estimates in these challenging areas. Fusion of MODIS with satellite sources that offer higher spatial resolution (Landsat and Sentinel-2) would improve estimates in regions with fragmented cropland, and fusion with cloud-penetrating radar data (Sentinel-1 Synthetic Aperture Radar) would offer higher robustness to clouds.

Irrigation detection

Because this work targeted only rainfed fields, the study period was limited by available irrigation maps (which were used to eliminate irrigated fields from consideration). The latest irrigation map available for Mato Grosso was produced in 2014, forcing the study to end at that year. While it is theoretically possible to train a classifier or image analysis algorithm to detect the locations of center pivot irrigation using the 2014 data as training information, it was difficult to do in practice. In the United States, an image analysis algorithm based on Google Earth Engine (GEE) detects center pivots by searching for circular patterns in remotely sensed images. However, unlike the standardized center pivots in the US, Brazil's center pivots are non-uniform in radius and layout, making it necessary to use different algorithms to detect differently arranged and sized center pivots. Additionally, while center pivots in the US consistently appear different (much greener) than their backgrounds at the beginning of spring, there is no season during which Brazilian center pivots consistently contrast with their background. Finally, the patchiness of natural vegetation in Brazil causes a high degree of falsely detected center pivots. The possibility of irrigation techniques beyond center pivot further complicates irrigation detection. These challenges could be tackled with deep learning algorithms or image analysis techniques.

Additional variables in regression model

The regression model that quantifies planting dates' sensitivity to onset explains only half of the total planting date variability. Its predictive power could be improved by including independent variables that are informed by ground-level surveys of growers, reflecting the economic, social logistical considerations involved in planting. Detailed surveys about plant-

ing decisions are not yet available for Mato Grosso, but would be valuable for understanding which adaptations are employed and how extensively they are used. This knowledge would also help define a region's adaptive capacity, highlight barriers to adaptation, and direct efforts to increase resilience.

New climatic drivers of planting date can also be explored. The model's assumption that wet season onset is the primary climatic driver of planting is reasonable in a tropical area such as Mato Grosso, but temperature and freezing/thawing dates should be included before applying a similar approach in colder regions: in northern China, wheat farmers responded to warming by planting 2.1 days earlier for each 1° C increase in maximum temperature [136]. A model with these adjustments could explain planting date behavior in a larger range of climates.

Including these additional variables can clarify the cognitive drivers behind adaptation, help us predict which adaptation strategies are more likely, and form the basis of better statistical or process-based models to understand adaptive behavior and future yields.

Bibliography

[1] G.M. Abrahao and M.H. Costa. "Evolution of rain and photoperiod limitations on the soybean growing season in Brazil: The rise (and possible fall) of double-cropping systems". In: *Agricultural and Forest Meteorology* 245 (2018), pp. 32–45.

[2] Agrosatelite. *Planted area and crop yield*. `https://agrosatelite.com.br/en/products/safras/#planted-area`. 2017.

[3] FM Akinseye et al. "Evaluation of the onset and length of growing season to define planting date—'a case study for Mali (West Africa)'". In: *Theoretical and Applied Climatology* 124 (2016), pp. 973–983.

[4] Peter Alexander et al. "Adaptation of global land use and management intensity to changes in climate and atmospheric carbon dioxide". In: *Global Change Biology* 24 (2018), pp. 1–10.

[5] Vesselin Alexandrov et al. "Potential impact of climate change on selected agricultural crops in north-eastern Austria". In: *Global Change Biology* 8 (2002), pp. 372–389.

[6] Luis Frenando Alliprandini, Claudiomir Abatti, Paulo Fernando Bertagnolli, et al. "Understanding Soybean Maturity Groups in Brazil: Environment, Cultivar Classifi cation, and Stability". In: *Crop Science* 49 (2009), pp. 801–808.

[7] Fernando Andrade et al. "Yield responses to narrow rows depend on increased radiation interception". In: *Agronomy Journal* 95 (2002), pp. 975–980.

[8] Muhuddin Rajin Anwar et al. "Adapting agriculture to climate change: a review". In: *Theoretical Applied Climatology* 113 (2013), pp. 225–245.

[9] Sotirios V Archontoulis, Fernando E Miguez, and Kenneth J Moore. "A methodology and an optimization tool to calibrate phenology of short-day species included in the APSIM PLANT model: Application to soybean". In: *Environmental Modelling Software* 62 (2014), pp. 465–477.

[10] Damien Arvor et al. "Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices". In: *Applied Geography* 32 (2012), pp. 702–713.

[11] Damien Arvor et al. "Mapping and spatial analysis of the soybean agricultural frontier in Mato Grosso, Brazil, using remote sensing data". In: *GeoJournal* 78 (2012), pp. 833–850.

[12] Damien Arvor et al. "Spatial patterns of rainfall regimes related to levels of double cropping agriculture systems in Mato Grosso (Brazil)". In: *International Journal of Climatology* 96 (2015), pp. 69–83.

[13] hamed Ashouri et al. "PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies". In: *American Meteorological Society* 32 (2012), pp. 702–713.

[14] Brian S Baldwin and Robert D Cossar. "Caster yield in response to planting date at four locations in the south-central United States". In: *Industrial Crops and Products* 29 (2009), pp. 316–319.

[15] AM Bastidas et al. "Soybean sowing date: the vegetative, reproductive, and agronomic impacts". In: *Agronomy and Horticulture* 99 (2008), pp. 1–14.

[16] A Begue et al. "Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI". In: *International Journal of Remote Sensing* 31 (2010), pp. 5391–5407.

[17] Douglas K Bolton and Mark A Friedl. "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics". In: *Agricultural and Forest Meteorology* 173 (2013), pp. 74–84.

[18] Alberte Bondeau et al. "Modelling the role of agriculture for the 20th century global terrestrial carbon balance". In: *Global Change Biology* 13 (2007), pp. 1–22.

[19] Allison Borchers et al. "Multi-cropping practices: recent trends in double-cropping". In: *United States Department of Agriculture Economic Information Bulletin* 125 (2014), pp. 1–22.

[20] M Boschetti et al. "Multi-year monitoring of rice crop phenology through time series analysis of MODIS images". In: *International Journal of Remote Sensing* 30 (2009), pp. 4643–4662.

[21] Sandra Brito. "Atlas da Irrigação mostra estudo do uso da água na agricultura nacional". In: *Agencia Nacional de Aguas, Embrapa* (2017).

[22] J Christopher Brown et al. "Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data". In: *Remote Sensing of the Environment* 130 (2013), pp. 39–50.

[23] Aline Bussmann et al. "Sowing date determinants for Sahelian rainfed agriculture in the context of agricultural policies and water management". In: *Land Use Policy* 52 (2016), pp. 316–328.

[24] Nathalie Butt, Paula Afonso de Oliveira, and Marcos Heil Costa. "Evidence that deforestation affects the onset of the rainy season in Rondonia, Brazil". In: *Journal of Geophysical Research* 116 (2011), pp. 1–8.

[25] AJ Challinor et al. "A meta-analysis of crop yield under climate change and adaptation". In: *Nature Climate Change* 4 (2014), pp. 287–291.

[26] Andrew Challinor et al. "Assessing the vulnerability of food crop systems in Africa to climate change". In: *Climate Change* 83 (2007), pp. 381–399.

[27] Yaoliang Chen, Dengsheng Lu, Emilio Moran, et al. "Mapping croplands, cropping patterns, and crop types using MODIS time-series data". In: *International Journal of Applied Earth Observation and Geoinformation* 69 (2018), pp. 133–147.

[28] Nick Clinton. *Time Series Analysis in Earth Engine*. `goo.gl/lMwd2Y`. Accessed: 2018-06-24. 2017.

[29] Avery S Cohn et al. "Cropping frequency and area response to climate variability can exceed yield response". In: *Nature Climate Change* 6 (2016), pp. 601–604.

[30] Paulo Correa and Cristiane Schmidt. "Public Research Organizations and Agricultural Development in Brazil: How Did Embrapa Get It Right?" In: *The World Bank* 145 (2014), pp. 1–10.

[31] M.H Costa and G.F Pires. "Effects of Amazon and Central Brazil deforestation scenarios on the duration of the dry season in the arc of deforestation". In: *International Journal of Climatology* 30 (2010), pp. 1970–1979.

[32] Marcos H Costa, Leonardo C Fleck, Avery S Cohn, et al. "Climate risks to Amazon agriculture suggest a rationale to conserve local ecosystems". In: *Frontiers in Ecology and the Environment* 17 (2019), pp. 1–7.

[33] Temesgen Tadesse Deressa et al. "Determinants of farmers' choice of adaptation methods to climate change in the Nile Basin of Ethiopia". In: *Global Environmental Change* 19 (2009), pp. 248–255.

[34] WRSS Dharmarathna, Srikantha Herath, and SB Weerakoon. "Changing the planting date as a climate change adaptation strategy for rice production in Kurunegala district, Sri Lanka". In: *Sustainability Science* 9 (2014), pp. 103–107.

[35] Laura Dobot et al. "Crop planting date matters: Estimation methods and effect on future yields". In: *Agricultural and Forest Meteorology* 223 (2016), pp. 103–115.

[36] I Dounias, C Aubry, and A Capillon. "Decision-making processes for crop management on African farms. Modelling from a case study of cotton crops in northern Cameroon". In: *Agricultural Systems* 73 (2003), pp. 233–230.

[37] Caroline M Dunning, Emily Black, and Richard P Allan. "Later Wet Seasons with More Intense Rainfall over Africa under Future Climate Change". In: *American Meteorological Society* 1 (2018), pp. 9719–9738.

[38] J Elliott et al. "The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0)". In: *Geoscientific Model Development* 8 (2015), pp. 261–277.

[39] EUROSTAT. *Regional statistics/Agriculture–Agricultural products–Crop products. Regional agriculture statistics: Areas harvested, yields, production.* `http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database`. 2008.

[40] FAO. *AQUASTAT Main Database, Food and Agriculture Organization of the United Nations (FAO).* `http://www.fao.org/land-water/databases-and-software/aquastat/en/`. 2016.

[41] Tom G Farr, Paul A Rosen, Edward Caro, et al. "The Shuttle Radar Topography Mission". In: *Reviews of Geophysics* 45 (2007), pp. 1–33.

[42] Marcelo Favarao. *Soybean planting, Brazil Style.* `https://www.farmprogress.com/soybeans/soybean-planting-brazil-style`. 2012.

[43] Philip M Fearnside. "Soybean cultivation as a threat to the environment in Brazil". In: *Environmental Conservation* 23 (2001), pp. 23–28.

[44] Giuseppe Feola et al. "Researching farmer behaviour in climate change adaptation and sustainable agriculture: Lessons learned from five case studies". In: *Journal of Rural Studies* 39 (2015), pp. 78–84.

[45] Jonathan A Foley, Navin Ramankutty, Kate A Brauman, et al. "Solutions for a cultivated planet". In: *Nature* 478 (2011), pp. 337–342.

[46] Food and Agriculture Organization of the United Nations (FAO). *AQUASTAT Review of Agricultural Water Use Per Country: Irrigation Cropping Calendar Per Country.* `http://www.fao.org/nr/water/aquastat/water_use/index.stm`. 2005.

[47] BY Fosu-Mensah, PLG Vlek, and DS MacCarthy. "Farmers' perception and adaptation to climate change: a case study of Sekyedumase district in Ghana". In: *Environment, Development and Sustainability* 14 (2012), pp. 495–505.

[48] R Fu, L Yin, W Li, et al. "Increased dry-season length over southern Amazonia in recent decades and its implication for future climate projection". In: *Proceedings of the National Academy of Sciences* 1073 (2013), pp. 1–6.

[49] Chris Funk et al. "The climate hazards infrared precipitation with stations – a new envrionmental record for monitoring extremes". In: *Nature Scientific Data* 2 (2015), pp. 1–21.

[50] G.L Galford et al. "Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil". In: *Remote Sensing of Environment* 112 (2007), pp. 576–587.

[51] Rachael D Garret, Eric F Lambin, and Rosamond L Naylor. "Land institutions and supply chain configurations as determinants of soybean planted area and yields in Brazil". In: *Land Use Policy* 31 (2013), pp. 385–396.

[52] Tagel Gebrehiwot and Anne van der Veen. "Farm Level Adaptation to Climate Change: The Case of Farmer's in the Ethiopian Highlands". In: *Environmental Management* 52 (2013), pp. 29–44.

[53] Roland A Geerken. "An algorithm to classify and monitor seasonal variations in vegetation phenologies and their inter-annual change". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (2009), pp. 422–431.

[54] Sharon M Gourdji, Adam M Sibley, and David B Lobell. "Global crop exposure to critical high temperatures in the reproductive period: historical trends and future project". In: *Environmental Research Letters* 8 (2013), pp. 1–10.

[55] Patricio Grassini, Lenny GJ van Bussel, et al. "How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis". In: *Field Crops Research* 177 (2015), pp. 49–63.

[56] Anna C Hampf et al. "Future yields of double-cropping systems in the Southern Amazon, Brazil, under climate change and technological development". In: *Agricultural Systems* 177 (2020), pp. 1–18.

[57] Celia A Harvey, Zo Lalaina Rakotobe, et al. "Extreme vulnerability of smallholder farmers to agricultural risks and climate change in Madagascar". In: *Philosophical Transactions of the Royal Society - Biological Sciences* 369 (2014).

[58] Rashid M Hassan and Charles Nhemachena. "Determinants of African farmers' strategies for adapting to climate change: multinomial choice analysis". In: *Research in Agricultural and Applied Economics* 3 (2008), pp. 83–104.

[59] Jerry L Hatfield and John H Prueger. "Temperature extremes: Effect on plant growth and development". In: *Weather and Climate Extremes* 10 (2015), pp. 4–10.

[60] Thomas W Hertel, Marshall B Burke, and David B Lobell. "The poverty implications of climate-induced crop yield changes by 2030". In: *Global Environmental Change* 20 (2010), pp. 577–585.

[61] Thomas Hilker et al. "Remote sensing of tropical ecosystems: Atmospheric correction and cloud masking matter". In: *Remote Sensing of Environment* 127 (2012), pp. 370–384.

[62] G Hmimina et al. "Evaluation of the potential of MODIS satellite data to predict vegetation phenology in different biomes: An investigation using ground-based NDVI measurements". In: *Remote Sensing of Environment* 132 (2013), pp. 145–158.

[63] S Mark Howden et al. "Adapting agriculture to climate change". In: *Proceedings of the National Academy of Sciences* 104 (2007), pp. 19691–19696.

[64] AR Huete et al. "A comparison of vegetation indices over a global set of TM images for EOS-MODIS". In: *Remote Sensing of Environment* 59 (1997), pp. 440–451.

[65] Boubacar Ibrahim et al. "Changes in rainfall regime over Burkina Faso under the climate change conditions simulated by 5 regional climate models". In: *Climate Dynamics* 42 (2014), pp. 1363–1381.

[66] Toshichika Iizumi, Wonsik Kim, and Motoki Nishimori. "Modeling the global sowing and harvesting windows of major crops around the year 2000". In: *Journal of Advances in Modeling Earth Systems* 11 (2019), pp. 99–112.

[67] IMEA. *Crop Progress Report*. `imea.com.br/imea-site/relatorios-mercado`. 2019.

[68] Meha Jain et al. "Understanding the causes and consequences of differential decision-making in adaptation research: Adapting to a delayed monsoon onset in Gujarat, India". In: *Global Change Biology* 23 (2017), pp. 2687–2704.

[69] SK Jalota et al. "Mitigating future climate change effects by shifting planting dates of crops in rice–wheat cropping system". In: *Regional Environmental Change* 12 (2012), pp. 913–922.

[70] Zhenong Jin et al. "The combined and separate impacts of climate extremes on the current and future US rainfed maize and soybean production under elevated CO2". In: *Global Change Biology* 23 (2017), pp. 2687–2704.

[71] J Joiner et al. "The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange". In: *Remote Sensing of Environment* 152 (2014), pp. 375–391.

[72] JW Jones et al. "The DSSAT cropping system model". In: *European Journal of Agronomy* 18 (2003), pp. 235–265.

[73] Peter G Jones and Philip K Thornton. "The potential impacts of climate change on maize production in Africa and Latin America in 2055". In: *Global Environmental Change* 13 (2003), pp. 51–59.

[74] Per Jonsson and Lars Edklundh. "TIMESAT – a program for analyzing time-series of satellite sensor data". In: *Computers & Geosciences* 30 (2004), pp. 833–845.

[75] Namrata Kala. "Ambiguity aversion and learning in a changing world: the potential effects of climate change from India agriculture". In: *Working Paper* (2015).

[76] Jude H Kastens et al. "Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil". In: *PLoS ONE* 12 (2017), pp. 1–21.

[77] Yoram J Kaufman, Didier Tanre, and Olivier Boucher. "A satellite view of aerosols in the climate system". In: *Nature* 419 (2002), pp. 215–223.

[78] Uttam Khanal et al. "Farmers' Adaptation to Climate Change, Its Determinants and Impacts on Rice Yield in Nepal". In: *Ecological Economics* 144 (2018), pp. 139–147.

[79] Jerry Knox et al. "Climate change impacts on crop productivity in African and South Asia". In: *Environmental Research Letters* 7 (2012), pp. 1–8.

[80] Christopher J Kucharik. "A multidecadal trend of earlier corn planting in the central USA". In: *Agronomy Journal* 98 (2006), pp. 1544–1550.

[81] Christopher J Kucharik and Shawn P Serbin. "Impacts of recent climate change on Wisconsin corn and soybean yield trends". In: *Environmental Research Letters* 3 (2008), pp. 1–10.

[82] Joseph G Lauer et al. "Corn hybrid response to planting date in the Northern Corn Belt". In: *Agronomy Journal* 91 (1999), pp. 834–839.

[83] P Laux, H Kunstmann, and A Bardossy. "Predicting the regional onset of the rainy season in West Africa". In: *International Journal of Climatology* 28 (2008), pp. 329–342.

[84] Patrick Laux et al. "Impact of climate change on agricultural productivity under rainfed conditions in Cameroon - A method to improve attainable crop yields by planting date adaptations". In: *Agricultural and Forest Meteorology* 150 (2010), pp. 1258–1271.

[85] Morgan C Levy et al. "Addressing rainfall data selection uncertainty using connections between rainfall and streamflow". In: *Nature Scientific Reports* 7 (2017), pp. 1–12.

[86] Brant Liebmann et al. "Onset and End of the Rainy Season in South America in Observations and the ECHAM 4.5 Atmospheric General Circulation Model". In: *American Meteorological Society Journal of Climate* 20 (2007), pp. 2037–2050.

[87] David B Lobell et al. "Prioritizing Climate Change Adaptation Needs for Food Security in 2030". In: *Science* 319 (2008), pp. 607–610.

[88] David B Lobell et al. "Satellite detection of earlier wheat sowing in India and implications for yield trends". In: *Agricultural Systems* 115 (2013), pp. 137–143.

[89] Juan Lopez-Sanchez, Shane R Cloude, and J. David Ballester-Berman. "Rice phenology monitoring by means of SAR polarimetry at X-band". In: *IEEE Transactions on Geoscience and Remote Sensing* 50 (2012), pp. 2695–2709.

[90] Marcos Javier de Luca, Marco Nogueira, and Mariangela Hungria. "Feasibility of lowering soybean planting density without compromising nitrogen fixation and yield". In: *Agronomy Journal* 106 (2014), pp. 1–7.

[91] Shaoxiu Ma, Galine Churkina, and Kristina Trusilova. "Investigating the impact of climate change on crop phenological events in Europe with a phenology model". In: *International Journal Biometeorology* 56 (2012), pp. 749–763.

[92] Project Mapbiomas. *Mapbiomas - Collection 3.1 of Brazilian Land Cover & Use Map Series*. `http://mapbiomas.org/`. 2019.

[93] Jose A Marengo et al. "The drought of 2010 in the context of historical droughts in the Amazon region". In: *Geophysical Research Letters* 38 (2011), pp. 1–5.

[94] Romain Marteau et al. "The onset of the rainy season and farmers' sowing strategy for pearl millet cultivation in Southwest Niger". In: *Agricultural and Forest Meteorology* 151 (2011), pp. 1356–1369.

[95] Camilla Mathison, Pete Deva Chetan adn Falloon, and Andrew J Challinor. "Estimating sowing and harvest dates based on the Asian summer monsoon". In: *Earth System Dynamics* 9 (2018), pp. 563–592.

[96] Ole Mertz et al. "Adaptation to climate change in developing countries". In: *Environmental Management* 43 (2009), pp. 743–752.

[97] Edward M Mugalavai et al. "Analysis of rainfall onset, cessation and length of growing season for western Kenya". In: *Agricultural and Forest Meteorology* 148 (2008), pp. 1123–1135.

[98] Ocean Biology Processing Group NASA Goddard Space Flight Center Ocean Ecology Laboratory. *Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua MYD09A1.006*. 2019.

[99] Ocean Biology Processing Group NASA Goddard Space Flight Center Ocean Ecology Laboratory. *Moderate Resolution Imaging Spectroradiometer (MODIS) Terra MOD09A1.006*. 2019.

[100] Rosamond L Naylor et al. "Assessing risks of climate variability and climate change for Indonesian rice agriculture". In: *Proceedings of the National Academy of Sciences* 104 (2007), pp. 7752–7757.

[101] Milad Nouri et al. "Towards shifting planting date as an adaptation practice for rainfed wheat response to climate change". In: *Agricultural Water Management* 186 (2017), pp. 108–119.

[102] J.I Ortiz-Monasterio, S.S Dhillon, and R.A Fischer. "Date of sowing effects on grain yield and yield components of irrigated spring wheat cultivars and relationships with radiation and temperature in Ludhiana, India". In: *Field Crops Research* 37 (1994), pp. 169–184.

[103] Henny Osbahr et al. "Supporting agricultural innovation in Uganda to respond to climate risk: linking climate change and variability with farmer perceptions". In: *Experimental Agriculture* 47 (2011), pp. 293–316.

[104] Zhuokun Pan et al. "Mapping crop phenology using NDVI time-series derived from HJ-1 A/B data". In: *International Journal of Applied Earth Observation and Geoinformation* 34 (2015), pp. 188–197.

[105] Michelle Picoli et al. "Big earth observation time series analysis for monitoring Brazilian agriculture". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (2018), pp. 328–339.

[106] Gabrielle F Pires et al. "Increased climate risk in Brazilian double cropping agriculture systems: Implications for land use in Northern Brazil". In: *Agricultural and Forest Meteorology* 228 (2016), pp. 286–298.

[107] Planet. *Planet Application Program Interface: In Space for Life on Earth*. `https://api.planet.com`. 2017.

[108] Felix T Portmann, Stefan Siebert, and Petra Doll. "MIRCA2000 – Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling". In: *Global Biogeochemical Cycles* 24 (2010), pp. 1–24.

[109] Dirk Raes et al. "Evaluation of first planting dates recommended by criteria currently used in Zimbabwe". In: *Agricultural and Forest Meteorology* 125 (2004), pp. 177–185.

[110] BC Reed, MD Schwartz, and X Xiao. *Phenology of Ecosystem Processes*. New York, new York: Springer, 2009.

[111] Pytrik Reidsma et al. "Adaptation to climate change and climate variability in European agriculture: The importance of farm level responses". In: 32 (2010), pp. 91–102.

[112] Jie Ren, James B. Campbell, and Yang Shao. "Estimation of SoS and EoS for midwestern US corn and soybean crops". In: *Remote Sensing* 9 (2017), pp. 1–14.

[113] Cynthia Rosenzweig, Joshua Elliott, et al. "Assessing agricultural risks of climate change in the 21st century in a global gridded crop model comparison". In: *Proceedings of the National Academy of Sciences* 111 (2014), pp. 3268–3273.

[114] D.P. Roy, M.A. Wulder, T.R Loveland, et al. "Landsat-8: Science and product vision for terrestrial global change research". In: *Remote Sensing of Environment* 145 (2014), pp. 154–172.

[115] W.J Sacks et al. "Crop planting dates: an analysis of global patterns". In: *Global Ecology and Biogeography* 19 (2010), pp. 607–620.

[116] Toshihiro Sakamoto et al. "A crop phenology detection method using time-series MODIS data". In: *Remote Sensing of Environment* 96 (2005), pp. 366–374.

[117] Toshihiro Sakamoto et al. "A Two-Step Filtering approach for detecting maize and soybean phenology with time-series MODIS data". In: *Remote Sensing of Environment* 114 (2010), pp. 2146–2159.

[118] Gilvan Sampaio et al. "Regional cliamte change over eastern Amazonia caused by pasture and soybean cropland expansion". In: *Geophysical Resarch Letters* 34 (2007), pp. 1–7.

[119] Carlos dos Santos, Montserrat Salmeron, and Larry Purcell. "Soybean phenology prediction tool for the US midsouth". In: *Agricultural and Environmental Letters* 4 (2019), pp. 1–4.

[120] Benoit Sarr. "Present and future climate change in the semi-arid region of West Africa: a crucial input for practical adaptation in agriculture". In: *Atmospheric Science Letters* 13 (2012), pp. 108–112.

[121] Wolfram Schlenker and David B Lobell. "Robust negative impacts of climate change on African agriculture". In: *Environmental Research Letters* 5 (2010), pp. 1–20.

[122] Wolfram Schlenker and Michael J Roberts. "Nonlinear temperature effects indicate severe damages to US crop yields under climate change". In: *Proceedings of the National Academy of Sciences* 37 (2009), pp. 15594–15598.

[123] Josef Schmidhuber and Francesco N Tubiello. "Global food security under climate change". In: *Proceedings of the National Academy of Sciences* 104 (2007), pp. 19703–19708.

[124] C.A Seifert and D.B Lobell. "Response of double cropping suitability to climate change in the United States". In: *Environmental Research Letters* 10 (2015), pp. 1–6.

[125] R Shumway and Stoffer D. *Time series analysis and its applications: with R examples*. 2017.

[126] Britaldo Soares-Filho et al. "Cracking Brazil's Forest Code". In: *Science* 25 (2014), pp. 363–364.

[127] A Soltani and T.R Sinclair. *Modeling physiology of crop development, growth and yield*. CAB International. 2012.

[128] Kamel Soudani et al. "Evaluation of the onset of green-up in temperate deciduous broadleaf forests derived from Moderate Resolution Imaging Spectroradiometer (MODIS) data". In: *Remote Sensing of Environment* 119 (2008), pp. 2643–2655.

[129] Keith R Spangler, Lynch Amanda H, and Stephanie A Spera. "Precipitation drivers of cropping frquency in the Brazilian Cerrado: evidence and implications for decision-making". In: *Weather, Climate and Society* 9 (2017), pp. 201–213.

[130] Stephanie A Spera et al. "Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics". In: *Environmental Research Letters* 9 (2014), pp. 1–12.

[131] Mike Staton. *Soybean planting depth matters*. `https://www.canr.msu.edu/news/soybean_planting_depth_matters`. 2016.

[132] RD Stern, MD Dennett, and DJ Garbutt. "The start of the rains in West Africa". In: *Journal of Climatology* 66 (1981), pp. 59–68.

[133] Claudio O Stockle, Marcello Donatelli, and Roger Nelson. "CropSyst, a cropping systems simulation model". In: *European Journal of Agronomy* 18 (2003), pp. 289–307.

[134] Benjamin Sultan et al. "Agricultural impacts of large-scale variability of the West African monsoon". In: *Agricultural and Forest Meteorology* 128 (2004), pp. 93–110.

[135] Eugene S Takle et al. "Climate forecasts for corn producer decision making". In: *Earth Interactions* 18 (2014), pp. 1–8.

[136] Tulu Tao et al. "Climate changes and trends in phenology and yields of field crops in China, 1981–2000". In: *Agricultural and Forest Meteorology* 138 (2006), pp. 82–92.

[137] Philip K Thornton et al. "Is agricultural adaptation to global change in lower-income countries on track to meet the future food production challenge?" In: *Global Environmental Change* 52 (2018), pp. 37–48.

[138] Philip K Thornton et al. "Spatial variation of crop yield response to climate change in East Africa". In: *Global Environmental Change* 19 (2009), pp. 54–65.

[139] D Urban, K Guan, and M Jain. "Estimating sowing dates from satellite data over the U.S. Midwest: A comparison of multiple sensors and metrics". In: *Remote Sensing of Environment* 211 (2018), pp. 400–412.

[140] David Urban et al. "Greater sensitivity to drought accompanies maize yield increase in the US Midwest". In: *Science* 344 (2014), pp. 516–519.

[141] USDA. *Crop Progress Report*. `https://www.nass.usda.gov/Publications/National_Crop_Progress/`. 2019.

[142] Jan Verbesselt et al. "Phenological change detection while accounting for abrupt and gradual trends in satellite image time series". In: *Remote Sensing of Environment* 114 (2010), pp. 2970–2980.

[143] Louis V Verchot et al. "Climate change: linking adaptation and mitigation through agroforestry". In: *Mitigation and Adaptation Strategies for Global Change* 12 (2007), pp. 901–918.

[144] Daniel de Castro Victoria et al. "Cropland area estimates using MODIS NDVI time series in the state of Mato Grosso, Brazil". In: *Pesquisa Agropecuaria Brasileira* 47 (2012), pp. 1270–1278.

[145] Edward K Vizy et al. "Projected changes in Malawi's growing season". In: *Climate Dynamics* 45 (2015), pp. 1673–1698.

[146] H Wagenseil and C Samimi. "Assessing spatio-temporal variations in plant phenology using Fourier analysis on NDVI time series: results from a dry savannah environment in Namibia". In: *International Journal of Remote Sensing* 27 (2006), pp. 3455–3471.

[147] K Waha et al. "Adaptation to climate change through the choice of cropping system and sowing date in sub-Saharan Africa". In: *Global Environmental Change* 23 (2013), pp. 130–143.

[148] K Waha et al. "Climate-driven simulation of global crop sowing dates". In: *Global Ecology and Biogeography* 21 (2012), pp. 247–259.

[149] Moussa Waongo et al. "A Crop Model and Fuzzy Rule Based Approach for Optimizing Maize Planting Dates in Burkina Faso, West Africa". In: *Journal of Applied Meteorology and Climatology* 53 (2014), pp. 598–613.

[150] Alyssa K Whitcraft, Inbal Becker-Reshef, and Christopher O Justice. "Agricultural growing season calendars derived from MODIS surface reflectance". In: *International Journal of Digital Earth* 8 (2015), pp. 173–197.

[151] Jeffrey W White et al. "Methodologies for simulating impacts of climate change on crop production". In: *Field Crop Research* 124 (2011), pp. 357–368.

[152] Harold Willis. *Soybean plant biology: History, plant structure growth cycles*. `https://www.ecofarmingdaily.com/grow-crops/grow-soybeans/soybean-crop-science/biology/`. 2019.

[153] Alexandre C Xavier, Carey W King, and Bridget R Scanlon. "Daily gridded meteorological variables in Brazil (1980–2013)". In: *International Journal of Climatology* 36 (2016), pp. 2644–2659.

[154] Viviana Zalles, Matthew C Hansen, Peter V Potapov, et al. "Near doubling of Brazil's intensive row crop area since 2000". In: *Proceedings of the National Academy of Sciences* 116 (2019), pp. 428–435.

[155] Linglin Zeng et al. "A hybrid approach for detecting corn and soybean phenology with time-series MODIS data". In: *Remote Sensing of Environment* 181 (2016), pp. 237–250.

[156] Xiaoyang Zhang et al. "Monitoring vegetation phenology using MODIS". In: *Remote Sensing of Environment* 84 (2003), pp. 471–475.

[157] Liheng Zhong et al. "Automated mapping of soybean and corn using phenology". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 119 (2016), pp. 151–164.

Appendix A

Supporting Information for Chapter 2

This Supporting Information details (1) the complex/nonlinear fitting algorithms that I tested against my linearized harmonic fitting algorithm and (2) the derivation of planting and harvest dates from Planet Labs imagery.

A.1 Complex/nonlinear fitting algorithms

In Chapter 2, I argued that my simple, linear timeseries analysis method extracts phenological parameters from MODIS images without significant loss of estimation accuracy compared to complex or nonlinear methods. The three nonlinear/complex methods are: (1) nonlinear 1st order harmonic, (2) linearized 3rd order harmonic, and (3) Savitsky-Golay filter. Here, I describe the calculation of peak date and quarter period from these alternative methods.

The Savitsky-Golay filter is a third-order polynomial fit over a manually adjusted moving average window ranging from 5 to 11 points, and is implemented with R's sgolayfilt() function. The linearized 3rd order harmonic function is fit over the whole growing season (of both the first and second crop if present) assuming a frequency equal to 6π/N (units of yr^{-1}), where N is the number of data points in the timeseries. It is necessary to fit over both peaks because a 3rd order harmonic contains too many estimated parameters to successfully fit the roughly 20 satellite observations in a single crop's EVI profile. Finally, the 1st order harmonic is fit with R's nls() nonlinear regression function, with an initial guess of frequency at 0.15 yr^{-1}. Each function is fit to cloud-filtered, unsmoothed, EVI timeseries.

After fitting each function to the EVI timeseries, phenological parameters are extracted from the smoothed EVI profile. For all fitting functions, the peak date is set as the date that fitted EVI reaches its maximum fitted value. The period is extracted from the 1st order harmonic curve as a simple inversion of the estimated frequency: $period = 365/2\pi\omega$, which is then divided by 4 to obtain the quarter period. However, because quarter period is not explicitly calculated in the Savitsky-Golay filter and linearized 3rd order harmonic methods in TIMESAT, I approximate it as half the distance between (1) the peak day and (2) the date of minimum fitted EVI to the left of the peak day.

When the greenup date is defined as the maximum EVI date minus a numerically estimated or fitted quarter period (as with 1st order and linearized 1st order harmonic), its location is skewed towards the date of the minimum EVI. This indicates that the EVI profile of soy does not stretch a full harmonic cycle, and the "quarter period" is equal to a quarter of the fitted harmonic cycle, not a quarter length of the crop cycle itself.

A.2 Planting and harvest dates derived from Planet Labs images

Planting and harvest dates derived from Planet Labs images are used to assess the pixel level accuracy and robustness of my estimation method over fields which are known to be soy. I upload these Planet Labs images to Google Earth Engine, delineate individual fields by hand, and record the earliest and latest possible planting and harvest dates for each field. Figure A.1 displays some Planet Labs imagery samples, highlighting planting and harvest detection and the influence of clouds on data creation. The final dataset is a set of images depicting the earliest and latest possible planting and harvest dates for soy and, if it exists, for the second crop. This dataset is used to evaluate the pixel-level accuracy of my estimation method, in the absence of additional error introduced by the land cover map at the regional level.

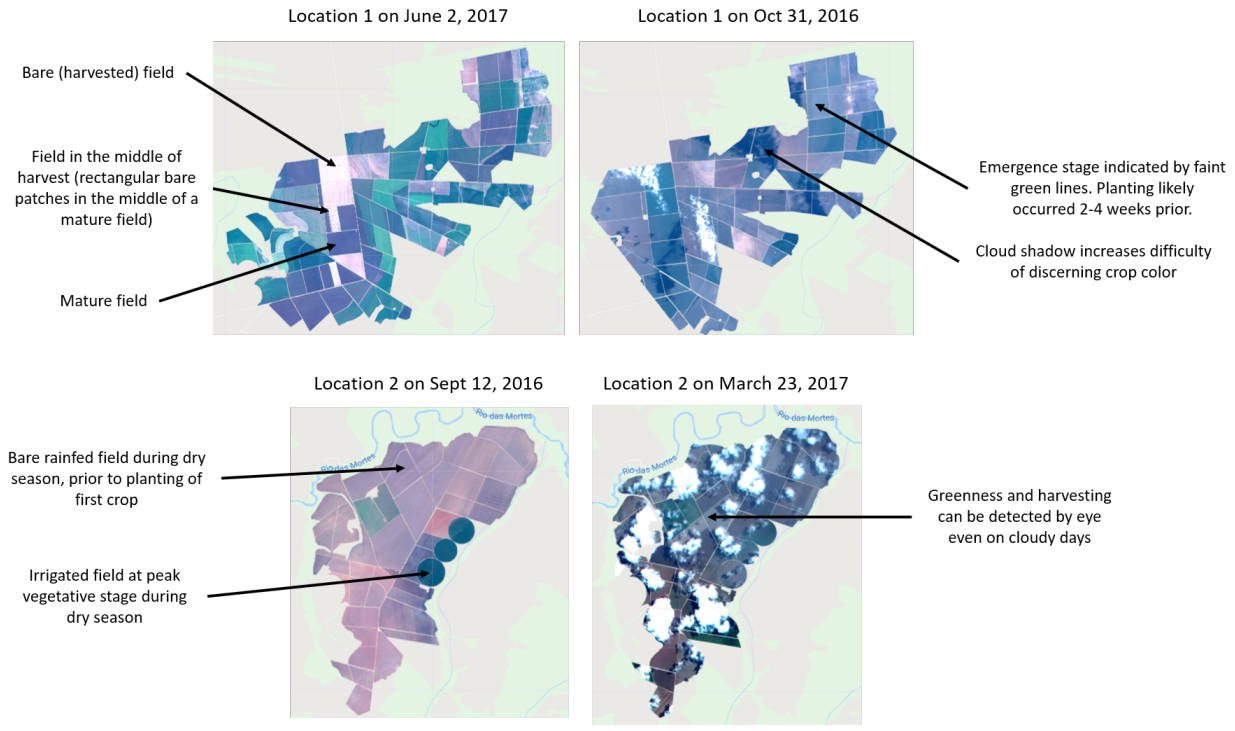


Figure A.1: Planet Labs images from two locations in Mato Grosso, ranging from the start of the growing season (September) to the end (June) illustrate the visual cues that were used to estimate planting and harvest dates for each field. Clouds and cloud shadows impacted the quality of the estimates. Locations are numbered following Figure 2.3.

Appendix B

Supporting Information for Chapter 3

A spatially explicit soy crop calendar requires not only a method that estimates planting and harvest dates at individual pixels, but also a crop cover map that specifies the location of soy, allowing us to speak about differences in management for the two cropping systems. In this Supporting Information, I provide details on the classification process.

B.1 Quality control of training points

The classification process begins with quality control and pooling of the crop training points. Since all training points are, in theory, located over soy and other agriculture, I mask out any points that fall outside of the Mapbiomas agriculture class. This reduces the risk that a few misplaced training points at the edge of agricultural patches would bias the classifier. While there are thousands of training points, the vast majority of them are classed as double cropped soy (a proportion that's reflective of the land cover of Mato Grosso in general), and only a minority are classed as single cropped soy and other agriculture. To increase the number of training points belonging to these other classes, I pool all crop points from 2003 to 2017 to train a single classifier. This classifier is then used to classify the crop cover in all years. I choose this pooled approach over the alternative of training separate classifiers for individual years because it increases the accuracy of the single cropped soy and other agriculture classes, giving an overall accuracy boost of 3% compared to training and classifying each year separately. Finally, irrigated pixels do not follow the same phenological patterns as rainfed pixels, and are masked out to maintain the accuracy of the land cover classification. Fields that experience high EVI during the dry season are not treated well in a classifier trained on phenospectral information of rainfed crops.

B.2 Crop cover classification

Next, I calculate the input data over the quality controlled and pooled training points. The set of input data used in classification, chosen from a stepwise process described in the

next section, are pictured in Figure 3.2. To retrieve phenological parameters, I perform the timeseries analysis from Step 2 (described in Chapter 2) on MODIS-derived EVI assuming that all pixels are double cropped, giving an estimate of the seasonality and cycle length of the first and (hypothetical) second crop. Attempting to calculate phenological parameters for the second crop on a pixel that is in reality single cropped produces phenological dates that are out of reasonable bounds, an indication that the pixels is, in fact, single cropped. Phenological information also helps to distinguish soy from crops with very different cycle lengths and seasonalities. However, phenological information alone may not be enough to separate soy from other agriculture: in Mato Grosso, it is likely that phenological and seasonality variations within soy varieties are as large as phenological variations among different crop species. Therefore I use spectral information, retrieved at specific crop developmental stages, to supplement the phenological information. I use median EVI calculated over 8 day windows surrounding the phenological stages highlighted in Figure 3.2 to separate crop types based on their physiological properties. I calculate a median EVI over a small time window instead of interpolating EVI to a specific phenological date in order to decrease the input data's sensitivity to noise. Similarly, the input data do not include surface reflectances between the first and second peak dates because this period is frequently cloudy, causing the classifier to be overly sensitive to the infill technique and noise.

A Cartesian classifier, chosen over a random forest classifier for its higher overall accuracy, is trained and tested in GEE using the training dataset. Classification accuracies are calculated using repeated cross-validation in which 30% of the data are randomly removed from the training points and used for testing. Using the phenospectral input dataset and Cartesian classifier, the crop cover map achieves an overall accuracy of 82.2 +/- 0.5%. Consumer's and producer's accuracies are displayed in Table B.1.

B.3 Input data selection

The set of input data used in classification is selected through a stepwise process. In addition to the pheno-spectral input dataset that is ultimately used to construct the soy land cover map, I consider other sets of input information. These fall under two categories: (1) spectral data reported based on calendar month; and (2) phenological information derived from timeseries analysis.

The first category, spectral data, consists of 8-day composite surface reflectances for a set of MODIS bands. I test several spectral band and date range combinations in order to select a combination that describes the full range of spectral differences among the crop cover classes without introducing unnecessary information and noise. For example, Figure B.1a shows that the spectral differences among the three crop classes are negligible during the dry season and beginning of the wet season, and Figure B.1b shows that bands 1 and 2 (red and NIR) appear most different among the three classes. Of the three spectral band combinations (Bands 1-7; Bands 1 and 2; and EVI), and two date ranges (all year from Aug 1 to July 31; and wet season from Dec 1 to June 1) tested, the EVI-wet season combination

produced the highest classification accuracy of 76.7%. All variations of spectral input data result in relatively low overall accuracy in the mid-70s because they all rely on a constant crop seasonality over time and space. To resolve this issue while continuing to use only spectral data would require classifiers specific to each region and year, which complicates the analysis and requires more training data than are available.

The second category, phenological data, allows more flexibility for crop seasonality by training the classifier on phenological stages derived from timeseries analysis. The peak EVI date, crop cycle length, fitted amplitude of the EVI curve, and peak EVI for the first and second crop are used as the input data, giving an improved classification accuracy of 81%. However, relying on phenological data alone assumes that crop species have pronounced differences in crop seasonality and cycle lengths; adding spectral information improves classification accuracy by allowing the physiological differences among crop species to be considered during classification. Therefore, I use both phenological and spectral input data to ensure that the classifier is flexible to variations in cropping seasonality and sensitive to physiological differences between soy and non-soy crops. Table B.1 summarizes classification performance for pheno-spectral, phenological, and spectral input data.

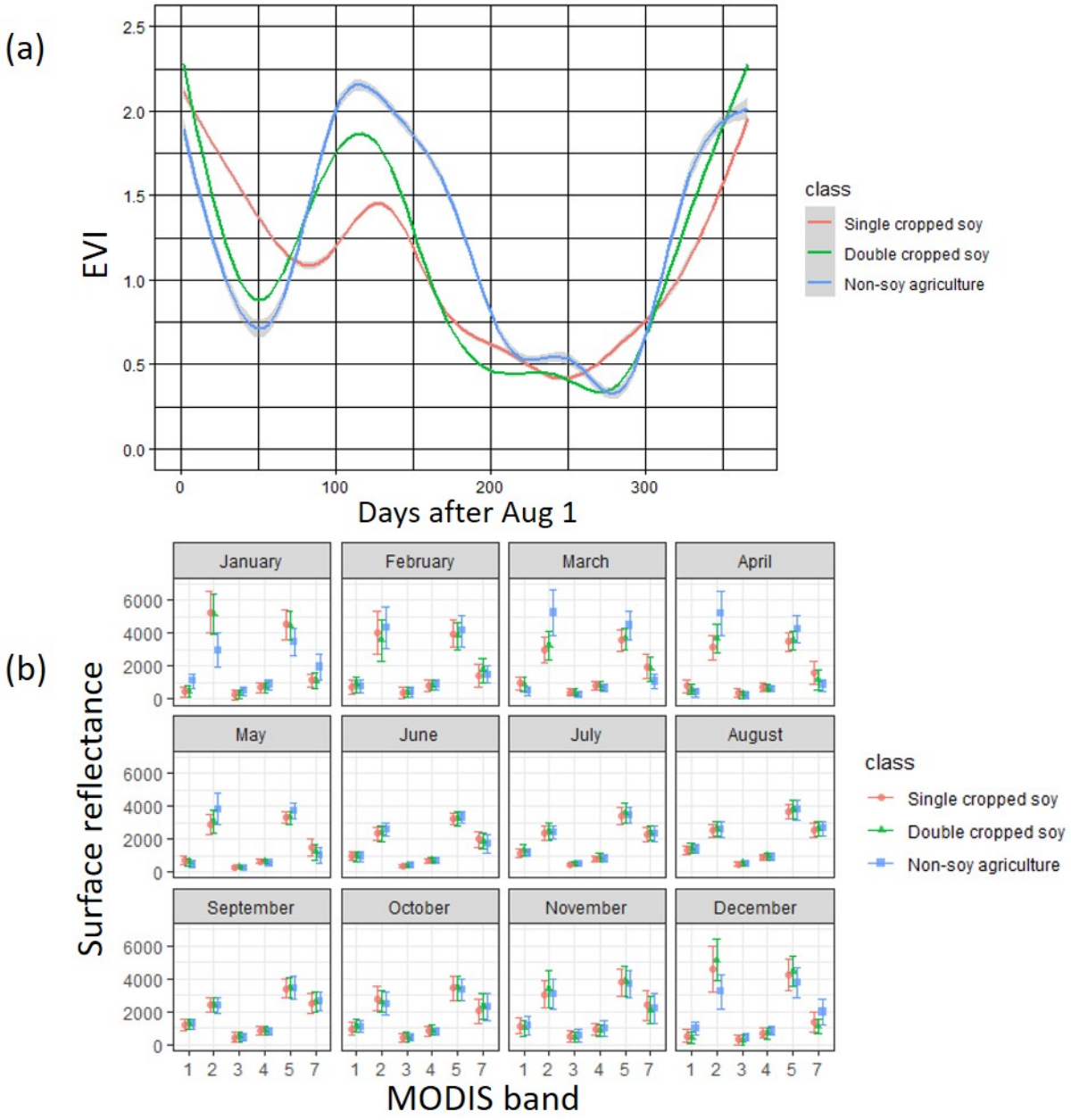


Figure B.1: (a) The majority of the variation among the classes occurs between December and June. Gray intervals represent standard deviation. (b) The majority of the variation among the classes is NIR (band 2) and red (band 1) during December to June. This indicates that EVI, which incorporates NIR and red, is a good multispectral index to separate the classes. Error bars represent standard deviation.

	Phenospectral input data	Phenological input data	Spectral input data
Phenological information	Fitted amplitude, peak date, crop cycle length, peak EVI value	Fitted amplitude, peak date, crop cycle length, peak EVI value	None
Spectral information	Cloud filtered EVI from MODIS at phenological stages	None	MODIS bands 1-5, 7
Classification accuracy	82.5	81.8	73.5
Producer's accuracy	0.44 (SC soy) 0.92 (DC soy) 0.44 (other agriculture)	0.44 (SC soy) 0.92 (DC soy) 0.4 (other agriculture)	0.35 (SC soy) 0.87 (DC soy) 0.3 (other agriculture)
Consumer's accuracy	0.59 (SC soy) 0.86 (DC soy) 0.65 (other agriculture)	0.58 (SC soy) 0.86 (DC soy) 0.57 (other agriculture)	0.45 (SC soy) 0.8 (DC soy) 0.35 (other agriculture)
Insight	Phenospectral data uses the largest amount of relevant information, giving it the highest accuracy.	Phenological information alone produces the largest jump in accuracy.	Spectral data, when not aligned with phenological stages, produces the lowest accuracy.

Table B.1: Pheno-spectral input data has higher accuracy than phenological or spectral input data alone.

Appendix C

Supporting Information for Chapter 4

This Supporting Information (1) details the model selection process, (2) provides additional model results, and (3) repeats the the model selection and model results using the $\text{frequency}_{8,\ \text{CHIRPS}}$ onset definition. The $\text{frequency}_{8,\ \text{CHIRPS}}$ definition performs best for single cropped soy for all except the 5th percentile. I show that the conclusions under this alternative onset definition are similar to those reported for the chosen definition, $\text{frequency}_{10,\ \text{PERSIANN}}$.

C.1 Model selection

Model specification choices such as the observation scale, predictor set, handling of residual autocorrelation, and model type are made through a series of exploratory regressions. The sections below describe the model selection steps, and report results using $\text{frequency}_{10,\ \text{PERSIANN}}$ as the onset definition.

C.1.1 Observation scale selection

I first choose the observation scale for the models tested. Four observation levels of planting date are available: pixel-scale, cell-scale, property-scale, and municipality-scale. In order of increasing size, pixel-scale planting dates are available at the raw resolution of MODIS, 500 m; cell-scale planting dates correspond to the resolution of onset estimates (5 km or 25 km cells for estimates derived from CHIRPS and PERSIANN, respectively); property-scale planting dates are aggregated to the level of individual farm properties (average area of 5 km^2); and municipality-scale planting dates are aggregated to the level of individual municipalities (average area of 3,192 km^2). In Chapter 3, I observed high variation in planting date within farm properties, suggesting the need for field-level observation. However, this is impossible given the much larger grid size of CHIRPS and PERSIANN (5 km and 25 km) precipitation data.

I choose the observation scale that appropriately balances noise in the planting date estimates (based on R^2 of the fitted linear model), leverages the spatial information available

in onset estimates, and reflects the expectation of the scope at which planting dates respond to onset. Because, at this point in model selection, I have not yet chosen the model type or predictors, the R^2 is calculated with exploratory pooled OLS regression with the predictors onset, year, longitude, and latitude. This OLS regression is intended to explore the proportion of variability at each observation scale that can be explained by available predictors; I do not use the results of the model for prediction or inference.

Figure C.1 shows planting date data aggregated to each of the observation scales considered. The scale of the onset estimates is the most appropriate observation scale because it reduces the variance in planting estimates that plague the smaller pixel scale, without eliminating important spatial details by aggregating to the scale of municipalities. Farm property areas are on the same order of magnitude as those of onset cells, and in theory would be the most suitable observation scale because planting dates are decided at the property level. Characteristics of a property, such as household size, property rights, insurance access, risk tolerance, and access to agricultural credit and equipment may all influence the planting date. Larger farms are more readily able to access to machinery and seeds, and therefore less susceptible to delayed planting. However, larger farms may also require more time to physically plant, which may cancel out the benefit of timely resource access. In Brazil, soy cropland may reach 10,000 ha in area and require 2 - 4 weeks to complete planting [106]. Additionally, farmers may mix crop varieties in order to distribute the risk of crop failure (planting early maturing varieties to hedge for a short wet season, and late maturing varieties for their higher yields), and larger farms may devote more land to experimental fields [68, 23]. These effects may be approximated by property size. However, an exploratory regression reveals no direct relationship between the size of the farm and farm-averaged planting dates. The range of planting dates within a property is also uncorrelated with the size of the property. Additionally, at the property scale, the model explains very little variability because planting date variability is extremely high - the R^2 of an OLS model is 0.25. The dominant controls on planting date at the farm scale are, therefore, too fine to be captured by the data. At the municipality scale, much more of the variability is explained in the model (R^2 of 0.47), but the aggregation eliminates much of the meaningful spatial variability in planting date observations and may hide important relationships that emerge at finer scales. A highly aggregated observation scale may be unable to capture real local differences in how growers respond to planting decision drivers.

Onset cells are therefore the most appropriate observation scale for planting dates, with R^2 of 0.56 at the 25 km scale and 0.31 at the 5 km scale. This scale exploits the full spatial variability of available onset data (the predictor of interest) without introducing unnecessary variability in the planting observations or eliminating real spatial patterns through over-aggregation. Because I explore two precipitation datasets of different spatial resolution, the scale of the onset is either 5 km (for estimates derived from CHIRPS) or 25 km (for estimates derived from PERSIANN).

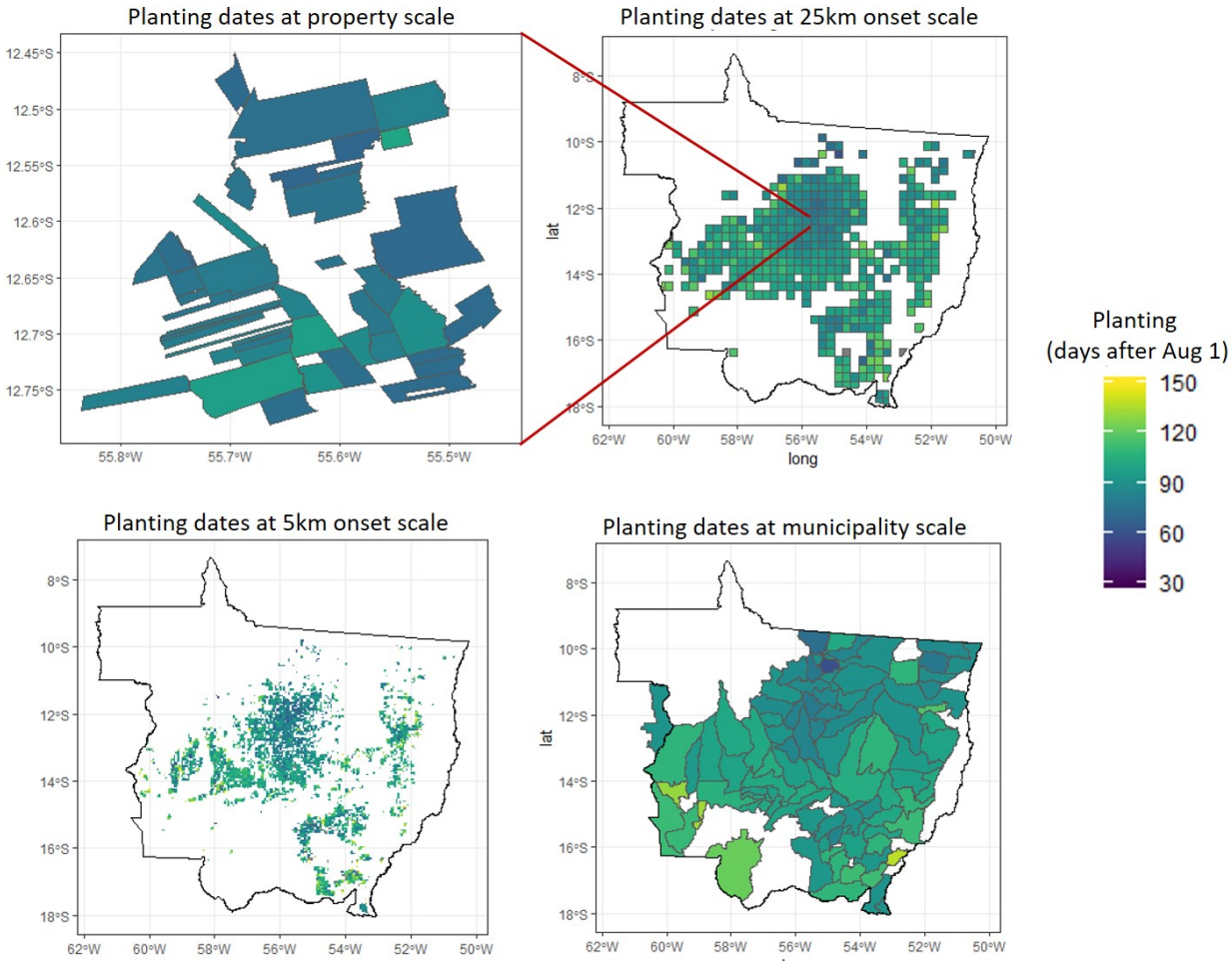


Figure C.1: Planting dates from 2014, aggregated to each of the three observation scales considered. The property scale map is shown over the area of a single onset cell. The onset data scale was chosen as most appropriate.

All subsequent modeling choices are made on the basis on the selected observation scale.

C.1.2 Predictor selection

The full set of predictors considered for model selection, each of which is a potential confounding factor that may influence the estimated onset sensitivity, are listed below. The list includes interaction terms, which are denoted with a colon (:) between the interacting predictors.

1. Wet season onset. Onset is reported as days after August 1 of planting year, and calculated using a variety of onset definitions and precipitation datasets.
2. Year
3. Cropping intensity (single or double cropped)
4. Latitude, longitude
5. Region: Mato Grosso was split into four "regions": central, west, east and south, according to broad spatial trends in planting date. The boundaries for these regions were chosen following exploratory analysis and initial regression, which suggested that the central region has an earlier "baseline" planting date, independent of the onset, compared to the other three regions. This nonlinear effect is not captured in the longitude term.
6. Percentile of planting date (5th, 25th, 50th, 75th, 95th percentile of planted pixels within each observation cell)
7. Onset:cropping intensity
8. Onset:percentile
9. Onset:latitude
10. Onset:longitude
11. Previous year's onset: Climate information in the previous year may be incorporated into farmers' memory, and impact planting date in the subsequent year.
12. Long-term total annual rainfall from 2004 to 2014: This is used as a proxy for spatial patterns in climate that aren't captured in other geographic predictors (latitude, longitude, and region).

I use a series of exploratory pooled OLS regressions to select among these predictors (though the estimated coefficients are not used for predictions or inference). With onset fixed as one of the predictors in all exploratory regressions, the other possible predictors are added stepwise to a pooled OLS model based on adjusted R^2. The top five predictors selected are onset, percentile, cropping intensity, year, and region.

The interaction terms onset:percentile and onset:intensity are also statistically significant, but not among the first five selected. While these interaction terms explain less of planting date variability, they are interesting to this work because they suggest that cropping intensity impacts farmers' sensitivity to onset, and that substantial variation onset sensitivity exists within each observation cell.

I choose to model single and double cropped soy separately because they they are expected to behave differently in response to the physical constraints. Though planting date for rainfed

crops is restricted by wet season onset, it is also restricted at wet season demise by the need to keep plants from water stress during the grain filling phenological stage. To ensure that the crop reaches maturity before the end of the wet season, planting cannot occur too late. This "too late to harvest" threshold is earlier for double cropped fields because the first crop, soy, must be planted and harvested in time for the second crop, usually corn, to reach maturity [1]. In contrast, single cropped soy may be planted under a more flexible timeline. Therefore it's expected that single and double cropped soy indeed have different levels of sensitivity to onset, an assumption confirmed by the statistical significance of the onset:intensity term.

Similarly, I choose to model each percentile of planting date separately because the statistically significant onset:percentile term indicates that fields planted earlier (5th percentile) are more reactive to onset than fields planted later (95th percentile). Separate models for each percentile will characterize their varying degrees of sensitivity.

In the final model specification, I explore ten versions of the planting date as the dependent variable: a combination of two cropping intensities and five percentiles. These separate models will quantify the differences in onset sensitivity for each intensity and percentile.

All subsequent modeling choices are made on the basis of the selected observation scale and predictor set. All five predictors were used in the OLS_{pooled} and random forest specifications; however, in the OLS_{FE} specification, the location-based predictor (region) was eliminated because spatial information is subsumed within the fixed effects term.

C.1.3 Autocorrelation of residuals

Independent residuals are an important assumption for OLS_{pooled} and OLS_{FE} models. Residuals that are spatially or temporally autocorrelated violates this independence assumption and produce a form of pseudo-replication, causing the p-values associated with the estimated coefficients to be artificially small and elevating the risk of inferring a statistically significant coefficient where there is none. A model whose dependent variable (planting date) has an intrinsic autocorrelation structure that cannot be captured in the predictors must explicitly account for that autocorrelation.

I quantify the temporal autocorrelation of the residuals using the Durbin-Watson test with Bonferroni correction, and the spatial autocorrelation with Moran's I. While no residuals are temporally autocorrelated for models of all cropping intensities and percentiles, they were spatially autocorrelated for OLS_{pooled} in all years, intensities and percentiles. Moran's I statistics ranged from 0.247 to 0.156, corresponding to p-values of 0.002 to 0.038. Spatial autocorrelation was not found for OLS_{FE} residuals for any years, intensities and percentiles, with Moran's I of -0.038 to 0.023, corresponding to p-values of 0.629 to 0.386.

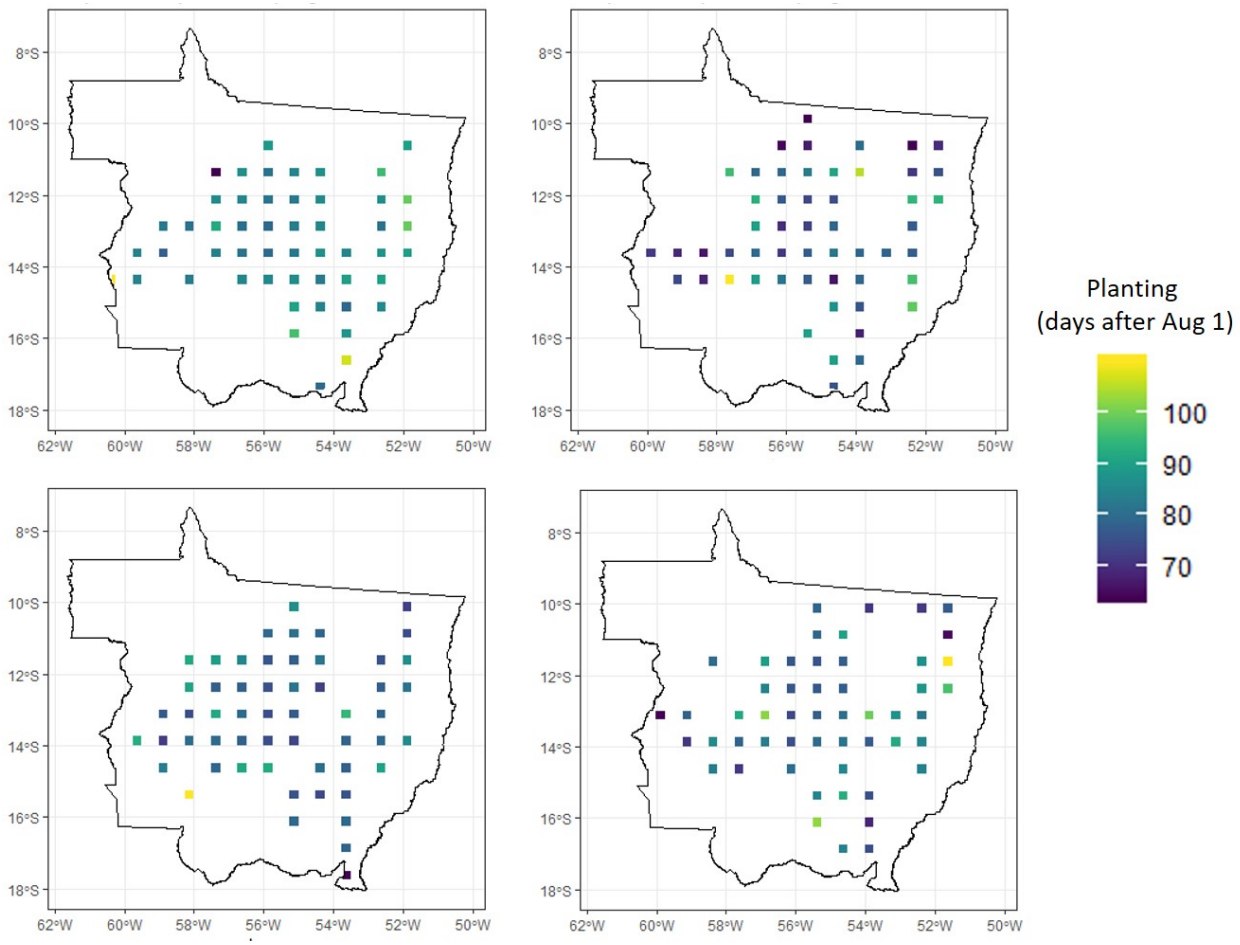


Figure C.2: Sampled planting dates for DC soy in 2014 based on different sampling grid positions.

To eliminate spatial autocorrelation in $\text{OLS}_{\text{pooled}}$ residuals, I sample the observations with a sampling grid so that spatially adjacent observations are never included in one model. The size of the intervals was chosen to: (1) avoid spatial autocorrelation in the residual, while (2) maximizing prediction accuracy in new years and locations, (3) maximizing the percent of total data used in each model, (4) minimizing the coefficients' sensitivity to sampling grid position. Figure C.2 shows examples of sampling grid locations tested.

The process of spatial sampling creates additional uncertainty that's associated with the location of the sampling grid, which affects the specific data points that are selected for the model. To account for this uncertainty, I offset the grid by 25 km increments in the latitude and longitude directions. I fit a separate model for each position of the sampling grid, and report the uncertainty in estimated coefficients due to both the sampling grid position and

the standard error of the individual estimates. Figure C.3 shows the four criteria that were used to select a sampling grid size for double cropped soy at 25th percentile, where error bars represent uncertainty due to sampling grid position. Because the observations chosen from the sampling grid depends on the location of the sampling grid itself, I choose the interval length that meets the four criteria for all possible shifts of the sampling grid position.

I select a sampling grid size of 75 km, an interval large enough to eliminate of spatial autocorrelation, but small enough to maximize the percent of total data points used in each model and therefore improve model consistency under changing sampling grid locations. Its high prediction accuracy is insensitive to sampling grid location, and its onset coefficients are insensitive to sampling grid location.

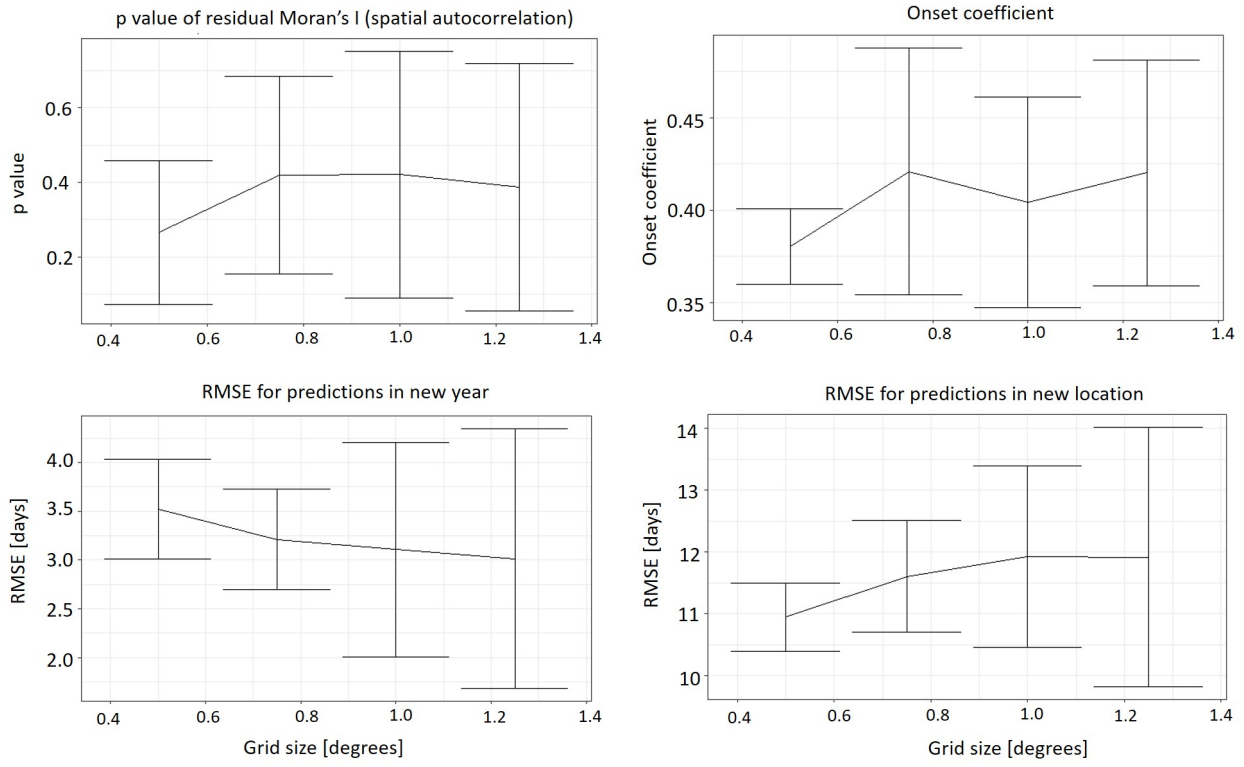


Figure C.3: The optimal sampling grid size is 75 km for the 25 km observation scale. As grid size increases, residual autocorrelation declines but uncertainty due to the sampling grid location increases. Results for double cropped soy at 25th percentile are shown here.

The final choice of model type is made on the basis of the selected observation scale, predictors, and autocorrelation handling.

C.1.4 Model type selection

I explore three model types: (1) pooled ordinary least squares regression (OLS_{pooled}), (2) OLS regression with fixed effects (OLS_{FE}), and (3) random forest.

OLS_{pooled} and OLS_{FE} models attempt to fit the response variable (planting date) as a linear function of the predictor variables, providing estimates of the intercept and slopes (coefficients). OLS_{FE} regression extends OLS_{pooled} to control for unobserved explanatory variables that are constant over time but variable over space.

Random forest is a machine learning model built on a collection of independently trained decision trees, whose nodes are split on a randomly chosen subset of predictors. Here, I train 600 individual trees, with each node split on two randomly chosen predictors. The decision tree structure allows random forest models to account for complex interactions and nonlinear relationships between planting date and its predictors, a significant advantage over OLS_{pooled} and OLS_{FE}. However, OLS_{pooled} and OLS_{FE} provide interpretable coefficients of how planting date changes with each of the predictors, while random forest operates like a black box. Due to lack of interpretability, a random forest model with missing predictors and low prediction accuracy cannot be used to draw conclusions about the onset-planting date relationship. In contrast, while missing predictors would deter OLS_{pooled} and OLS_{FE} predictions, the models can still produce interpretable, unbiased estimates the relationship between onset and planting date if a set of assumptions are met. Random forest models are unable to produce these generalized insights.

Because the model will be used to predict how growers will change planting date in response to climate change, predictive ability is an important model selection criterion. Three prediction accuracy metrics informed the selection of model type: (1) validation RMSE following a randomized 70%-30% train-test split of the spatiotemporal data; (2) prediction RMSE at individual locations whose data was eliminated (for all years) from training; and (3) prediction RMSE during individual years whose data was eliminated (for all locations) from training. These cross-validation metrics allow me to evaluate the models' predictive accuracy not only under standard validation conditions, but also under previously unseen locations and years. Errors associated with specific years help to define the models' predictive ability in early-onset versus late-onset years. High predictive ability in late-onset years is crucial for estimating how planting date will respond to biophysical climate change. Additionally, I consider interpretability and robustness to missing predictors.

I test the predictive ability of the OLS_{pooled}, OLS_{FE} and random forest models under the chosen observation scale of 25 km; the $\text{frequency}_{10,\,\text{PERSIANN}}$ onset; and predictor set of onset, year, region, percentile and cropping intensity. Because the OLS_{FE} specification includes separate intercepts for each observation, I eliminate the location-based predictors for the OLS_{FE} model. The prediction accuracy for the ten cropping intensity $\times$ percentile combinations are calculated for all three model types; the OLS_{pooled} specification additionally has varying prediction accuracy depending on the position of the sampling grid. The prediction accuracy for each model type, summarized for double cropped soy at the 25th percentile for all sampling grid positions, is reported in Figure C.4.

The random forest has lower prediction accuracy during new years, a pattern especially pronounced in 2010, which experienced a much earlier onset than the other years. However, as shown in Table C.1, the random forest model also performs best for predictions in new locations. While random forest's prediction accuracy in new locations improves when it is trained using more data (by not sampling the onset grid, and using all spatially adjacent data points for training), more data does not help it predict in new years. The inability of random forest to predict in new years makes it less relevant for this work.

	Random Forest (RMSE)	OLS_{pooled} (RMSE)	OLS_{FE} (RMSE)	Random Forest (Error)	OLS_{pooled} (Error)	OLS_{FE} (Error)
Prediction in new cells	7.91 +/- 3.56	11.66 +/- 3.11	N/A	2.52 +/- 4.47	7.81 +/- 6.01	N/A
Prediction in randomly selected test set	13.25	13.57 +/- 0.09	12.61	0.31	-0.08 +/- 0.82	-0.18

Table C.1: Prediction error for each model type, for double cropped soy at 25th percentile. Standard deviation for eliminated cells represent variation in prediction error across different eliminated cells, and standard deviation for OLS_{pooled} represent variation in prediction error for different sampling grid locations. It is not possible to predict in new cells for the FE specification.

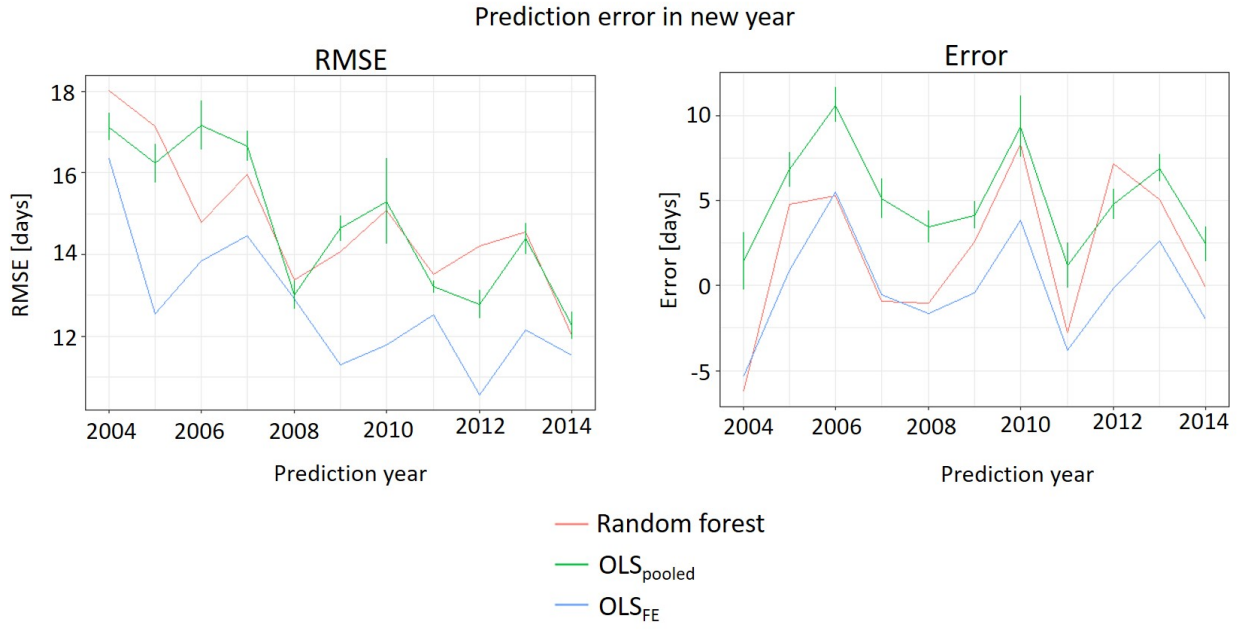


Figure C.4: Prediction accuracy for double cropped soy at 25th percentile. For OLS_{pooled} models, error bars include the effect of shifting sampling grid locations.

I choose the OLS_{FE} model because it balances the simple interpretability of OLS_{pooled} models while accounting for the nonlinear spatial pattern in baseline planting dates. Random forest models could have captured this spatial nonlinearity, but due to the likely incomplete set of predictors, they have no predictive advantage over OLS_{pooled} and OLS_{FE} models. While random forest models predict better under the randomized train-test split scenario, they perform more poorly during new years and locations compared to OLS_{pooled} and OLS_{FE} specifications. I choose to prioritize performance accuracy under new years because the model is used for prediction in Chapter 5. While missing predictors would impact predictive accuracy for all three model types, the ability of OLS_{pooled} and OLS_{FE} to isolate the effect of onset on planting date makes it more useful for this work.

C.2 Estimated fixed effects

The fixed effects estimated from the final model specification are shown in Figure C.5. There is a complex spatial pattern in the fixed effects, which indicates earlier baseline planting in the center of the state and later planting towards the edges. This is a nonlinear pattern that would not have been captured in OLS_{pooled} with spatial predictors like latitude, longitude,

and region. The fixed effects capture geographically complex baseline differences in planting date across space.

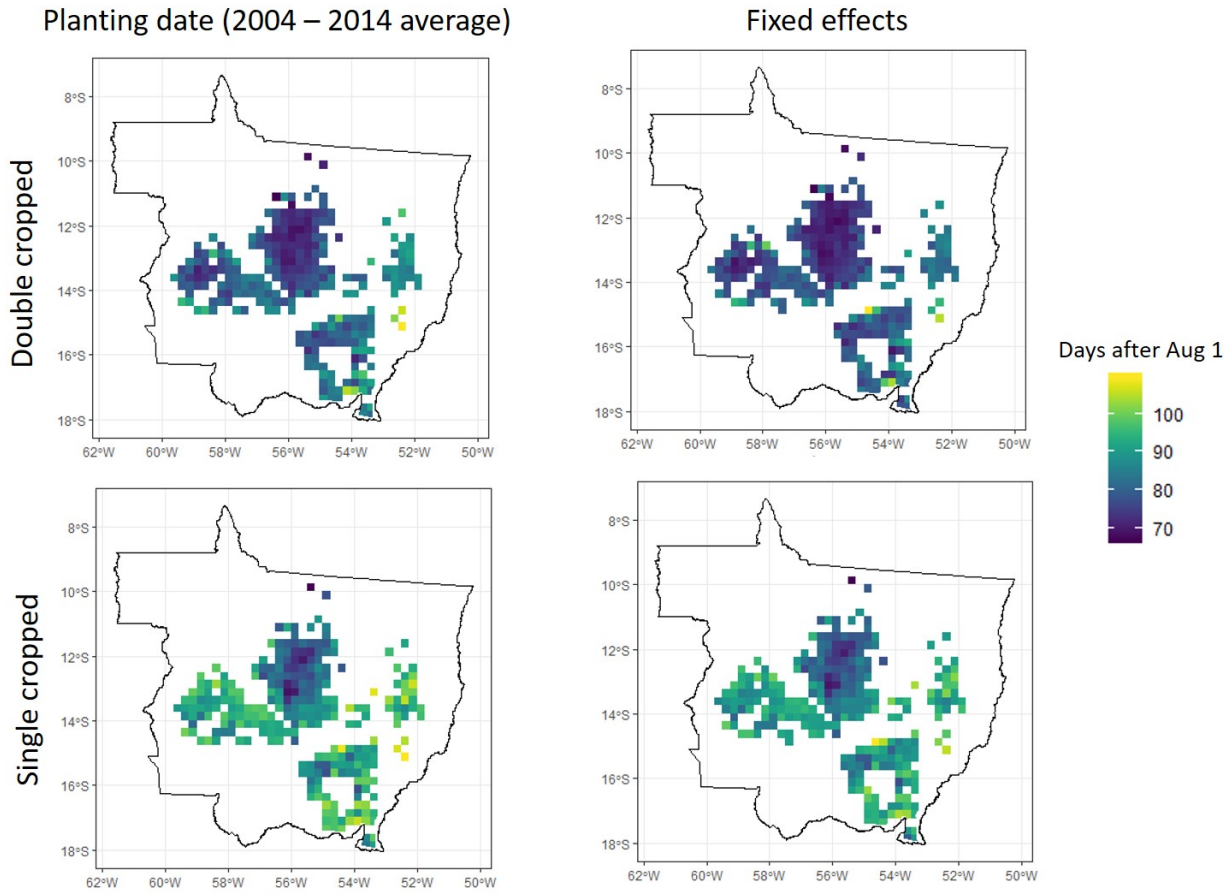


Figure C.5: The fixed effects for both single and double cropped soy have a complex spatial pattern. Here, the OLS_{FE} model specification was run for all observations to allow fixed effects to be fit for every observation. The 25th percentile of planting (averaged across 2004 to 2014) is shown here, but a similar pattern is observed for other percentiles.

The statistical significance of fixed effect terms is confirmed with an F test for individual effects, for each intensity and percentile. The test confirms that planting dates are affected by a time-invariant, location-varying factors which are not captured in simple location indicators such as latitude, longitude, and region. However, the OLS_{FE} specification still only has an R^2 between 0.42 and 0.56 (depending on the intensity and percentile), suggesting the possibility of spatiotemporally varying predictors not captured in the model.

C.3 Methods and results under an alternative onset definition

Chapter 4 and the sections above report the outcome of the exploratory regressions, statistical tests, and modeled results for the $\text{frequency}_{10,\ \text{PERSIANN}}$ onset definition, but similar results are obtained for all other onset definitions. The choice of onset definition had no impact on the choice of observation scale, predictor set, and model type. Results for the onset definition most highly correlated to single cropped soy planted in the 25th percentile or later, $\text{frequency}_{8,\ \text{CHIRPS}}$, are shown in this section.

C.3.1 Model specification

Predictor set

The R^2 of OLS_{pooled} models are 0.34, 0.31, and 0.46 at the property, 5 km cell, and municipality scales, respectively. While the R^2 for the property and 5 km scales are similar, I choose the 5 km onset cell as the observation scale because the advantage of using properties as the scale is unclear: properties vary greatly in size, and there is no relationship between planting dates and property size. Using the 5 km cell as the observation scale, the same five predictors (onset, percentile, cropping intensity, year, and region) explain the most variation in planting dates.

Residual autocorrelation

Spatial autocorrelation of residuals for OLS_{pooled} is present (Moran's I of 0.28 to 0.35, corresponding to p-value of 0.002) and not present for OLS_{FE} (Moran's I of -0.04, corresponding to a p value of 0.79 to 0.95, depending on the intensity and percentile). Temporal autocorrelation of residuals for OLS_{pooled} and OLS_{FE} was not present, with maximum 0.002 percent of cells with temporal autocorrelation under Bonferroni correction.

To avoid spatial autocorrelation in OLS_{pooled}, I again sample the data according to a sampling grid. Because the $\text{frequency}_{8,\ \text{CHIRPS}}$ definition is calculated at the higher CHIRPS resolution of 5 km, I test a finer set of sampling grid sizes compared to the PERSIANN-derived onset (25 km). Figure C.6 shows that the optimal sampling grid size is 25 km. This size eliminates autocorrelation in the residuals without the loss of precision that occurs when the sampling grid size increases.

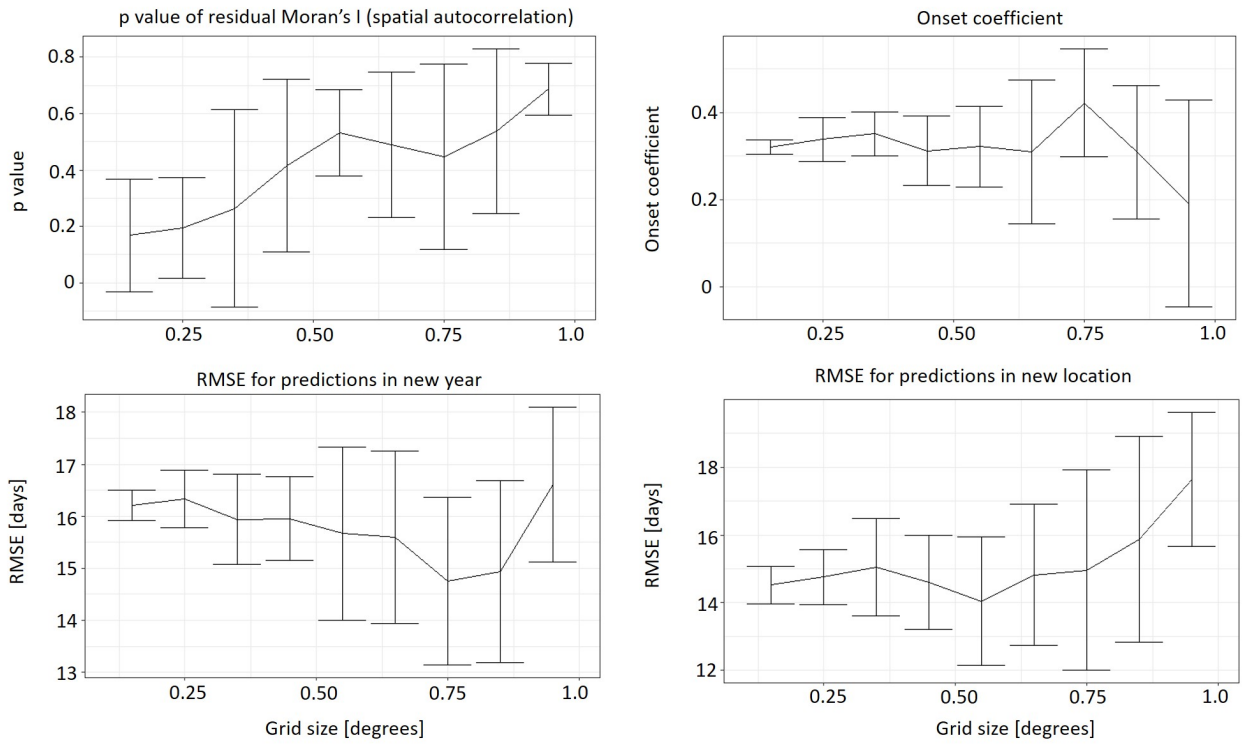


Figure C.6: The optimal sampling grid size is 25 km for the 5 km observation scale. Results for double cropped soy at 25th percentile are shown.

Model type selection

The prediction accuracy for each model type, summarized for double cropped soy at the 25th percentile, is reported in Table C.2 and Figure C.7. The OLS_{FE} model again performs best for predictions in new years, and was selected as the model type. The F test indicates statistically significant fixed effects for each intensity and percentile.

C.3.2 Model results

Model evaluation

I confirm that the fitted models (one for each cropping intensity, percentile and sampling grid location) satisfy the linear regression assumptions. Residual plots confirm that the residuals have zero mean, not correlated with the fitted value, and are homoscedastic; the QQ plot confirms that the residuals are normally distributed (Figure C.8); Durbin-Watson and Moran's I, reported in Section C.3.1, show that the residuals are not temporally or spatially

	Random Forest (RMSE)	OLS_{pooled} (RMSE)	OLS_{FE} (RMSE)	Random Forest (Error)	OLS_{pooled} (Error)	OLS_{FE} (Error)
Prediction in new cells	11.68 ± 4.91	14.79 ± 3.40	N/A	2.07 ± 4.38	6.67 ± 5.68	N/A
Prediction in randomly selected test set	14.98	16.02	15.59	-0.24	0.004	-0.20

Table C.2: Prediction error for each model type, for double cropped soy at 25th percentile. Standard deviation for eliminated cells represent variation in prediction error across different eliminated cells, and standard deviation for OLS_{pooled} represent variation in prediction error for different sampling grid locations. It is not possible to predict in new cells for the OLS_{FE} specification.

autocorrelated; and a correlation matrix shows that the predictors are not multicollinear and that residuals are exogeneous (Table C.3).

	Onset	Year	Residual
Onset	1	0.033	-1.3×10^{-3}
Year	0.033	1	4×10^{-5}
Residual	-1.3×10^{-3}	4×10^{-5}	1

Table C.3: Correlations show that predictors are not multicollinear and that residuals are exogenous. Results are reported for double cropped soy at 25th percentile.

Robustness and sensitivity tests

Table C.4 shows that the onset coefficient from OLS_{pooled} is robust to the elimination of known predictors. Even with all predictors but onset eliminated, the onset coefficient changes by a maximum of 0.06 compared to a model with the full set of known predictors. The OLS_{FE}

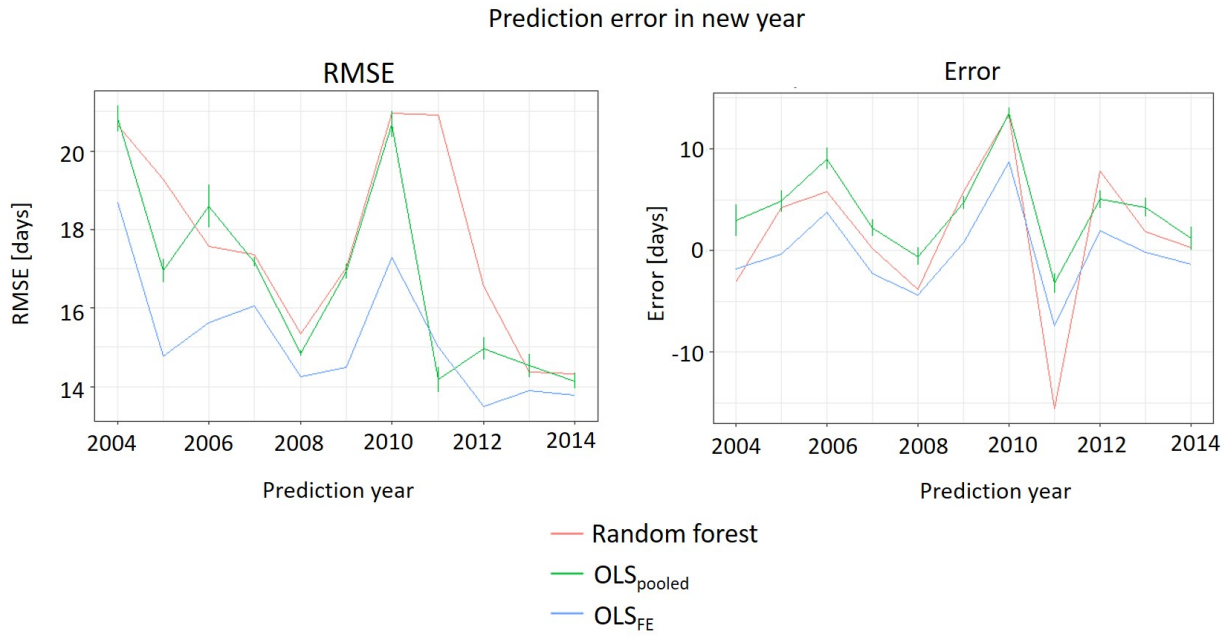


Figure C.7: Prediction accuracy for double cropped soy at 5th percentile. For OLS_{pooled} models, error bars include the effect of shifting sampling grid locations.

specification is similarly robust when year (the only other predictor) is eliminated: Figure C.9 shows that the OLS_{FE} specification across all intensities and percentiles is also robust when year is eliminated.

While these tests cannot guarantee that the onset coefficient is robust to the absence of unknown predictors, it is encouraging that the residuals are uncorrelated to onset (Table C.3), suggesting that the onset coefficient is unbiased.

	SC onset coefficient	DC onset coefficient
FE, all predictors	0.27	0.32
FE, only onset	0.25	0.30
OLS, all predictors	0.26	0.31
OLS, only onset	0.31	0.37

Table C.4: Onset coefficients estimated by OLS_{pooled}. "All predictors" means that onset, year, latitude, longitude, and region were used. The 25th percentile is modeled here.

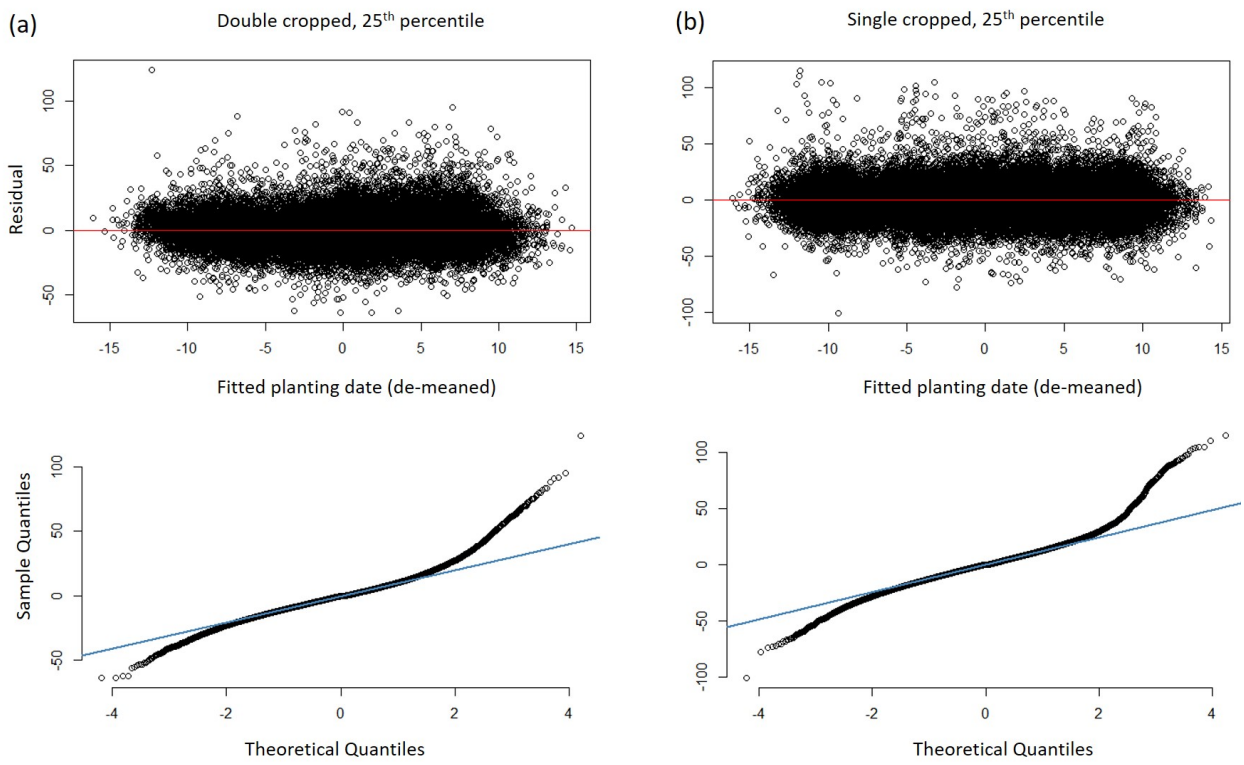


Figure C.8: Residual plots confirm normal, homoscedastic residuals.

Fitted model results

The onset coefficients for the chosen OLS_{FE} specification and onset definitions are summarized in Figure C.10. The error bars represent bootstrapped uncertainties in planting date estimates and the standard error of individual coefficient estimates. Similar to the $\text{frequency}_{10,\ \text{PERSIANN}}$ definition, the onset coefficient changes with the cropping intensity and planting date percentile. Onset coefficient is higher for soy that is planted early (double cropped soy and soy in the 5th percentile) compared to soy that is planted later (single cropped soy and soy in the 95th percentile). Though the spread in onset coefficients among percentiles for single cropped soy is less obvious for $\text{frequency}_{8,\ \text{CHIRPS}}$ than for $\text{frequency}_{10,\ \text{PERSIANN}}$, the differences in bootstrapped onset coefficients are still statistically significant.

Additionally, the results indicate that planting date became earlier with each successive year, independently of the onset. The year coefficients, shown in Figure C.10, are statistically significant at all cropping intensities and planting percentiles, indicating that the trend to earlier planting dates affects all soy growers.

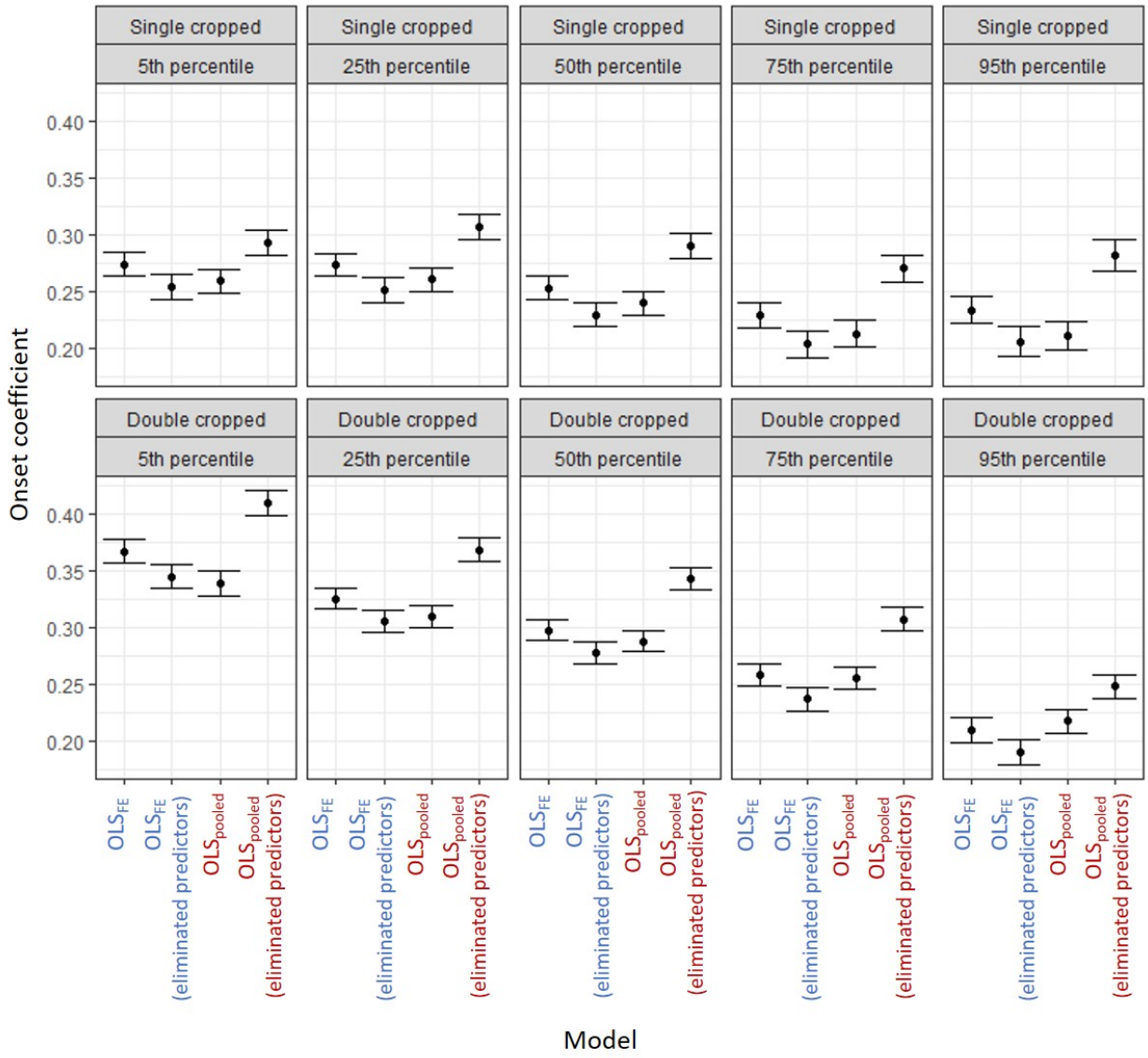


Figure C.9: Onset coefficient is robust to eliminated predictors in both the OLS_{FE} and OLS_{pooled} specifications. Error bars represent standard error.

Figure C.10 shows that despite uncertainty from errors in planting date data and standard error of the fitted coefficient, significant differences in the onset coefficient for different cropping intensities and eliminated predictors can be observed.

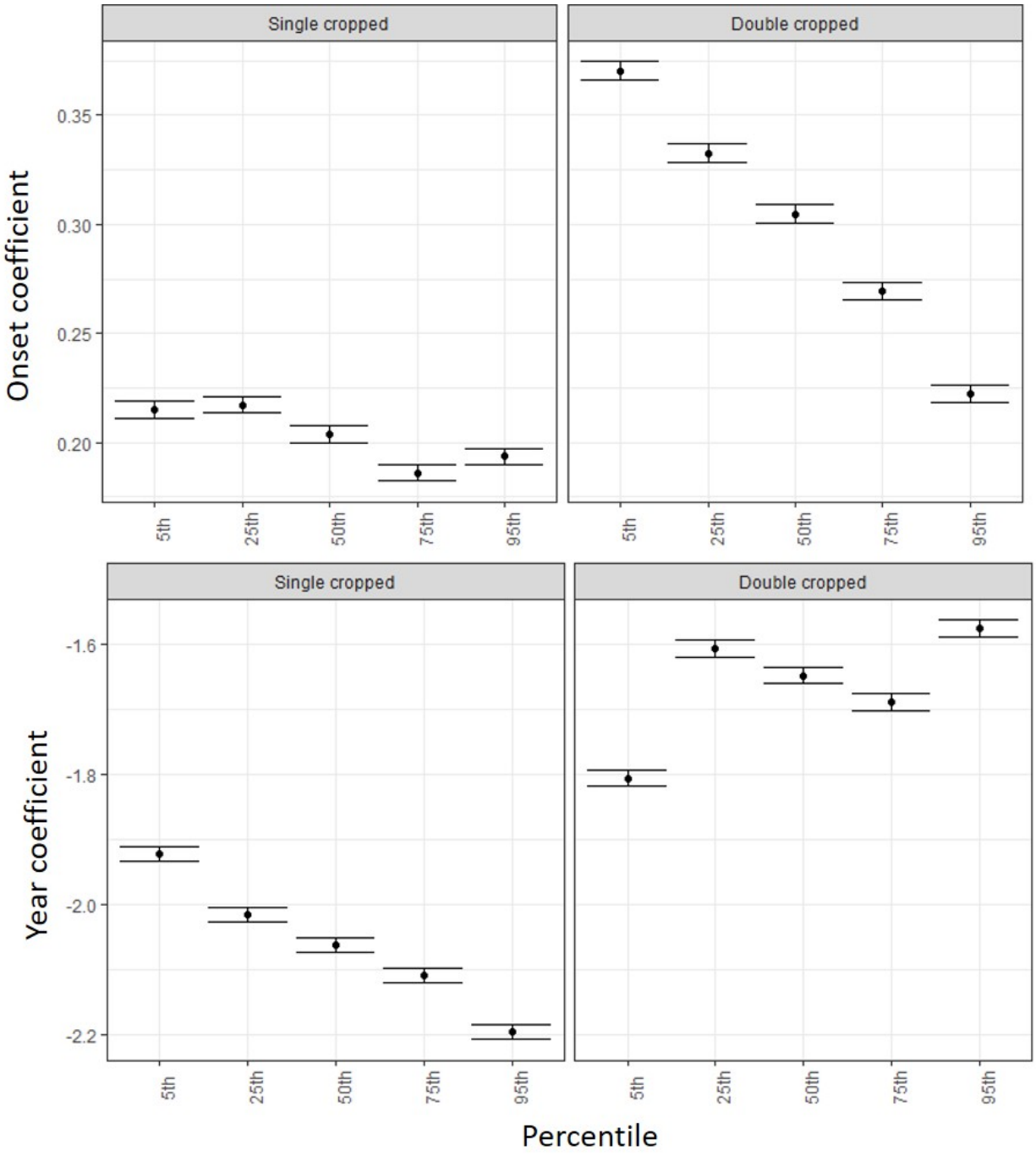


Figure C.10: Onset coefficients appear statistically different among cropping intensities and percentiles, despite uncertainty. Error bars represent the standard deviation of 1,000 bootstrapped coefficients, reflecting planting date estimation error.

I use a t-test to determine if the onset coefficients are different among onset definitions, cropping intensities, and percentiles under uncertainty arising from planting date estimation error and standard error of the coefficients. The p-values of all unpaired, two-sided t-tests were below the threshold of 10^{-15}. These tests confirm that the different cropping intensities and planting percentiles generally have different sensitivities to onset, which can be detected above the noise in the estimated planting date. While it is clear that single cropping experiences some exceptions to the trend toward decreasing coefficients for later-planted fields, the magnitude of these exceptions are minor.