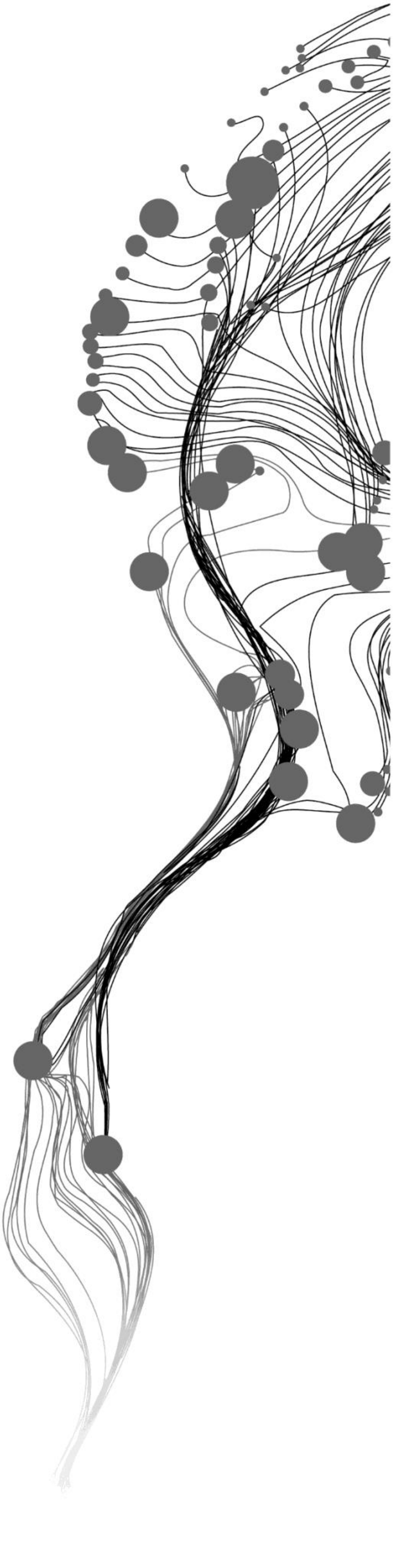
REMOTE SENSING BASED PRE-SEASON YELLOW RUST EARLY WARNING IN ETHIOPIA

CHINATSU ENDO

December, 2020

SUPERVISORS:

Dr. Kees de Bie, ITC, The University of Twente, The Netherlands
Dr. David E. Tenenbaum, Lund University, Sweden



REMOTE SENSING BASED PRE-SEASON YELLOW RUST EARLY WARNING IN ETHIOPIA

CHINATSU ENDO

Enschede, The Netherlands, December, 2020

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Natural Resource Management

SUPERVISORS:
Dr. Kees de Bie, ITC, The University of Twente, The Netherlands
Dr. David E. Tenenbaum, Lund University, Sweden

THESIS ASSESSMENT BOARD:
Dr. Ir. Anton Vrieling (Chair)
Dr. Ulrik Mårtensson (External Examiner, Director of Studies, Lund University, Sweden)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author and do not necessarily represent those of the Faculty.

ABSTRACT

Yellow rust (*Puccinia striiformis f. sp. Tritici*) is a crop disease of wheat that regularly causes yield loss in Ethiopia. The disease has significant consequences for the country's crop production, food security, health, and socioeconomic well-being. Anticipating yellow rust epidemics can help better manage them and mitigate their adverse impacts. This study explores the potential of remote sensing-based early prediction of yellow rust in the Oromia region in Ethiopia. The research focuses on modeling the incidence of yellow rust among young wheat in the region by looking at unique environmental conditions that enable off-season survival of the rust pathogen.

Tiller and boot-level yellow rust incidence data from 2016-2018 in Oromia was analyzed together with the environmental variables generated through AgERA5 (temperature), CHIRPS (precipitation), ProbaV-NDVI, and SRTM-DEM (terrain characteristics). Univariate Area Under ROC Curve analysis and Classification Tree analysis were used to understand the influential environmental variables and filter those with high relevance to the early-stage rust infection. Subsequently, General Additive Model and Boosted Regression Tree were applied to fit and test the early warning models and their prediction capacity. The models were built for three data sets: data with all available observations; tiller-level observations; and data that share the same climate zone.

As a result, the climate zone-based GAM model performed at a 78% accuracy level with Kappa 0.44 (moderate). The tiller-only GAM model performed at a 72% accuracy level with Kappa 0.44 (moderate). The all-observation BRT model had a 71% accuracy level with Kappa 0.34 (fair agreement). Rain characteristics served as particularly strong predictors in these models. Especially, excessive rain had a strong relationship with a lower probability of yellow rust cases among young wheat. The models also suggest that terrain characteristics serve as the static environmental conditions that expose certain locations to the disease. The study demonstrated the potential of yellow rust early warning solely based on remote sensing. The models could be further tested with a larger volume of data set to confirm the strength. Consideration of the probability of varying rust severity (low, moderate, high) and types of wheat cultivars would further add value to the models. Lastly, additional field and laboratory-based knowledge on the off-season rust survival would be a vital step towards a more accurate configuration of early warning models.

ACKNOWLEDGEMENTS

This study was realized with the support from the International Maize and Wheat Improvement Center (CIMMYT) in Ethiopia through the timely provision of rust survey data. Special appreciation goes to Dr. David Hodson, Mr. Yoseph Alemayehu, and Dr. Gerald Blasch at CIMMYT for their time and guidance during the data collection phase.

I want to thank my supervisors, Associate Professor Dr. Kees de Bie at Twente University and Associate Professor Dr. David E. Tenenbaum at Lund University, for guiding the design and execution of my final project. Special thanks to Ir. Louise Van Leeuwen at Twente University for encouraging me to pursue the on-campus study in early 2020. Despite the limitation and challenges during the COVID-19 pandemic, this visit enriched my learning experience as an iGEON distance student. Finally, completion of the degree would have been impossible without the understanding of my husband. Thank you, Dan, for your unconditional support for my journey of learning.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AUC	Area Under the ROC Curve
BRT	Boosted Regression Tree
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CIMMYT	International Maize and Wheat Improvement Center
CT	Classification Tree
DN	Digital Number
ECMWF	European Center for Medium-range Weather Forecast
AgERA5	Agriculture ERA5
ERA5	European Center for Medium-range Weather Forecast Reanalysis 5th Generation
FAO	Food and Agriculture Organization
FEWS NET	Famine Early Warning Systems Network
FN	False Negative
FP	False Positive
GAM	General Additive Model
GEE	Google Earth Engine
GIS	Geographic Information System
GLM	Generalized Linear Model
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
Pst.	Puccinia Striiformis f. sp. Tritici
ROC	Receiver Operating Characteristic
RQ	Research Question
RS	Remote Sensing/Remotely Sensed
SRTM	Shuttle Radar Topography Mission
TN	True Negative
TP	True Positive
USAID	United States Agency for International Development
USDA	United States Department of Agriculture
USGS	United States Geological Survey
VIF	Variance Inflation Factor
WGS83	World Geodetic System 1984

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Yellow (stripe) rust, a wheat disease caused by the fungus *Puccinia striiformis f. sp. Tritici (Pst)* (Zadoks, 1961) is common in Ethiopia, causing frequent crop failure and resulting in economic loss (Jaleta et al., 2019). Ethiopia's agriculture sector accounts for 37% of the country's GDP, employing 72% of the total population, of which 74% are small-scale farmers (FAO, 2018). Ethiopia is a leading wheat producer in sub-Saharan Africa (FAO, 2018), but the country's wheat production has been continuously undermined by rust epidemics such as in 1977, 1980-83, 1986, 1993, 2010, and 2013-2014 (Badebo et al., 1990, Jaleta et al., 2019, Olivera et al., 2015). Ethiopia's average wheat yield capacity is about 1.83$t\ ha^{-1}$, which is much lower than the world average of 3.47$t\ ha^{-1}$ (Mengesha, 2020). Disruptive rust epidemics are compounded by a new norm of extreme weather, droughts, and floods, generating additional pressure on wheat production, contributing to food insecurity in a country with a growing population (Alemu and Mengistu, 2019) and where a quarter of the population still live under US$ 1.9 a day (WB, 2020). Food insecurity can fuel other long-term complications such as malnutrition (Humphries et al., 2015), conflicts over food resources, and social instability (Martin-Shields and Stojetz, 2019).

The study of wheat rust started as early as 1767, and over the centuries, the wheat rust pathology has been better understood. As a result, its management has been somewhat successful through the introduction of fungicides and disease-resistant wheat varieties (Martinelli et al., 2015). Despite improved rust management techniques, the fungi evolve and new races of yellow rust emerge and continue to impact up to 5% of the crop across the wheat-producing countries today (Wellings, 2011).

Yellow Rust mainly spreads in the form of *urediniospores* through the wind, and it can disperse over large areas (Beest et al., 2008, Chen and Kang, 2017, Eriksson, 1894). The rust can propagate aggressively depending on atmospheric conditions, such as temperature, humidity, and sunlight (Zadoks, 1961). Efforts have been made to estimate the severity of future rust epidemics and potential wheat production loss based on climate data (Beest et al., 2008, Coakley et al., 1987, Grabow et al., 2016, Park, 1990). A similar approach with climate data has been applied in Ethiopia's early detection and communication of wheat rust outbreaks during the season, with the aim to help rust control measures (Allen-Sader et al., 2019).

Many of the yellow rust prediction models rely on rust incidence observations from the middle of a wheat season or the records of epidemics that come at the end of the season to facilitate more effective fungicide use. Meanwhile, projections made based on rust cases from the middle of the season imply that the disease is already happening and there is production loss inevitably expected for farmers. While complete avoidance of rust damage is impossible, such loss can be costly for the many small farmers of Ethiopia.

This research explores the possibility of identifying the signs of yellow rust outbreak earlier than the planting season in Ethiopia by looking at the conditions that enable yellow rust to survive the off-season period. Yellow rust can spread through dormant spores on volunteer wheat after the harvest (Rapilly, 1979, Zadoks, 1961). If the wheat-growing sites meet certain environmental conditions known to enable off-season survival of the pathogen, yellow rust outbreaks in the surrounding wheat field could be anticipated earlier before the crop season. This research will test this hypothesis in the context of Ethiopia's Oromia region by examining the relationship between yellow rust cases at the early stage of wheat growth and various environmental conditions observable with remote sensing.

Objectives and Research Questions

This research's overall objective is to develop and test a model for pre-season early warning of yellow rust in the Oromia Region of Ethiopia based on the environmental conditions favorable for off-season survival of yellow rust by maximizing the use of remotely sensed (RS) environmental data. The study designed the following sub-objectives and research questions to guide various steps of the research.

Sub-objective 1:

To examine the relationship between the past yellow rust incidence and the relevant RS-based environmental conditions during the pre-planting season.

> *Research Question 1.a*
>
> What are the associations between the yellow rust incidence and off-season environmental conditions captured by RS-derived indicators?
>
> *Research Question 1.b.*
>
> What are the most relevant or important yellow-rust inducing environmental parameters detected before planting season?

Sub-objective 2:

To develop a functional yellow rust prediction model based on the off-season environmental conditions in the Oromia region.

> *Research Question 2*
>
> How reliably can yellow rust incidence be predicted before planting season by the environmental conditions captured by RS-derived indicators?

2. BACKGROUND

2.1 Yellow Rust

Yellow rust starts with yellow or light-orange colored smooth surface flecks of varying sizes on primary leaves, lower leaves, transition leaves, or even on stem leaves (Zadoks, 1961). Over several days, these freckles transform into lesions with little bumps of pustules that could eventually cover leaf surfaces (Zadoks, 1961). Figure 1 below shows the progression of yellow rust infection on leaves.

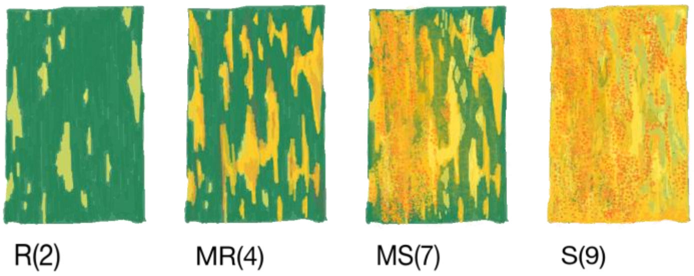


Figure 1: Adult plant host response to yellow rust

The graphic was adapted from Roelfs et al. (1992)

Puccinia striiformis requires a host plant to survive on, and these plants are categorized into the primary host that is wheat, and alternate hosts that are non-wheat plants (Grabow, 2016, Aime et al., 2017). On the primary host, pathogen reproduces asexually in the form of urediniospores (Figure 2), one of the five spore stages (Grabow, 2016).

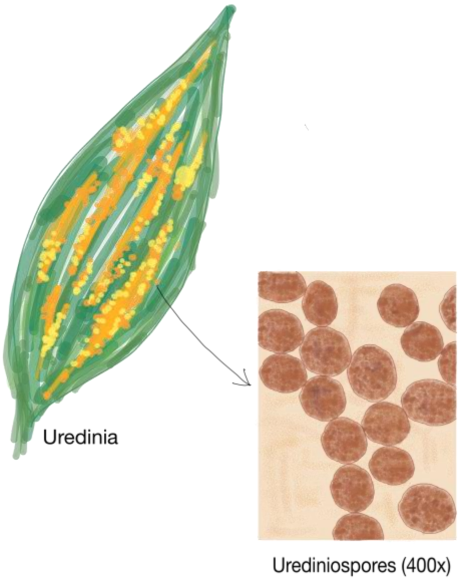


Figure 2: Symptoms and spore (urediniospore) morphology of yellow rust disease

The graphic was adapted from Roelfs et al. (1992)

Weeds and other local plants have been thought to serve as an alternate host (Rapilly, 1979). However, so far, only a limited number of plants such as *Berberis spp* (Yue Jin et al, 2010) and *Mahonia aquifolium* (Oregon grape) (Wang and Chen, 2013) are proven to be alternate hosts that can contribute to increased pathogen variability. Between the primary host and alternate host, yellow rust completes five distinct spore stages: uredinial, telial, basidial, pycnial, and aecial stages (Mehmood et al., 2020).

Figure 3 below is the schematic illustration of the lifecycle of *Puccinia striiformis f. sp. tritici (Pst),* which occurs on the primary host (wheat) and alternate hosts throughout different stages of the rust's life cycle. Yellow rust starts as an infection by urediniospores (A). The yellow patches of urediniospores become dark spots of teliospores (B) and basidiospores (C), which could infect alternate hosts. In the process of infecting the alternate hosts, the disease propagates in the form of pycnispores and aeciospores (Sexual cycle). Urediniospores can continue infecting wheat without advancing to teliospores (Asexual cycle). The final aecial stage can disperse aeciospores to infect wheat as well.

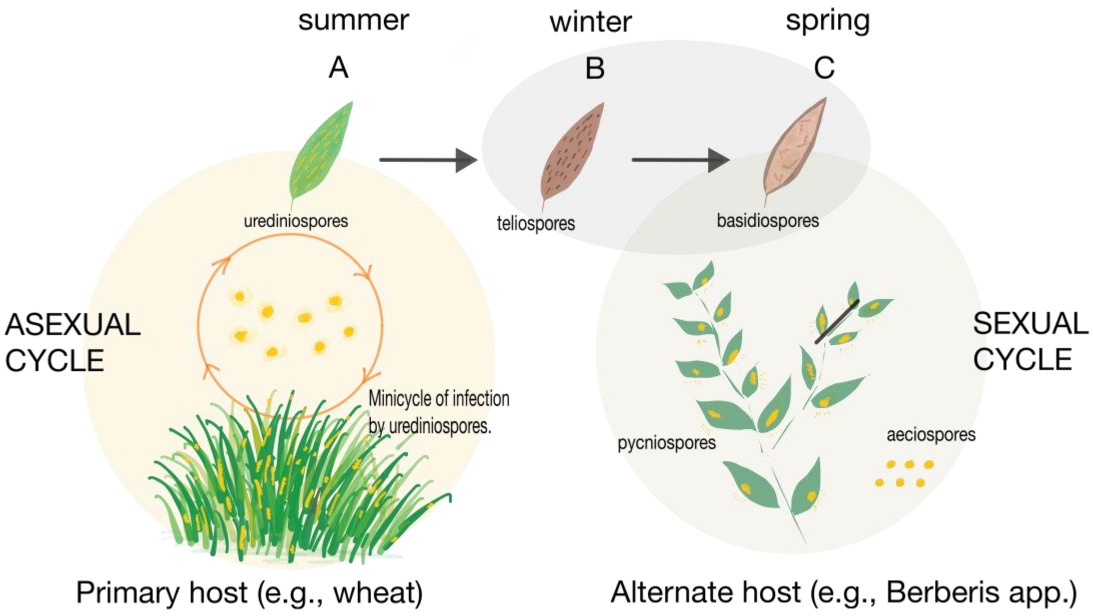


Figure 3: Lifecycle of Puccinia striiformis f. sp. tritici (Pst) on primary host and alternate host Adapted and modified from Mehmood et al. (2020)

Figure 4 demonstrates how the infection propagates on wheat from a single piece of urediniospore. The infection starts by arrival and adhesion of a urediniospore. The spore extends the germination tube to form an appressorium and penetrate through the leaves' tissues where rust colonization and reproduction occurs (Kumar et al., 2018). Deposition of a urediniospore on leaves and subsequent germination and appressoria formation depends on various climate factors such as temperature, rainfall, humidity, and sunlight (Park, 1990, Rapilly, 1979, de Vallavieille-Pope et al., 2018, Zadoks, 1961). The same climate factors also influence the speed, termination, and latency (being inactive but live infection after germination and before pustulation) (Rapilly, F, 1979). Rain and wind can be rust spreading factors but also have adverse effects on spore survival (Rapilly, 1979, Chen, 2005).

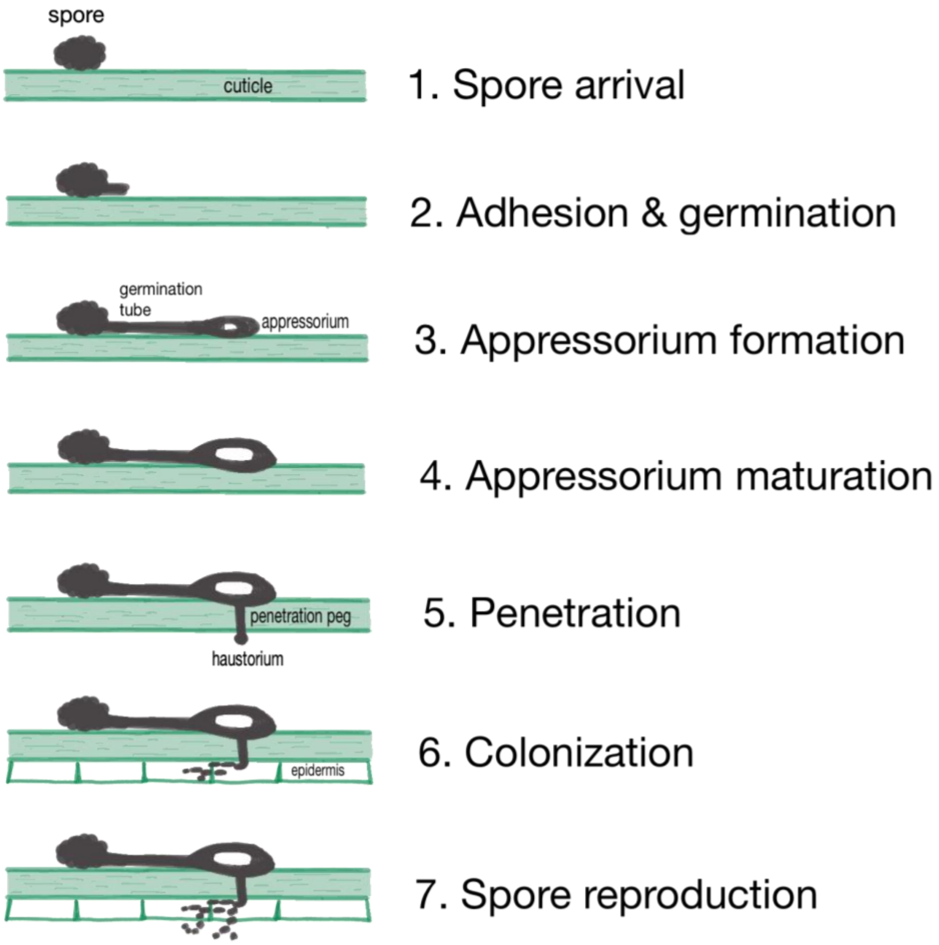


Figure 4: Propagation of infection on leave by a urediniospore

The graphic was adapted from Kumar et al. (2018)

The disease can originate from distant locations through spores traveling in the air for hundreds of kilometers (Rapilly, 1979, Zadoks, 1961). It can also spread from the spores that remained dormant on the voluntary wheat (primary host) or alternate host nearby wheat fields after the harvesting (Rapilly, 1979, Zadoks, 1961). Such dormant rust infections are the result of so-called rust *oversummering* or *overwintering*.

2.2 Oversummering and Overwintering of Yellow Rust

Frequent outbreaks of yellow rust are partly explained by the pathogens surviving the season when wheat crops are not grown (Sharma-Poudyal et al., 2014). Oversummering is the survival of rust pathogens during summer as a latent or dormant infection between harvest and the next season, and it occurs on self-grown volunteer wheat from the grain shed during harvest and late-tillers that grew out of the roots left after harvest (Zadok, 1961) (Figure 5). Plowing before planting does not entirely remove volunteer wheat with oversummering rusts, leading to infecting autumn-sown wheat, some of which carry the pathogen until the following spring by overwintering (Zadoks, 1961). Temperature and precipitation can determine the effectiveness of volunteer wheat and later spread the yellow rust pathogens. Under a stable high temperature and lack of rainfall, oversummering of yellow rust can be interrupted (Zadoks and Bouwman, 1985). However, warm/cool weather with sufficient water available creates a conducive environment for the pathogen to survive.

On the other hand, overwintering is the survival of rust infection on winter wheat planted in autumn that goes through a slow vegetative phase during winter and continues growing in the following spring (Zadoks and Bouwman, 1985). Overwintering occurs as urediniomycelium (not necessarily visible on the leaves but germination and appressorium penetration occurred already) in the wheat plants exposed to yellow rust at one point and endures winter climate (Zadoks, 1961). The pathogen can die at a temperature below about - 4 ℃, but so long as the host plant is alive, it can survive as a latent infection for 118 to 150 days in a growth conducive environment, such as snow cover, which provides insulation and allows the pathogen to survive (Zadoks, 1961).

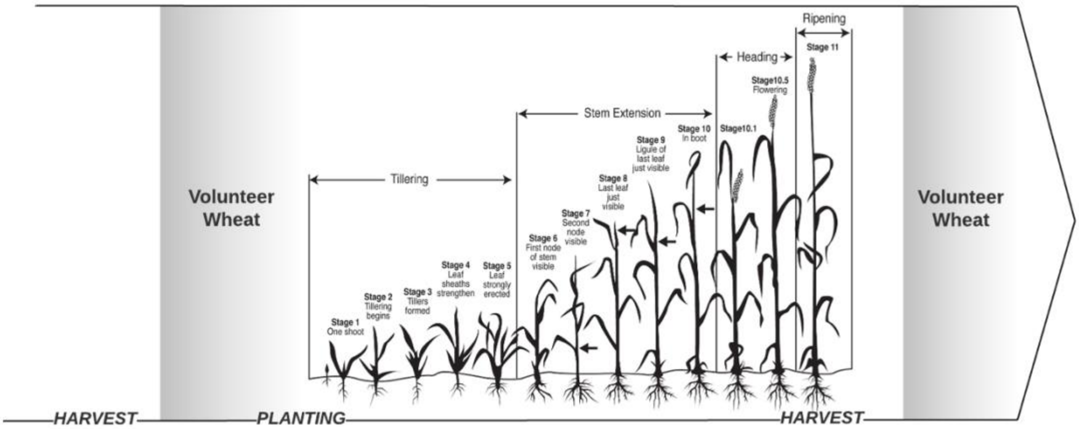


Figure 5: Growth cycle of wheat influenced by the rust infection on volunteer wheat

After harvest, some grains remain in the field and end up growing as new young wheat. When this young wheat gets infected by yellow rust and survives as dormant/latent infection during the off-season, it can infect the new wheat during the following wheat season. The graphic was made based on the Feekes Scale of Wheat Development adapted from Large (1954) and Marsalis and Goldberg (2016).

Off-season survival of rust on volunteer wheat increases the chance of local infection of young wheat in the following season and the severity of overall rust incidence later on. Eversmeyer and Kramer (1998) observed that there was a significant difference in the leaf rust severity between the field with the prevalence of rust-infected volunteer wheat plants (80-100% severity) and the fields of the same wheat with no volunteer plants around (10-30%). According to Zadoks and Bouwman (1985), one lesion of yellow rust overwintering is sufficient to cause a rust epidemic in the upcoming spring. As such, while distant spore dispersal is a common way of rust spreading, proximity to the infected volunteer wheat matters a great deal. Anticipating the areas where off-season rust survival occurs has drawn attention to promoting better control of yellow rust (Sharma-Poudyal et al., 2014).

2.3 Yellow Rust Prediction Models

Over the years, the epidemiology of yellow rust has advanced. It led to various rust management techniques such as fungicide application, continuous improvement of rust-resistant cultivars, and adjusting farming practices (Chen and Kang, 2017). Efforts have also been made to predict yellow rust epidemics to mitigate the potential loss from the disease.

Coakley et al. (1987) presented one of the earlier predictive models of yellow rust severity of a few winter wheat varieties. The model applied a stepwise regression to analyze the critical meteorological factors associated with disease severity to develop the so-called Disease Index. Rust data used in this model was from milk and dough wheat growth stages for three wheat cultivars. Parameters such as temperatures in October and November, the days of maximum temperature above 25℃, precipitation in June played a crucial role in this model. The model intended to facilitate a more effective fungicide application.

Grabow et al. (2016) worked on yellow rust epidemic models for winter wheat in Kansas in the United States. A combination of Classification Tree and Generalized Estimating Equation (GEE) selected the key predictors and modeled the epidemics. Soil moisture from October to December (planting season for winter wheat) was strongly associated with yellow rust epidemics. Several other environmental predictors such as temperature, relative humidity, and precipitation from March to May (the period during which nodes development to boot, heading, and flowering occur) were applied to classify the predicted severity of epidemics based on the yield loss data. This model's novelty was in consideration of soil moisture, which was assumed to provide sufficient wet microenvironments that are favorable for the yellow rust pathogen to grow on leaves. Soil moisture can influence canopies' growth, where yellow rust thrives under wet conditions (Grabow et al., 2016).

In Canada, Newlands (2018) proposed an integrated model-based framework for forecasting disease risk in Southern Alberta. The two models were built based on the Coffee Leaf Rust models developed in Colombia (temperatures and leaf wetness duration) and the multivariate spatiotemporal endemic-epidemic model (leaf wetness duration, temperature, relative humidity). In addition to weather station data and RS-based environmental data, the study was supported by airborne inoculum samples (spores traveling in the air) collected in the region.

Sharma-Poudyal et al. (2014)'s work is one of the few studies available that looked at predicting off-season survival of the yellow rust pathogen with temperature, humidity, and precipitation. The model projected the extent to which climates of different US regions are favorable for oversummering or overwintering of yellow rust. Similarly, Xu et al. (2019) pursued a yellow rust overwintering model for northwestern China based on the cultivars' empirical field observations with different hardiness levels related to temperature. The logistic models demonstrated that overwintering of *Puccinia striiformis f. sp.* is mainly influenced by the duration of low temperature in the coldest period in December and January. Overwintering probability had different thresholds for the cultivars with different hardiness. For example, the probability declined under the average temperature below -2 ℃ for the cultivar with weak winter hardiness, but the cultivar with moderate and strong winter hardiness saw the decline in overwintering probability only below -4 ℃.

In recent years, learnings from the past rust modeling have been applied in the context of Ethiopia as well. Allen-Sader et al. (2019) present an overview of the novel early warning system of wheat rust in Ethiopia, where a synthetic rust predictive model feeds into the network of last-mile communication with the farmers about the risk of rust infection. Canopy temperature, free moisture, and solar radiation derived from the UK meteorological Unified Model serve as the model's critical environmental parameters. What makes the model unique compared to other models is that this considers atmospheric spore dispersion (based on the Numerical Atmospheric dispersion Modeling Environment (NAME) model), which influences the long-distance dispersal of urediniospores.

2.4 Knowledge Gap

The earlier yellow rust prediction models mostly rely on yellow rust incidence data from the middle of the growing season or yield loss data at the end of the season to promote efficient fungicide-based rust control. Meanwhile, these models naturally imply that the disease is already occurring, and there is some level of production loss inevitably expected. Such loss can be very costly for many smallholder farmers in Ethiopia. As Eversmeyer and Kramer (1998) observed in their study, infection and survival of yellow rust during the off-season tends to lead to severe rust infection during the actual wheat season. Rust infection in young wheat also has a considerable impact on the quality of the grains produced later (Wellings, 2011). Observing the conditions of oversummering and overwintering is one way to promote earlier warning of yellow rust, but this is not a commonly adopted approach and yet to be examined in Ethiopia.

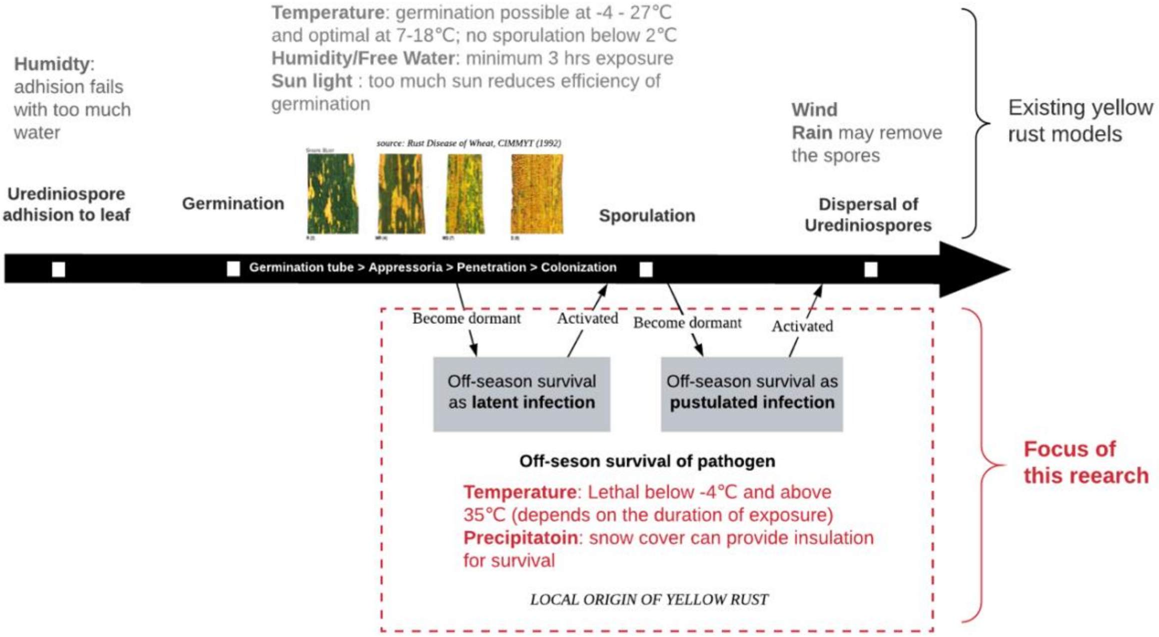

Humidty: adhision fails with too much water
Temperature: germination possible at -4 - 27°C and optimal at 7-18°C; no sporulation below 2°C
Humidity/Free Water: minimum 3 hrs exposure
Sun light : too much sun reduces efficiency of germination
source: Rust Disease of Wheat, CIMMYT (1992)
Wind
Rain may remove the spores
Existing yellow rust models
Urediniospore adhision to leaf
Germination
Sporulation
Dispersal of Urediniospores
Germination tube > Appressoria > Penetration > Colonization
Become dormant
Activated
Become dormant
Activated
Off-season survival as latent infection
Off-season survival as pustulated infection
Off-seson survival of pathogen
Temperature: Lethal below -4°C and above 35°C (depends on the duration of exposure)
Precipitatoin: snow cover can provide insulation for survival
LOCAL ORIGIN OF YELLOW RUST
Focus of this reearch

Figure 6: Focus of this study in relation with the existing yellow rust models and stages of Pts infection

This study focuses on the environmental conditions that enable the pathogen to survive in latent infection (become dormant before sporulating) or pustulated infection during the off-season.

Cold and hot weather usually terminate the yellow rust pathogen (Zadoks, 1961). However, if the climate is warm or cold enough, *urediniomycelia*, which is the critical inoculum for yellow rust, can remain latent on the host plants, prolonging the incubation time of the rust before the actual wheat season begins (Rapilly, 1979, Sharma-Poudyal et al., 2014, Tollenaar and Houston, 1967, Zadoks, 1961). As crop season begins, the surviving urediniospores are dispersed through the air to nearby or distant crop fields to infect the newly planted wheat (Rapilly, 1979). Tollenaar and Houston (1967), Eversmeyer and Kramer (1996), and Sharma-Poudyal et al. (2014) looked into the potential of off-season fungi survival in relation to meteorological parameters that are similar to the ones used in rust epidemic prediction. These studies were done in the US, where a robust network of weather stations is available. In-situ weather stations are not widely available in Ethiopia. However, remote sensing technology can feed the necessary environmental data at a high spatial and temporal resolution.

3. Method

3.1 Study Area

The research will focus on Ethiopia's Oromia region. The Oromia region spreads through the center to the western and southern parts of Ethiopia. The region is situated between the latitude of 3°30' N and 10°23' N and the longitude of 34°7' E and 42°55' E (Figure 7), covering a total area of 353,690 square kilometers that is split into several climate zones by the Great African Valley offering abundant agricultural cropland including that for wheat production (Mohammed et al., 2020).

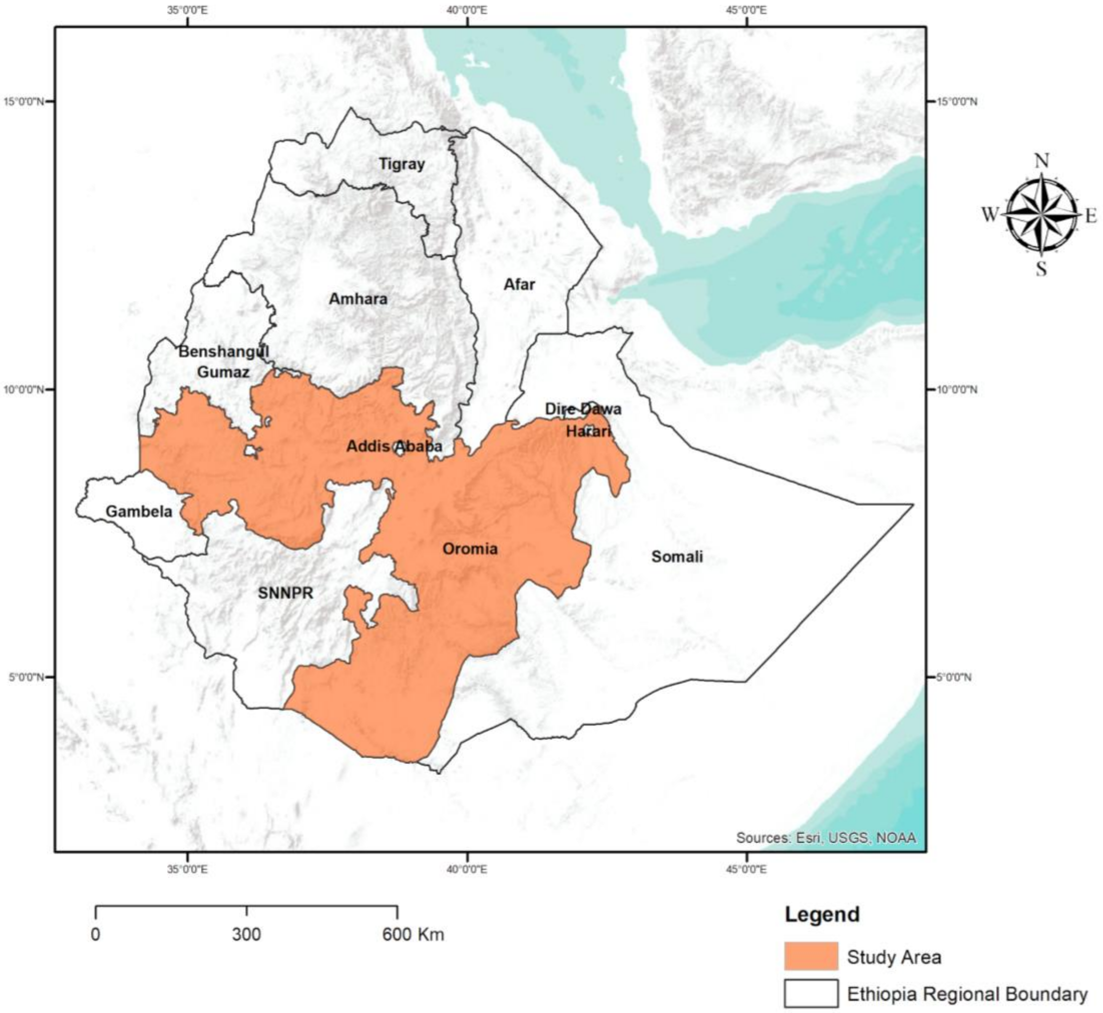


Figure 7: Map of Study Area - Oromia region, Ethiopia

3.2 Methodological Flowchart

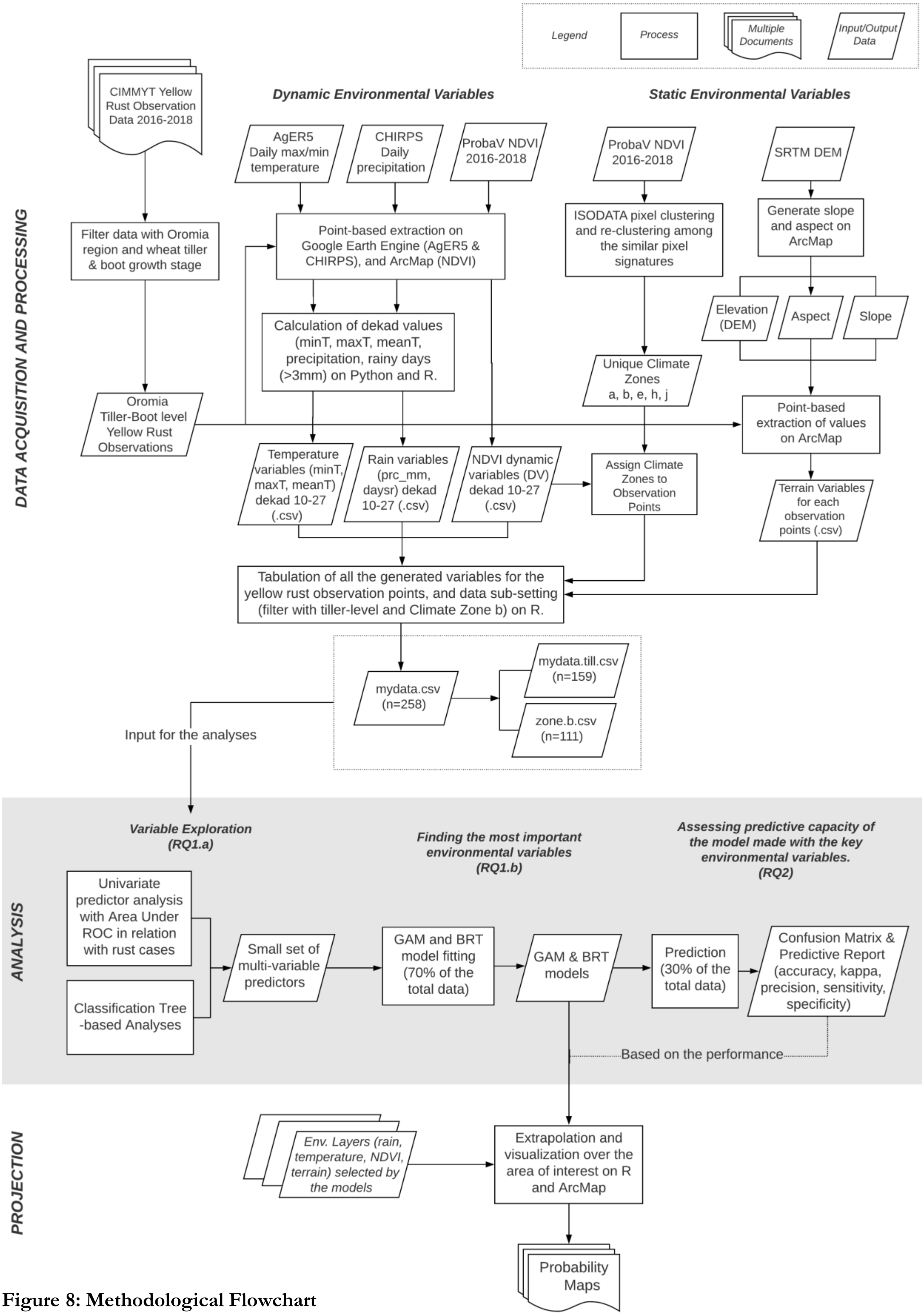

Legend
Process
Multiple Documents
Input/Output Data
DATA ACQUISITION AND PROCESSING
CIMMYT Yellow Rust Observation Data 2016-2018
Dynamic Environmental Variables
Static Environmental Variables
AgER5 Daily max/min temperature
CHIRPS Daily precipitation
ProbaV NDVI 2016-2018
ProbaV NDVI 2016-2018
SRTM DEM
Filter data with Oromia region and wheat tiller & boot growth stage
Point-based extraction on Google Earth Engine (AgER5 & CHIRPS), and ArcMap (NDVI)
ISODATA pixel clustering and re-clustering among the similar pixel signatures
Generate slope and aspect on ArcMap
Elevation (DEM)
Aspect
Slope
Calculation of dekad values (minT, maxT, meanT, precipitation, rainy days (>3mm) on Python and R.
Unique Climate Zones a, b, e, h, j
Oromia Tiller-Boot level Yellow Rust Observations
Point-based extraction of values on ArcMap
Temperature variables (minT, maxT, meanT) dekad 10-27 (.csv)
Rain variables (prc_mm, daysr) dekad 10-27 (.csv)
NDVI dynamic variables (DV) dekad 10-27 (.csv)
Assign Climate Zones to Observation Points
Terrain Variables for each observation points (.csv)
Tabulation of all the generated variables for the yellow rust observation points, and data sub-setting (filter with tiller-level and Climate Zone b) on R.
mydata.csv (n=258)
mydata.till.csv (n=159)
zone.b.csv (n=111)
Input for the analyses
ANALYSIS
Variable Exploration (RQ1.a)
Finding the most important environmental variables (RQ1.b)
Assessing predictive capacity of the model made with the key environmental variables. (RQ2)
Univariate predictor analysis with Area Under ROC in relation with rust cases
Classification Tree -based Analyses
Small set of multi-variable predictors
GAM and BRT model fitting (70% of the total data)
GAM & BRT models
Prediction (30% of the total data)
Confusion Matrix & Predictive Report (accuracy, kappa, precision, sensitivity, specificity)
Based on the performance
PROJECTION
Env. Layers (rain, temperature, NDVI, terrain) selected by the models
Extrapolation and visualization over the area of interest on R and ArcMap
Probability Maps

Figure 8: Methodological Flowchart

Figure 8 is the methodological flow chart outlining the process of data acquisition, processing, and analysis. The following section describes the data and steps more in detail.

3.3 Data

Yellow Rust Incidence Data

Yellow rust incidence data was attained from the International Maize and Wheat Improvement Center (CIMMYT) Ethiopia. The original data set contained 4,342 yellow rust observation points over three years (2016–2018) across the country recorded at different wheat growth stages - Tiller, Boot, Heading, Flowering, Milk, Dough, and Maturity. This study's geographic scope is limited to the Oromia Region, and the analysis is on the yellow rust cases in the early stage of wheat rust in relation to pre-seasonal environmental conditions. Therefore, only the observations from the Oromia Region at Tiller and Boot stage were filtered. In total, 258 observations from 2016 to 2018 were made available for the analysis (Table 1).

Table 1: Yellow rust incidence data used in the study

Yellow Rust Incidence (Tiller and Boot)	*None (0)*	*Low (less than 20%)*	*Moderate (20-40%)*	*High (more than 40%)*	*Total (n=258)*
2016	35	83	9	3	108
2017	40	26	1	5	71
2018	32	45	3	3	79

RS-based Data on Environmental Condition

Weather (temperature and precipitation) and Normalized Difference Vegetation Index (NDVI) were used in the analysis as dynamic environmental parameters. In addition, in this study, elevation, slope, and aspects were also considered as the static environmental conditions that could influence the off-season survival of yellow rust pathogens. While NDVI is regarded as a dynamic environmental parameter, the study also used this to identify a static characteristic of climate zones based on a unique range of NDVI values. The dynamic and static RS-based products are summarized in Table 2 and Table 3.

Table 2: List of RS-based products for dynamic environmental conditions

RS Product	Spatial res.	period	Use
AgERA5 (Temperature)	11 km	2016-2018	Temperature predictor variables
CHIRPS (Precipitation)	5.55 km	2016-2018	Rain-based predictor variables
NDVI 10-day maximum composite data (ProbaV)	1 km	2016-2018	NDVI-based predictor variables

Table 3: List of RS-based products for static environmental conditions

RS Product	Spatial res.	period	Use
NDVI 10-day maximum composite data (ProbaV)	1 km	2016-2018	Common climate zone
SRTM-DEM	30m	2000	Elevation, slope, and aspect

AgERA5 (Temperature)

A collection of daily surface meteorological data prepared for environmental and agricultural modeling. Temperature data is among the multiple parameters made available. The temperature data (Kelvin) is available from 1979 to 2018 at the resolution of 0.1° grid (about 11 km) with global coverage. The product is the aggregation and correction of ECMWF (European Center for Medium-range Weather Forecast) ERA5 data. ERA stands for ECMWF Re-Analysis, a deterministic climatic, land, and oceanic climate data at surface level with 30km (0.28215°) spatial resolution. ERA5 derives from historical observations by multiple satellite sensors into global estimates using advanced modeling and data assimilation systems. ECMWF ERA5 data went through spatial scaling down to 0.1° grid with Nearest Neighborhood algorithm, temporal aggregation to daily time steps, and bias correction based on the finer topography, land use pattern, and land-sea delineations to arrive at AgERA5. Source: ECMWF (2020)

CHIRPS (Precipitation)

A Quasi-global rainfall data set is available over 30 years at 0.05° grid (5.55 km) resolution. CHIRPS has been developed since 1999 by the U.S. Geological Survey Earth Resources Observation and Science Center, initially to support the United States Agency for International Development (USAID)'s Famine Early Warning System Network (FEWS NET) in collaboration with the

National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA). The product is derived through: rainfall estimates by the infrared Cold Cloud Duration (the measurement of the threshold at which clouds become precipitation), long-term historical in-situ observation data, and existing gauge observations for bias correction. In Ethiopia, CHIRPS products are commonly used in the analysis of precipitation anomalies, drought, and food insecurity. Reference: Funk et al. (2015)

Note: Relative humidity is another commonly used climate parameter in the study of rust propagation. However, currently available humidity data was at the resolution of 27km (Global Forecast Systems) or 17km (UK Met Unified Model). This spatial resolution was considered not adequate for the analysis since many rust observations spread within the space of 1 to 5 km (many observation points would end up having the same relative humidity value). The literature review suggests that relative humidity becomes essential at the time of germination to sporulation of the rust but not concerning the off-season survival of the already germinated or sporulated lesion of yellow rust (Tollenaar and Houston, 1967, Eversmeyer and Kramer, 1998). Hence this variable was not included in this study.

Normalized Difference Vegetation Index (NDVI)

A vegetation index is calculated by comparing the visible and near-infrared sunlight reflected by the surface. NDVI layers entail the maximum value (range: -0.08 - 0.9) out of 10 individual images taken over ten sequential days at 1km spatial resolution with the geographic projection WGS84 (EPSG:4326). The data were generated by the Global Land Service of Copernicus, the Earth Observation program of the European Commission in Digital Number (DN) through PROBA-V daily top-of-atmosphere orbit reflectance values (BRDF-adjusted; Release-Candidate #3 produced by VITO). The retrieved images were processed to obtain their long-term median data by dekad between 1999 and 2018 (20 years) to create the 36 dekad specific "normal" data series.

NDVI physical values (PhyVal) are usually generated using DN value, scale factor, and offset (VITO, 2019).

$$PhyVal = DN * 0.004\ (scale\ factor) - 0.08\ (offset)$$

In this study, NDVI was used as an environmental variable potentially associated with yellow rust incidence. Also NDVI was used to subset the yellow rust observation data by a unique climate zone. (See 3.4 Data Processing, Data Sub-setting)

DEM

Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) from the NASA was used as altitudes and to generate additional terrain characteristics such as slope and aspect (orientation of slope). SRTM DEM comes in WGS84 Datum and 30m/90m (USGS) spatial resolution.

3.4 Data Processing

Yellow Rust Incidence Categories

Yellow rust incidence was recorded in four levels: None (0), Low (1), Moderate (2), and High (3). The study initially aimed to assess all four levels of incidence to compare the probability of different yellow rust incidence levels. However, among the observations, the Moderate and High incidence was minimal. Thus, the study used a binary category of yellow rust (0, absent) and yellow rust (1, present) (which includes low to high incidence).

Retrieval and Processing of Weather Data

AgER5 and CHIRPS data were accessed through Google Earth Engine (GEE) using the point feature (.shp) generated with the yellow rust observation data from CIMMYT. The daily values for the period of April-September, 2016-2018, were extracted through GEE and tabulated using Python and R to calculate the dekad (10-day) maximum, minimum, and mean temperature; dekad sum of precipitation; and dekad number of rainy days (>3mm). The Javascript used for the point-based extraction of AgERA5 and CHIRPS data is available in the Appendices.

A 'dekad' approach for dynamic environmental variables

Of all the RS-based data retrieved for temperature, precipitation, and NDVI, variables for modeling were generated for the period between April and September. April is a few months before the wheat cropping season begins in Ethiopia. September is where some tiller-level observations were still observed in rust data each year. Earlier rust prediction models typically applied monthly intervals or rolling averages over 10, 20, 30, and 60 days for those dynamic variables. However, in this study, 10-day (dekad) was applied to align the NDVI data interval and was prepared as a 10-day maximum composition. A dekad is a period of ten days typically used in weather and vegetation analysis. For example, the first dekad of January is from 1st to 10th January. The second dekad is from 11th to 20th January, and the third dekad is from 21st to 31st January (the third dekad in the month with the 31st day contains 11 days). In this study, the dekad numbering was done annually from the beginning of January till the end of December (dekad 1 to 36). The analyses focused on the data from the dekad 10 (1-10 April) to the dekad 27 (21-30 September). Each dekad measure was considered as a dynamic environmental condition that represents a certain point in time.

Precipitation, temperature, and NDVI are dynamic variables that change over time. Meanwhile, elevation (DEM), slope, and aspect are considered as static environment variables. Table 4 below is the list of dynamic and static variables prepared based on the data from AgERA5 (temperature), CHIRPS (precipitation), NDVI, and SRTM (elevation, slope, aspect). A total of 111 variables were initially taken into consideration.

Table 4: Environmental variables and description

Variable Code	Description
prc_mm_10 ~ prc_mm_27	Accumulated precipitation (mm) per dekad
daysr_10 ~ daysr_27	The number of rainy days with more than 3mm precipitation
maxT_10 ~ maxT_27	Maximum temperature in dekad (℃)
minT_10 ~ minT_27	Minimum temperature in dekad (℃)
meanT_10 ~ meanT_27	Average temperature in dekad (℃)
ndvi_10 ~ ndvi_27	10-day vegetation density (DN-value)
DEM	Elevation (m)
Slope	Degree of slope
Aspect	Compass direction that slope faces. 0 = North, 90 = East, 180 = South, 270 = West

Data sub-setting

The rust observation data prepared for this study contains the observations from tiller and boot level from all over Oromia. Different climate zones, growth levels (tiller or boot), or even observation timing could have distinct characteristics in the relationship with environmental variables, hence yields a better model. Therefore, the original data was further subset into the tiller-only data set, and Climate Zone b data set.

The variability of climate zones was determined based on the unique characteristics of NDVI propagation over time (through ISODATA pixel clustering), shared across the observation points. Of five major climate zones (a, b, e, h, j) identified (Figures 9 and 10), the study used the Climate Zone b data set, which had more than 100 observations.

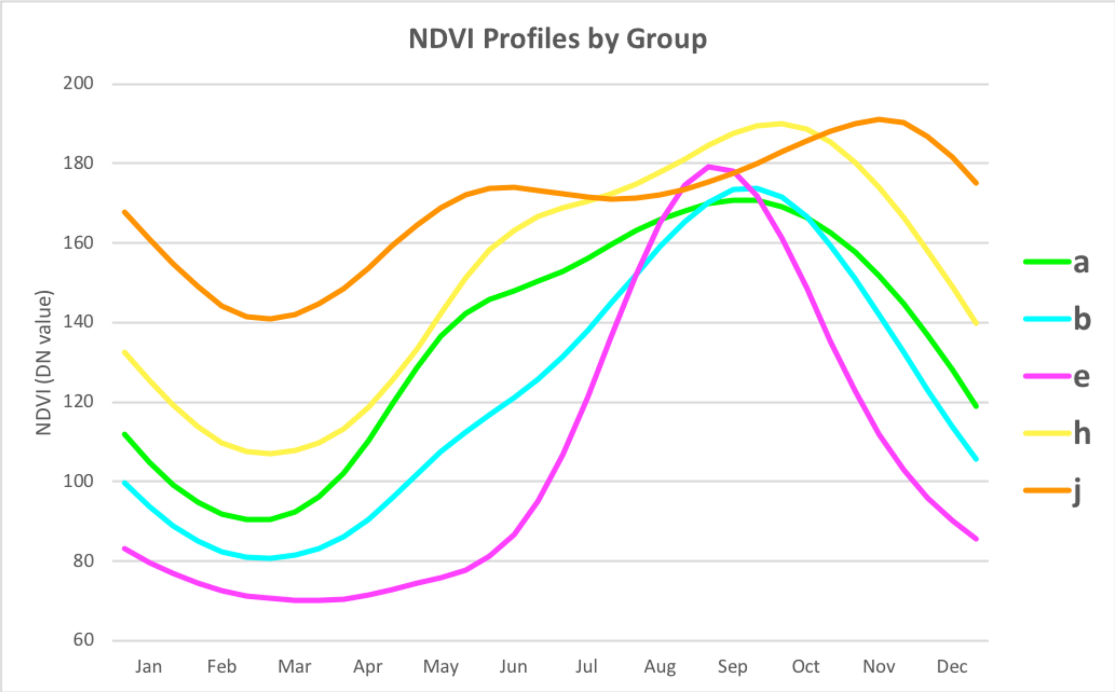


Figure 9: NDVI profile by group (climate zone)

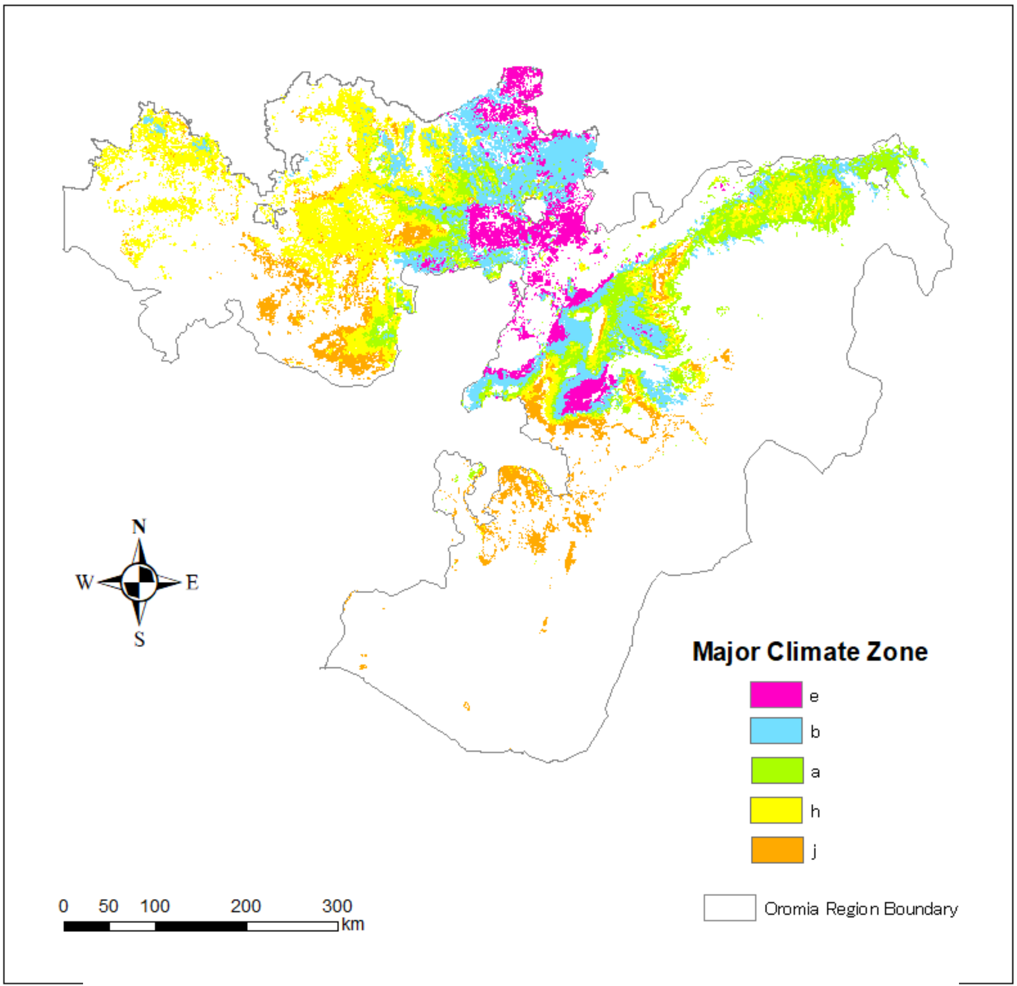


Figure 10: Map of major climate zones in Oromia Region

After all, three sets of data: *mydata*, *mydata.till,* and *zone.b* were used in the analysis (Table 5).

Table 5: Three data sets prepared for the analysis

Dataset	Description	Observations (n= total, [0] = no rust, [1] = rust)
mydata	Tiller and boot level observations.	n = 258 [0] 95, [1]163
mydata.till	Tiller-level yellow rust observations.	n = 159 [0] 75, [1] 84
zone.b	Climate Zone b, tiller and boot level observations.	n = 111 [0] 28 , [1]83

3.5 Analysis

Initially, the weather data was explored to understand the seasonal weather variability and crop growing seasons around the locations of rust observations in the Oromia Region. Subsequently, using the tabulated data sets, analyses were conducted to address the three Research Questions (RQs) designed for this study (Figure 8: Methodological Flowchart - Analysis). The scripts used in the analyses are available in Appendices.

3.5.1 Variable Exploration (RQ1.a)

Research Question 1.a (RQ1.a) probes the association between yellow rust incidence and environmental conditions. The RS-based 111 environmental variables (temperature, precipitation, NDVI, and terrain characteristics) were examined against yellow rust observations in the three data subsets: *mydata*, *mydata.till*, and *zone.b*. The objective here was to understand what types of variables are more associated with yellow rust and narrow down the number of related variables. A combination of univariate correlation analysis and Classification Tree (CT) analysis were applied.

Univariate correlation:

When there are multiple variables in hand, Area Under ROC Curve (AUC) helps identify the more relevant ones than the others. Especially when the response variable (rust incidence) is categorical (i.e., incidence or no-incidence), AUC quantifies the extent to which the respective variable can separate these two categories. An AUC score of around 0.5 is an indication of a completely irrelevant variable. The R. package *'caret'* was used to calculate AUC for each variable. AUC values were calculated by k-fold cross validation that enabled several repetitions of AUC value calculation. By averaging multiple cycles of AUC calculation, the AUC values presented were made more reliable. AUC helps identify the variables that are individually associated with yellow rust infection. However, this does not address interactions between different variables that may create an environment conducive to potential off-season survival of pathogen and impact early infection at the tiller/boot-level.

Classification Tree (CT):

CT was applied to identify a small number of variables that serve as good predictors. It categorizes the observation data into smaller and homogeneous groups by repeating a binary splitting based on the influential predictors (Hastie et al., 2009). This splitting aims to categorize the observation data, for example, in the case of yellow rust, into *infected* or *not infected* based on the influencing factors such as temperature and precipitation. Initially, the data categorized as *infected* may contain some uninfected observations, but this 'impurity' minimizes as splitting is repeated multiple times to better categorize the classes. The resulting summary of all the splitting forms a tree-like shape. Practically, CT is a modeling process on its own, but this was used purely for variable exploration and reducing the number of potential environmental variables in this part of the analysis. CT was undertaken using the *'cart'* package in R.

The variables were analyzed group-wise: maxT, minT, meanT, prc_mm, daysr, ndvi, and terrain (DEM, slope, aspect). The top-performing variables from each variable group were put together to find out the combination variables that achieved the lowest relative errors, and cross-validation errors were grouped as the variables most associated with the rust and forwarded to the next step to address RQ1.b.

3.5.2 Finding the most critical variables (RQ1.b)

The study applied General Additive Model (GAM) and Boosted Regression Model (BRT) to understand more about the critical variables associated with the early yellow rust incidence. Datasets were randomly split into training data (70%) and test data (30%) using the R package *'caret'*. This part of the analysis essentially builds models that explain the interaction of different environmental variables related to yellow rust incidence. The best performing models were forwarded to address the subsequent Research Question 2.

GAM

GAM (Hastie et al., 2009) is a progression of the Generalized Linear Model (GLM, Nelder and Wedderburn (1972)) which had considered the response variable that are not-normally distributed. GAM enhances GLM by considering nominal/categorical and ordinal predictors in their characteristics and maximizing a model's prediction capacity (Ravindra et al., 2019). While the ordinary regression model fits simple least-squares as function, GAM model fitting is based on the 'smoothing' function using a scatterplot smoother such as cubic smoothing spline or kernel smoother (Hastie et al., 2009). The smoothing function takes into consideration the nature of predictive variables that are not normally distributed. Thus, GAM is a flexible statistical method for identifying and characterizing nonlinear regression effects (Hastie et al., 2009).

With a random variable Y and a set of the predictor variable $X_1,\ X_2,\ \ldots\ ,\ X_p$, a regression model estimates $E(\ Y \mid X_1,\ X_2,\ \ldots\ ,\ X_p)$. The formula for a traditional regression model like GLM is expressed as:

$$E(\ Y \mid X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

where $\beta_0,\ \beta_1,\ \ldots\ , \beta_p$ are generated by least squares. Meanwhile, GAM assumes the following formula:

$$E(\ Y \mid X_1, X_2, \ldots, X_p) = s_0 + \sum_{j=1}^{p} s_j(X_j)$$

where $s_j(.)$'s are smooth functions that are estimated through a scatterplot smoother. The details of how scatter smoothers work are available by Hastie and Tibshirani (1986).

To the best of the author's knowledge, GAM has not been applied in yellow rust modeling. However, this has been widely used in many other fields, such as in plant ecology (Yee and Mitchell, 1991), species habitat study (Suárez-Seoane et al., 2002), and environmental health (Bouzid et al., 2014). The R package *'mgcv'* was used in GAM analysis. GAM has a function called 'smoothing' or 'splines' to realize flexible non-linear expression. In R. package *'mgcv'*, this smoothness can be defined by the user or automatically suggested by setting the method with Restricted Maximum Likelihood (REML).

Some of the methods to understand model convergences are:

i. *summary()* for model statistics to check parametric coefficient and significance of smooth terms

ii. *plogis()* to transform the model outcome to the log-odds scale to assess the extent of the model's prediction of a positive outcome (i.e., yellow rust infection).

iii. *plot()* to visualize the partial effect of the concerned variables with a confidence interval.

iv. *gam.check()* to check the random distribution of residuals for each predictor variable. "*Basis function"* (this influences the smooth parameter) for variables is adjustable to improve the model performance.

v. Collinearity and Concurvity check

Collinearity is the correlation among the predictors that potentially influence the model convergence. *ggpairs()* in the *'GGally'* package was used to plot the variable interactions. Variance Inflation Factor (VIF) calculation was also conducted to decide which variable to drop.

Concurvity is when one variable smooth term in GAM is approximated by one or more other variable smooth terms. Even though the variables are not collinear, concurvity can occur. The function *concurvity()* was used to check and rule out potential concurvity.

BRT

BRT (Friedman, 2001) is a combination of statistics and machine learning, guided by an algorithm to achieve the most optimal model (Youssef et al., 2016). BRT's rule sets are two-fold: *"classification/regression trees"* to find the most influential predictors; and *"boosting"* to synthesize many possible models to build the best performing model (Elith et al., 2008, Schapire, 2003).

There are four parameters that need to be set and adjusted to maximize the resulting model performance.

i. *learning rate (lr)*: signifies the contribution of each tree to the final fitted model;

ii. *tree complexity (tc)*: the number of total nodes (split point) in the tree;

iii. *number of trees (nt)*: the result of *lr* and *tr*; and

iv. *bag fraction (bf)*: the portion of data to be used for each iteration.

BRT can be applied to data that is not-normally distributed, and it is widely used in ecological and environmental model building (Naghibi et al., 2016, Pittman and Brown, 2011, Zellweger et al., 2013). BRT can select only relevant variables and ignore non-informative predictors. However, as Elith et al. (2008) point out, for small datasets where redundant predictors may degrade performance by increasing variance, it is better to simplify the list of predictor variables in advance instead of putting them all at once into the model. Thus, only the pre-selected list of predictor variables from RQ1.a was used in BRT. The R package *'dismo'* and *'gbm.step'* were used in BRT analysis.

In R, *gbm.step()* uses cross-validation (default k=10) to estimate the optimal number of trees. Considering the relatively small sample size (number of observations) used in the analysis, tree complexity (tc) 2 and learning rate (lr) 0.001 were used, unless a smaller *lr* yielded better models.

Model statistics in *summary()* and *gbm.plot()* report were examined to understand the relative importance/influence of key environmental variables. Tree Complexity *(tc)* and Learning Rate *(lr)* were as necessary to achieve better model statistics.

gbm.interactions() was used to understand the interactions between the critical environmental variables, and *gbm.perspec()* was used to visualize the interactions.

3.5.3 Assessing Model Predictive Capacity (RQ2)

Research Question 2 (RQ2) examines how reliably the RS-based environmental predictors can project yellow rust incidence among young wheat. The trained GAM and BRT models were used to predict yellow rust incidence using the test data (30% of observations). The R. package *'mgcv'* and *'gbm'* were used to conduct prediction.

The output of model prediction is in the form of probability with the values ranging from 0 to 1. This value was classified into 0 (probability < 0.5) and 1 (probability >= 0.5) in order to compare with the actual incidence of yellow rust.

A confusion matrix (Figure 11) was created with the prediction and actual observation.

Prediction

0	1	Actual Observation
True Negative (TN)	False Positive (FP)	0
False Negative (FN)	True Positive (TP)	1

Figure 11: Confusion Matrix

On R., a function *ModelPerformance()* was used to examine the key statistics to assess the GAM and BRT models' predictive performance. These statistics include Accuracy, Kappa Statistic, Sensitivity, Specificity, and Precision.

a. **Accuracy** is the ratio of correct predictions calculated by the true positive (TP) and true negative (TN) divided by the total number of events.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b. **Kappa Statistic** (Cohen, 1960): the extent to which prediction and observations agree with the actual yellow rust incidence

$$Kappa\ Statistics = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the relative observed agreement, andP_e is the hypothetical probability of chance agreement.

$$P_o = \frac{TP + TN}{TP + TN + FP + FN}$$

$$P_e = \left(\frac{TN + FN}{TP + TN + FP + FN} * \frac{TN + FP}{TP + TN + FP + FN}\right) + \left(\frac{FP + TP}{TP + TN + FP + FN} * \frac{FN + TP}{TP + TN + FP + FN}\right)$$

Kappa statistic	Level of agreement
≦0	no agreement
0.01 – 0.20	none to slight agreement
0.21 – 0.40	fair agreement
0.41 – 0.60	moderate agreement
0.61 – 0.80	substantial agreement
0.81 – 1.00	almost perfect agreement

c. **Sensitivity** is a true positive (TP) rate. It measures the rate of actual yellow rust cases corrected (predicted yellow rust case is an incidence in the observation)

$$Sensitivity = \frac{TP}{TP + FN}$$

d. **Specificity** is a true negative (TN) rate. It measures the rate of the negatives correctly predicted (the predicted no-yellow rust case is the no-yellow rust in the real observation data)

$$Specificity = \frac{TN}{TN + FP}$$

e. **Precision** is how accurately the model predicted the positive cases

$$Precision = \frac{TP}{TP + FP}$$

3.5.4 Model Extrapolation over Oromia Region

The models with good predictive performance were used to extrapolate the yellow rust probabilities (in the scale of 0 – 1) over wider areas of interest. The maps were generated for 2016, 2017, and 2018, respectively. The key environmental variables from the identified dekad period (for dynamic variable) were re-generated as raster layers from the respective sources (AgER5, CHIRPS ProbaV NDVI, and STRM DEM) of RS products. The R scripts used in the extrapolation are available in Appendices.

4. Results

4.1 Understanding Oromia's Wheat Growing Environment

In Ethiopia, wheat is generally grown at a high elevation of around 1500 – 3200 meters above sea level in cool weather, and sowing happens during Meher, which is the primary crop growing season with rain lasting from June till September, and harvesting is from October through January (USDA, 2015). The period between wheat harvest and sowing (i.e., from March to May) is a minor growing season called Belg with lesser rain, suitable for growing potatoes and yams (Alemayehu et al., 2012, Mohammed et al., 2020).

Figure 12 shows the average maximum/minimum temperature and precipitation (mm) across all the rust observation locations throughout the three years from 2016 to 2018. Meher season (June to September) generally shows an increased amount of rain with moderate temperatures ranging from 10 to 22℃. Rainfall drops drastically from around October till January, which is the season for harvesting. The minimum temperature during this season drops, but stays above 0℃, while the maximum temperature is slowly on the rise, which leads to an increased day-night temperature difference. It is also noticeable that every year there is a rise of rainfall in April-May, right before the Meher season (i.e., the end of Belg season).

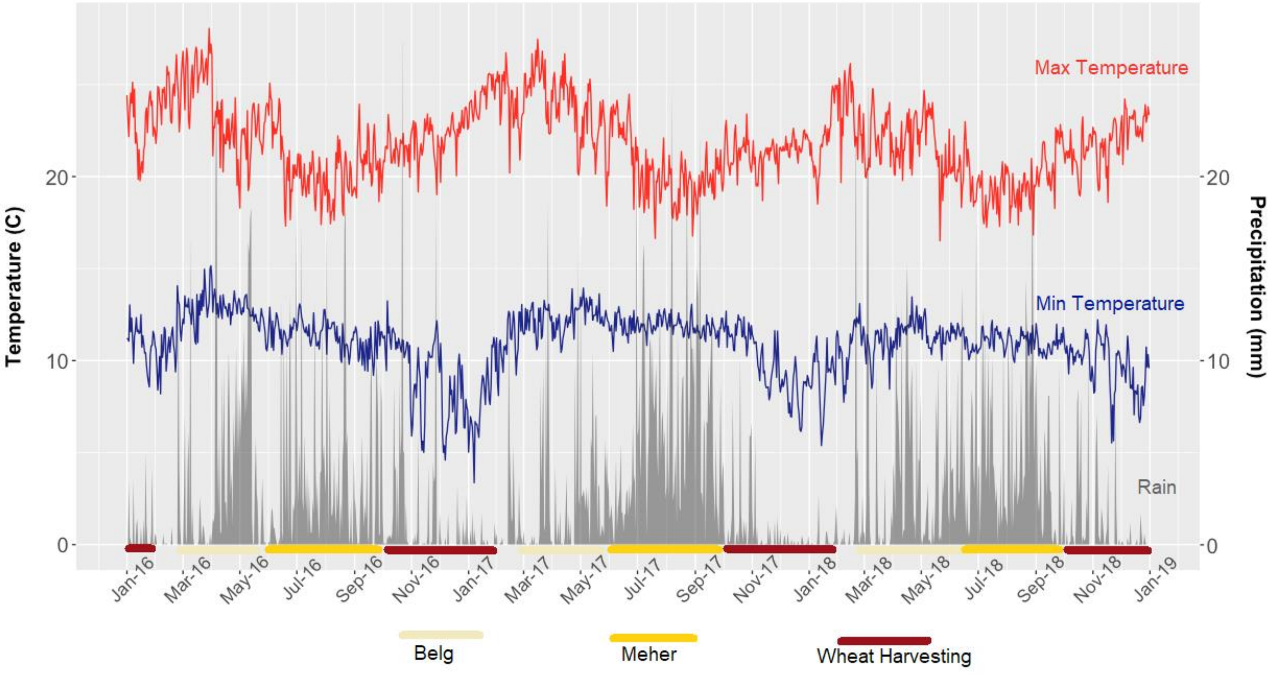


Figure 12: Temperature (max, min) and precipitation at the rust observation points in Oromia 2016–2018

The map below (Figure 13) represents the yellow rust observation locations in the Oromia region with elevation spanning from 1620 to 2978 meters above the sea level.

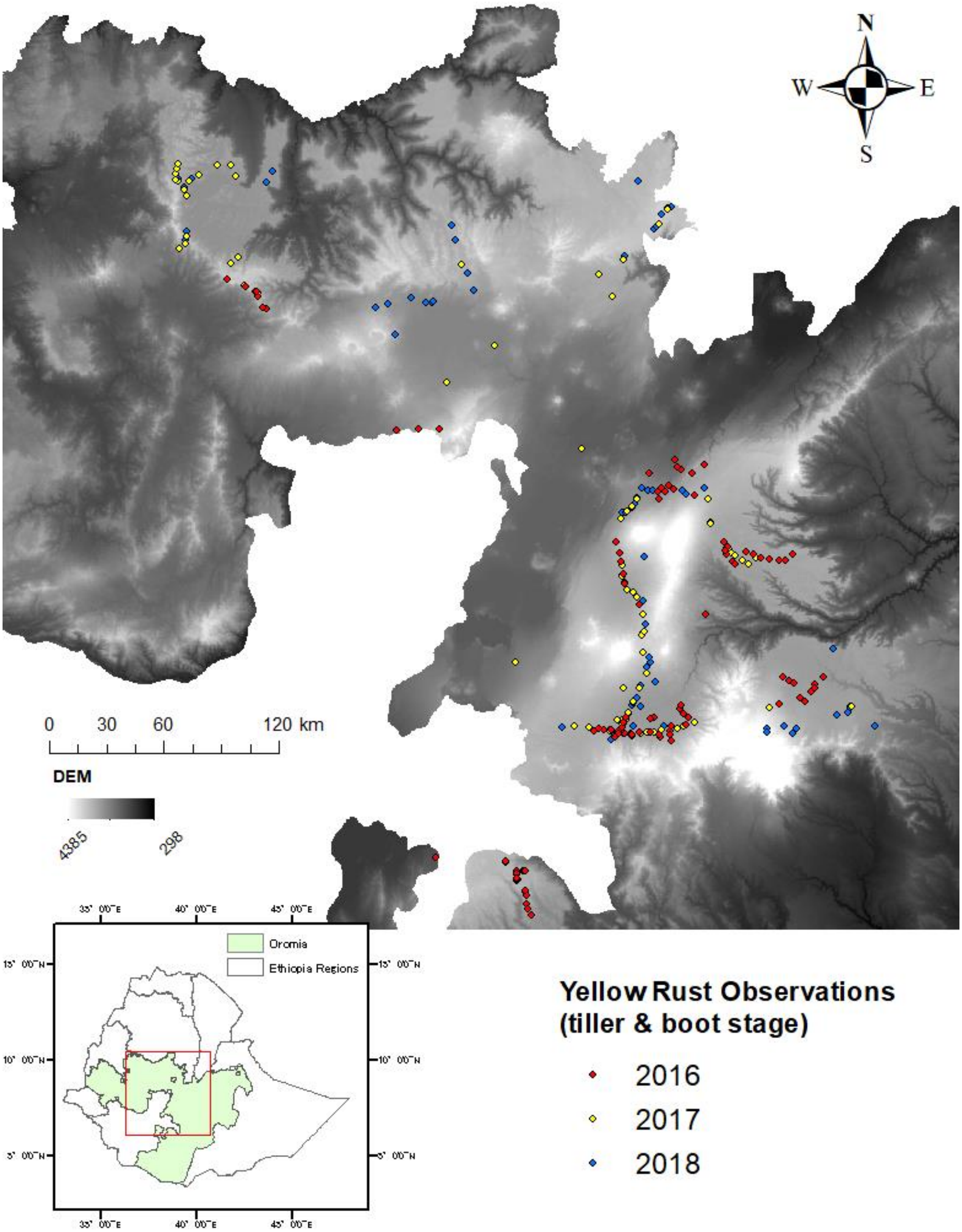


Figure 13: Distribution of yellow rust observation points in Oromia Region

According to the yellow rust observation data from CIMMYT, there is a wide variety of wheat cultivars grown in the region. About one-quarter of them are Digelu or Hidase variety (Figure 14).

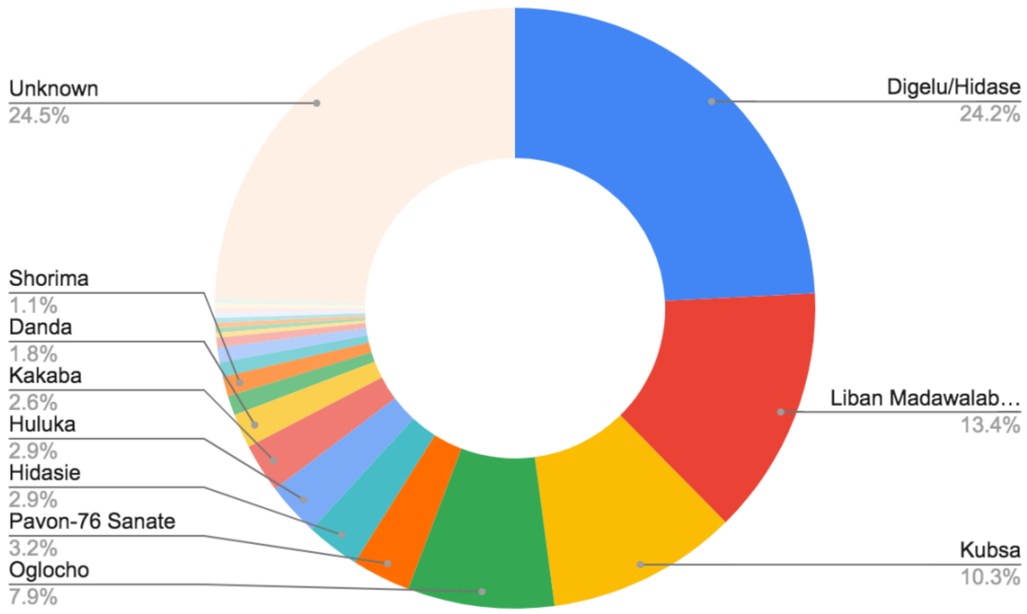


Figure 14: Proportion of different wheat varieties grown in Oromia Region

The table below shows the distribution of yellow rust observation data across different periods in dekad. The data used in this study are limited to the ones recorded at the tiller and boot stage of wheat growth, and most of the observations were recorded in August and September. Based on the growth stage and observation time, the planting period for most of the wheat field observed is anticipated somewhere between June and August, depending on the location. This trend concurs with major crop growing Meher season in Ethiopia.

Table 6: Distribution of rust observation across dekad periods

Month	April			May			June			July			August			September			
Date	1-10	11-20	21-30	1-10	11-20	21-31	1-10	11-20	21-31	1-10	11-20	21-31	1-10	11-20	21-31	1-10	11-20	21-30	
Dekad	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
Tiller											1		71	10	36	10	31		159
Boot			1								3		12	10	51	5	15	2	99

□ Anticipated period of pre-planting/planting season □

Earlier, in the preparation of dynamic environmental variables (precipitation, temperature, and NDVI), variables were prepared for the dekad period from 10 (April 1-11) to 27 (September 21-31). Considering the distribution of rust observations in Table 6, most of the locations are entirely in the wheat-growing season from August (dekad 22). As this study's focus is on pre-season environmental conditions, eventually, dynamic environmental variables were considered only up to dekad 21. Dekad 21 is the end of July and possibly still a pre-planting season for some locations for which yellow rust cases were recorded in late August and September. It reduces the initially prepared 111 environmental variables to 74 variables.

Overwintering or Oversummering?

Based on the rust observation data and the knowledge of general wheat-growing practice in Ethiopia, it seems that the Oromia region does not grow so-called winter wheat, which is usually sown during autumn to yield over the following spring. Temperatures maintain above 0℃ and below 30℃ throughout the year as well as during the off-season.

Past studies examined pathogen termination temperature. At the low-end, under a temperature as low as -4℃ without snow cover, the pathogen can perish together with the host plants (Zadoks, 1961). At the high-end, it is known that urediniospores diminish at the temperature of more than 25℃ for a certain number of days (Dennis, 1987). Tollenaar and Houston (1967) give a variation of temperatures to be considered as pathogen termination points as a 10-day average minimum temperature of 22.3℃ or 10-day average maximum temperature of 32.4℃.

The temperature trend in the Oromia region suggests that the climate is never too cold or too hot for yellow rust to die out, and rather conducive temperature range for the pathogen to oversummer unless there are other conditions to terminate the infection. Meanwhile, in Ethiopia, the term *oversummering* may be somewhat confusing because their off-season (i.e., February to May/June) is not summer as the term is used in the context of Europe or North America. This season is a combination of the post-harvesting period, Belg (short rain period), and early-Meher (primary rain season). The study considered this as *off-season survival of pathogen* instead of using the term *oversummering* to avoid such confusion.

4.2 Association between rust incidence and pre-season environmental condition

In this section, the results of variable exploration are introduced for each of the three data sets below.

a. All observations (data frame: *mydata*)

b. Only tiller-level observations (data frame: *mydata.till*)

c. Observations from Climate Zone B (data frame: *zone.b*)

a. All observations (data frame: mydata)

Figure 15 ranks the variables with univariate Area Under ROC Curve (AUC) scores of more than 0.55. The pattern observed here is that rain-based variables (i.e., precipitation and number of rainy days) dominate higher AUC scores above 0.6. NDVI and temperature variables seem to have less strong univariate association compared to that of rain variables. Terrain characteristics such as slope and aspect show some relevance as an individual variable related to yellow rust cases.

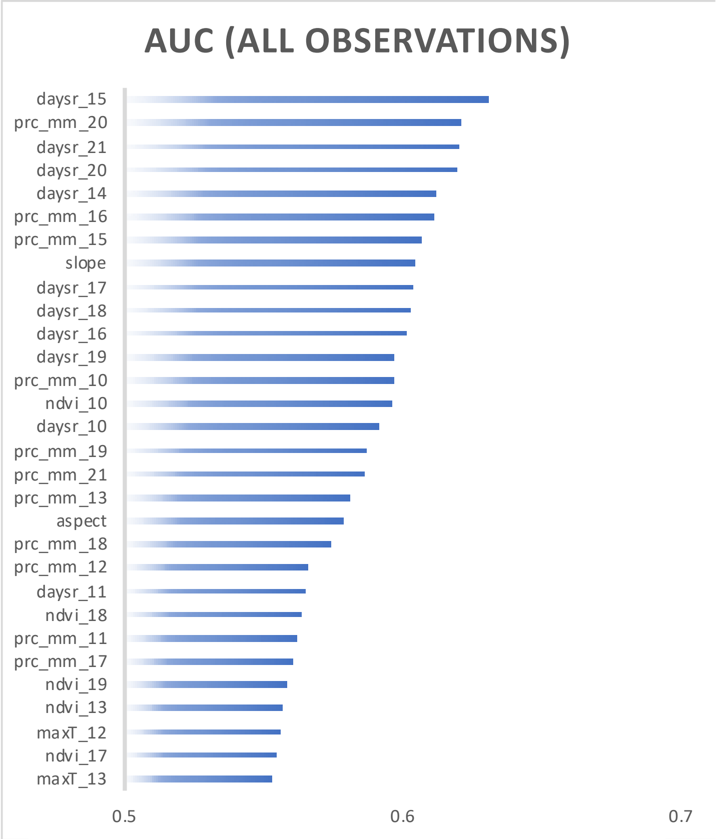


Figure 15: AUC values (all observations)

The result of Classification Tree (CT)-based analysis was a small set of multi-variables to classify yellow rust incidence at the tiller and boot level. A total of 11 variables were selected as a set of multi-variables associated with yellow rust among young wheat. These are:

- Number of days during dekad 20 and 21;
- Precipitation during dekad 10 and 14;
- Maximum temperature from dekad 19;
- Minimum temperature from dekad 14;

- NDVI from dekad 10, 14, and 20;
- Elevation (DEM); and
- Aspect

The table below highlights the selected variables with variable importance (values highlighted) corresponding to the dekad period.

Table 7: CT selected variables and variable importance (all observations)

		Dekad period corresponding to calendar month/date																	
			April			May			June			July			August			September	
		10	20	30	10	20	31	10	20	30	10	20	31	10	20	31	10	20	
		10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*
# of rust observation												4		83	20	87	15	46	2
Dynamic variables	maxT										11								
	minT					8													
	meanT																		
	prc_mm	8				8													
	daysr											5	16						
	ndvi	6				12						9							
Static variables	DEM	6																	
	slope	10																	
	aspect	1																	

Figure 16 below shows the 11 environmental variables' interactions that achieved the best classification result (relative error 0.39, cross-validation error 0.76, and cross-validation standard deviation 0.076). As briefly explained in the methodology section, CT analysis was not for modeling but to explore the variables and narrow down the number of influential variables.

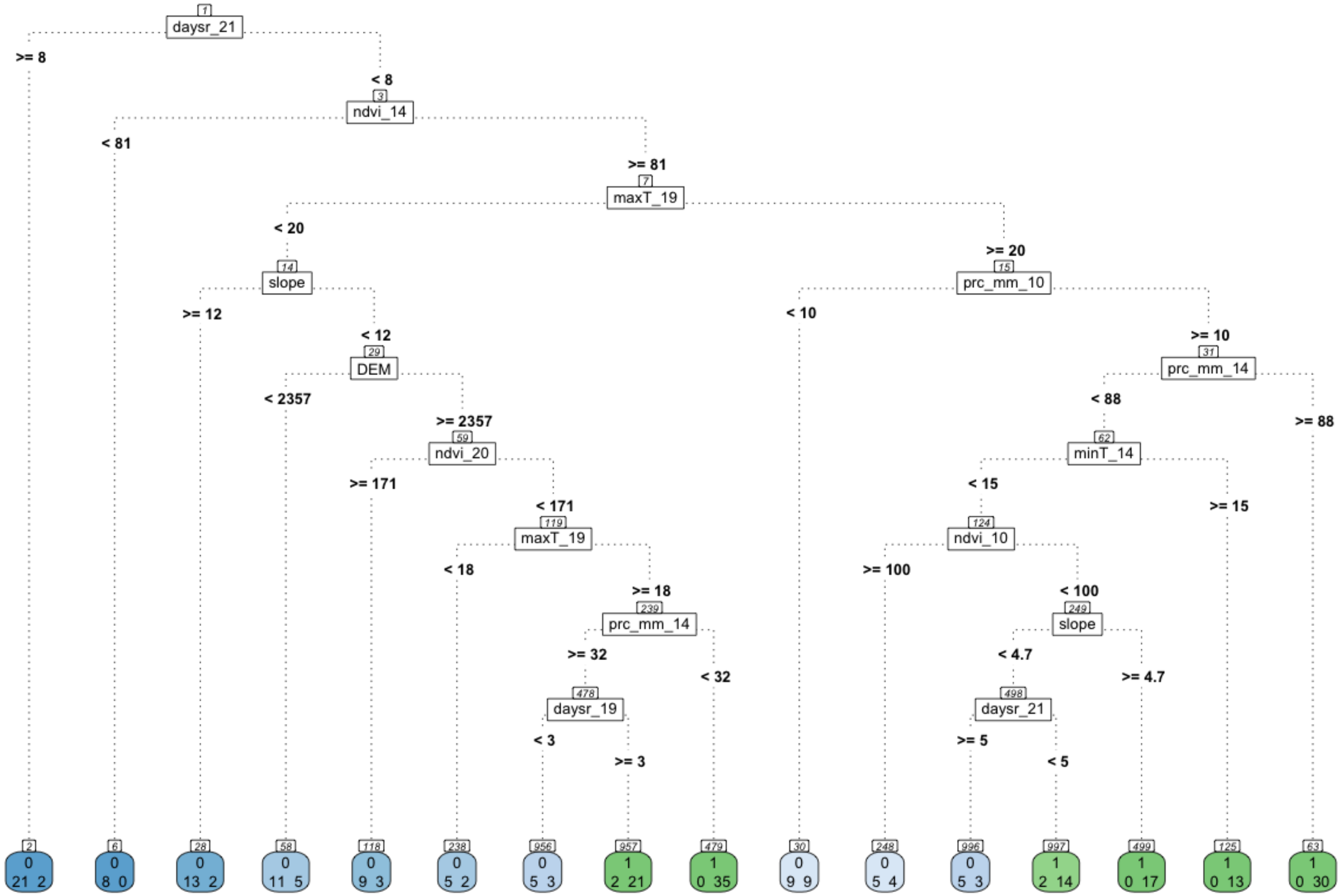


Figure 16: Classification Trees (all observations)

b. Tiller-level dataset (mydata.till)

In the tiller-only data set, the rain-based variables were again more associated with the yellow rust cases (Figure 17). It is particularly the case for the variables with an AUC score above 0.6.

Slope stands out in the tiller-level dataset as highly associated univariate with the early incidence of yellow rust. More NDVI variables appear to have an AUC score of more than 0.55 compared to the previous data set (all-observation data set). There are no temperature-related variables that showed a significant univariate association with a positive rust observation.

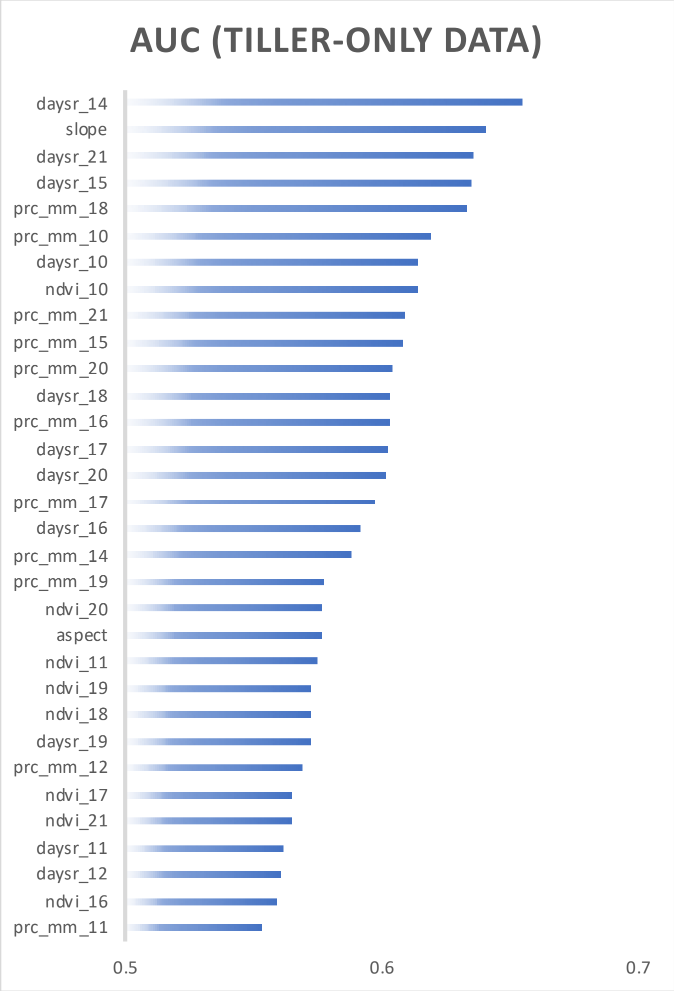


Figure 17: AUC values (tiller-only)

Meanwhile, CT analysis identified a combination of 8 variables that are a mixture of high-AUC variables and low-AUC variables from univariate analysis. Those are:

- Number of rainy days in dekad 21;
- Maximum temperature during dekad 12;
- Minimum temperature during dekad 11;

- Mean temperature during dekad 11;
- NDVI during dekad 19;
- Slope;
- Elevation (DEM); and
- Aspect

The matrix below highlights the selected variables with variable importance (values with highlight) corresponding to the dekad period.

Table 8: CT selected variables and variable importance (tiller-only)

Dekad period corresponding to calendar month/date

		April			May			June			July			August			September		
		10	20	30	10	20	31	10	20	30	10	20	31	10	20	31	10	20	
		10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*
# of rust observation												1		71	10	36	10	31	
Dynamic variables	maxT			21															
	minT		9																
	meanT			19															
	prc_mm																		
	daysr												15						
	ndvi										19								
Static variables	DEM	10																	
	slope	10																	
	aspect	9																	

The Classification Tree (Figure 18) shows that the eight variables' interaction is less complex than the all-observation data set. The splits made by max_12 and min_11 (right side of the tree) show an interaction of maximum temperature and minimum temperature in April. The areas where the temperature ranges from a minimum of 15℃ and a maximum of 22℃ had a clear association with yellow rust cases later on at tiller-level growth. One of the critical variables, daysr_21, classified 13% of the data into 'no yellow rust incidence' when there are more than eight days of rainy days during the dekad.

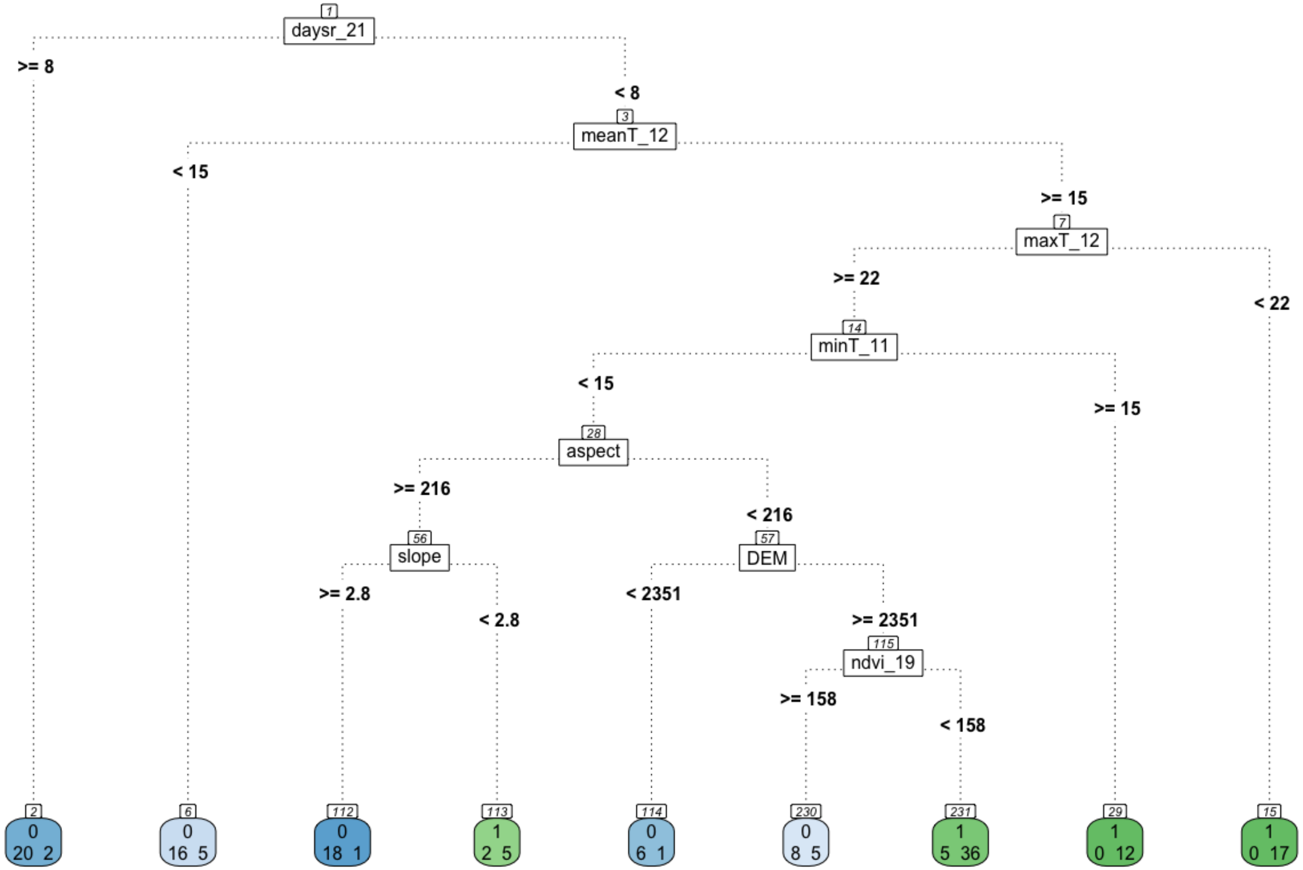


Figure 18: Classification Trees (tiller-only)

c. Observations from the Climate Zone b (data frame: zone.b)

Climate Zone b shares the similar characteristics of the propagation of vegetation in the region (NDVI). The Figure 19 shows those variables scored AUC more than 0.5. In this dataset, three top variables had AUC score more than 0.7 and those are precipitation in dekad 18 and dekad 19; and the number of rainy days during dekad 15. While rain-based variables again show more substantial univariate relevance, the ranking also indicates that several NDVI-based variables and terrain-based variables also have higher AUC values (>0.6), unlike the other two data sets.

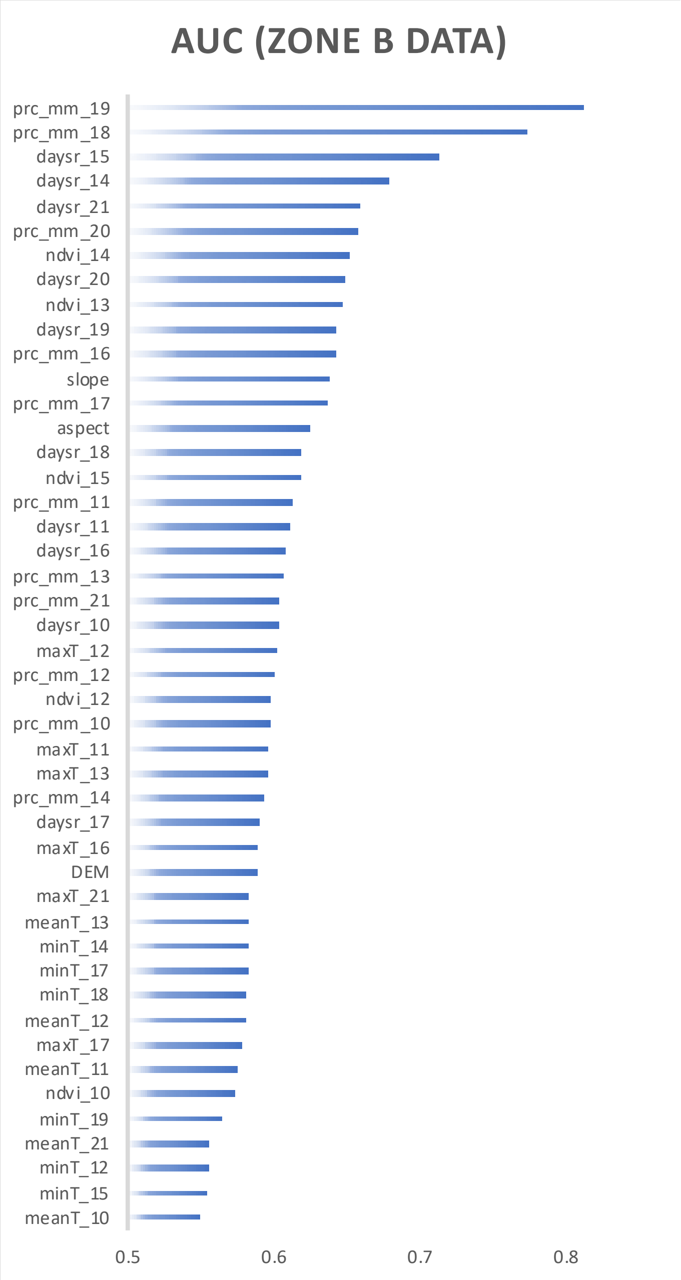


Figure 19: AUC values (Climate Zone b)

The result of CT analysis was a combination of 3 variables:

- Precipitation during dekad 16;
- Precipitation during dekad 18; and
- Aspect

The table below highlights the selected variables with variable importance and corresponding to the dekad period. It indicates that the precipitation during dekad 18 has the highest variable importance, followed by the precipitation during dekad 16 and aspect.

Table 9: CT selected variables and variable importance (Climate Zone b)

Dekad period corresponding to calendar month/date

			April			May			June			July			August			September	
		10	20	30	10	20	31	10	20	30	10	20	31	10	20	31	10	20	
		10	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*
# of rust observation												1		38	5	47	1	19	
Dynamic variables	maxT																		
	minT																		
	meanT																		
	prc_mm							37		45									
	daysr																		
	ndvi																		

Static variables	DEM	
	slope	
	aspect	18

The tree generated from this analysis (Figure 20) shows the interaction of the three variables selected. The tree splits made by the two strong precipitation variables indicate that the rainfall above a certain amount has an association with the absence of yellow rust infection.

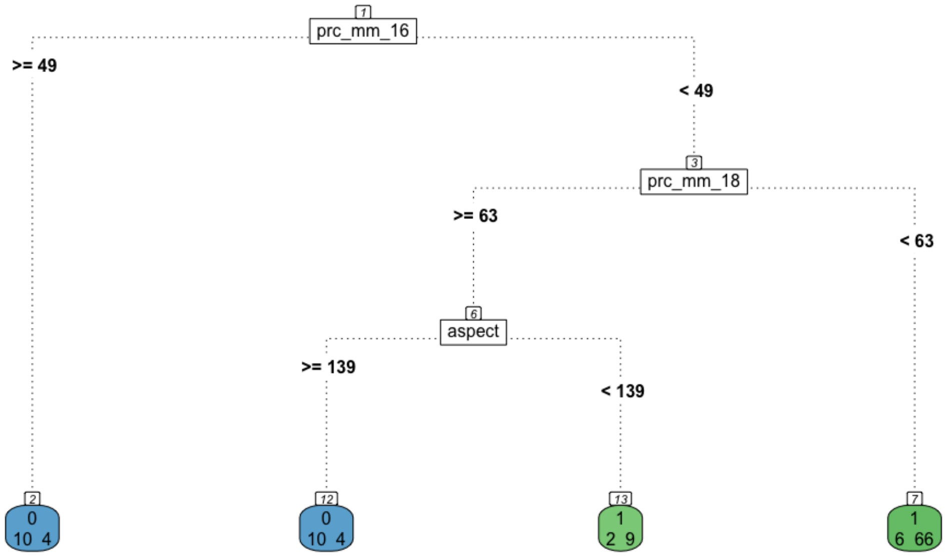


Figure 20: Classification Trees (Climate Zone b)

Summary of variable exploration

Higher relevance of rain-based variables with yellow rust incidence was observed in the simple evaluation of univariate Area Under ROC Curve (AUC) scores across all data sets. The same was observed in the variable importance under the Classification Tree (CT) analysis to identify combinations of variables associated with the yellow rust incidence. The larger the number of total observations, the more the associated variables were identified across a broad spectrum of environmental variables. For example, the dataset mydata (n=258), which entails all the tiller-boot level yellow rust observations from the Oromia region from 2016 to 2018, has 11 variables that seem highly associated with the rust cases. Tiller-only data set, mydata.till (n=159), had 8 multi-variables suggested. Climate zone-based dataset, Zone.b (n=111), had only three variables highly associated with the rust. The set of multi-variable identified was forwarded to the next step to analyze the most critical parameters further.

4.3 Most Influential Environmental Parameters

The previous section selected a small number of multi-variables as rust associated off-season environmental variables. Based on these variables, GAM and BRT models were fit to understand the most influential yellow rust inducing environmental parameters and their interactions. This section addresses the Research Question 1.b (RQ1.b) and uses 70% of the total observations in each data sets described below.

1. All observations (*mydata, training n=182*)
2. Only tiller-level observations (*mydata.till, training n=112*)
3. Observations from Climate Zone B - tiller and boot mixed (*zone.b, training n=79*)

a. All observations (data set: mydata)

The training data set with 182 observations were used in fitting GAM and BRT. In the GAM model, the approximate significance of smooth terms indicates that the number of rainy days in dekad 21 (daysr_21) and maximum temperatures in dekad 19 (maxT_19) are the most critical variables. They are followed by slope, altitude (DEM), and precipitation during dekad 14 (prc_mm_14).

```
> summary(gam.mydata2)
Family: binomial
Link function: logit
Formula:
rust ~ s(maxT_19) + s(prc_mm_10) + s(prc_mm_14) + s(daysr_21)
+ s(ndvi_10) + s(DEM) + s(slope)
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6720     0.1994    3.37 0.000752 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(maxT_19)   1.000  1.000  8.689 0.00320 **
s(prc_mm_10) 2.674  3.272  4.830 0.22636
s(prc_mm_14) 2.590  3.273  6.754 0.09672 .
s(daysr_21)  2.968  3.700 14.597 0.00546 **
s(ndvi_10)   1.000  1.000  1.962 0.16137
s(DEM)       1.000  1.000  2.718 0.09919 .
s(slope)     1.000  1.000  4.586 0.03224 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.285   Deviance explained = 27.9%
-REML = 98.599  Scale est. = 1         n = 182
```

Partial effect plots (Figure 21) show the effect of the respective predictor's smooths that makes up the model. The x-axis of each plot is the value range of the respective predictor. The y-axis indicates the probability of rust incidence occurring according to the fitted model on the scale of 0 to 1. The short, sometimes dense, tick marks on the x-axis are from the observations, and the circles around the smooth lines are partial residuals. The gray shaded area indicates a 95% confidence interval, and the narrower shade indicates improved confidence.

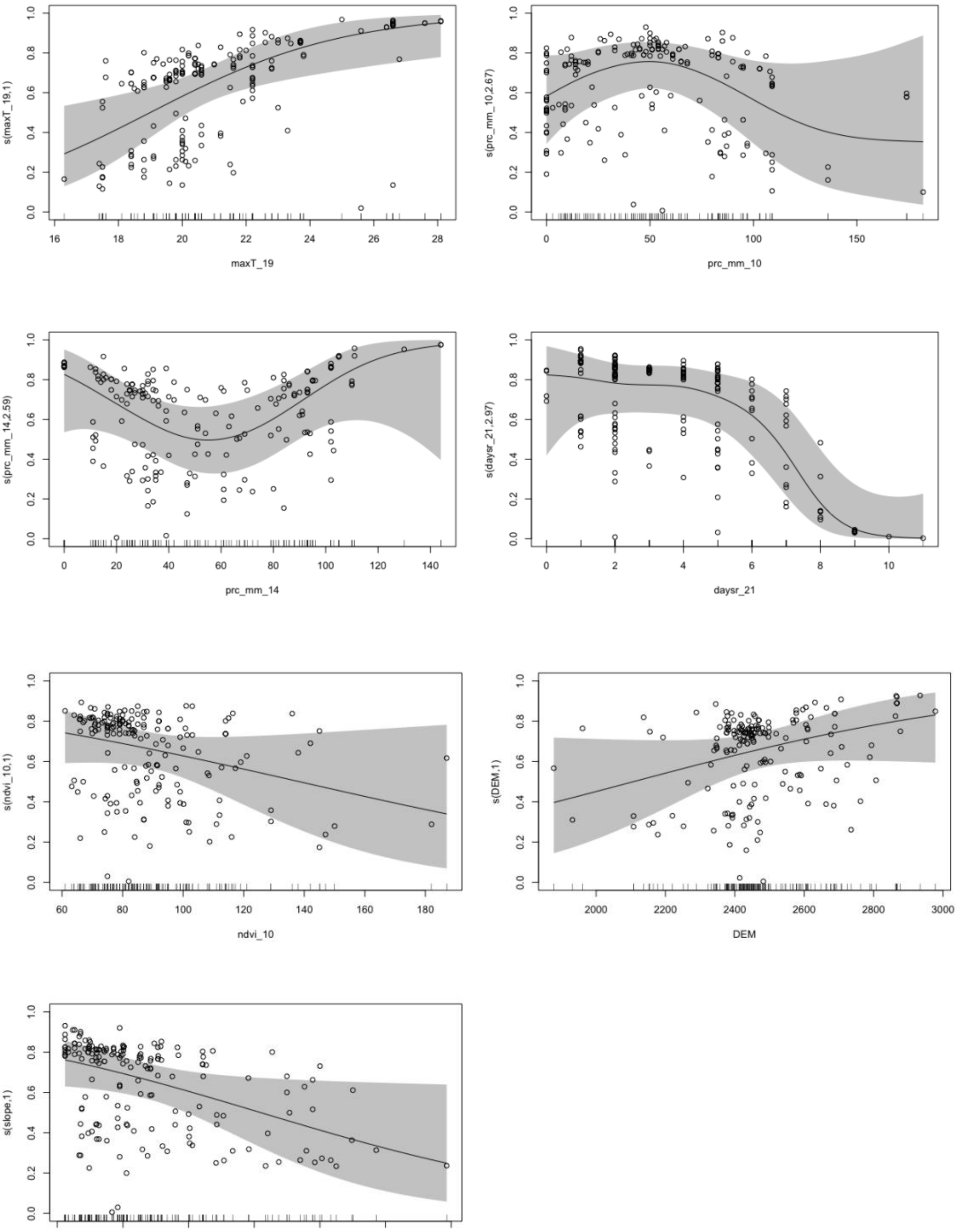


Figure 21: GAM Partial Effect Plot (all observations)

The partial plot for maximum temperature in the early-July (dekad 19) shows that the probability of yellow rust incidence increases when the temperature is warm about 20℃ or above up to about 28℃. Two precipitation periods have contributed to the model with a different smooth line. One is the precipitation during early April (dekad 10), and the other is the precipitation during the middle of May (dekad 14). The scatter plot for the number of rainy days during late July (daysr_21) shows that more than six days of rainy days contributed to a drastic reduction of the probability of rust incidence.

The two static environmental parameters, altitude (DEM) and slope, show that the probability of early-stage yellow rusts incidence tends to increase in the places with milder slope and higher elevation.

Finally, this model's only NDVI parameter comes from early April (ndvi_10, upper-left corner). The scatter plot shows that the NDVI (DN value) between 60 and 90 is related to yellow rust incidence with a higher confidence interval (narrow gray shadow area).

The boxplot below (Figure 22) shows the trend of NDVI in this data set. Early April contributed to this model, and this is when the vegetation indices are at the lowest among the dekad periods.

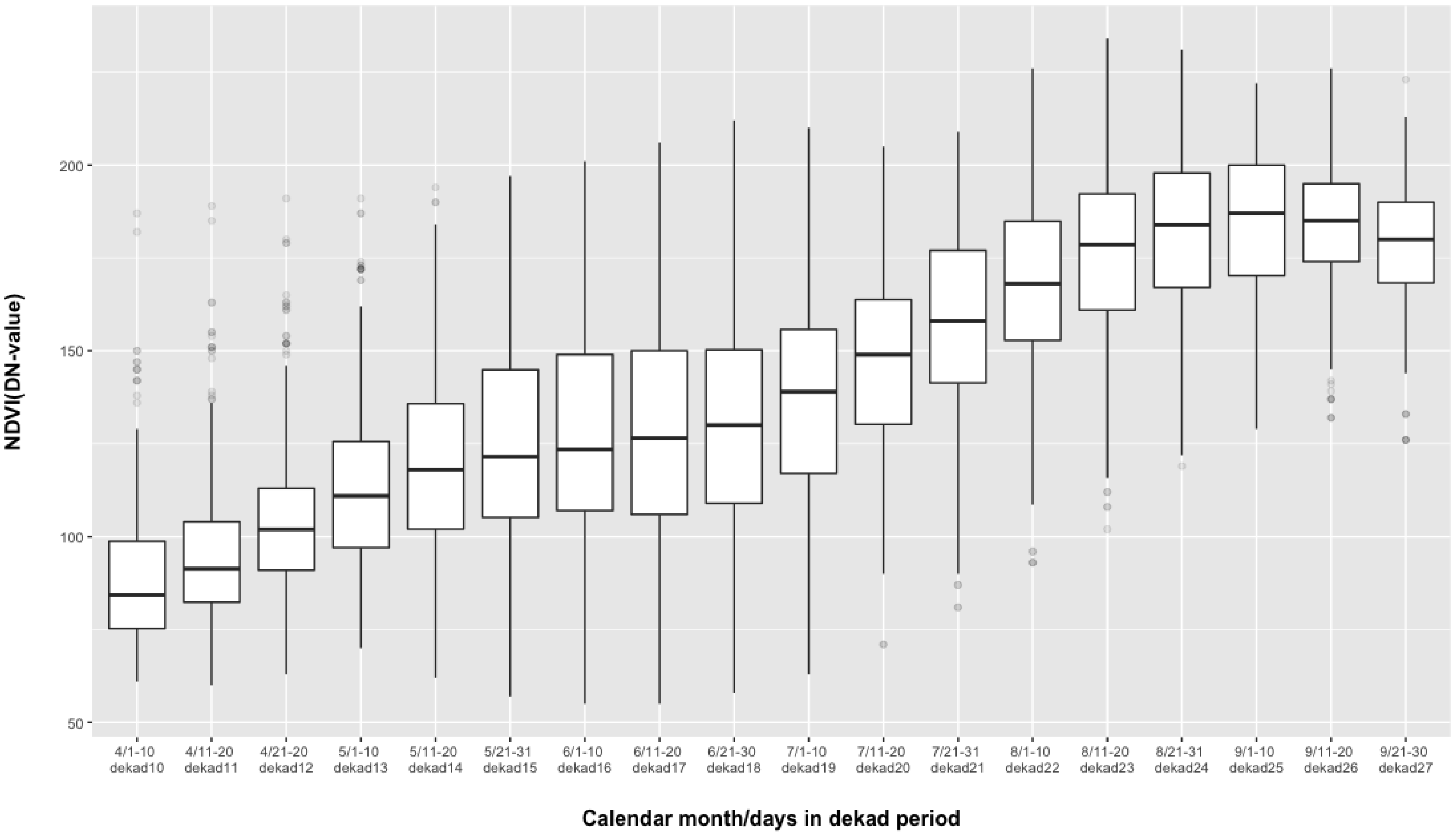


Figure 22: Box chart of NDVI trend (all observations)

Meanwhile, in the BRT model, the number of rainy days in late-July (daysr_21) had the highest variable importance. This is followed by slope, precipitation during early April (prc_mm_10), NDVI during mid-May (ndvi_14), the maximum temperature during early-July (maxT_19), and elevation (DEM).

```
> summary(yrust.tc2.lr001.3)
var         rel.inf
daysr_21    19.391407
slope       18.621347
prc_mm_10   17.009422
ndvi_14     15.945774
maxT_19     14.345800
DEM         10.448235
```

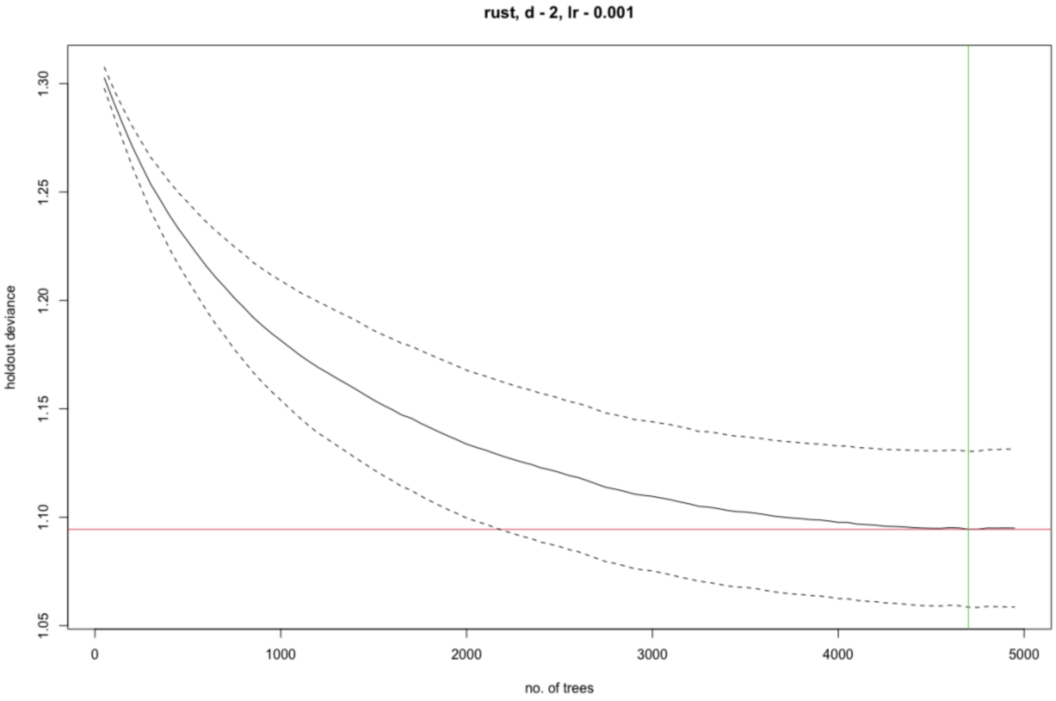


Figure 23: BRT optimal number of trees (all observations)

The black curve line is the mean, and dotted curves indicate one standard error for holdout deviance. The red horizontal line (minimum of the mean holdout deviance) and the green vertical line crosses at the number of trees at which minimum deviance occurs.

There are two noteworthy interactions between the key variables. Those are the interaction among altitude (DEM), slope, and the number of rainy days in dekad 21. Figure 24 visualizes the most critical variable interaction, which is between altitude and slope. It indicates a higher probability of early-stage rust infection (vertical axis in the 3D figure) at an elevation higher than 2400m. The probability is even higher when combined with a slope of less than 10 degrees.

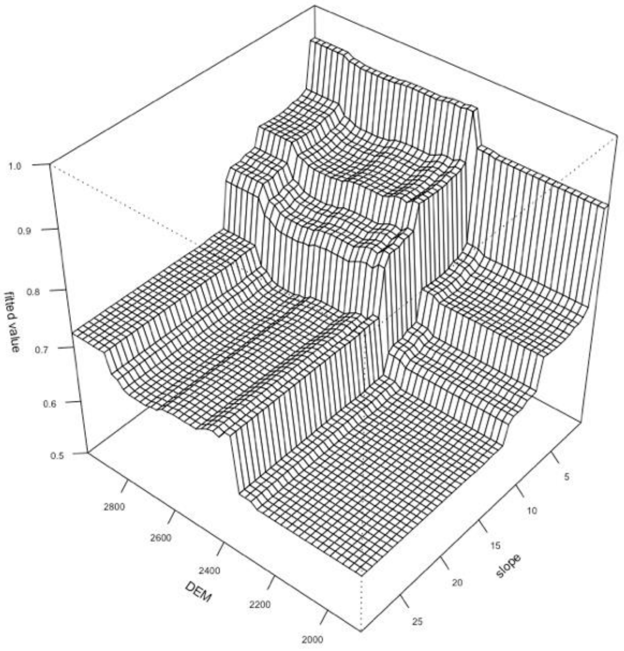


Figure 24: BRT variable interaction between altitude and slope (all observation)

The interaction between altitude and daysr_21 (Figure 25) indicates a higher probability of rust incidence at tiller at the places where elevation is more than 2400m with less than seven days of rain during the off-season period of around 21-31 July.

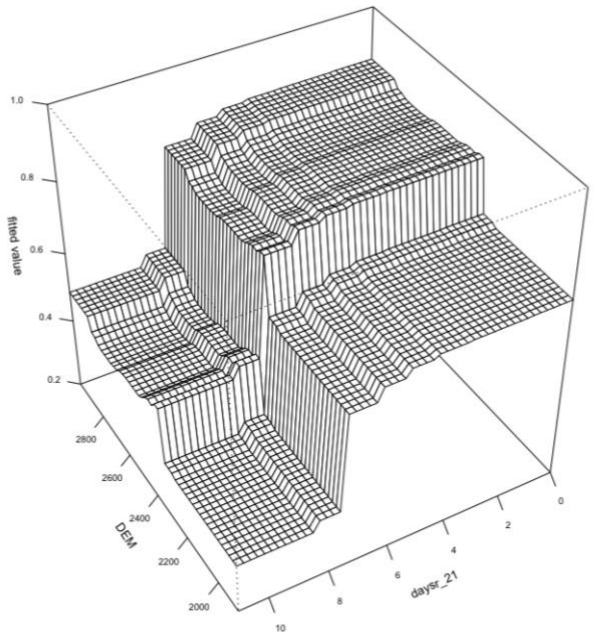


Figure 25: BRT variable interaction between altitude and daysr_21 (all observation)

b. Tiller-level observations (data frame: mydata.till)

The test data set with 112 observations were used to fit GAM and BRT for tiller-level observation data.

The best performing GAM model was based on the mix of variables the number of rainy days during end-July (daysr_21), the minimum temperature during mid-April (minT_11), aspect, slope, altitude (DEM), and NDVI at the beginning of July (ndvi_19).

```
> summary(gam.tiller)
Family: binomial
Link function: logit
Formula:
rust ~ s(daysr_21) + s(minT_11) + s(aspect) + s(slope) + s(DEM)
+ s(ndvi_19)
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.01917    0.26716   0.072    0.943
Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(daysr_21) 3.019  3.791  8.437 0.07825 .
s(minT_11)  4.143  5.138  8.612 0.13311
s(aspect)   1.000  1.000  9.649 0.00189 **
s(slope)    1.048  1.093  4.262 0.05010 .
s(DEM)      1.000  1.000  1.612 0.20425
s(ndvi_19)  2.493  3.149  2.599 0.49131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.384   Deviance explained = 38.6%
-REML = 59.683  Scale est. = 1         n = 112
```

From the summary report, the most significant smooth term is the aspect. Its partial plot (Figure 26) below shows that the aspect of 10-150 degrees (ranging from North-East to South-East) has a higher probability of early-stage yellow rust incidence. The number of rainy days during dekad 21(upper-left corner of Figure 26) and slope are also significant in this model, and they exhibit similar characteristics as the other GAM model fit with the data set with all observations. The number of rainy days of more than six days during this period drastically decreases the probability of rust and a higher probability for the wheat grown on a milder slope.

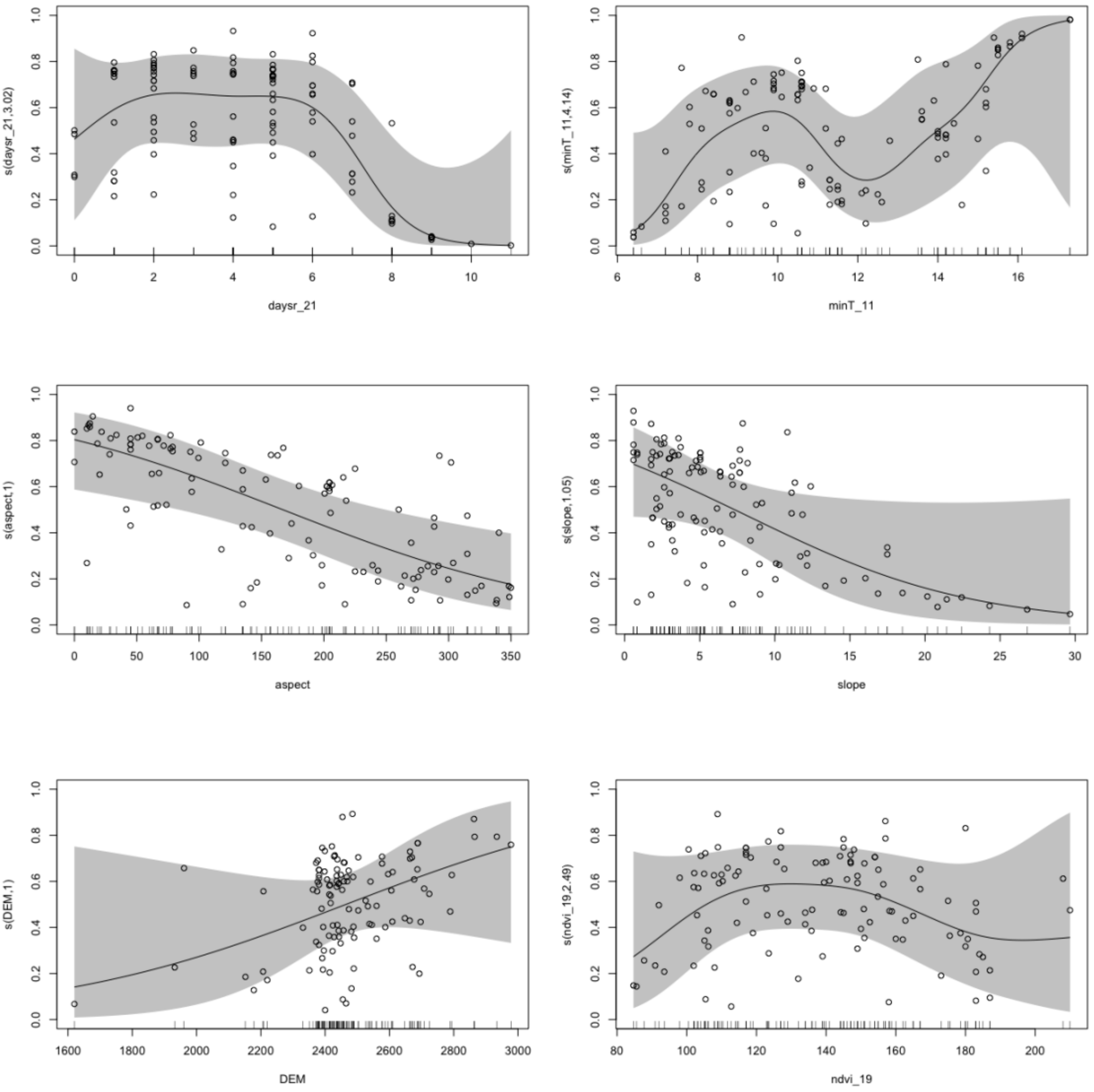


Figure 26: GAM Partial Effect Plot (tiller-only)

It is observed from the width of the partial plots' confidence interval that the static variables such as aspect, slope, and elevation (DEM) tend to show a more distinct relationship with yellow rust incidence than the dynamic variables such as temperature and NDVI.

In the BRT model for tiller-only data, the same variables with higher significance as the GAM model (i.e., aspect, daysr_21, slope) play the top most significant variables.

```
> summary(yrust.tc2.lr.001.c)
var         rel.inf
aspect      25.335903
daysr_21    22.798807
slope       15.913423
maxT_12     14.615917
```

```
ndvi_19     14.203403
DEM         7.132548
```

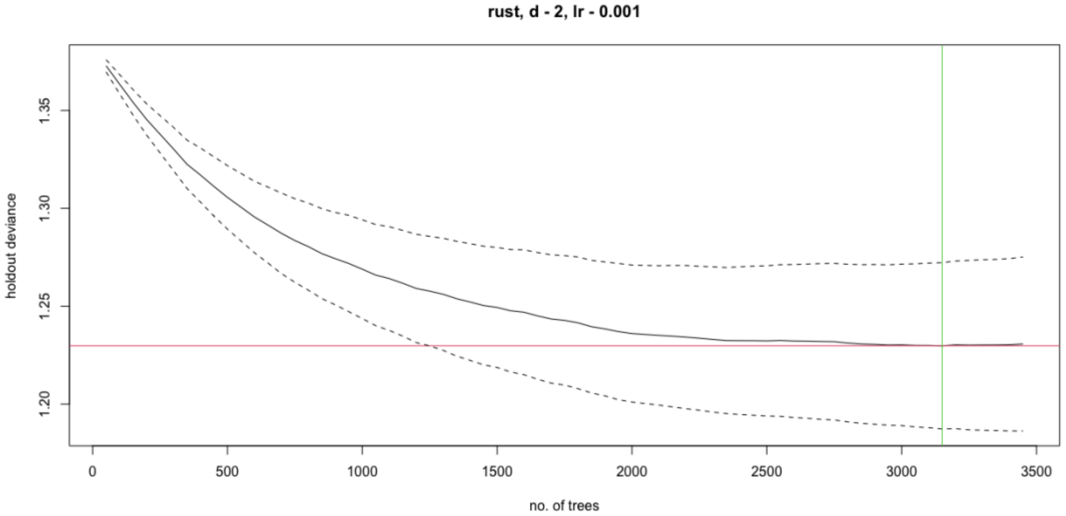


Figure 27: BRT optimal number of trees (tiller-only)

The BRT model for tiller-level observation yielded about 3200 trees as the tree's optimal size, and the model identified two critical variable interactions. One is between maxT_12 and daysr_21, and the other one is between daysr_21 and aspect.

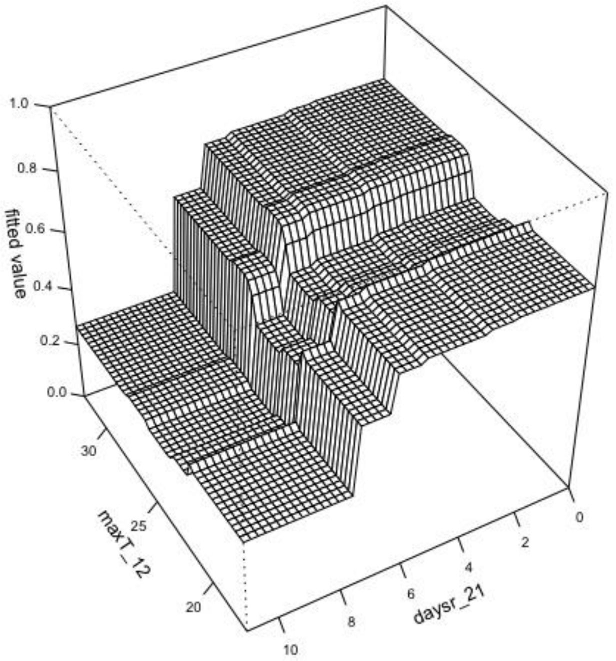


Figure 28: BRT variable interaction between maxT_12 and daysr_21 (tiller-only)

Figure 28 is a 3D visualization of the most critical variable interaction between maximum temperature in dekad 12 and the number of rainy days in dekad 21. While the general trend is that a high number of rainy days lowered the probability, some probabilities variations depend on the level of maximum temperature.

Another 3D variable interaction between aspect and number of rainy days during dekad 21 (Figure 29) shows that the areas with less than 50 degrees aspect (North-East direction) receiving less than five days of rain during this period had the highest probability of yellow rust.

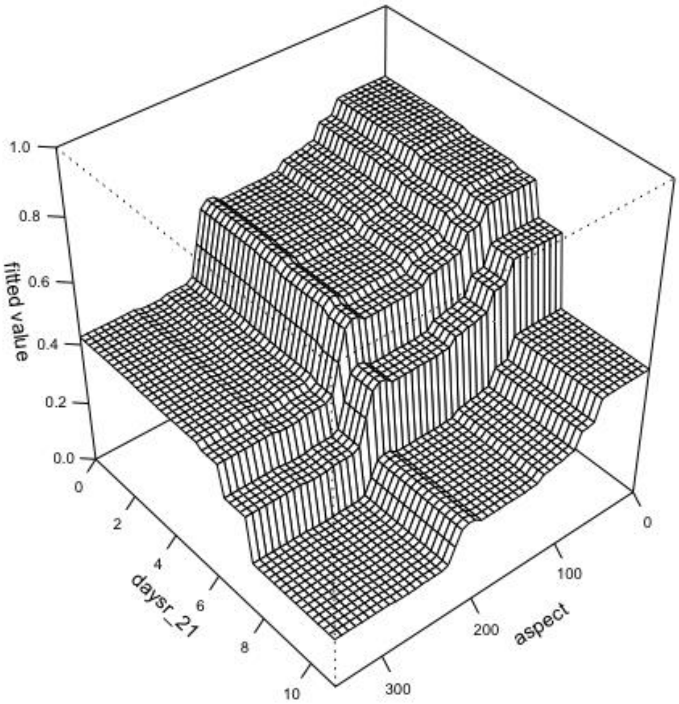


Figure 29: BRT variable interaction between daysr_21 and aspect (tiller-only)

c. Observations from Zone b - tiller and boot mixed (data frame: zone.b)

Zone b training data with 79 observations were used to fit GAM and BRT. Due to the small volume of training data, the BRT model did not fully converge in this data set. Thus, this section presents only the GAM model outcome. The resulting GAM model had precipitation during mid-June (prc_mm_18) and aspect as the only and most important variables.

```
> summary(gam.zoneb)
Family: binomial
Link function: logit
Formula:
rust ~ s(prc_mm_18) + s(aspect)
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4357     0.3357   4.277 1.89e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(prc_mm_18) 1.000  1.000  9.084 0.00258 **
s(aspect)    3.211  3.997 10.952 0.02622 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) =  0.256   Deviance explained = 25.8%
-REML = 37.248  Scale est. = 1         n = 79

> plogis(coef(gam.zoneb)[1])
(Intercept)
  0.8077877
```

The partial effect plots (Figure 30) indicate that the accumulated precipitation in dekad 18 (left) less than 60mm has a higher probability of yellow rust among young wheat later during the season. Further, the places with aspects from 30 to 250 degrees (North-East ~ South ~ South-West) show a higher probability.

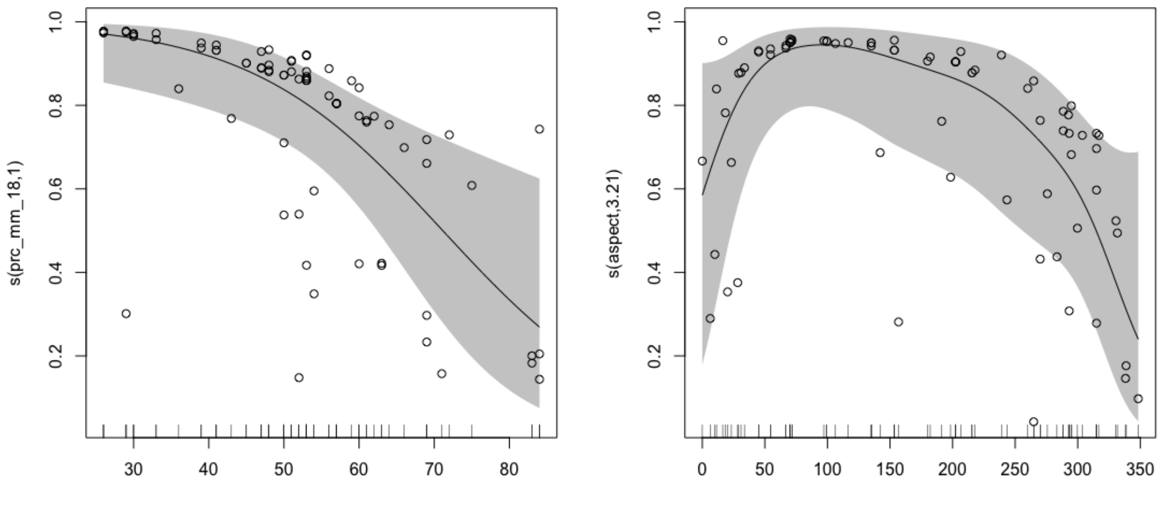


Figure 30: GAM Partial Effect Plot (zone.b)

4.4 Prediction and Accuracy

In the previous section, several yellow rust models were fitted based on the off-season environmental conditions. One GAM model and one BRT model was trained for the data set with all-observations (n=258) and tiller-only data set (n=159). For the Climate Zone b data set (n=111), only the GAM model was trained. The models' predictive capacity was assessed with the 30% test data, based on accuracy, kappa statistics, precision, sensitivity, and specificity. Together with the confusion matrix, the statistics were summarized in the table below.

Table 10: Confusion matrix and model predictive performance statistics

Data	all observation		tiller-only		climate zone b
Observation	n = 258		n = 159		n = 111
Model type	GAM	BRT	GAM	BRT	GAM
Confusion Matrix	FN 9 FP 17 TN 11 TP 39	FN 7 FP 15 TN 13 TP 41	FN 5 FP 8 TN 14 TP 20	FN 6 FP 9 TN 13 TP 19	FN 4 FP 3 TN 5 TP 20
Accuracy Kappa Precision Sensitivity Specificity	0.6579 0.2184 0.6964 0.8125 0.3929	0.7105 0.3386 0.7321 0.8542 0.4643	0.7234 0.44 0.7143 0.8 0.6364	0.6809 0.3538 0.6786 0.76 0.5909	0.7812 0.44 0.8696 0.8333 0.625

Overall, the Climate Zone b GAM model performed the best at a 78% accuracy level with Kappa 0.44 (moderate agreement). The tiller-only GAM model achieved a 72% accuracy level and Kappa 0.44 (moderate agreement). Finally, the all-observation BRT model performed at a 71% accuracy level with Kappa 0.34 (fair agreement). The models performed better in predicting positive yellow rust cases than in predicting no-yellow rust cases. It is observed from the higher values for sensitivity (true-positive rate) than specificity (true-negative rate).

4.5 Model Extrapolation

Based on the predictive capacity assessed in the previous section, the GAM model for tiller-only observations and Climate Zone b were extrapolated over a wider area to visualize the probability of yellow rust incidence at an early stage of wheat growth.

Figure 31 below represents the probability of yellow rust incidence at the early stage of wheat growth (tiller-level) over the Oromia Region. The tiller-only GAM model (accuracy 72%, Kappa 0.44) was used for this. The probability is expressed on the scale of 0-1, and the value closer to 1 indicates a higher probability of yellow rust infection.

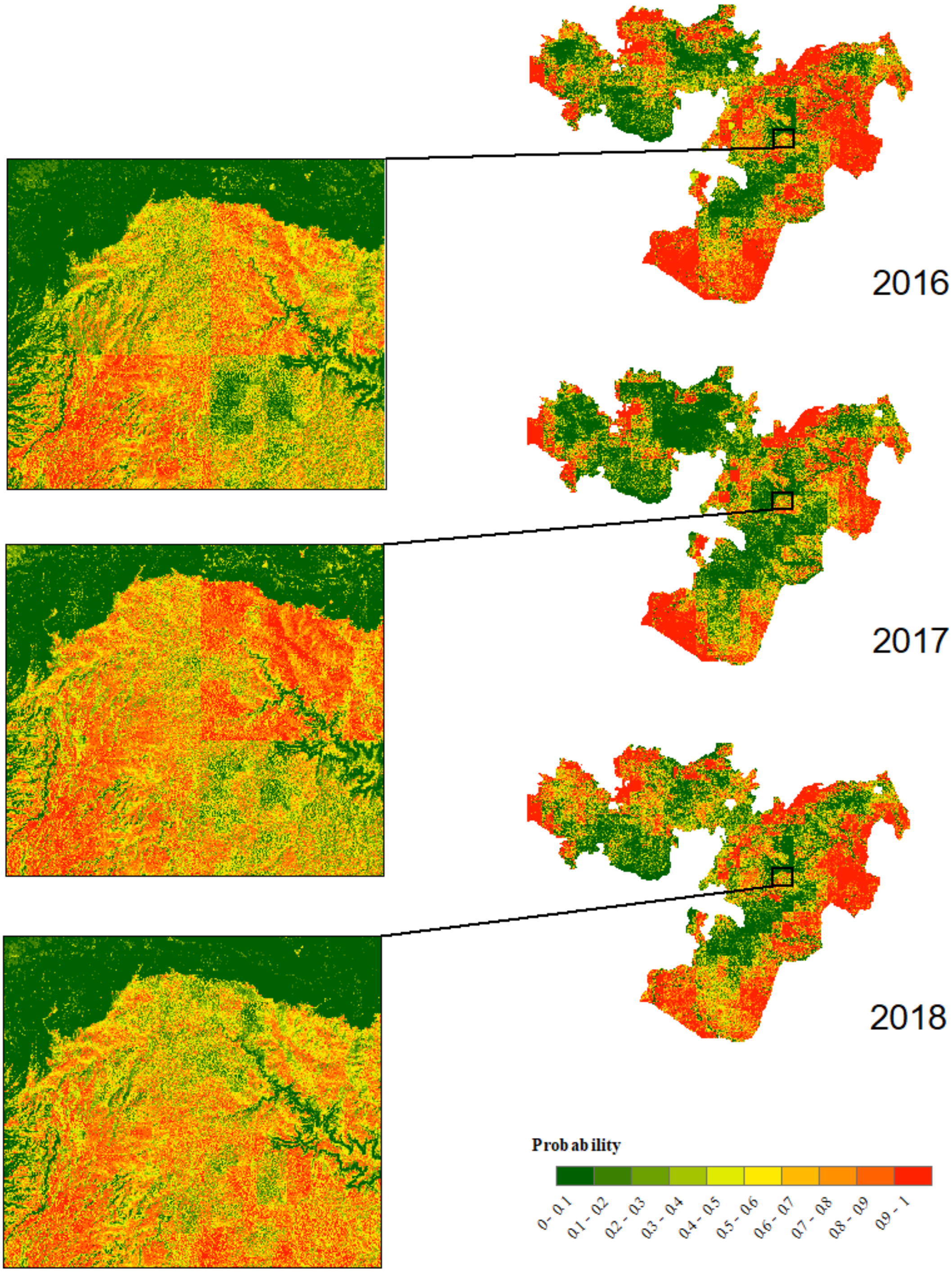


Figure 31: Yellow rust probability maps with the tiller-only GAM model

Similarly, Figure 32 is an extrapolation of the zone.b GAM model (accuracy 78%, Kappa 0.44) over the Climate Zone b. The higher probability of yellow rust incidence is represented by orange and red color.

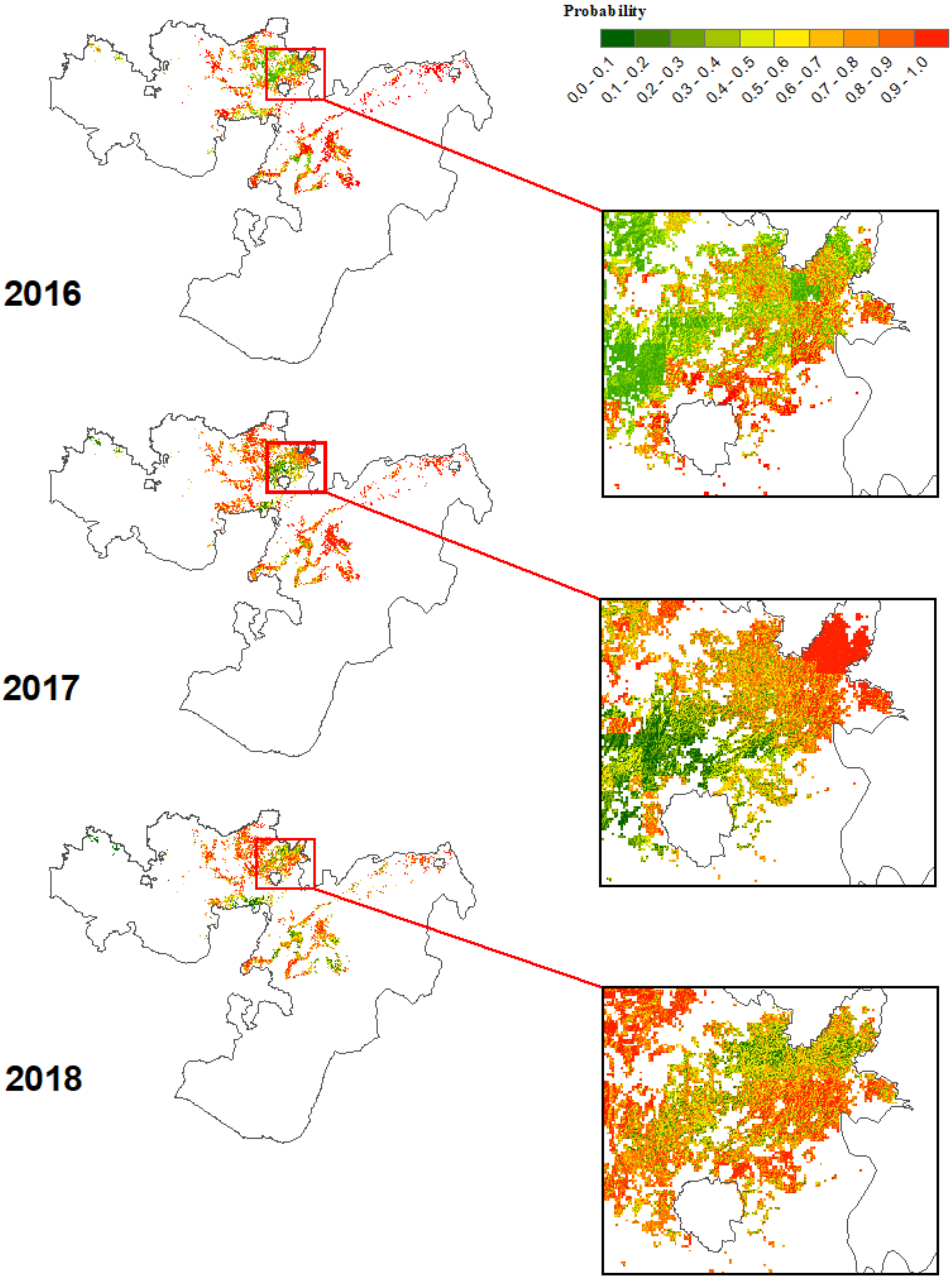


Figure 32: Yellow rust probability map for Climate Zone based on the zone.b GAM model

5. Discussion

This study examined RS-based environmental conditions in Oromia's wheat growing area during the off-season and their relationship with early-stage yellow rust incidence. The study is based on the assumption that yellow rust survives in the field during the off-season after every harvesting, contributing to early local infection of young wheat, which increases the risk of repetitive yellow rust epidemics in the field. This section presents some reflections on the results of the analyses in line with the research questions and recommendations for future research.

Rainfall as a critical parameter

The variable exploration and model training process demonstrated higher importance of rain-based variables during the off-season than temperature, which is often regarded as the critical parameter in the past yellow rust modeling and exploration of pathogen survival during the off-season (Dennis, 1987, Tollenaar and Houston, 1967). The models for all-observation data (n=258) and tiller-observation data (n=159) indicated that more than 6-7 days of rainy days in late July decreased the probability of rust at the tiller and boot stage. Similarly, the model for the Climate Zone b observation data (n=111) indicated higher probabilities of yellow rust incidence in the places rainfall was less at the end of June. This trend recalls that long periods of rain are typically not conducive to the survival of rust during the off-season as it washes the spores from the plants, and low-placed infected leaves can be covered by mud, hence terminating the pathogens (Zadoks and Bouwman, 1985). In the Oromia region, maximum and minimum temperatures tend to remain at around 20-27℃ and 3-13℃, respectively, throughout the off-season. As per the epidemiological studies conducted in the past, these temperatures are not hot or cold enough to terminate the pathogens and provide a stable conducive environment for pathogens' survival. Under such conditions, it is possible that rainfall stood as the critical determiner and the parameter that brings the impact of off-season rust survival on the following wheat season in the Oromia region.

How early is early enough?

The periods that are deemed to be essential for the fitted models are June and July. These months are the beginning of Meher, the primary rain season. Considering the dates of rust observations concentrated around August and September (tiller/boot growth level), June and July are estimated as the period right before or around sowing could happen at many locations. In other words, monitoring weather conditions toward the end of the off-season instead of right after harvesting could lead to an effective early warning of the potential impact of off-season survival of yellow rust pathogen on young wheat plants at the beginning of a new wheat season.

Role of static environmental characteristics

The rust observations used for this study are located at a relatively higher altitude between 1620m and 2978m. According to Tollenaar and Houston (1967), off-season survival (in their study, *oversummering*) of yellow rust is less likely at the elevation below 1829m (6000ft) because of unfavorable summer temperatures at these altitudes. The models trained for all-observation data set and tiller-observation data set were featured partially by altitude (DEM) parameter and confirmed this trend by showing that rust probability increased as the altitude increased. What was unique about this study was the findings around the roles that other static terrain characteristics can play in modeling. For example, milder slope (less than 10 degrees) and aspect less than 200 degrees (facing somewhere in the rage of North-East to South) increased the probability of yellow rust incidence in the new season. It triggers a thought around exposure to sunlight that may or may not influence prudence of pathogen survival on the host plants during the off-season.

Strength of Models

While the comparison of GAM and BRT as modeling method is not the focus of this study, applying different modeling methods helped confirm the most critical environmental variables and arrive at a better-performing model depending on the concerned data set. For the all-observation data set, BRT performed better with 71% accuracy. As for the tiller-observation data set and Climate Zone b data set, GAM models achieved 72% and 78 % accuracy, respectively. Looking at the Kappa statistics, the models for tiller-only observation data and the Climate Zone b data set are moderately reliable, while the all-observation data set requires further improvement. All the models were more robust in sensitivity than their specificity. This indicates that the models predict rust cases better than predicting no-rust cases. For this, the threshold to classify rust case or no-rust (0.5) could be adjusted and see if the performance changes.

One emerging hypothesis from the result of the predictive capacity assessment is that when the observation data is more homogeneous than not, the models could be performing better. For example, the tiller-observation data set is limited to the observations at the tiller growth stage, which are about 2-3 weeks from sowing. Boot-level observations have more time since the time of sowing, and they can be as matured as one month or even two months into the growing stage. The more mature the observed wheat is, the more possible it would be that the rust infection at that time is influenced by additional factors such as in-season rust propagation including longer-distance pathogen infection through the dispersal of urediniospores via wind. An early warning model based on off-season environmental conditions may be more effective if it limits the observations to tiller-level rust cases.

Climate Zone b data performed the best in prediction. This data set has common climate conditions based on the vegetation trend. This model's predictive capacity is the most promising among the three models. However, at this point, it is not clear if the model performed better (accuracy) because of the homogeneity of the data set (i.e., the same climate zone), or it was simply because the data set was small. At the time of data preparation, other Climate Zones were identified based on the NDVI profiles, but due to the number of observations available, they were not included in the modeling process. If a larger volume of observations is made available, climate Zone-based predictive models can be tested further to confirm its strength.

Opportunities for RS-based 'earlier' warning of yellow rust

The study demonstrated the possibility of using solely RS-based freely available data to analyze and model the relationship between yellow rust incidence on young wheat and off-season environmental conditions in Ethiopia's Oromia region. Amid a limited number of country-specific studies on yellow rust epidemic prediction, this study's outcome highlights some opportunities that could pave the way for a functional earlier warning of yellow rust incidence in Ethiopia.

For example, it would be more useful to understand what makes Moderate and High incidence at a very early wheat growth stage. Due to the smaller number of Moderate-High incidence in the sample data, this study's analysis was limited to binary categories. Similarly, yellow rust effects based on cultivar types would add values to modeling and make sense to understand the susceptibility of different wheat variety to yellow rust when certain environmental conditions are met. It was impossible to include this analysis in this study, as one-quarter of the observations did not have a cultivar name assigned to the observation. These aspects would be important in order to prioritize mitigation actions on the ground. If certain areas are prone to more severe incidence of yellow rust than the others, or if some cultivars seem to be more vulnerable to particular environmental conditions than the others, the resources and guidance should be directed to those areas with priority. With a larger volume of data beyond this study's time-scale, some of these additional analyses may be possible.

In this study, the off-season period was identified based on the general crop calendar and the estimated from the date of rust observation recorded at the tiller and boot stage in the absence of planting date information. However, it is worth noting that planting dates could play an important role in better understanding the relationship between the pre-planting environmental conditions and early-stage yellow rust incidence.

The modeling approach is one way to understand and analyze the complexity of the system of yellow rust infection. However, it requires strong empirical knowledge of how biology and physiology behave in a particular set of environment. This study drew the knowledge of yellow rust epidemiology from the work of Zadoks (1961), Tollenaar and Houston (1967), Coakley and Line (1981), Dennis (1987), Rapilly (1979), and Devallavieillepope et al. (1995). While their work is still referred to by many recent rust modeling initiatives, empirical studies around off-season survival of yellow rust are still very limited. Laboratory-based observation of actual survival of yellow rust on specific wheat cultivars in the Ethiopian highland would be highly beneficial for fine-tuning of the environmental parameters used in the modeling. For example, in this, study the rain parameters were designed with the accumulation of precipitation and the number of rainy days (>3mm) over ten days. It may well be the number of consecutive rainy days over a much shorter or longer period of time that matters more.

Similarly, the study was undertaken with the assumption that pathogen's off-season survival on volunteer wheat influences the local infection in the upcoming season. While some unique characteristics were found on this relationship through this study, some field studies of off-season rust survival and local infection in the context of the Oromia region would be beneficial to confirm or modify the model configurations. For example, the presence of volunteer wheat or alternative hosts and their yellow rust infection status in the wheat field could be monitored during the off-season and linked with the nearby observation of yellow rust infection among young wheat in the upcoming wheat season.

6. Conclusion

The research explored the possibility of earlier forecasting of yellow rust infection by looking at the RS-based environmental conditions unique to off-season survival of yellow rust in the Oromia region of Ethiopia. While the epidemiology of yellow rust typically indicates the importance of temperature in the survival of pathogens, the study highlighted additional factors of rain and terrain characteristics that play a key role in the relationship between the off-season environmental conditions and yellow rust incidence on young wheat in the next season. Climate zone-based observations and tiller-only observations generated moderately reliable predictive models (Accuracy > 70%, Kappa > 0.40).

Further analysis is recommended using a larger volume of observation data to confirm the model's general strength and allow for more specific categorical analysis based on different severity of rust incidence and unique cultivars. Little is known or documented empirically about the ground reality of off-season yellow rust survival, especially in the context of Ethiopia. With such additional analysis and empirical knowledge, the approaches tested in this study could be enhanced for its practical application for RS-based early warning of yellow rust in the future.

LIST OF REFERENCES

AIME, M. C., MCTAGGART, A. R., MONDO, S. J. & DUPLESSIS, S. 2017. Chapter Seven - Phylogenetics and Phylogenomics of Rust Fungi. *In:* TOWNSEND, J. P. & WANG, Z. (eds.) *Advances in Genetics.* Academic Press.

ALEMAYEHU, S., PAUL, D. & SINAFIKEH, A. 2012. Crop Production in Ethiopia: Regional Patterns and Trends Summary of ESSP II Working Paper 16,"Crop Production in Ethiopia: Regional Patterns and Trends". International Food Policy Research Institute. Research Note 11.

ALEMU, T. & MENGISTU, A. 2019. Impacts of Climate Change on Food Security in Ethiopia: Adaptation and Mitigation Options: A Review: Soil-Water-Plant Nexus.

ALLEN-SADER, C., THURSTON, W., MEYER, M., NURE, E., BACHA, N., ALEMAYEHU, Y., STUTT, R., SAFKA, D., CRAIG, A. P., DERSO, E., BURGIN, L. E., MILLINGTON, S. C., HORT, M. C., HODSON, D. P. & GILLIGAN, C. A. 2019. An early warning system to predict and mitigate wheat rust diseases in Ethiopia. *Environmental Research Letters,* 14.

BADEBO, A., STUBBS, R. W., VAN GINKEL, M. & GEBEYEHU, G. 1990. Identification of resistance genes to Puccinia striiformis in seedlings of Ethiopian and CIMMYT bread wheat varieties and lines. *Netherlands Journal of Plant Pathology,* 96**,** 199-210.

BEEST, D. E. T., PAVELEY, N. D., SHAW, M. W. & VAN DEN BOSCH, F. 2008. Disease-weather relationships for powdery mildew and yellow rust on winter wheat. *Phytopathology,* 98**,** 609-617.

BOUZID, M., COLÓN-GONZÁLEZ, F. J., LUNG, T., LAKE, I. R. & HUNTER, P. R. 2014. Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever. *BMC Public Health,* 14**,** 781.

CHEN, X. & KANG, Z. 2017. Integrated Control of Stripe Rust. *In:* CHEN, X. & KANG, Z. (eds.) *Stripe Rust.* Dordrecht: Springer Netherlands.

CHEN, X. M. 2005. Epidemiology and control of stripe rust [Puccinia striiformisf. sp.tritici] on wheat. *Canadian Journal of Plant Pathology,* 27**,** 314-337.

COAKLEY, S. M. & LINE, R. F. 1981. Quantitative Relationships Between Climatic Variables and Stripe Rust Epidemics on Winter Wheat. *Phytopathology,* 71**,** 461-467.

COAKLEY, S. M., LINE, R. F. & MCDANIEL, L. R. 1987. Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data.

COHEN, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* 20**,** 37-46.

DE VALLAVIEILLE-POPE, C., BAHRI, B., LECONTE, M., ZURFLUH, O., BELAID, Y., MAGHREBI, E., HUARD, F., HUBER, L., LAUNAY, M. & BANCAL, M. O. 2018.

Thermal generalist behaviour of invasive Puccinia striiformis f. sp tritici strains under current and future climate conditions. *Plant Pathology,* 67**,** 1307-1320.

DENNIS, J. I. 1987. Effect of high temperatures on survival and development of Puccinia striiformis on wheat. *Transactions of the British Mycological Society,* 88**,** 91-96.

DEVALLAVIEILLEPOPE, C., HUBER, L., LECONTE, M. & GOYEAU, H. 1995. Comparative effects of temperature and interrupted wet periods on germination, penetration, and infection of Puccinia-Recondita F SP and P-Striiformis on wheat seedlings. *Phytopathology,* 85**,** 409-415.

ECMWF. 2020. *Agrometeorological indicators from 1979 to 2018 derived from reanalysis* [Online]. Available: https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-agrometeorological-indicators?tab=overview [Accessed 10 April 2020].

ELITH, J., LEATHWICK, J. R. & HASTIE, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology,* 77**,** 802-813.

ERIKSSON, J. 1894. *Ueber die Specialisirung des Parasitismus bei den Getreiderostpilzen*, G. Borntraeger.

EVERSMEYER, M. G. & KRAMER, C. L. 1996. Modeling winter and early spring survival of Puccinia recondita in wheat nurseries during 1980 to 1993. *Plant Disease,* 80**,** 490-493.

EVERSMEYER, M. G. & KRAMER, C. L. 1998. Models of early spring survival of wheat leaf rust in the central Great Plains. *Plant Disease,* 82**,** 987-991.

FAO 2018. Forecasting threats to the food chain affecting food security in countries and regions (July-Sept 2018). *Food Chain Crisis Early Warning Bulletin.* Food and Agriculture Organization.

FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* 29**,** 1189-1232.

FUNK, C., PETERSON, P., LANDSFELD, M., PEDREROS, D., VERDIN, J., SHUKLA, S., HUSAK, G., ROWLAND, J., HARRISON, L., HOELL, A. & MICHAELSEN, J. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data,* 2**,** 150066.

GRABOW, B. 2016. *Environmental Conditions Associated with Stripe Rust and Leaf Rust Epidemics in Kansas Winter Wheat (An Abstract of a Dissertation).* PhD, Kansas State University.

GRABOW, B. S., SHAH, D. A. & DEWOLF, E. D. 2016. Environmental Conditions Associated with Stripe Rust in Kansas Winter Wheat. *Plant Disease,* 100**,** 2306-2312.

HASTIE, T. & TIBSHIRANI, R. 1986. Generalized Additive Models. *Statistical Science,* 1**,** 297-310.

HASTIE, T., TIBSHIRANI, R., SPRINGERLINK & FRIEDMAN, J. 2009. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer New York.

HUMPHRIES, D. L., DEARDEN, K. A., CROOKSTON, B. T., FERNALD, L. C., STEIN, A. D., WOLDEHANNA, T., PENNY, M. E. & BEHRMAN, J. R. 2015. Cross-Sectional and Longitudinal Associations between Household Food Security and Child

Anthropometry at Ages 5 and 8 Years in Ethiopia, India, Peru, and Vietnam. *Journal of Nutrition,* 145**,** 1924-1933.

JALETA, M., HODSON, D., ABEYO, B., YIRGA, C. & ERENSTEIN, O. 2019. Smallholders' coping mechanisms with wheat rust epidemics: Lessons from Ethiopia. *PloS one,* 14**,** e0219327-e0219327.

KUMAR, M., BRAR, A., YADAV, M., CHAWADE, A., VIVEKANAND, V. & PAREEK, N. 2018. Chitinases-Potential Candidates for Enhanced Plant Resistance towards Fungal Pathogens. *Agriculture,* 8.

LARGE, E. C. 1954. GROWTH STAGES IN CEREALS ILLUSTRATION OF THE FEEKES SCALE. 3**,** 128-129.

MARSALIS, M. A. & GOLDBERG, N. P. 2016. *Leaf, Stem, and Stripe Rust Diseases of Wheat* [Online]. College of Agricultural, Consumer and Environmental Sciences, New Mexico State University. Available: https://aces.nmsu.edu/pubs/_a/A415/welcome.html [Accessed].

MARTIN-SHIELDS, C. P. & STOJETZ, W. 2019. Food security and conflict: Empirical challenges and future opportunities for research and policy making on food security and conflict. *World Development,* 119**,** 150-164.

MARTINELLI, F., SCALENGHE, R., DAVINO, S., PANNO, S., SCUDERI, G., RUISI, P., VILLA, P., STROPPIANA, D., BOSCHETTI, M., GOULART, L. R., DAVIS, C. E. & DANDEKAR, A. M. 2015. Advanced methods of plant disease detection. A review. *Agronomy for Sustainable Development,* 35**,** 1-25.

MEHMOOD, S., SAJID, M., ZHAO, J., HUANG, L. & KANG, Z. 2020. Alternate Hosts of Puccinia striiformis f. sp. tritici and Their Role. *Pathogens,* 9**,** 434.

MENGESHA, G. G. 2020. Management of yellow rust (Puccinia striiformis f.sp. tritici) and stem rust (Puccinia graminis f.sp tritici) of bread wheat through host resistance and fungicide application in Southern Ethiopia. *Cogent Food & Agriculture,* 6.

MOHAMMED, I., MARSHALL, M., DE BIE, K., ESTES, L. & NELSON, A. 2020. A blended census and multiscale remote sensing approach to probabilistic cropland mapping in complex landscapes. *ISPRS Journal of Photogrammetry and Remote Sensing,* 161**,** 233-245.

NAGHIBI, S. A., POURGHASEMI, H. R. & DIXON, B. 2016. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental Monitoring and Assessment,* 188.

NELDER, J. A. & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General),* 135**,** 370.

NEWLANDS, N. K. 2018. Model-Based Forecasting of Agricultural Crop Disease Risk at the Regional Scale, Integrating Airborne Inoculum, Environmental, and Satellite-Based Monitoring Data. *Frontiers in Environmental Science,* 6.

OLIVERA, P., NEWCOMB, M., SZABO, L. J., ROUSE, M., JOHNSON, J., GALE, S., LUSTER, D. G., HODSON, D., COX, J. A., BURGIN, L., HORT, M., GILLIGAN, C. A., PATPOUR, M., JUSTESEN, A. F., HOVMOLLER, M. S., WOLDEAB, G., HAILU, E., HUNDIE, B., TADESSE, K., PUMPHREY, M., SINGH, R. P. & JIN, Y. 2015. Phenotypic and Genotypic Characterization of Race TKTTF of Puccinia graminis f. sp tritici that Caused a Wheat Stem Rust Epidemic in Southern Ethiopia in 2013-14. *Phytopathology,* 105**,** 917-928.

PARK, R. F. 1990. The role of temperature and rainfall in the epidemiology of Puccinia striiformis f.sp. tritici in the summer rainfall area of eastern Australia. *Plant Pathology,* 39**,** 416.

PITTMAN, S. J. & BROWN, K. A. 2011. Multi-Scale Approach for Predicting Fish Species Distributions across Coral Reef Seascapes. *PLoS ONE,* 6**,** e20583.

RAPILLY, F. 1979. Yello Rust Epidemiology. *Annual Review of Phytopathology,* 17**,** 59-73.

RAVINDRA, K., RATTAN, P., MOR, S. & AGGARWAL, A. N. 2019. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International,* 132**,** 104987.

ROELFS, A. P., SINGH, R. P. & SAARI, E. E. 1992. *Rust Diseases of Wheat: Concepts and methods of disease management,* Mexico, CIMMYT.

SCHAPIRE, R. E. 2003. The Boosting Approach to Machine Learning: An Overview. Springer New York.

SHARMA-POUDYAL, D., CHEN, X. & RUPP, R. A. 2014. Potential oversummering and overwintering regions for the wheat stripe rust pathogen in the contiguous United States. *International Journal of Biometeorology,* 58**,** 987-997.

SUÁREZ-SEOANE, S., OSBORNE, P. E. & ALONSO, J. C. 2002. Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalized additive models. *Journal of Applied Ecology,* 39**,** 755-771.

TOLLENAAR, H. & HOUSTON, B. R. 1967. A Study on the Epidemiology of Stripe Rust, Puccinia Striiformis West., in California. *Canadian Journal of Botany,* 45**,** 291-307.

USDA 2015. Commodity Intelligence Report. *In:* UNITED STATES DEPARTMENT OF AGRICULTURE, F. A. S. (ed.). United States Department of Agriculture, Foreign Agriculture Service.

USGS. *Digital Elevation - Shuttle Radar Topography Mission* [Online]. Available: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1-arc?qt-science_center_objects=0#qt-science_center_objects [Accessed 23 July 2020].

VITO 2019. Product User Manual (draft) Normalized Difference Vegetation Index Collection 1km Vr.3. *In:* CONSORTIUM, C.-G. L. (ed.). Copernicus Global Land Service (GCLS).

WANG, M. N. & CHEN, X. M. 2013. First Report of Oregon Grape (Mahonia aquifolium) as an Alternate Host for the Wheat Stripe Rust Pathogen (Puccinia striiformis f. sp tritici) Under Artificial Inoculation. *Plant Disease,* 97, 839-839.

WB 2020. Ethiopia Poverty Assessment: Harnessing Continued Growth for Accelerated Poverty Reduction. Washington DC: World Bank.

WELLINGS, C. R. 2011. Global status of stripe rust: a review of historical and current threats. *Euphytica,* 179, 129-141.

XU, X., MA, L. & HU, X. 2019. Overwintering of Wheat Stripe Rust Under Field Conditions in the Northwestern Regions of China. *Plant Disease,* 103, 638-644.

YEE, T. W. & MITCHELL, N. D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science,* 2, 587-602.

YOUSSEF, A. M., POURGHASEMI, H. R., POURTAGHI, Z. S. & AL-KATHEERI, M. M. 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides,* 13, 839-856.

ZADOKS, J. C. 1961. Yellow rust on wheat studies in epidemiology and physiologic specialization. *Tijdschrift Over Plantenziekten,* 67, 69-256.

ZADOKS, J. C. & BOUWMAN, J. J. 1985. 11 - Epidemiology in Europe. *In:* ROELFS, A. P. & BUSHNELL, W. R. (eds.) *Diseases, Distribution, Epidemiology, and Control.* Academic Press.

ZELLWEGER, F., BRAUNISCH, V., BALTENSWEILER, A. & BOLLMANN, K. 2013. Remotely sensed forest structural complexity predicts multi species occurrence at the landscape scale. *Forest Ecology and Management,* 307, 303-312.

APPENDICES

Chapter 3. Method

3.4 Data Processing

Retrieval of precipitation and temperature data on GEE

Google Earth Engine (GEE), point-data extraction of precipitation (CHIRPS), and temperature (AgERA5). Below is the example of precipitation data extracted for the month of April 2018. (Javascript reference)

```
var aoi: Table users/ccendoo/OromiaYellowRust2018_TB
print(aoi);
Map.addLayer(aoi);

// Script to extract CHIRPS precipitation values with point data on
Google Earth Engine

var start = ('2016-04-01');
var end = ('2016-05-01');

// Daily precipitation - load in image collection and filter by area and
date
var era5_prec = ee.ImageCollection('ECMWF/ERA5/DAILY')
                  .select('total_precipitation')
                  .filter(ee.Filter.date(start, end))
                  .map(function(image){return image.clip(aoi)});
//Clips data based on "aoi"

print('collection', era5_prec);

//Create variables and extract data
var scale = era5_prec.first().projection().nominalScale().multiply(0.5);
print(scale);
era5_prec = era5_prec.filter(ee.Filter.listContains('system:band_names',
era5_prec.first().bandNames().get(0)));

var ft = ee.FeatureCollection(ee.List([]));
//Function to extract values from image collection based on point file
and export as a table
var fill = function(img, ini) {
  var inift = ee.FeatureCollection(ini);
  var ft2 = img.reduceRegions(aoi, ee.Reducer.first(), scale);
  var date = img.date().format("YYYYMMdd");
  var ft3 = ft2.map(function(f){return f.set("date", date)});
return inift.merge(ft3);
};

// Iterates over the ImageCollection
var profile = ee.FeatureCollection(era5_prec.iterate(fill, ft));
print(profile,'profile');

// Export
Export.table.toDrive({
  collection : profile,
  description : "ERA5_prec-"+start+"-"+end,
  fileNamePrefix : "ERA5_prec-"+start+"-"+end,
  fileFormat : 'CSV',
  folder: 'ERA5',
  selectors: ["date","first"]
});
```

3.5 Analysis

3.5.1 Variable Exploration (RQ1.a)

Univariate AUC values

```
library(tidyverse)
set.seed(123)
folds <- createFolds(mydata2$rust, k=10)
# The result of this is a list of vectors storing the row numbers for
each of the k=10 requested folds.

library(caret)
# Use lapply() to conduct identical steps to calculate the Area Under
ROC curve (AUC) for each fold

rocVal <- lapply(folds, function(x){
  test <- mydata2[x, ]
  train <- mydata2[-x, ]
  rocVal <- filterVarImp(x = train[ , -1], y = train$rust)
})

#Combine list of all 10-fold AUC data frames into one to calculate a
mean
library(data.table) # to activate rbindlist()
rocVal_comb <- Map(cbind, rocVal, predictor = lapply(rocVal, rownames))
rocVal.mydata2 <- rbindlist(rocVal_comb, idcol = TRUE) %>%
  group_by(predictor) %>%
  summarise_at(vars(X0, X1), list(mean_ROC = mean))
```

Classification Tree (‘cart’ package)

```
library(rpart)
library(rpart.plot)
cart.model <- rpart(rust ~ var1 + var2 + ... + varX,
                     method = 'class', #classification
                    data = dataset,
                    parms = list(split='information'),
                    control = rpart.control(cp=0.001))

               #cp = complexity parameter
              #The parameter ‘information’ is a splitting criterion, and
              it is also called entropy index that forms the category
              groups by minimizing the within-group diversity.

rpart.plot(cart.model, type = 5, extra = 1, branch.lty = 3, box.palette
= "auto", nn=TRUE)

summary(cart.model)
plotcp(cart.model) # plot complexity parameter
```

3.5.2 Finding the most critical variables (RQ1.b)

General Additive Model (GAM)

```
library(mgcv)
gam.model <- gam(rust ~ s(var1) + s(var2) + ... + s(varX),
                   data = dataset,
                   family = binomial, #Classification
                   method = "REML")

# "REML": Restricted Maximum Likelihood method: automatic smooth
parameter selection

summary(gam.mydata2)
plogis(coef(gam.mydata2)[1])
gam.check(gam.mydata2)

par(mfrow = c(2, 2))
plot(gam.mydata2, pages = 2,
     trans =plogis, # transform y-axis to 0-1 scale
     shift = coef(gam.mydata2)[1], # adding model intercept
     seWithMean = TRUE, # consider intercept uncertainty
     residuals = TRUE, pch = 1, cex = 1,
     shade = TRUE)
```

```
GAM concurvity report
> concurvity(gam.tiller, full = TRUE)
                 para s(daysr_21) s(minT_11) s(aspect)  s(slope)    s(DEM) s(ndvi_19)
worst    1.358463e-18   0.7813170  0.7784903 0.6610410 0.7400053 0.7767376  0.7792632
observed 1.358463e-18   0.6680082  0.5369568 0.4489580 0.6543160 0.5753127  0.5027777
estimate 1.358463e-18   0.5712507  0.5238092 0.4527114 0.5995162 0.5561714  0.6337436
```

In the convurvity report when the values on 'worst' is above 8, there is a possibility of concurvity and adjustment in predictors may be required. Thus, the details should be checked for each variable against the other variables to find out which combination of variables have concurvity.

```
library(GGally)
library(tidyverse)
# Checking potential collinearity
dataset %>% ggpairs(columns = c("maxT_19","minT_14", "prc_mm_10",
                             "prc_mm_14", "daysr_20", "daysr_21",
                             "ndvi_10", "ndvi_14", "ndvi_20", "DEM",
                             "slope"),
                upper = list(continuous = wrap('cor', size =4)),
                lower = list(combo = wrap("facethist", bins = 30)))

# VIF calculation
VIFcalc(data.frame(dataset$maxT_19, dataset $minT_14, dataset
$prc_mm_10,
                   dataset $prc_mm_14,dataset$daysr_20,dataset$daysr_21,
                  dataset $ndvi_10, dataset $ndvi_14, dataset$ndvi_20,
                  dataset$DEM, dataset$slope))


# GAM PREDICTION
predictions <- predict(gam.model, newdata = dataset,
                       type = "link", se.fit = TRUE)
```

Boosted Regression Tree (BRT)

```
library(gbm)
library(dismo)
source("brt.functions.R")
yrust.tc2.lr.001 <- gbm.step(data=mydata2train,
                             gbm.x =c(53,10,99,3,14,69,84,38,25,39,19),
                             gbm.y = 1,
                             family = "bernoulli",  #binomial model
                             tree.complexity = 2,
                             learning.rate = 0.001,
                             bag.fraction = 0.75)
# bag fraction specifies proportion of data to be selected at each step
(In this case, 75% of the data is drawn at random)


# Identify important interactions of variables(pair-wise interactions)
find.int <- gbm.interactions(yrust.tc2.lr001.3)
find.int$interactions
find.int$rank.list

# Visualizing the identified key interactions (example)
gbm.perspec(yrust.tc2.lr001.3, 7, 6, theta = 150)

gbm.perspec(yrust.tc2.lr001.3, 6, 7, z.range = c(0.5, 1), theta = 220 ,
cex.lab = 0.8, cex.axis = 0.6)

gbm.perspec(yrust.tc2.lr001.3, 6, 4, z.range = c(0.2, 1), theta = 240 ,
cex.lab = 0.8, cex.axis = 0.6)


# BRT prediction
predictions <- predict.gbm(yrust.tc2.lr001.3, mydata2test,
                         n.trees = yrust.tc2.lr001.3$gbm.call$best.trees,
                        type = "response")
```

3.5.3 Assessing Model Predictive Capacity (RQ2)

Generating Confusion Matrix

```
# Creating the table with probability, prediction (0 or 1), actual
observation (0 or 1), and accuracy (TP, FP, FN, TN)

pred.table <- as.data.frame(predictions2)  %>%
  rename(probability=predictions2) %>%
  mutate(prediction=if_else(probability >= 0.5, '1', '0')) %>%
  cbind(mydata2test$rust) %>%
  rename_at(3, ~'observation') %>%
  mutate(accuracy = case_when(
    prediction == 1 & observation == 1 ~ "TP",
    prediction == 1 & observation == 0 ~ "FP",
    prediction == 0 & observation == 1 ~ "FN",
    prediction == 0 & observation == 0 ~ "TN"))

pred.table # print prediction table

# create confusion matrix by counting TP, FP, FN, TN
ConfusionMatrix <- as.data.frame(table(pred.table$accuracy))
ConfusionMatrix

fn <- ConfusionMatrix[ConfusionMatrix$Var1 == "FN", "Freq"]
fp <- ConfusionMatrix[ConfusionMatrix$Var1 == "FP", "Freq"]
tn <- ConfusionMatrix[ConfusionMatrix$Var1 == "TN", "Freq"]
tp <- ConfusionMatrix[ConfusionMatrix$Var1 == "TP", "Freq"]
```

ModelPerformance()

```
#calculate statistics for accuracy, kappa, precision, sensitivity, and
specificity.
ModelPerformance = function(tp, tn, fp, fn){
  { # Accuracy
    correct = tp+tn
    total = tp+tn+fp+fn
    print(paste0("Accuracy: ", round(correct/total, digits= 4)))
  }
  { # Kappa
    total=tp+tn+fp+fn
    observed_acc=(tp+tn)/total
    expected_acc=((tn+fn)/total)*((tn+fp)/total) +
((fp+tp)/total)*((fn+tp)/total)
    print(paste0("Kappa: ", round((observed_acc - expected_acc)/(1 -
expected_acc), digits = 4)))
  }

  { # Precision
    print(paste0("Precision: ", round(tp/(tp+fp), digits = 4)))
  }

  { # Sensitivity
    print(paste0("Sensitivity: ", round(tp/(tp+fn), digits = 4)))
  }

  { # Specificity
    print(paste0("Specificity: ", round(tn/(tn+fp), digits = 4)))
  }
```

3.5.4 Model Extrapolation

An example based on the GAM model for tiller-only data, 2018 map.

```
# Generate extrapolation map based on the GAM model fit with the tiller-
only dataset

setwd("WORKING FOLDER LOCATION")
tiller.train <- read.csv("tiller.train.csv", header=TRUE)
```

```
# Below is the GAM model that performed the best - Accuracy 0.72, Kappa
0.44 (moderate agreement)

library(mgcv)
gam.tiller <- gam(rust ~ s(daysr_21) + s(minT_11) + s(aspect) + s(slope)
+ s(DEM) + s(ndvi_19),
                  data = tiller.train,
                  family = binomial,
                  method = "REML")

summary(gam.tiller)
plogis(coef(gam.tiller)[1])
gam.check(gam.tiller)
plot(gam.tiller, pages = 2,
     trans =plogis, # transform y-axis to 0-1 scale
     shift = coef(gam.tiller)[1], # adding model intercept
     seWithMean = TRUE, # consider intercept uncertainty
     residuals = TRUE, pch = 1, cex = 1,
     shade = TRUE)


# Preparation for extrapolation using the raster layers of the key
predictor variables.

# Step 1: Load maps into R.
require(raster) # Enabling R to read and write maps
require(rgdal)

# Load maps that are relevant for the analysis.
# Make sure to change the year of the folder (2016, 2017, 2018)
depending on the year of interest! (DEM, slope and aspect come from the
same location)
daysr_21.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/2018/daysr_21.ovr")
minT_11.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/2018/minT_11.ovr")
ndvi_19.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/2018/ndvi_19.ovr")
DEM.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/DEM.ovr")
slope.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/slope.ovr")
aspect.rs <- raster("/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/aspect.ovr")

# Step 2: Converting raster image to data frame (.df)
daysr_21.df <-as.data.frame(daysr_21.rs)
minT_11.df <-as.data.frame(minT_11.rs)
ndvi_19.df <-as.data.frame(ndvi_19.rs)
DEM.df <-as.data.frame(DEM.rs)
slope.df <-as.data.frame(slope.rs)
aspect.df <-as.data.frame(aspect.rs)

# Collate all the maps into one data frame
gam.till.df <- data.frame(daysr_21 = daysr_21.df, minT_11 =
minT_11.df,ndvi_19 = ndvi_19.df, DEM = DEM.df, slope = slope.df, aspect
= aspect.df)

# Column head for minT_11 remained "mint_11". Change this to the exact
variable name "minT_11" so that model extrapolation works well.
colnames(gam.till.df)[2] = "minT_11"

# Step 3: calculate prediction for the data frame generated
gam.till.df$predict <- predict.gam(gam.tiller, gam.till.df, type =
"response") # the new data frame (in matrix)
# "response" indicates probability in the scale of 0-1. Do not use the
Link function!
```

```
# Step 4: converting predictions into map
gam.till.matrix <- matrix(gam.till.df$predict,
                      nrow=DEM.rs@nrows, ncol=DEM.rs@ncols, byrow=TRUE)
gam.till.rs <- raster(gam.till.matrix,crs=DEM.rs@crs,
                     xmn=DEM.rs@extent@xmin,
                     ymn=DEM.rs@extent@ymin,
                     xmx=DEM.rs@extent@xmax,
                     ymx=DEM.rs@extent@ymax)

# check the result raster and save.
library(RColorBrewer)
coul <- colorRampPalette(c("white", "yellow", "orange","brown"))
plot(gam.till.rs, col = coul(100), axes = FALSE)

#exporting the image to file
writeRaster(gam.till.rs,"/Volumes/My Passport for Mac/MSc
iGEON/Extrapolation/GAM_Tiller/2018/gam.till.2018.img")
```