

# MULTI-TEMPORAL MULTI-MODAL PREDICTIVE MODELLING OF PLANT PHENOTYPES

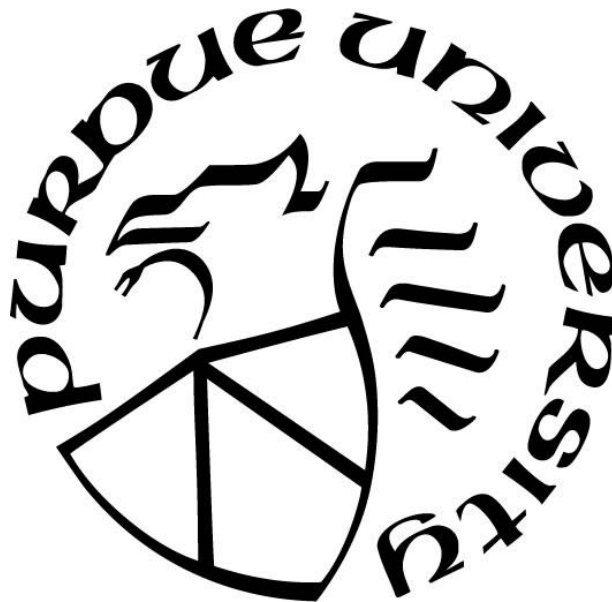
by  
**Ali Masjedi**

**A Dissertation**

*Submitted to the Faculty of Purdue University*

*In Partial Fulfillment of the Requirements for the degree of*

**Doctor of Philosophy**



Lyles School of Civil Engineering

West Lafayette, Indiana

May 2020

**THE PURDUE UNIVERSITY GRADUATE SCHOOL**  
**STATEMENT OF COMMITTEE APPROVAL**

**Dr. Melba M. Crawford, Chair**  
Lyles School of Civil Engineering

**Dr. Ayman F. Habib**  
Lyles School of Civil Engineering

**Dr. Mitchell R. Tuinstra**  
Department of Agronomy

**Dr. Xiao Wang**  
Department of Statistics

**Approved by:**  
Dr. Dulcy Abraham

*To my family and friends.*

## **ACKNOWLEDGMENTS**

I would like to express my sincere gratitude to my advisor Dr. Melba Crawford for all her invaluable guidance, and encouragement. Without her valuable suggestions on my research and providing excellent support, this dissertation would not have been possible. Besides my advisor, I would like to thank the rest of my committee: Dr. Ayman Habib, Dr. Mitchell Tuinstra, and Dr. Xiao Wang, for their insightful comments, encouragement, and valuable inputs throughout this work.

This work was supported by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy under Grant DE-AR0000593, and I would like to thank all the members of Purdue TERRA team for their exceptional work on this project since 2015.

My gratitude also goes to my family for their support, encouragement, patience, and love. In addition, many thanks to my great friends; Evan, Meghdad, Tamer, Magdy, Jill, Karoll, Taojun, Neal, Tian, Radhika, Jieqiong, Kai-Wei, Zhou, Ehsan, Amir, Mohsen, and Behnam.

# TABLE OF CONTENTS

TABLE OF CONTENTS .....	5
LIST OF TABLES .....	9
LIST OF FIGURES .....	11
LIST OF ABBREVIATIONS.....	15
ABSTRACT .....	16
1. INTRODUCTION .....	18
1.1 Remote Sensing Predictive Models.....	18
1.2 Remote Sensing Based Phenotyping .....	18
1.3 Research Challenges for Remote Sensing Based Phenotyping.....	19
1.3.1 Complex Relationship Between Biomass and Remote Sensing Data .....	19
1.3.2 Limited Number of Samples .....	20
1.3.3 High Variability Within the Samples .....	20
1.4 Organization of the Dissertation .....	20
2. MULTI-TEMPORAL PREDICTIVE MODELLING OF SORGHUM BIOMASS USING UAV-BASED HYPERSPECTRAL AND LIDAR DATA.....	22
2.1 Abstract.....	22
2.2 Introduction.....	23
2.3 Materials and Methods .....	25
2.3.1 Ground Reference Data .....	25
2.3.2 Remote Sensing Data.....	28
2.3.3 Feature Extraction .....	29
Hyperspectral-Based Features .....	29
Spectral Reflectance .....	29
Vegetation Indices .....	29
Integration Features .....	30
Derivative Features.....	31
LiDAR-based Features.....	32
Height Percentile .....	32
Canopy Volume.....	33

Canopy Cover.....	33
Height Statistics.....	33
2.3.4 Regression-Based Modeling Approaches .....	33
2.4 Results and Discussion .....	34
2.4.1 Data Screening .....	34
Time Series of Biomass Data .....	34
Time Series of Remote Sensing Data .....	36
Relationship between Features and Biomass .....	41
2.4.2 Biomass Predictive Models.....	46
Impact of the data source and regression method on the prediction results.....	46
Predictions in time .....	52
Multi-temporal predictions of end-of-season biomass .....	53
Impact of the number of samples on biomass prediction.....	54
2.5 Conclusions .....	56
3. ACCURATE PHENOTYPING USING UAV-BASED REMOTE SENSING DATA, GENOMIC MARKERS, AND WEATHER DATA .....	57
3.1 Abstract.....	57
3.2 Introduction.....	57
3.3 Materials and Methods .....	60
3.3.1 Field Design .....	60
3.3.2 Data Description.....	61
Ground Sampling of Biomass .....	61
Remote Sensing Data.....	62
RS Feature Extraction.....	64
Genotype Data.....	64
Genomic Feature Extraction .....	64
Features-Biomass Correlation.....	66
Weather Data.....	67
3.4 Results and Discussion .....	69
3.4.1 Genomic-based Predictive Models.....	69
3.4.2 Comprehensive Predictive Models.....	71

Models trained on one year of data.....	71
Models trained on two previous years of data.....	76
Genomic-driven feature evaluation .....	79
Multi-temporal Predictions .....	80
3.5 Conclusions .....	81
4. PREDICTION OF CROP YIELD USING UAV TIME SERIES DATA AND RECURRENT NEURAL NETWORKS.....	82
4.1 Abstract.....	82
4.2 Introduction.....	82
4.3 Feature Learning.....	85
4.3.1 Feature Reduction by Auto-encoders .....	86
4.3.2 Feature Extraction by Auto-encoders .....	88
4.3.3 Supervised Feature Learning by Auto-encoders .....	92
4.4 Predictive Models.....	92
4.4.1 Support Vector Regression .....	92
4.4.2 Recurrent Neural Network for Prediction.....	92
4.5 Data and Experimental Setup.....	94
4.5.1 Field Ground Reference Data.....	94
4.5.2 Remote Sensing Data.....	95
4.5.3 Experimental Setup.....	96
4.6 Results.....	97
4.6.1 Feature Learning Results .....	97
4.6.2 Feature Extraction Results .....	100
Feature Extraction from Hyperspectral Data Using CAEN-2d.....	100
Feature Extraction from LiDAR Data Using RAEN .....	103
4.6.3 Multi-year Biomass Prediction.....	108
4.6.4 Multi-time Predictive Models .....	110
4.7 Conclusions.....	112
5. CONCLUSIONS AND RECOMMENDATIONS .....	113
5.1 Summary.....	113
5.2 Key Objectives and Findings .....	113

5.3 Future Work .....	116
REFERENCES .....	118

## LIST OF TABLES

Table 2.1. Experiment designs for the 2017 and 2018 growing seasons. ....	26
Table 2.2. Commercial varieties planted in the hybrid calibration panel. ....	26
Table 2.3. Remote Sensing Data Sets .....	29
Table 2.4. Vegetation indices extracted from each HSI spectrum.....	30
Table 2.5. Integration features extracted from each HSI spectrum .....	30
Table 2.6. Derivative features were extracted from each first derivative (FDR) and second derivative (SDR) spectrum.....	32
Table 2.7. Summary of the prediction results of the experimental trials .....	49
Table 2.8. Analysis of variance of the $R^2$ respective to the data source, regression method, and their interaction .....	51
Table 2.9. Analysis of variance of the $R^2$ respective to the data source, regression method, cultivar type, and their interactions .....	51
Table 3.1. Experimental design for the 2017, 2018, and 2019 growing seasons. ....	63
Table 3.2. Grid search parameters for regression methods .....	70
Table 4.1. CAEN-2d details .....	89
Table 4.2. Experiment designs for the 2017-2019 growing seasons. ....	95
Table 4.3. Remote Sensing Data Sets Details .....	96
Table 4.4. $R^2$ values of SVR-based biomass prediction (of field-based results) using latent features extracted by the proposed <b>FAEN</b> trained on different experimental data sets and tested on all experiments. The maximum $R^2$ of the predictions for each experiment is in bold.....	99
Table 4.5. $R^2$ values of SVR-based biomass prediction (of field-based results) using latent features extracted by the proposed <b>CAEN</b> trained on different experimental data sets and tested on all experiments. The maximum $R^2$ of the predictions for each experiment is bold.....	100
Table 4.6. $R^2$ values of SVR-based biomass prediction (of field-based results) using latent LiDAR-based features extracted by the proposed <b>CAEN-2d</b> trained on different experimental data sets and tested on all experiments. Blue and green colors show unsupervised and supervised, respectively. The maximum $R^2$ of each row is highlighted. ....	103
Table 4.7. List of experiments for feature extraction using RAEN-1 and RAEN-2.....	103
Table 4.8. $R^2$ values of SVR-based biomass prediction (of field-based results) using latent LiDAR-based features extracted by the proposed <b>RAEN-1</b> and <b>RAEN-2</b> trained on different experimental data sets and tested on all experiments. Blue and green colors show unsupervised and supervised, respectively. The maximum $R^2$ value for each row is highlighted. ....	107

Table 4.9. Maximum $R^2$ values of SVR-based biomass prediction using latent features extracted from hyperspectral and LiDAR data by the proposed networks in this study. The maximum $R^2$ value for each row and from each data type is highlighted. ....	107
Table 4.10. The $R^2$ value (relative to field based measurements) of biomass predictions using SVR and RNN-1 with GRU and LSTM units trained on three experimental data sets and tested on all experiments. ....	108

## LIST OF FIGURES

Figure 2.1. Distribution of the fresh biomass data in the nine trails in the 2017 and 2018 growing seasons.....	27
Figure 2.2. Example spectra, FDR, and SDR of two sorghum varieties in the HyCal-18 panel on Aug 2 <sup>nd</sup> , 2018.....	31
Figure 2.3. Ground reference data collected during the 2017 and 2018 growing seasons in the HyCal panels. For each variety, the data samples are sorted based on the day after sowing (DAS). .....	35
Figure 2.4. Tukey’s pairwise multi-comparison test for the fresh biomass data collected in the 2017 and 2018 growing seasons. Blue indicates that the two varieties are significantly different ( $\alpha=0.05$ ). .....	36
Figure 2.5. Spectra of the 18 Sorghum Varieties in HyCal-18 panel on July 18 <sup>th</sup> , 2018. The varieties are very similar in the visible range of the spectrum, but substantial variability in the NIR portion of the spectrum.....	37
Figure 2.6. Example of reflectance of one of the varieties in the HyCal-18 experiment (“ <i>Trudan 8</i> ”) during the 2018 growing season. ....	37
Figure 2.7. The 8 cm resolution DSM for multiple dates in the 2018 growing season for the SbDivTc experiment. ....	38
Figure 2.8. Point cloud data for rows two and three of two plots of the HyCal-18 experiment for multiple dates in the 2018 growing season. “341x10” and “Trudan Headless” achieve a maximum height of 1.4m and 3.2m, respectively. ....	38
Figure 2.9. Average and standard deviation of the height of all the plots in each experimental trail in the 2017 and 2018 growing seasons.....	39
Figure 2.10. Height histogram for “341x10” and “Trudan Headless” in the HyCal panel across the 2017 growing season.....	40
Figure 2.11. Correlation matrix calculated using all the remotely sensed hyperspectral data over the HyCal-17 experiment on different dates in the 2017 growing season. Note that as the LiDAR and hyperspectral sensors were flown on separate platforms, the number of available data sets differs, and the data were not always collected on the same day. ....	40
Figure 2.12. Multiple-comparison Tukey’s test for the OSAVI index collected over the HyCal-17 experiment. Green shows the two varieties are significantly different from each other ( $\alpha=0.05$ ). .....	41
Figure 2.13. Multiple-comparison Tukey’s test for the LiDAR-based volume collected over the HyCal-18 experiment. Green shows the two varieties are significantly different from each other ( $\alpha=0.05$ ). ....	41

Figure 2.14. Comparison of the 90 percentile height, Intg_NIR1, and the biomass data for the dwarf grain sorghum 341x10 and the photoperiod sensitive Trudan Headless varieties in the 2017 and 2018 growing seasons.....	42
Figure 2.15. $R^2$ values of the linear regression-based models developed for the end of season fresh biomass using LiDAR-based features at four stages of growth. Feature 1 to 8 represent: #1: 30 <sup>th</sup> percentile height, #2: 50 <sup>th</sup> percentile height, #3: 95 <sup>th</sup> percentile height, #4: coefficient of variation of height, #5: volume, #6: canopy cover (threshold=0.1), #7: canopy cover (threshold=0.3), #8: canopy cover (threshold=0.5). ....	44
Figure 2.16. $R^2$ values of the linear regression-based models developed for the end of season fresh biomass using VNIR features at four stages of growth. Feature 1 to 8 represent: #1: FDR_min, #2: Intg_NIR1, #3: SDR_slope, #4: Intg_NIR1, #5: NDVI, #6: SR <sub>800,680</sub> , #7: OSAVI, and #8: MCARI. ....	44
Figure 2.17. The average $R^2$ values of the linear models developed for all the dates and all the experiments for each feature type from hyperspectral and LiDAR data.....	45
Figure 2.18. Average and maximum $R^2$ values of the linear models developed for all the dates and all the experiments for each band from hyperspectral VNIR and SWIR data. ....	45
Figure 2.19. $R^2$ values of the end of season fresh biomass predictions, using six data sources, and the three PLSR, SVR, and RF methods for all the experiments conducted in the 2017 and 2018 growing seasons. ....	47
Figure 2.20. The box plot for the fresh biomass data for all the experiments conducted in the 2017 and 2018 growing seasons. ....	48
Figure 2.21. Box plot of the prediction results for various data sources and regression methods. ....	50
Figure 2.22. Prediction results of PLSR, SVR, and RF models developed for the HyCal-17 and HyCal-18 experiments, using all the data sources and leave-one-out cross validation strategy. ...	52
Figure 2.23. $R^2$ values for biomass predictions during the 2017 and 2018 growing seasons using the SVR models developed based on VNIR hyperspectral and LiDAR data for the hybrid calibration panels. ....	53
Figure 2.24. $R^2$ for end of season predictions using hyperspectral and LiDAR data collected on different dates for the SbBAP-17, SbDiv-17, and SbDivTc-18 experiment. ....	54
Figure 2.25. Impact of the number of training samples on $R^2$ of the predictive models using SVR and PLSR models. ....	55
Figure 3.1. RGB bands of the hyperspectral data and the LiDAR point cloud of the SbDivTcCal and SbDivTc panels on August 2 <sup>nd</sup> , 2019. ....	61
Figure 3.2. An example area of the InCal-17 experiment. Two plots (black rectangles) and rows two and three (yellow rectangles) are highlighted. ....	62
Figure 3.3. Box plot for the fresh biomass data for all the experiments conducted in the 2017, 2018, and 2019 growing seasons. ....	62

Figure 3.4. M600 platform with the RGB, LiDAR, and VNIR sensors flown in the 2019 growing season. A GNSS/INS unit (APX-15) was also mounted on the platform providing accurate position and orientation information required for geo-referencing of the captured data. ....	64
Figure 3.5. Feature extraction using wavelet transform in 6, 8, and 10 levels. The X and Y axes are associated with the number of features and number of genotypes, respectively. ....	65
Figure 3.6. Feature extraction using PCA. The top row contains plots of combinations of the first 3 PCs for the combined hybrid and combined inbred genotypes. The plots on the bottom row show the first two principal components of the InCal-17, InCalTc-18, and SbDtcCal-19 genotypes are located with respect to those of the combined hybrid and inbred genotypes. ....	66
Figure 3.7. $R^2$ values of the linear regression-based models developed for fresh biomass using genomic markers and wavelet-6 features for three experiments, SbDiv17, SbDtc18, and SbDtc19 conducted in 2017-2019. ....	67
Figure 3.8. Weather conditions in the 2017-2019 growing seasons at ACRE. ....	68
Figure 3.9. Weather-biomass relationship in the four experiments, and two pairs of similar experiments conducted in two years: InCal-17 vs. InCal-18 and SbDTc-18 vs. SbDTc-19. ....	69
Figure 3.10. $R^2$ of predictions using derived genomic features in Ridge, PLSR, SVR, and RF regression models. ....	71
Figure 3.11. Results of biomass and rank prediction for example experiments in 2018 and 2019 from SVR models trained on the 2017 experiments using RS, genotype, and weather data. ....	73
Figure 3.12. Results of biomass and rank prediction for example experiments in 2019 from SVR models trained on the 2018 experiments using RS, genotype, and weather data. ....	74
Figure 3.13. Results of biomass and rank prediction for example experiments in 2018 and 2019 from predictive models trained on the 2017 experiments using genomic (G), genomic and weather data (G+W), RS (RS), RS and weather data (RS+W), RS and genomic data (RS+G), and RS, genomic, and weather data (RS+G+W). ....	75
Figure 3.14. Biomass and rank prediction for example experiments in 2019 from predictive models trained on the 2018 experiments using various data sources. ....	76
Figure 3.15. Biomass and rank prediction for example experiments in 2019 from SVR models train on the 2017 and 2018 experiments using RS, genotype, and weather data ....	77
Figure 3.16. Results of biomass and rank prediction for experiments in 2019 from predictive models train on the 2017 and 2018 experiments using various data sources. ....	78
Figure 3.17. $R^2$ of predictions based principal components and wavelet-based genomic features in combination with RS and weather inputs in Ridge Regression, PLSR, SVR, and RF models ....	79
Figure 3.18. Impact of number of RS datasets on $R^2$ of predictions using Ridge, PLSR, SVR, and RF regression models. ....	80
Figure 4.1. Unsupervised feature learning using a fully connected auto-encoder network (FAEN). ....	87

Figure 4.2. Unsupervised feature learning using a convolutional auto-encoder network (CAEN). .....	87
Figure 4.3. Unsupervised feature learning using 2d convolutional auto-encoder network (CAEN-2d). .....	88
Figure 4.4. One-to-many vs. many-to-one RNN network.....	90
Figure 4.5. RNN auto-encoder network (RAEN-1) proposed for feature extraction from LiDAR data. .....	91
Figure 4.6. RNN auto-encoder network (RAEN-2) proposed for feature extraction from LiDAR data. .....	91
Figure 4.7. Recurrent neural network (RNN-1) developed in this study for biomass prediction, and its input vector at time $t_i$ . .....	93
Figure 4.8. Recurrent neural network (RNN-2) developed in this study for multi-temporal biomass prediction, and its input vectors at time $t_i$ and $t_j$ . .....	94
Figure 4.9. Original, reconstructed, and learned features for one variety of the HyCal-17 experiment using FAEN trained on the (a) HyCal-17 and (b) SbDiv-17 hyperspectral data sets.	98
Figure 4.10. CAEN-2d reconstruction results trained on SbDivTc-18.....	101
Figure 4.11. Original, reconstructed, and learned features for one variety of the HyCal-17 experiment using CAEN-2d trained on the (a) HyCal-17 and (b) SbDiv-17 experiments. ....	102
Figure 4.12. Error of reconstruction in X, Y, and Z for the points in an example plot with 1024 point as well as extracted features for one variety of the SbDivTv-18 experiment using RAEN-1 trained on the same experiment data.....	104
Figure 4.13. Error of reconstruction in X, Y, and Z for the points in an example plot with 1024 points as well as extracted features for one variety of the SbDivTv-18 experiment using RAEN-1 and RAEN-2 trained on the same experiment data.....	105
Figure 4.14. Reconstruction X, Y, and Z RMSE for all the plots of SbDivTv-18 experiment using RAEN-1 trained on the same experiment data. ....	105
Figure 4.15. Reconstruction X, Y, and Z RMSE for all the plots of SbDivTv-18 experiment using RAEN-2 trained on the same experiment data. ....	106
Figure 4.16. Biomass prediction results of RNN-1 with LSTM units trained and tested on different experimental data sets. ....	109
Figure 4.17. Biomass prediction results for the three trials using the proposed RNN with LSTM and GRU cells. ....	111
Figure 4.18. nRMSE of biomass estimation for individual dates for HyCal-2017 and HyCal-2018 experiments using the proposed network with LSTM cells. ....	111

## LIST OF ABBREVIATIONS

ACRE	Agronomy Center for Research and Education
ANOVA	Analysis of Variance
CAEN	Convolutional Auto-encoder Network
CNN	Convolutional Neural Network
DAS	Days After Sowing
DL	Deep Learning
DTM	Digital Terrain Model
FAEN	Fully Connected Auto-encoder Network
GDD	Growing Degree Day
GRU	Gated Recurrent Unit
GS	Genomic Selection
GSD	Ground Sampling Distance
LAI	Leaf Area Index
LiDAR	Light Detection and Ranging
LSTM	Long Short Term Memory
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
PCA	Principal Component Analysis
PLSR	Partial Least Squares Regression
RAEN	Recurrent Auto-encoder Network
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RS	Remote Sensing
SVR	Support Vector Regression
SWIR	Short-wave Infrared
UAV	Unmanned Aerial Vehicle
VNIR	Visible and Near Infrared

## ABSTRACT

High-throughput phenotyping using high spatial, spectral, and temporal resolution remote sensing (RS) data has become a critical part of the plant breeding chain focused on reducing the time and cost of the selection process for the “best” genotypes with respect to the trait(s) of interest. In this study, the potential of accurate and reliable sorghum biomass prediction using hyperspectral and LiDAR data acquired by sensors mounted on UAV platforms is investigated. Experiments comprised multiple varieties of grain and forage sorghum, including some photoperiod sensitive varieties, providing an opportunity to evaluate a wide range of genotypes and phenotypes.

Feature extraction is investigated, where various novel features, as well as traditional features, are extracted directly from the hyperspectral imagery and LiDAR point cloud data and input to classical machine learning (ML) regression based models. Predictive models are developed for multiple experiments conducted during the 2017, 2018, and 2019 growing seasons at the Agronomy Center for Research and Education (ACRE) at Purdue University. The impact of the regression method, data source, timing of RS and field-based biomass reference data acquisition, and number of samples on the prediction results are investigated.  $R^2$  values for end-of-season biomass ranged from 0.64 to 0.89 for different experiments when features from all the data sources were included. Using geometric based features derived from the LiDAR point cloud and the chemistry-based features extracted from hyperspectral data provided the most accurate predictions. The analysis of variance (ANOVA) of the accuracies of the predictive models showed that both the data source and regression method are important factors for a reliable prediction; however, the data source was more important with 69% significance, versus 28% significance for the regression method. The characteristics of the experiments, including the number of samples and the type of sorghum genotypes in the experiment also impacted prediction accuracy.

Including the genomic information and weather data in the “multi-year” predictive models is also investigated for prediction of the end of season biomass. Models based on one and two years of data are used to predict the biomass yield for the future years. The results show the high potential of the models for biomass and biomass rank predictions. While models developed using one year of data are able to predict biomass rank, using two years of data resulted in more accurate models, especially when RS data, which encode the environmental variation, are included. Also, the possibility of developing predictive models using the RS data collected until mid-season, rather

than the full season, is investigated. The results show that using the RS data until 60 days after sowing (DAS) in the models can predict the rank of biomass with  $R^2$  values of around 0.65-0.70. This not only reduces the time required for phenotyping by avoiding the manual sampling process, but also decreases the time and the cost of the RS data collections and the associated challenges of time-consuming processing and analysis of large data sets, and particularly for hyperspectral imaging data.

In addition to extracting features from the hyperspectral and LiDAR data and developing classical ML based predictive models, supervised and unsupervised feature learning based on fully connected, convolutional, and recurrent neural networks is also investigated. For hyperspectral data, supervised feature extraction provides more accurate predictions, while the features extracted from LiDAR data in an unsupervised training yield more accurate prediction.

Predictive models based on Recurrent Neural Networks (RNNs) are designed and implemented to accommodate high dimensional, multi-modal, multi-temporal data. RS data and weather data are incorporated in the RNN models. Results from multiple experiments focused on high throughput phenotyping of sorghum for biomass predictions are provided and evaluated. Using proposed RNNs for training on one experiment and predicting biomass for other experiments with different types of sorghum varieties illustrates the potential of the network for biomass prediction, and the challenges relative to small sample sizes, including weather and sensitivity to the associated ground reference information.

# **1. INTRODUCTION**

## **1.1 Remote Sensing Predictive Models**

Remote Sensing (RS) data are being used in a wide variety of scientific fields and applications, necessitating advances in sensors and capability for analysis. Classification and regression modeling are the two most frequent types of analysis currently being utilized for remote sensing image data. In classification, the goal is to assign a label to each pixel via supervised or unsupervised algorithms. In regression modeling, however, reference or “ground truth” data are used in conjunction with RS data to estimate or predict a variable(s) of interest. This study focuses both on the estimation of current season phenotypes and their prediction later in the season using multi-temporal multi-modal remote sensing data.

## **1.2 Remote Sensing Based Phenotyping**

With the advances in science and technology, it has become possible to generate thousands of genotypes of a plant and produce the seeds at low cost (Davey et al., 2011). To evaluate the performance of the plant varieties, numerous physical and agronomic traits are measured in the field during the growing season (phenotyping), and the varieties with the most desirable traits are selected. Planting thousands of varieties of a crop and measuring dozens of phenotypic traits for each would be impossibly time-consuming, laborious, and expensive. Thus, breeders plant only a subset of the varieties and collect the required measurements multiple times during the growing season to develop predictive models which can then be used to infer traits for the broader population. Field-based phenotyping is still required, which is very expensive and time-consuming with traditional methods, even for a small field. Further, some of the traits such as biomass, nitrogen content, water content, and LAI can only be measured through destructive sampling. To monitor these traits during the growing season, a large area is required, which increases the cost of the breeding studies. Recently, researchers have begun to explore the use of RS data to replace traditional in-field phenotyping.

Satellites and manned aircraft can provide data with high spectral resolution, but the spatial and temporal resolutions are inadequate for agricultural breeding plots. Remote sensing via Unmanned Aerial Vehicles (UAV), which is being investigated because of its capability to acquire

both high temporal and high spatial resolution data required for high throughput phenotyping over relatively limited areas, can potentially close the gap. UAVs can collect huge quantities of data “on demand” over experimental plots, providing opportunities for estimation and prediction of a wide range of agronomic traits such as water content, biomass, LAI, canopy cover, plant height and counts, nitrogen content, and disease. In this study, remotely sensed phenotyping of varieties of sorghum, with a focus on biomass prediction, is investigated using sensors on UAV platforms.

Sorghum has attracted attention in recent years, both for its broad-based potential usage, and its drought and heat tolerance. The grain of some varieties is now used for human consumption and animal feed in developed, as well as developing countries. Recently, some varieties of sorghum have been developed as an energy crop that can produce reasonable amounts of ethanol. Sorghum has an annual growth cycle, high calorific value, and low management cost, making it an efficient biofuel crop (Fernandes et al., 2018). Many researchers are working on enhanced genotypes that are more resilient in unfavorable environmental conditions, and/or varieties that can produce more energy-rich plant material (biomass). Thus, it is important to predict the final biomass of the planted varieties as soon as possible, so that genotype selection can be accomplished in a shorter period of time, and thus, less expensively.

The goal of this study is to develop a framework for remote phenotyping of plants using three types of UAV-based data; hyperspectral, and LiDAR. Many phenotypic traits that can be estimated and/or predicted, but biomass is the focus of this study because it is indicative of both the crop condition and the quantity of ethanol that can be produced.

### **1.3 Research Challenges for Remote Sensing Based Phenotyping**

#### **1.3.1 Complex Relationship Between Biomass and Remote Sensing Data**

The assumption is that the remote sensing data products, for example, height data or canopy cover estimates from LiDAR, are related to the traits such as biomass or LAI. Intuitively, the taller the plants - the greater the biomass, or the more extensive the canopy cover - the more the leaves and thus, a higher the LAI. However, the relationship between the traits and the remote sensing data is often more complex than for height. For example, increasing (or decreasing) the value of a red channel of hyperspectral data does not imply that a directly measurable change is occurring relative to the biomass. Also, for some features, the relationship between the traits and the remote

sensing data can be very weak. Finally, the relationships between the features and traits vary during the growing season. This is discussed in greater detail in the following sections.

### **1.3.2 Limited Number of Samples**

The key to the success for machine learning algorithms is their ability to learn the solution to the problem by analyzing the sample, or ground reference data. In general, the greater the number of the samples, the better the training of the algorithm, and as a result, the more accurate the classification/prediction. As a rule of thumb, it is recommended to have around  $10 \times d$  ( $d$  = number of the features) sample data. From the hyperspectral, LiDAR, and RGB data, hundreds of features can be extracted. However, as the output samples in this study are obtained by destructive sampling in the field, which is very time-consuming and expensive, the number of samples for the study is much fewer than recommended. The sample data for different fields where experiments were conducted in 2017-19 are described in Section 2. Developing systems of algorithms that can leverage the limited number of samples, reducing RS data dimensionality, and data augmentation are possible solutions to this issue.

### **1.3.3 High Variability Within the Samples**

Another challenge in this study is that sorghum varieties in some of the fields are very different from others, which creates challenges for the machine learning algorithms to learn patterns in the data, and further reduces the available quantity of data from a given pattern. A key to the success here is to extract the appropriate features that are predictors of the biomass (or any other trait of interest) so they can support the machine learning algorithms.

## **1.4 Organization of the Dissertation**

The dissertation is organized such that each chapter represents an individual paper associated with a specific investigation. As such, there is some common material associated with the description of the experiments. In Chapter 2, feature extraction from hyperspectral and LiDAR data is investigated, then the extracted features are used in traditional regression models for biomass prediction, and results are presented and discussed. The impact of the regression method,

data source, timing of RS and field-based biomass reference data acquisition, and the number of samples on the prediction results are investigated.

In Chapter 3, adding the genomic information and weather data to the “multi-year” predictive models is investigated for prediction of the end of season biomass. Models based on one year and two years of data are used to predict the biomass yield for the future years. Also, the possibility of developing predictive models using the RS data collected until mid-season, rather than the full season, is investigated.

In Chapter 4, supervised and unsupervised feature learning based on fully connected, convolutional, and recurrent neural networks are also investigated. Predictive models based on Recurrent Neural Networks (RNNs) are designed and implemented to accommodate high dimensional, multi-modal, multi-temporal data. RS data and weather data are incorporated in RNN models.

Chapter 5 provides a summary discussion of the results from the experiments detailed in the individual chapters and a critical review of the results of the conducted experiments as well as recommendations for the follow-on research.

## **2. MULTI-TEMPORAL PREDICTIVE MODELLING OF SORGHUM BIOMASS USING UAV-BASED HYPERSPECTRAL AND LIDAR DATA**

### **2.1 Abstract**

High-throughput phenotyping using high spatial, spectral, and temporal resolution remote sensing (RS) data has become a critical part of the plant breeding chain focused on reducing the time and cost of the selection process for the “best” genotypes with respect to the trait(s) of interest. In this paper, the potential of accurate and reliable sorghum biomass prediction using visible and near infrared (VNIR) and short-wave infrared (SWIR) hyperspectral data as well as light detection and ranging (LiDAR) data acquired by sensors mounted on UAV platforms is investigated. Predictive models are developed using three regression methods for nine experiments conducted during the 2017 and 2018 growing seasons at the Agronomy Center for Research and Education (ACRE) at Purdue University. Experiments comprised multiple varieties of grain and forage sorghum, including some photoperiod sensitive varieties, providing an opportunity to evaluate a wide range of genotypes and phenotypes. Feature extraction is investigated, where various novel features, as well as traditional features, are extracted from the hyperspectral imagery and LiDAR point cloud data. The impact of the regression method, data source, timing of RS and field-based biomass reference data acquisition, and the number of samples on the prediction results are investigated.  $R^2$  values for end-of-season biomass ranged from 0.64 to 0.89 for different experiments when features from all the data sources were included. Using geometric based features derived from the LiDAR point cloud and the chemistry-based features extracted from hyperspectral data provided the most accurate predictions. Evaluation of the impact of the time of data acquisition during the growing season on the prediction results indicated that although the most accurate and reliable end of season predictions were achieved using remotely sensed data from mid-season to end-of-season, predictions in mid-season provided adequate results to differentiate between promising varieties for selection. The analysis of variance (ANOVA) of the accuracies of the predictive models showed that both the data source and regression method are important factors for a reliable prediction; however, the data source was more important with 69% significance, versus 28% significance for the regression method. The characteristics of the

experiments, including the number of samples and the type of sorghum genotypes in the experiment also impacted prediction accuracy.

## **2.2 Introduction**

Biomass yield is an important trait of biofuel crops such as sorghum, as it is a key factor in determining the amount of biofuel that can be produced. With recent advances in science and technology surrounding genotyping, it has become possible to create numerous genotypes of a plant (Davey et al., 2011) and then select the genotypes with the maximum biomass production. However, traditional methods of biomass measurement involving labor-intensive and time-consuming destructive sampling do not meet the requirements for timely evaluation of the genotypes in large-scale breeding programs. Recently, remote sensing (RS) data have been explored for estimation of many phenotypic traits, including Leaf Area Index (LAI) (Liang et al., 2015; Potgieter et al., 2017), canopy height (Chu et al., 2018; Pugh et al., 2018), nitrogen content (Maimaitijiang et al., 2017), and biomass (Eitel et al., 2016; Jiating Li et al., 2018; Tattaris et al., 2016) to replace traditional in-field phenotyping.

Sensors on satellites and manned aircraft can provide data with high spectral resolution, but the spatial and temporal resolutions are inadequate for agricultural breeding programs that are based on small plots. Remote sensing via Unmanned Aerial Vehicles (UAV) is currently being investigated as a means to close the gap because of its capability to acquire the high temporal and spatial resolution data required for high throughput phenotyping over relatively limited areas. UAVs can collect huge quantities of data “on demand”, providing opportunities for estimation and prediction of a wide range of agronomic traits (S. B. Duan et al., 2014; T. Duan et al., 2017; Gouache et al., 2016; Lewis et al., 2017; Masjedi et al., 2018; Stanton et al., 2017; Z. Zhang et al., 2017).

In this study, remotely sensed biomass prediction of varieties of sorghum is investigated using the data acquired by RGB, hyperspectral and LiDAR sensors mounted on UAV platforms. Sorghum has attracted attention in recent years, both for its broad-based potential usage and its drought and heat tolerance. The grain of some varieties is now used for human consumption and animal feed in developed, as well as developing countries. Recently, some varieties of sorghum have been developed as an energy crop that can produce reasonable quantities of ethanol (Fernandes et al., 2018). Sorghum has an annual growth cycle, high calorific value, and low

management cost (Fernandes et al., 2018), making it an efficient biofuel crop. Many studies focus on developing enhanced genotypes that can produce more energy-rich plant material (biomass) (Ogbaga et al., 2019). It is important for these breeding studies to predict the end-of-season yield biomass of the planted varieties as soon as possible in the growing season so that genotype selection can be accomplished in a shorter period of time, and thus, less expensively.

Researchers have developed predictive models based on various remote sensing inputs and modeling approaches. Foster et al. (Foster et al., 2017) compared the performance of Partial Least Squares Regression (PLSR) and linear regression models in estimating biomass of a high-biomass sorghum variety. They concluded that PLSR can provide more accurate predictions using the normalized difference vegetation index (NDVI) calculated from field spectrometer measurements collected in July (three months after sowing). PLSR provided the best results in (Yue et al., 2018) as well, where eight regression techniques for wheat biomass estimation were investigated. Fassnacht et al. (2014) investigated the importance of the prediction method, as well as sample size and sensor type for biomass predictions, with the best results being obtained with a Random Forest (RF) model. Using airborne LiDAR and spaceborne hyperspectral data, the authors concluded that for their experiments, the sensor type was the most important factor in the prediction accuracy (Fassnacht et al., 2014). Multiple studies investigated biomass prediction using LiDAR data (Eitel et al., 2016; McGlinchy et al., 2014; Phua et al., 2017; Shao et al., 2018; Vaglio Laurin et al., 2016; Vastaranta et al., 2014; K. Zhao et al., 2018). Some studies have shown that fusion of airborne-based hyperspectral and LiDAR data provided better results than those achieved using data from either individual sensor type (Vaglio Laurin et al., 2014). Most of the studies have developed predictive models for a limited number of experiments, each including only a few genotypes of a crop, although in breeding programs, hundreds or thousands of genotypes with high variability in biomass, as well as spectral and structural characteristics, are included in each experiment. In this study, predictive models are developed for nine distinct experimental trials conducted over two years and include thousands of genotypes of sorghum.

Biomass prediction based on data analytics models and RS data is challenging for multiple reasons, including 1) the complex relationship between biomass and RS data, 2) limited number of ground reference samples for developing and validating models for an experiment, and 3) high variability between the samples in an experiment. The relationships between the RS-based features and traits vary across the growing season. A key to good predictive models is to extract the

appropriate features that are predictors of the trait of interest so they can support the machine learning algorithms. A small number of samples relative to the number of features (and thereby potentially the number of parameters to estimate) is a critical issue for remote sensing-based phenotyping, and reference sampling data in agricultural fields is time-consuming and expensive to collect. For this study, the diversity of the sorghum varieties and their associated physical characteristics are particularly challenging for machine learning algorithms, which typically learn from patterns. Developing systems of algorithms that can leverage the limited number of samples, reduce the dimensionality of RS data, and incorporate data augmentation approaches are possible solutions to this issue.

In this study, the objective is to provide a robust framework for predicting sorghum biomass which is suitable for plant breeding research and industrial applications. To accomplish this, we: 1) evaluate multiple prediction models, 2) investigate the impact of the timing of the RS data acquisition on prediction relative to days since sowing, and 3) evaluate the importance of the features extracted from the data.

## **2.3 Materials and Methods**

### **2.3.1 Ground Reference Data**

The field experiments were conducted over two years in approximately 2.8 ha sorghum breeding trials in different fields at the Purdue University Agronomy Center for Research and Education (ACRE) farm. There were four distinct trials in 2017: the hybrid calibration (HyCal-17), the inbred calibration (InCal-17), the sorghum biodiversity (SbDiv-17), and the sorghum bioenergy (SbBAP-17) panels. In 2018, the field experiments consisted of five distinct trials: the hybrid calibration (HyCal-18), the inbred calibration (InCal-18), the inbred calibration test cross (InCalTc-18), the sorghum biodiversity test cross (SbDivTc-18), and the sorghum nitrogen test (SNitTs-18). The experiments were conducted using randomized complete block designs and planted at 220,000 plants per ha. Details of the experiment trials are provided in Table 2.1. For the HyCal-17 and HyCal-18 panels, biomass data were destructively collected multiple times during each growing season. For all other experiments, the biomass data were collected only once at the end of each growing season using a two-row combine harvester. Table 2.2 lists the commercial hybrid varieties planted in the HyCal-17 and HyCal-18 experiments.

Table 2.1. Experiment designs for the 2017 and 2018 growing seasons.

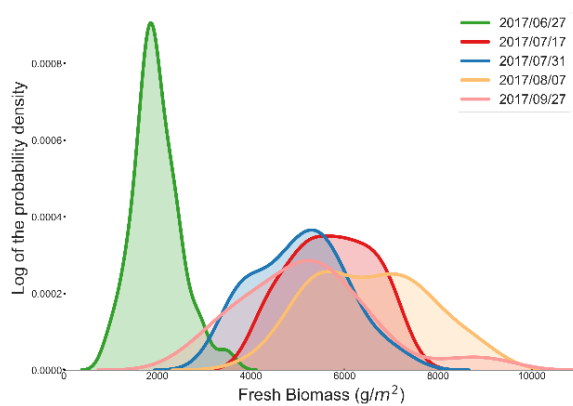
Trial	Year	Genotype	# of plots	# of genotypes	Sowing Date	Harvest Date	Available Biomass Data
HyCal-17	2017	Hybrid	72	18	May 16	Sep 27	6/27, 7/17, 7/31, 8/08, 9/27
InCal-17	2017	Inbred	120	60	May 16	Sep 27	9/27
SbBAP-17	2017	Inbred	760	350	May 16	Sep 28	9/28
SbDiv-17	2017	Inbred	1800	840	May 17	Nov 9	11/09
HyCal-18	2018	Hybrid	72	18	May 8	Aug 9	6/27, 7/12, 8/09
InCal-18	2018	Inbred	108	54	May 8	Aug 9	8/09
InCalTc-18	2018	Hybrid	108	54	May 8	Aug 6	8/06
SbDivTc-18	2018	Hybrid	1260	630	May 8	Aug 2	8/2 and 8/14
SNitTs-18	2018	Inbred	112	4	June 4	Oct 2	10/02

Table 2.2. Commercial varieties planted in the hybrid calibration panel.

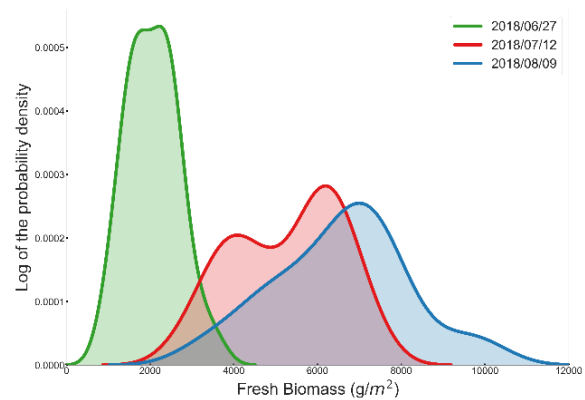
Variety	Variety name	Sorghum Type	Company
1	849F	Forage	Pioneer
2	877F	Forage	Pioneer
3	327x36 BMR	Forage	Richardson
4	341x10	Forage	Richardson
5	366x58	Food grain	Richardson
6	374x66	Food grain	Richardson
7	392x105 BMR	Forage	Richardson
8	400x38 BMR	Sudangrass	Richardson
9	400x82 BMR	Sudangrass	Richardson
10	HIKANE II	Forage	Sorghum Partners
11	NK300	Forage	Sorghum Partners
12	NK5418	Grain	Sorghum Partners
13	NK8416	Grain	Sorghum Partners
14	SS405	Forage	Sorghum Partners
15	Sordan 79	Forage	Sorghum Partners
16	Sordan Headless	Forage (photoperiod sensitive)	Sorghum Partners
17	Trudan 8	Forage	Sorghum Partners
18	Trudan Headless	Forage (photoperiod sensitive)	Sorghum Partners

Figure 2.1 shows the distribution of the biomass values for the experimental trials in the 2017 and 2018 growing seasons. Figure 2.1a and b show the fresh biomass distribution of the HyCal-17 and HyCal-18 panels, respectively, during the growing seasons. In 2017, the biomass data on June 27th, July 17th, and August 7th were collected by hand harvesting one meter sections

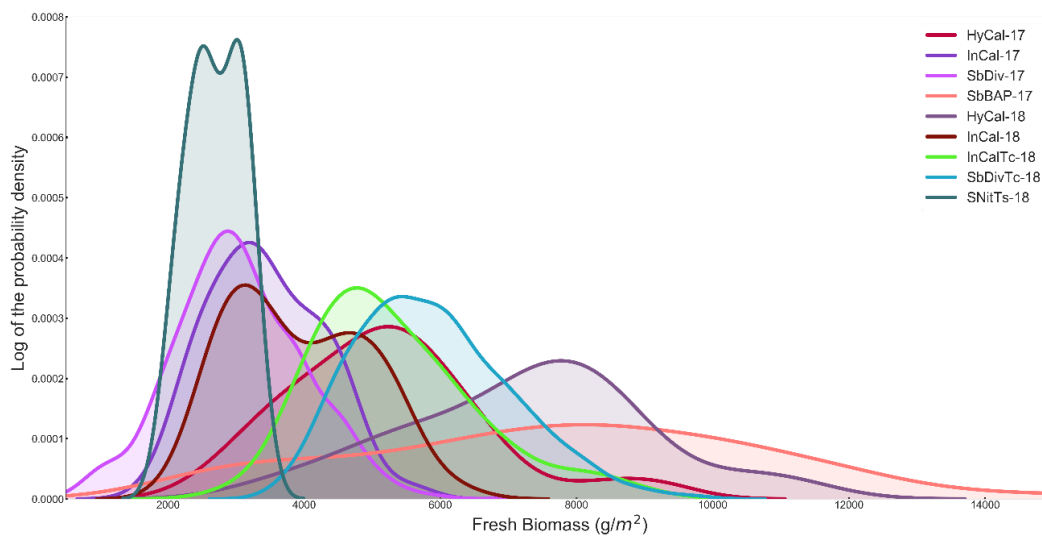
of three rows from each plot. Harvesting for all other dates was performed with a two-row combine harvester. These figures indicate that the genotypes have similar biomass early in the season, and they become more different at the end of the season. Figure 2.1c shows the distribution of the end of the season biomass of the nine trials over both years; 1) the InCal-17 and SbDiv-17 are similar, 2) the HyCal-17 and HyCal-18 have similar shapes but different ranges of values, 3) SbDivTc-18 and HyCal-17 are similar in both shape and range of values.



(a) HyCal panel in 2017



(b) HyCal panel in 2018



(c) The nine trials in 2017 and 2018

Figure 2.1. Distribution of the fresh biomass data in the nine trials in the 2017 and 2018 growing seasons.

### 2.3.2 Remote Sensing Data

This study includes RGB, hyperspectral, and LiDAR remote sensing data collected by custom designed UAV platforms. All remote sensing data acquisition platforms were flown with Global Navigation Satellite System/Inertial Navigation System (GNSS/INS) units for direct georeferencing. RGB data for this study were collected using a Sony Alpha ILCE-7R RGB camera with a Sony 35mm lens. The Sony Alpha 7R features a full-frame 36.4MP sensor delivering high-resolution UAV-based aerial imagery. LiDAR data were collected with a Velodyne VLP-16 3D LiDAR sensor operating in the strongest return mode and 600 rotations per minute (RPM). The VLP-16 is a 16 beam LiDAR unit with a 360-degree horizontal FOV, and a maximum range of 100 m. Both the RGB camera and the VLP-16 sensor are mounted on a DJI Matrice 600 Pro (M600P) platform. Spatial and temporal system calibration for the datasets used in this study were conducted using the approaches described in (Ravi et al., 2018) and (LaForest et al., 2019), respectively. Also, the georeferenced orthomosaics were generated using the structure from motion strategies introduced in (Hasheminasab et al., 2020; He et al., 2018).

Visible Near Infrared (VNIR) and Short Wave Infrared (SWIR) hyperspectral data were collected with two Headwall Photonics push-broom scanners. The VNIR data were acquired by a Nano-Hyperspec imaging sensor in 272 spectral bands at 2.2 nm/band from 400 nm to 1000 nm with 640 spatial channels at 7.4  $\mu\text{m}/\text{pixel}$ . In 2017, the sensor was flown at an altitude of 60 m with a 12 mm Schneider lens, resulting in a Ground Sampling Distance (GSD) of  $\sim 4$  cm. An 8 mm lens was used in 2018, and the flying height was 40 m to maintain the GSD at  $\sim 4$  cm. SWIR data were collected with a Headwall Photonics Micro-Hyperspec, a hyperspectral pushbroom scanner, which acquires 166 spectral bands at 10 nm/band from 900 nm to 2500 nm. The Micro-Hyperspec has 384 spatial channels at 24  $\mu\text{m}/\text{pixel}$ . In both years, the sensor was flown with a 25 mm lens at 40 m, resulting in approximately a 4 cm GSD. In 2018, the VNIR and SWIR sensors were integrated and flown together on a single UAV platform. A rigorous boresight calibration process was applied (Habib et al., 2018), yielding simultaneously collected co-aligned VNIR and SWIR data. All the hyperspectral data were converted to reflectance using the empirical line method to relate the spectra collected from the UAV to data acquired by an SVC 1024i field spectrometer over the calibration targets placed in the field for each acquisition. Table 2.3 lists the data acquired by the sensors in 2017 and 2018.

Table 2.3. Remote Sensing Data Sets

Year	Data Type	Field	Dates
2017	RGB and LiDAR	InCal, HyCal, SbDiv, and SbBAP	6/16, 6/21/6/27, 7/5,7/11, 7/14, 7/17, 7/25, 8/2, 8/8, 8/23, 8/30
	VNIR	HyCal and InCal SbDiv SbBAP	6/21, 6/28, 7/4, 7/12, 7/18, 7/25, 8/8, 8/23, 8/30, 9/10, 9/15 6/21, 6/27, 7/4, 7/18, 7/25, 7/30, 8/8, 8/14, 8/23, 9/10, 9/24, 9/30 6/21, 6/27, 7/4, 7/18, 7/25, 7/30, 8/8, 8/14, 8/23, 9/10, 9/24
	SWIR	InCal and HyCal SbDiv SbBAP	8/23, 8/30, 9/10, 9/15 8/2, 8/8, 8/14, 8/23, 8/30, 9/10, 9/30 8/2, 8/8, 8/14, 8/23, 8/30, 9/10
2018	RGB and LiDAR	HyCal, InCal, InCalTc, and SbDivTc SNitTs	5/22, 5/29, 6/4, 6/11, 6/20, 6/27, 7/2, 7/11, 7/18, 7/23, 8/1, 8/6 6/28, 7/3, 7/11, 7/17, 7/23, 8/1, 8/6, 8/16, 8/25, 9/5, 9/19
	VNIR and SWIR	HyCal, InCal, InCalTc SbDivTc SNitTs	6/4, 6/8, 6/14, 6/29, 7/3, 7/6, 7/11, 7/18, 7/25, 8/2, 8/9 6/4, 6/8, 6/14, 6/29, 7/3, 7/6, 7/10, 7/11, 7/25, 8/2 6/28, 7/3, 7/11, 7/18, 7/25, 8/2, 8/13, 8/28, 9/4, 9/12, 9/18

### 2.3.3 Feature Extraction

As discussed earlier, it is necessary to extract features that are related to the specific trait of interest and are preferably not redundant. In this study, both traditional and new candidate features focused on the relevance to biomass prediction were extracted from rows 2 and 3 of the 4- or 12-row plots to minimize the border effect.

#### *Hyperspectral-Based Features*

##### *Spectral Reflectance*

From the HSI data, the average reflectance values of the plots were calculated from rows 2 and 3 of each plot after masking the shadow and soil pixels.

##### *Vegetation Indices*

Vegetation Indices (VIs) obtained from HSI data have been widely used in different applications, as they are computationally simple and representative of the relevant chemically interpretable absorption and reflectance features in the spectrum. In this study, 13 vegetation indices, listed in Table 2.4 were extracted and used in the predictive models.

Table 2.4. Vegetation indices extracted from each HSI spectrum

Data Type	Index Name	Formulation	References
VNIR	NDVI	$(R750-R705) / (R750+R705)$	(A. Gitelson & Merzlyak, 1994)
	NDCI	$(R762-R527) / (R762+R527)$	(Marshak et al., 2000)
	Carte1	$R695 / R420$	(Carter, 1994)
	SR800,680	$R800 / R680$	(Sims & Gamon, 2002)
	SR675,700	$R675 / R700$	(A. A. Gitelson et al., 1996)
	SR700,670	$R700 / R670$	(McMurtrey et al., 1994)
	OSAVI	$(1+0.16) (R800-R670) / (R800+R670+0.16)$	(Rondeaux et al., 1996)
	MCARI	$[(R700-R670)-0.2(R700-R550)](R700 / R670)$	(Daughtry et al., 2000)
	REP	$700+40[(R670+R780) / 2-R700]/(R740-R700)($	(Clevers, 1994)
	PRI	$(R531-R570) / (R531+R570)$	(Gamon et al., 192 C.E.)
SWIR	NDWI	$(R860-R1240) / (R860+R1240)$	(Gao, 1996)
	NDLI	$[\log(1 / R1754) - \log(1 / R1680)] / [\log(1 / R1754) + \log(1 / R1680)]$	(Serrano et al., 2002)
	NDNI	$[\log(1 / R1510) - \log(1 / R1680)] / [\log(1 / R1510) + \log(1 / R1680)]$	(Serrano et al., 2002)

### Integration Features

NIR bands are particularly important for representing plant physiology, but are subject to the time during the growing season and environmental conditions. The area under the spectral curve for a given range from  $\lambda_a$  to  $\lambda_b$  is defined as  $Intg(\lambda_a, \lambda_b) = \int_{\lambda_a}^{\lambda_b} S(\lambda) d\lambda$ , where  $S(\lambda)$  is the reflectance at  $\lambda$  nm. Using different ranges of spectral values, six features were extracted from each HSI spectrum as listed in Table 2.5.

Table 2.5. Integration features extracted from each HSI spectrum

Data Type	Feature Name	$\lambda_a$	$\lambda_b$
VNIR	Intg_rededge	685	745
	Intg_NIR1	770	910
	Intg_NIR2	910	1000
SWIR	Intg_SWIR_r1	920	1353
	Intg_SWIR_r2	1430	1800
	Intg_SWIR_r3	1952	2385

## Derivative Features

The spectral derivatives, which quantify slope, curvature, and higher-order aspects of reflectance spectra, can be useful by revealing spectral features that may not be apparent in reflectance data alone (Thorp et al., 2017). For example, the “red-edge” position (between 680 nm and 750 nm) in crop reflectance data can be easily identified in the derivative spectra, and has been related to crop biomass (Demetriades-Shah et al., 1990). Feng et al. analyzed 20 spectral derivative features near the red edge area to estimate wheat leaf nitrogen concentration (Feng et al., 2014). In this study, the polished spectra were calculated using a Savitzky–Golay filter (Savitzky & Golay, 1964), then the first derivative (FDR) and second derivative (SDR) of the spectra were extracted. Figure 2.2 shows the spectra, FDR, and SDR of two sorghum varieties in the HyCal-18 panel on Aug 2<sup>nd</sup>, 2018. From FDR and SDR, 11 features were extracted and used in the additional analysis as described in Table 2.6. These features were selected at wavelengths where spectra of the varieties differed, and were also uncorrelated.

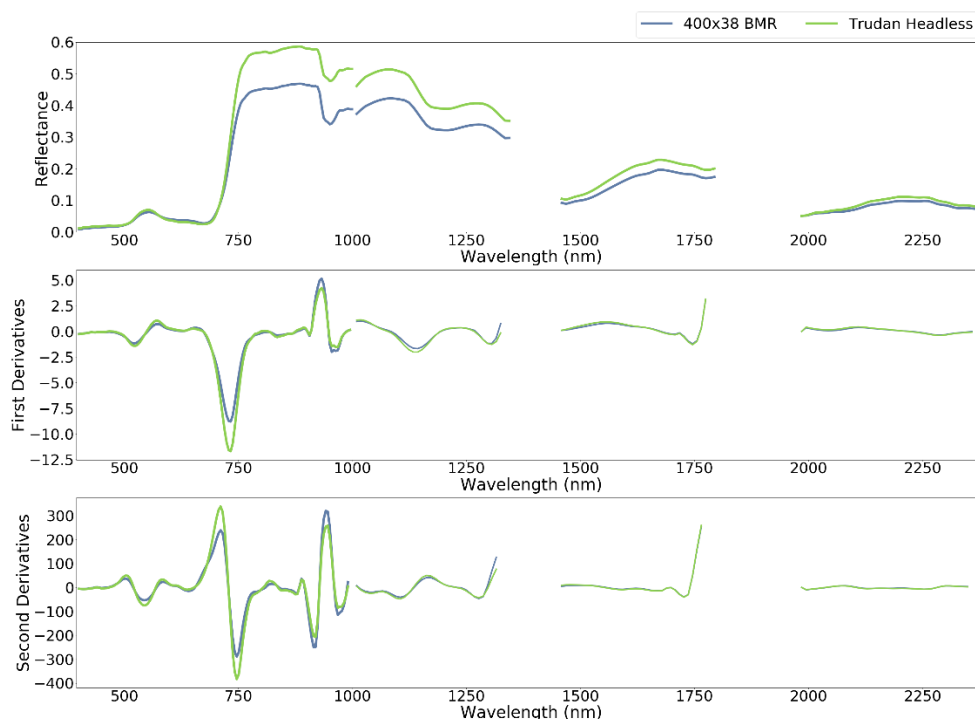


Figure 2.2. Example spectra, FDR, and SDR of two sorghum varieties in the HyCal-18 panel on Aug 2<sup>nd</sup>, 2018.

Table 2.6. Derivative features were extracted from each first derivative (FDR) and second derivative (SDR) spectrum

Data Type	Feature Name	Description
VNIR	FDR_slope	slope of the line that passes through the minimum of FDR and the maximum of FDR in 660-690 nm range
	FDR_min	minimum of FDR
	FDR_intg_NIR1	integration of FDR in bands between 670 and 780 nm
	FDR_intg_NIR2	integration of FDR in bands between 910 and 1000 nm
	SDR_slope_reddge	slope of the line passing through the maximum and the minimum of SDR
	SDR_intg	integration of SDR in all bands
SWIR	FDR_slope_r1	slope of the line that passes through the maximum of FDR in 1000-1050 nm range and the minimum of FDR in 1100-1200 nm range
	FDR_slope_r2	slope of the line that passes through the maximum of FDR in 1475-1525 nm range and the minimum of FDR in 1675-1725 nm range
	FDR_slope_r3	slope of the line that passes through the maximum of FDR in 2000-2050 nm range and the minimum of FDR in 2200-2300 nm range
	FDR_intg-r1	integration of FDR in bands between 920 and 1353 nm
	SDR_slope_r1	slope of the line that passes through the maximum of SDR in 1100-1200 nm range and the minimum of SDR in 1000-1100 nm range

### ***LiDAR-based Features***

The 3D structural characteristics of the plants in a plot can be described using various features extracted from LiDAR data. The Digital Terrain Model (DTM) derived from UAV-based LiDAR point clouds of the fields, which represents the bare earth height information, was derived using LiDAR data acquired before the emergence of the plants in each field, and assumed to be constant throughout the growing season. The height of points in the point cloud was estimated by subtracting the DTM from the “z” coordinate of each point. Then, the following features were extracted from the point cloud of each plot:

#### ***Height Percentile***

To capture the vertical distribution of the LiDAR points in each plot, the 30<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup> percentile height values from the point cloud of each plot were calculated.

### *Canopy Volume*

To estimate volume related characteristics of the canopy in each plot, a grid with cells of size  $8 \times 8$  cm was assigned to each plot, and then the associated height was calculated from the points located in each cell, multiplied by the size of the cell to estimate the volume of the canopy within each cell. The aggregate “volume” in each plot is referred to as the volume of the vegetation within a plot. The height of each cell in this study was calculated as the average of the height of the lowest point and the height of the highest point in each cell.

### *Canopy Cover*

Canopy cover can be estimated from LiDAR data as the ratio of above-ground points (or, canopy points) to the total number of LiDAR points in a given area. The following approach was used in this study for canopy cover estimation for each plot. First, the field is divided into grid cells of a user-defined dimensions ( $8 \times 8$  cm in this study, consistent with the canopy volume calculation). Then, for each grid cell, the LiDAR points are split into two groups, canopy points and bare earth based on their height using a user-defined threshold. The points above the threshold are considered as canopy points. The canopy cover is estimated as the ratio of the number of canopy points to the total number of LiDAR points in each cell. The average of the canopy cover estimated for the cells located in each plot is assigned as the canopy cover for that plot. In this study, candidate threshold values were 0.1, 0.2, 0.3, 0.4, 0.5, and 0.75 multiplied by the 95<sup>th</sup> percentile height of each plot, resulting in six height-dependent canopy cover related features.

### *Height Statistics*

The spatial distribution of the height of the LiDAR points in each plot can also be represented using statistical moments of the distribution. These statistics can also be included as the input features for predictive models.

## **2.3.4 Regression-Based Modeling Approaches**

Common regression-based approaches such as Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), and Random Forests (RF) are widely utilized to build predictive models with remote sensing based inputs. PLSR reduces a potentially large number of

measured collinear input variables to a few uncorrelated latent variables while seeking to explain the maximum multi-dimensional variance of the dependent variable via a linear model. PLSR has been used by researchers for developing predictive models to estimate leaf biochemical and biophysical properties (Asner & Martin, 2008), chlorophyll content (Y. R. Zhao et al., 2016), carotenoid content (Féret et al., 2011), relative water content (Ullah et al., 2014), protein, lignin, and cellulose (Thulin et al., 2014), leaf nitrogen content (Ecarnot et al., 2013), LAI (X. Li et al., 2014), and biomass (Vaglio Laurin et al., 2016).

SVR is a supervised non-parametric regression technique, and therefore, no assumptions regarding the underlying data model are required. SVR transforms the original feature space into a higher dimensional space (T. Zhang, 2001), with the goal of finding a hyperplane to predict the training data set. The optimal values of the kernel function parameters were obtained in this study by a general k-fold cross-validation in a grid search.

Random Forest (RF) modeling is an ensemble learning technique which uses a large set of classification and regression trees (CART) to make a prediction (Breiman, 2001). In random forest regression, each tree is built by randomly choosing a set of variables and a subset of training samples with replacement. The selected samples are used for training, and the remaining observations are used in an internal cross-validation process to determine the performance of the RF model. The outputs of all trees are aggregated to produce a final prediction. A review of RF modeling in remote sensing applications is available in (Belgiu & Drăgu, 2016). Similar to SVR, the parameter optimization was accomplished by a general k-fold cross-validation in a grid search method.

## **2.4 Results and Discussion**

### **2.4.1 Data Screening**

#### ***Time Series of Biomass Data***

In the 2017 and 2018 growing seasons, the destructive biomass data were collected multiple times (approximately every month) from the Hybrid calibration panels. Figure 2.3 shows the fresh weight and moisture content of the 18 sorghum varieties planted in the HyCal-17 and HyCal-18 experiments. The moisture content increases at the beginning of the season until it reaches its

maximum at 50 to 60 days after sowing. The fresh weight of the plants also increases rapidly at the beginning of the season, while at the end of the season, it decreases as the plants senesce. Among the varieties shown in Figure 2.3, those that are photoperiod sensitive (“*Sordan Headless*” and “*Trudan Headless*”) did not flower in the environment in which these experiments were conducted, and continued to add plant material until the end of the season.

Each variety has four replicates in the Hybrid Calibration panels, providing adequate sample data to compare the relationship across the varieties and associated changes during the growing season. For each date that the biomass data were collected, an ANOVA test was conducted on the fresh biomass, and if it indicated variability among the varieties was highly significant, a Tukey’s multi-comparison test was performed. Figure 2.4 shows the results of Tukey’s pairwise multi-comparison test ( $\alpha=.05$ ) for the fresh biomass data collected in the 2017 and 2018 growing seasons. In general, at the beginning of each season, only a few varieties were significantly different, while the variability among the varieties at the end of each growing season was greater. From Figure 2.4 it is also clear that the two photoperiod sensitive varieties were significantly different than the other varieties at the end of both 2017 and 2018 growing seasons.

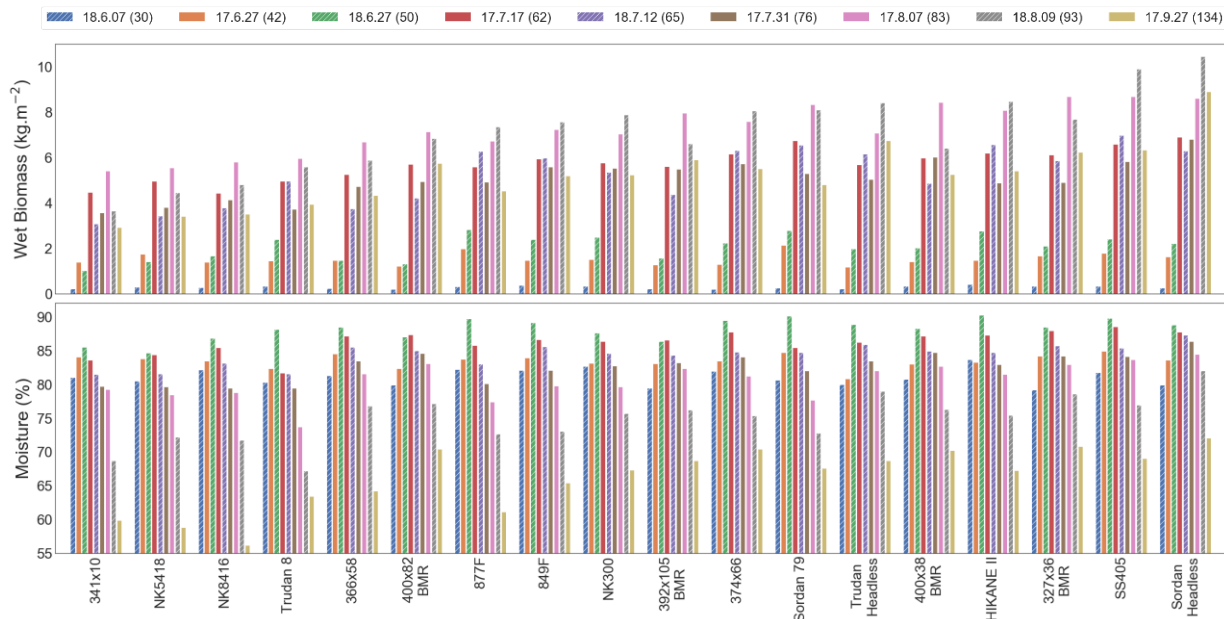


Figure 2.3. Ground reference data collected during the 2017 and 2018 growing seasons in the HyCal panels. For each variety, the data samples are sorted based on the day after sowing (DAS).

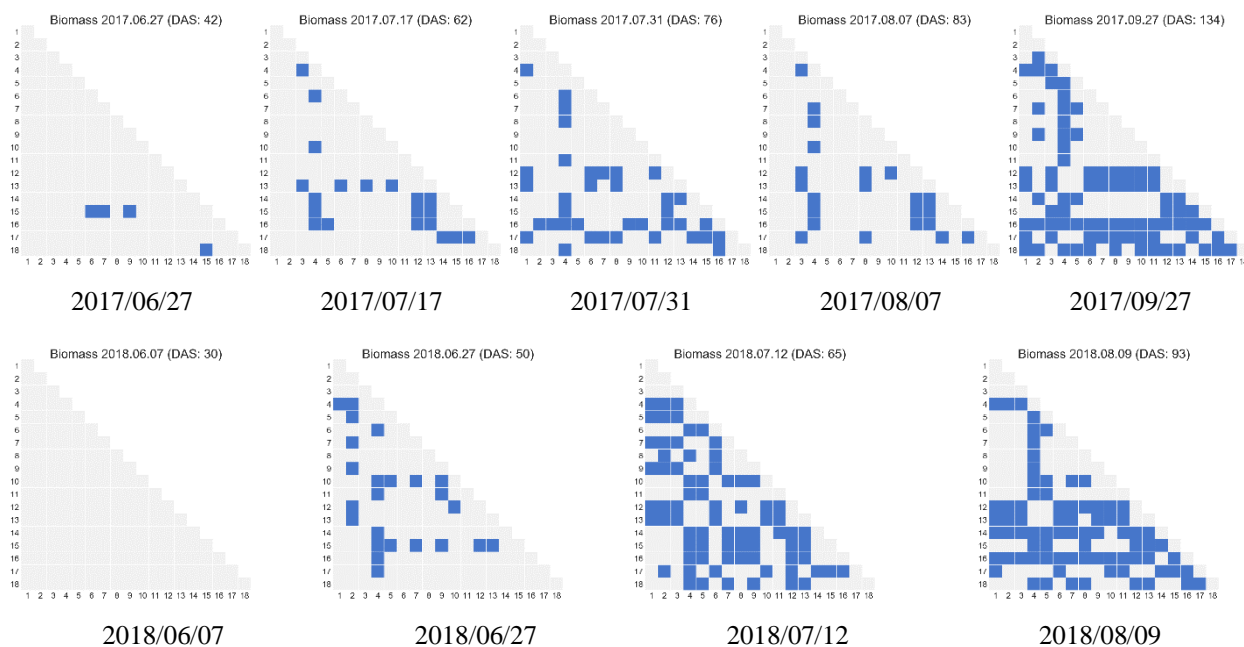


Figure 2.4. Tukey's pairwise multi-comparison test for the fresh biomass data collected in the 2017 and 2018 growing seasons. Blue indicates that the two varieties are significantly different ( $\alpha=0.05$ ).

### ***Time Series of Remote Sensing Data***

RS spectral signatures varied both across phenotypes and during the growing season. Figure 2.5 shows the average spectra of the 18 varieties of sorghum in the HyCal-18 panel on July 18<sup>th</sup>, 2018 (DAS=71). On that day, the signatures of the 18 varieties were very similar in the visible range of the spectrum, but there was more variability in the NIR portion of spectrum that may reflect variation in biochemical features, including lignin type and composition associated with the brown-midrib (bmr) traits. Figure 2.6 shows the reflectance of one of the varieties from June to September which shows there is very little change in the visible range of the spectrum, and especially in the blue and green bands, while the reflectance in NIR bands changes from about 35% to 60% on average. The maximum reflectance values were observed in the range of 750-850 nm on July 3<sup>rd</sup> (DAS=56). One of the reasons for changes in the reflectance for sorghum is the appearance of the panicles, usually a few days before the flowering date (the date on which 50% of panicles in a plot are flowered). Field notes indicate that the flowering date for “*Trudan 8*” was July 10<sup>th</sup>, 2018.

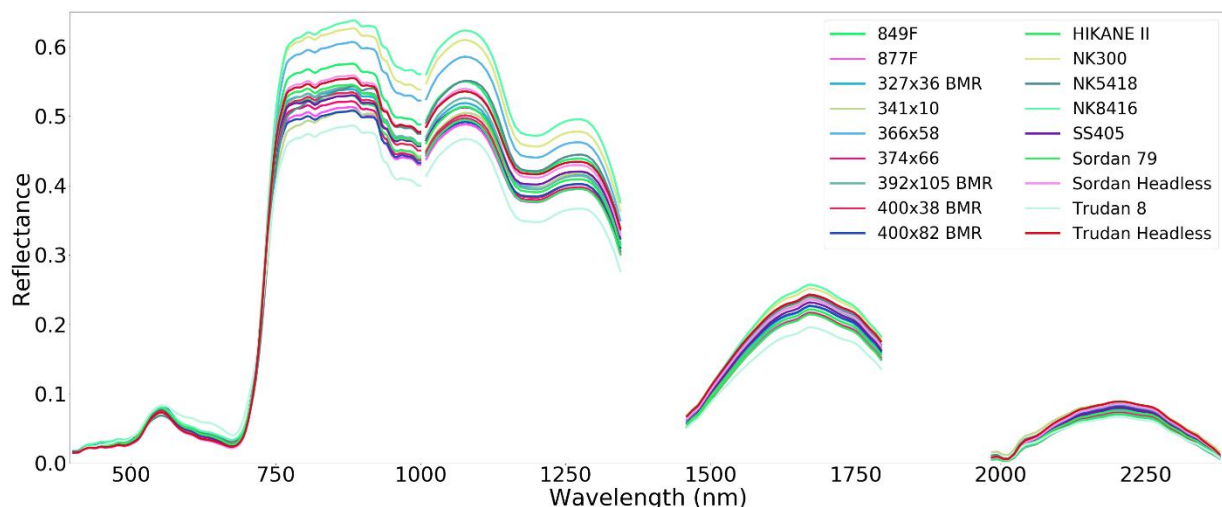


Figure 2.5. Spectra of the 18 Sorghum Varieties in HyCal-18 panel on July 18<sup>th</sup>, 2018. The varieties are very similar in the visible range of the spectrum, but substantial variability in the NIR portion of the spectrum.

Figure 2.7 shows the DSM generated from the LiDAR point cloud for the SbDivTc-18 panel from multiple dates in the 2018 growing season. From Figure 2.7, the plots are more similar at the beginning of the season, with greater differences later in the season. Figure 2.8 shows the point cloud data for two plots (rows 2 and 3) of the HyCal-18 experiment for multiple dates in the 2018 growing season. “341x10” is a dwarf grain sorghum variety (Figure 2.3), while “Trudan Headless” is a photoperiod sensitive forage sorghum with high biomass accumulation.

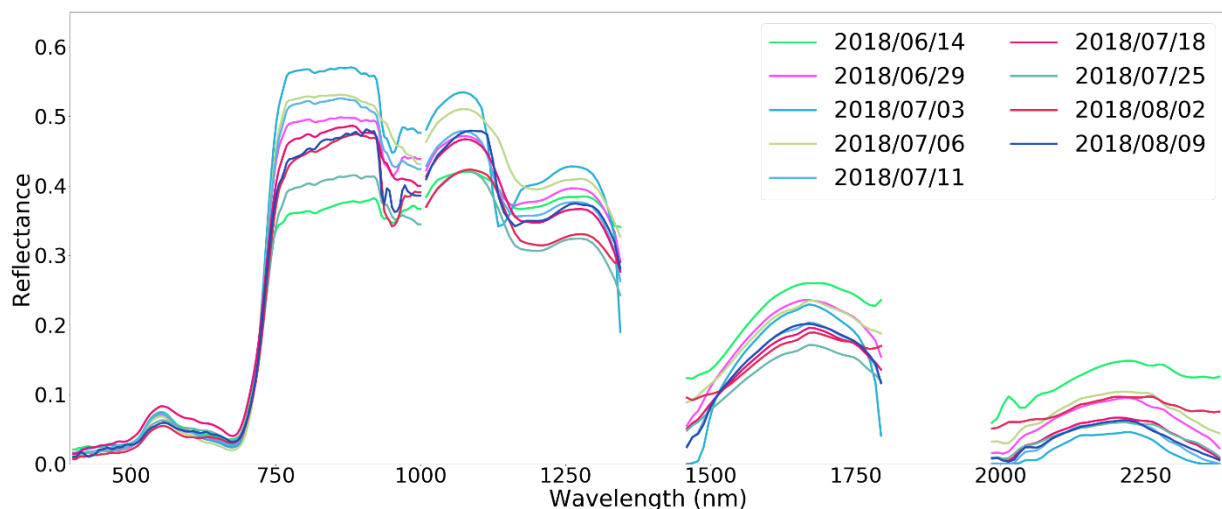


Figure 2.6. Example of reflectance of one of the varieties in the HyCal-18 experiment (“Trudan 8”) during the 2018 growing season.

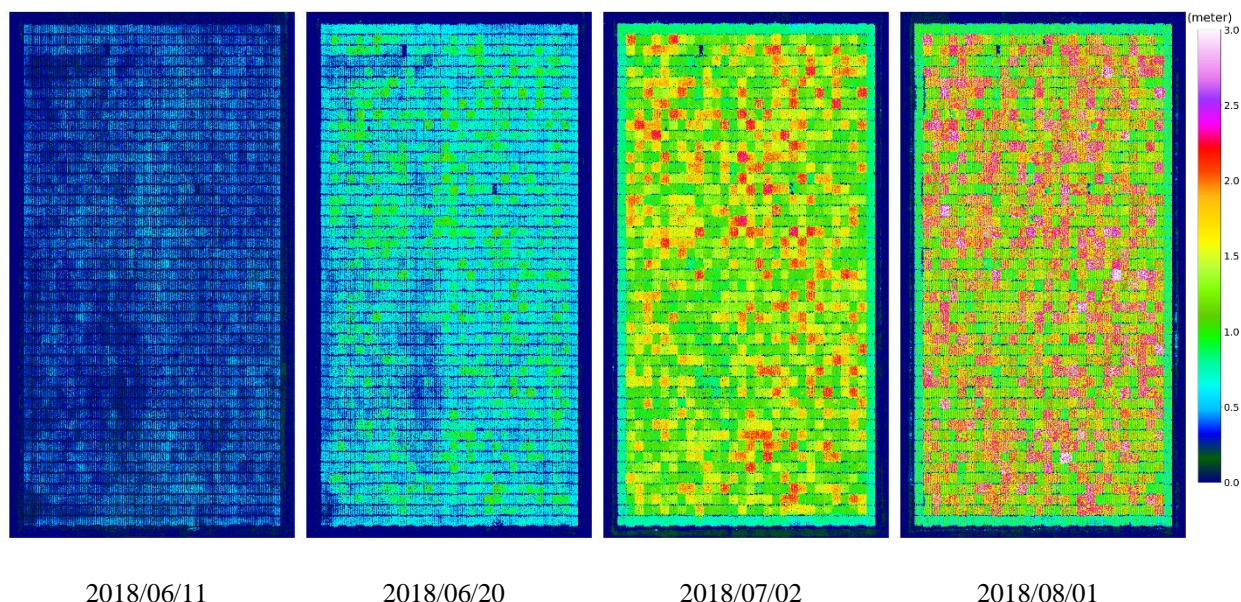


Figure 2.7. The 8 cm resolution DSM for multiple dates in the 2018 growing season for the SbDivTc experiment.

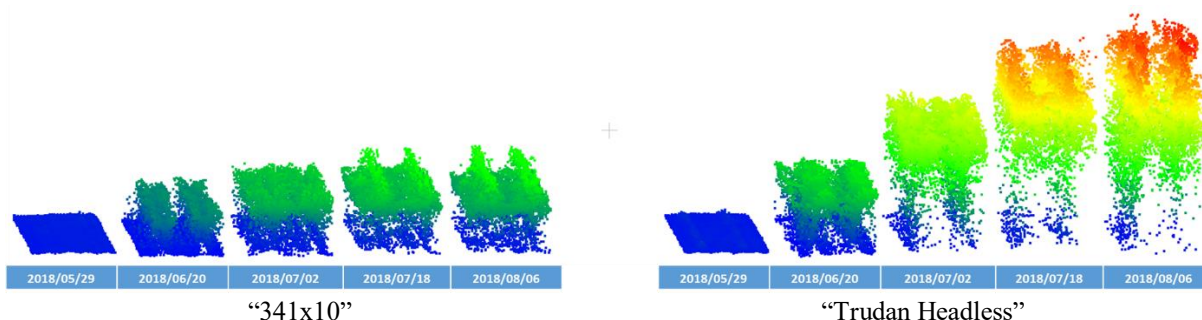


Figure 2.8. Point cloud data for rows two and three of two plots of the HyCal-18 experiment for multiple dates in the 2018 growing season. “341x10” and “Trudan Headless” achieve a maximum height of 1.4m and 3.2m, respectively.

Figure 2.9 shows the average and standard deviation of the height of all the plots for all the experiments in the 2017 and 2018 growing seasons extracted from the LiDAR point clouds, providing structural characteristics of the varieties planted in each experiment. For both HyCal-17 and HyCal-18, the height increases as the headless varieties continue to grow until the end of the season. The SbBAP-17 experiment also has the maximum average height at the end of the season, as it includes many plots of photoperiod sensitive genotypes that do not flower. The InCal-17, InCal-18, and SbDiv-17 include similar inbred varieties; thus, they have a very similar average height (also the lowest height values among the experiments). As was noted earlier, a histogram of the height of the points from the LiDAR point cloud provides information about the distribution

of different height values in a plot. This genotype dependent information may be discriminating in predictive models. Figure 2.10 shows the histograms for the dwarf grain sorghum 341x10 and the photoperiod sensitive Trudan Headless varieties in the HyCal-17 experiment.

The multi-temporal and cross-correlations during the growing season can also be used to screen for redundant features. Figure 2.11 shows the correlation matrices for the OSAVI and LiDAR-based canopy cover calculated using the combined data acquired over the HyCal-17 experiment in the 2017 growing season. It illustrates the rapid changes at the beginning of the season, especially prior to the flowering time (second week of July for this experiment) which is associated with the rapid growth of the plants. From Figure 2.11, OSAVI changed more than the canopy cover during the early season, and end of season OSAVI values have lower inter-temporal correlation compared to the correlation between the canopy cover values on corresponding dates.

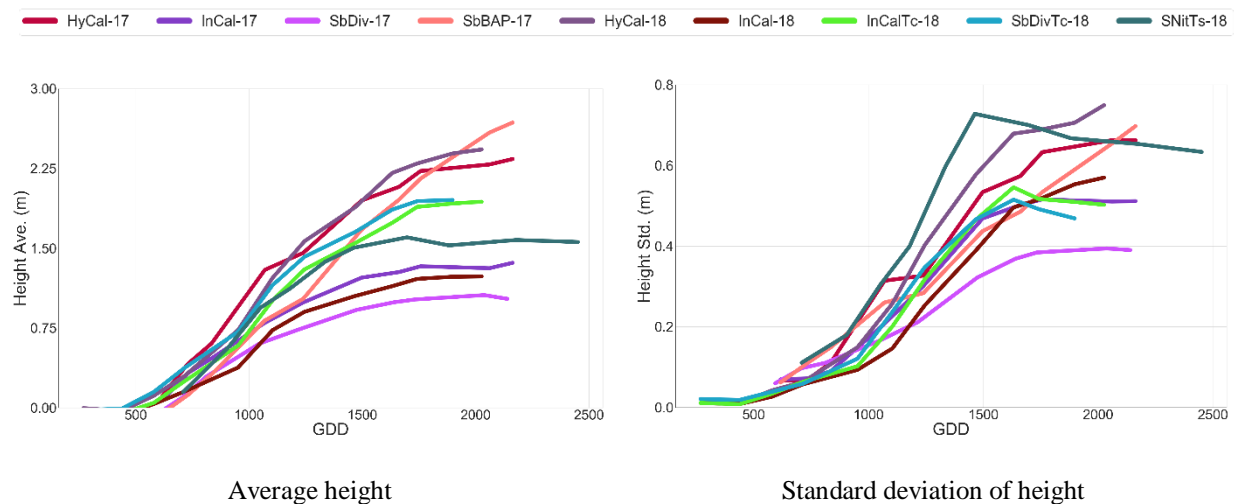


Figure 2.9. Average and standard deviation of the height of all the plots in each experimental trail in the 2017 and 2018 growing seasons.

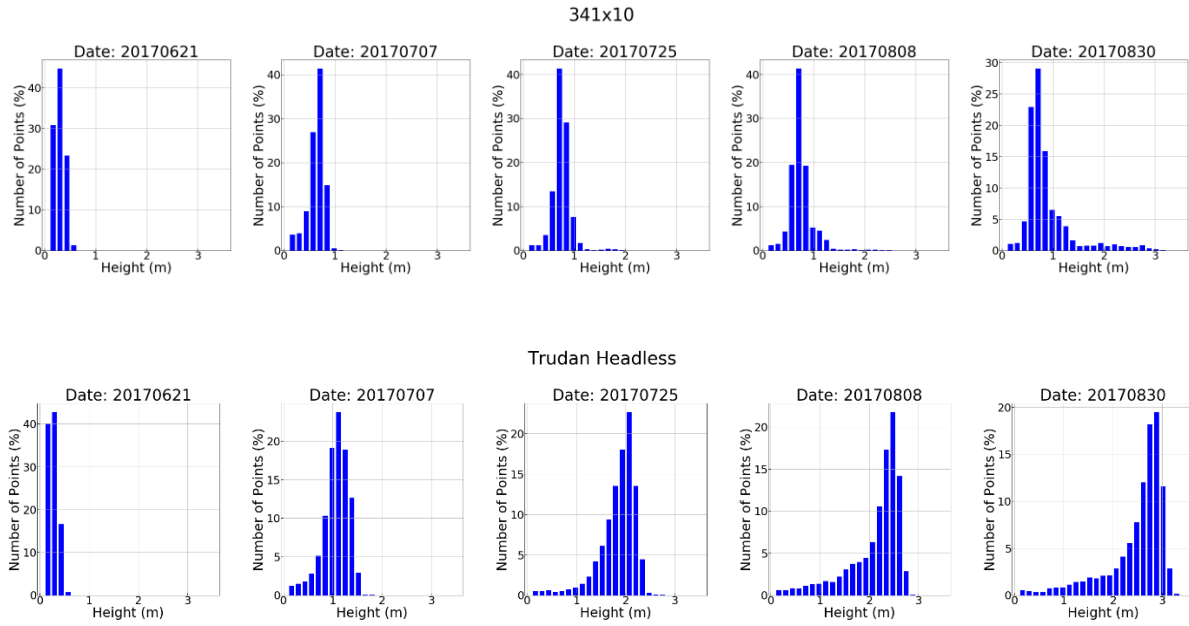


Figure 2.10. Height histogram for “341x10” and “Trudan Headless” in the HyCal panel across the 2017 growing season.

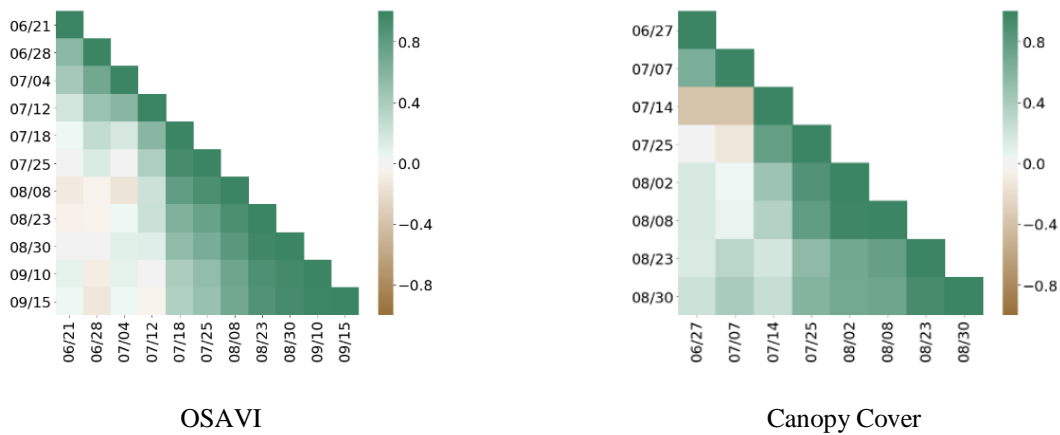


Figure 2.11. Correlation matrix calculated using all the remotely sensed hyperspectral data over the HyCal-17 experiment on different dates in the 2017 growing season. Note that as the LiDAR and hyperspectral sensors were flown on separate platforms, the number of available data sets differs, and the data were not always collected on the same day.

Similar to the last section, Figure 2.12 and Figure 2.13 show the results of the multiple-comparison Tukey’s test for the OSAVI and volume features for the HyCal-17 and HyCal-18 experiments on multiple dates during the 2017 and 2018 growing seasons. These results are

consistent with the results of Tukey's test conducted on biomass data in the previous section, which greater variability among the varieties at the end of each growing season.

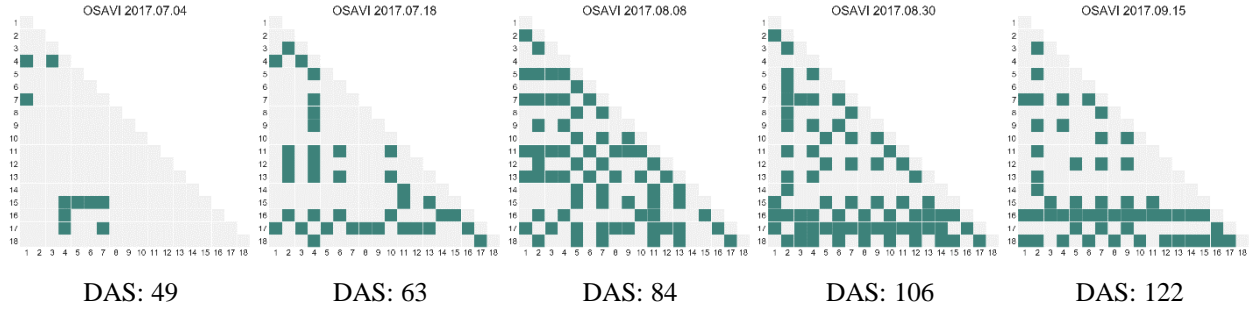


Figure 2.12. Multiple-comparison Tukey's test for the OSAVI index collected over the HyCal-17 experiment. Green shows the two varieties are significantly different from each other ( $\alpha=0.05$ ).

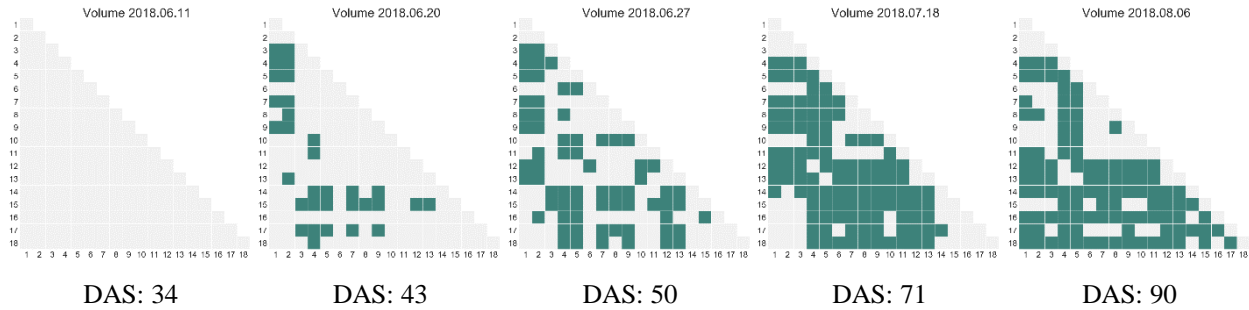


Figure 2.13. Multiple-comparison Tukey's test for the LiDAR-based volume collected over the HyCal-18 experiment. Green shows the two varieties are significantly different from each other ( $\alpha=0.05$ ).

### ***Relationship between Features and Biomass***

In the last two sections, the changes in the biomass and RS data during a growing season were discussed. In this section, the relationship between the biomass data and RS features, as well as the change in their relationship during the growing season, are discussed. Figure 2.14 shows one feature from LiDAR (90<sup>th</sup> percentile height of the plants), one feature from hyperspectral data (Intg\_NIR1), and the biomass data for the dwarf grain sorghum 341x10 and the photoperiod sensitive Trudan Headless varieties, both from the HyCal-17 and HyCal-18 experiments (one with low and one with high biomass values). To compare the data from the two years at the same stage

of growth, the data are plotted versus Growing Degree Day (GDD), a heat index calculated from temperature data for each day that is used to determine the growth stages of sorghum (Gerik et al., 2003). At GDD of 2100, the biomass in 2018 for both varieties was slightly higher than in 2017 (as noted in Table 2.1, the HyCal-18 was planted two weeks earlier than HyCal-17). The height data, Intg\_NIR1, and biomass data follow the same pattern of change over time for each variety in both growing seasons. The height for the photoperiod sensitive variety (“*Sordan Headless*”) always increases while other varieties stop increasing in height around flowering time (GDD of 1500); the Intg\_NIR1 increases rapidly earlier in the season, and then gradually decreases at around GDD of 1500 until the end of the season; the biomass continues to increase, and especially for the photoperiod sensitive variety. Inter-annual differences also inherently include the impact of the timing and quantity of rainfall.

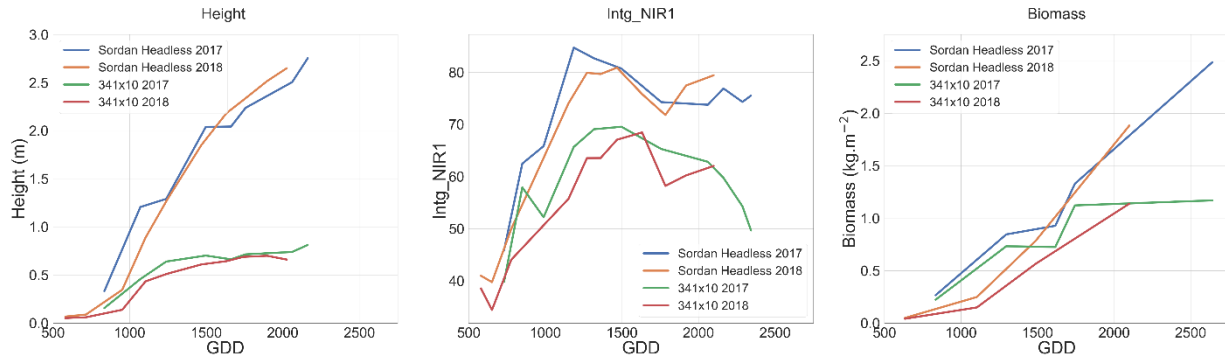


Figure 2.14. Comparison of the 90 percentile height, Intg\_NIR1, and the biomass data for the dwarf grain sorghum 341x10 and the photoperiod sensitive Trudan Headless varieties in the 2017 and 2018 growing seasons.

For each feature extracted from the RS data, the simple prediction potential ( $R^2$ ) and associated changes during a season were investigated using linear regression-based models for the end of season biomass prediction. Robust features should be applicable across the varieties, at least for common experiments. Figure 2.15 and Figure 2.16 show the  $R^2$  values of the models for each feature extracted from LiDAR and hyperspectral VNIR data, at four stages of growth and for all nine experiments conducted in the 2017 and 2018 growing seasons. From Figure 2.15, the 30<sup>th</sup> percentile height and volume features provided the highest  $R^2$  values for the end of season biomass among the LiDAR-based features for both HyCal-17 and HyCal-18 experiments, as the varieties in those experiments were more diverse in their structural characteristics, providing strong potential for biomass prediction using geometric-related features. The  $R^2$  values for different

LiDAR-based features in the InCal-17, InCal-18, and SbDiv-17 experiments are very similar, which is consistent with Figure 2.9; they all have the lowest average height and lowest variability in height compared to the other experiments, resulting in lower  $R^2$  values for these experiments compared to the experiments with hybrid cultivars. For both InCal-17 and InCal-18 experiments, the highest  $R^2$  values were obtained from feature #5 (coefficient of variation of height) which is representative of the distribution of the points in the canopy point cloud. The  $R^2$  values of the models developed for the SbDivTc-18 and InCalTc-18 are lower compared to the HyCal-17 and HyCal-18 experiments, but the same features (features #1, #2, and #5) provided the maximum  $R^2$  for all of these experiments, which include hybrid cultivars. The SbBAP-17 also includes hybrid cultivars; however, the  $R^2$  values for all the features are lower compared to all other experiments, mainly because the last LiDAR data were collected on Aug. 30<sup>th</sup>, and included many photoperiod sensitive cultivars which grew until the final biomass data were collected at the harvest (Sep. 28<sup>th</sup>). Other varieties did not grow during this time, which impacted the biomass-height relationships. Generally, for the experiments with the hybrid cultivars (refer to Table 2.1), the late season data sets provided the highest  $R^2$ , while for the experiments that included inbred cultivars, the data sets of GDDs yielded the lowest  $R^2$  values.

For the hyperspectral features shown in Figure 2.16, for both InCal-17 and InCal-18, the highest  $R^2$  values are associated with the dataset collected on ~80 DAS, while for other experiments, the dataset of ~95 DAS yielded the highest  $R^2$  values. Also, the same pattern for  $R^2$  values for the features of the InCal-17, InCal-18, and SbDiv-17 was observed. For these panels, the  $R^2$  is generally higher than the panels that include hybrid cultivars.

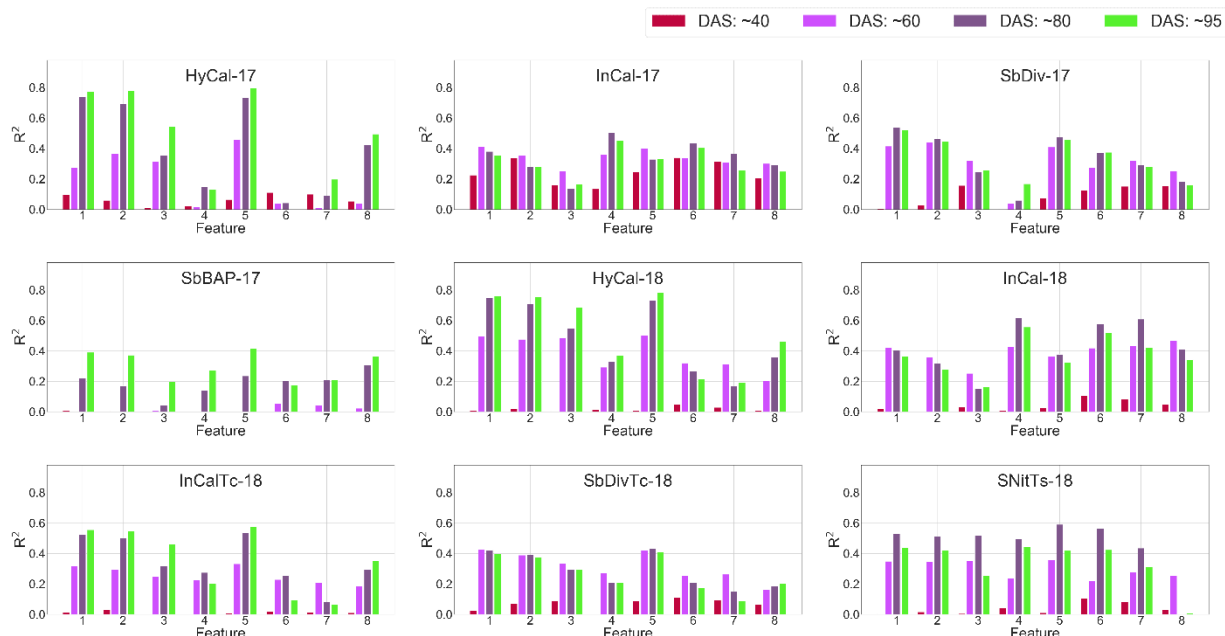


Figure 2.15.  $R^2$  values of the linear regression-based models developed for the end of season fresh biomass using LiDAR-based features at four stages of growth. Feature 1 to 8 represent: #1: 30<sup>th</sup> percentile height, #2: 50<sup>th</sup> percentile height, #3: 95<sup>th</sup> percentile height, #4: coefficient of variation of height, #5: volume, #6: canopy cover (threshold=0.1), #7: canopy cover (threshold=0.3), #8: canopy cover (threshold=0.5).

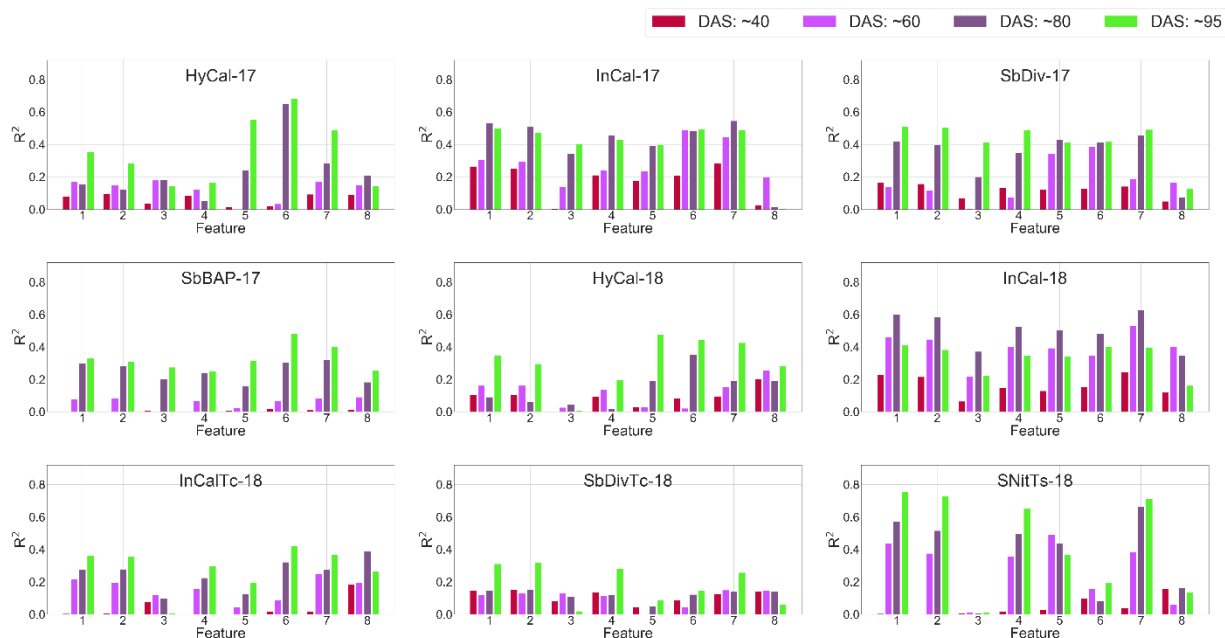


Figure 2.16.  $R^2$  values of the linear regression-based models developed for the end of season fresh biomass using VNIR features at four stages of growth. Feature 1 to 8 represent: #1: FDR\_min, #2: Intg\_NIR1, #3: SDR\_slope, #4: Intg\_NIR1, #5: NDVI, #6: SR<sub>800,680</sub>, #7: OSAVI, and #8: MCARI.

Given similar trends of the regression models shown in Figure 2.15 and Figure 2.16, the models were developed across all the experiments for each of the features, and all the available dates to investigate the potential for using a common set of features for all experiments and times for the multiple input predictive models. The average  $R^2$  for each feature, from all the dates and all the experiments is provided in Figure 2.17, which shows volume, 30<sup>th</sup> percentile height, OSAVI, FDR-min, and NDWI features had the highest average  $R^2$  from LiDAR, VNIR, and SWIR data sets, respectively. Linear regression models were also developed for the individual band values from both hyperspectral VNIR and SWIR data. The average and maximum  $R^2$  for each band from all the dates and all the experiments are shown in Figure 2.18, which shows that the area of spectrum between 750 and 1100 nm provided the highest  $R^2$  for the linear regression models. While the  $R^2$  values for some experiments and some dates for the bands in 2000-2300 nm are relatively high (30-60%), the average  $R^2$  values are much lower in comparison to the 750-1100 nm range.

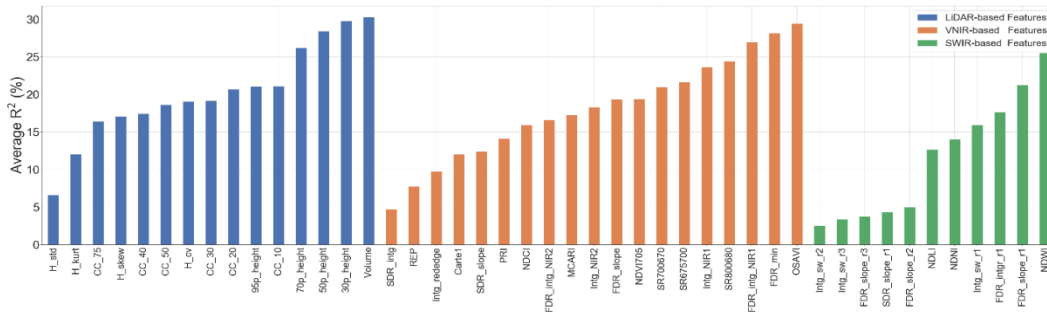


Figure 2.17. The average  $R^2$  values of the linear models developed for all the dates and all the experiments for each feature type from hyperspectral and LiDAR data.

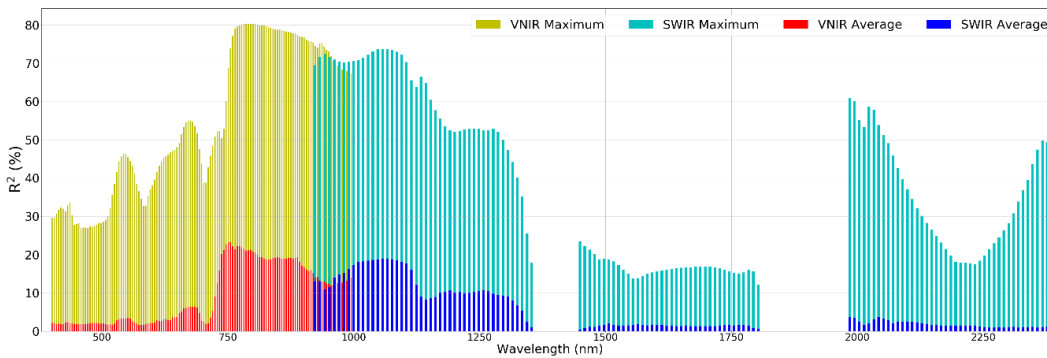


Figure 2.18. Average and maximum  $R^2$  values of the linear models developed for all the dates and all the experiments for each band from hyperspectral VNIR and SWIR data.

### 2.4.2 Biomass Predictive Models

In this section, the results related to the impact of different regression methods, the time of biomass sampling and remote sensing data acquisition, and the number of samples on the prediction results are provided.

#### *Impact of the data source and regression method on the prediction results*

To evaluate the performance of different regression-based modeling approaches, PLSR, SVR, and RF were implemented for end of season biomass prediction using the LiDAR and hyperspectral data collected in each growing season. Figure 2.19 shows the  $R^2$  values of the predictions relative to the reference data for all the experiments, using six data sources (LiDAR, VNIR, SWIR, and combinations), and the three methods. For each prediction with a data source, all available data sets over the whole season were used for training and validation of the models, where two thirds of the sample data (or a maximum of 200 samples) were randomly selected 100 times for the training of the algorithm, and the remaining samples were used for cross validation via the hold-out method. For all the experiments except the SNiTs-18, all the replicates of a variety were assigned to either the training or test sets to avoid any impact from the number of replicates on the prediction results. SNiTs-18, however, included only four varieties, and a different number of replicates for each variety; thus, the training and test sets were assigned randomly from the plots regardless of their varieties for this experiment. Potential reasons for differences in predictions include:

- i. Diversity in the samples: the regression models are better able to learn the pattern in the data when the samples are more diverse.
- ii. Number of data samples: the larger the number of data points in an experiment, the higher accuracies are typically achieved for the prediction.
- iii. Similarity between the training and test data sets; if the training and test data sets are very different, then overfitting can occur for the training data set, resulting in decreased accuracy of the predictions. Note that this can happen when the number of data samples is limited, which causes unlike training and test sets even when the samples are selected randomly. Also, if there is a significant range of biomass values in one experiment, there is more chance to have dissimilar training and test sets.

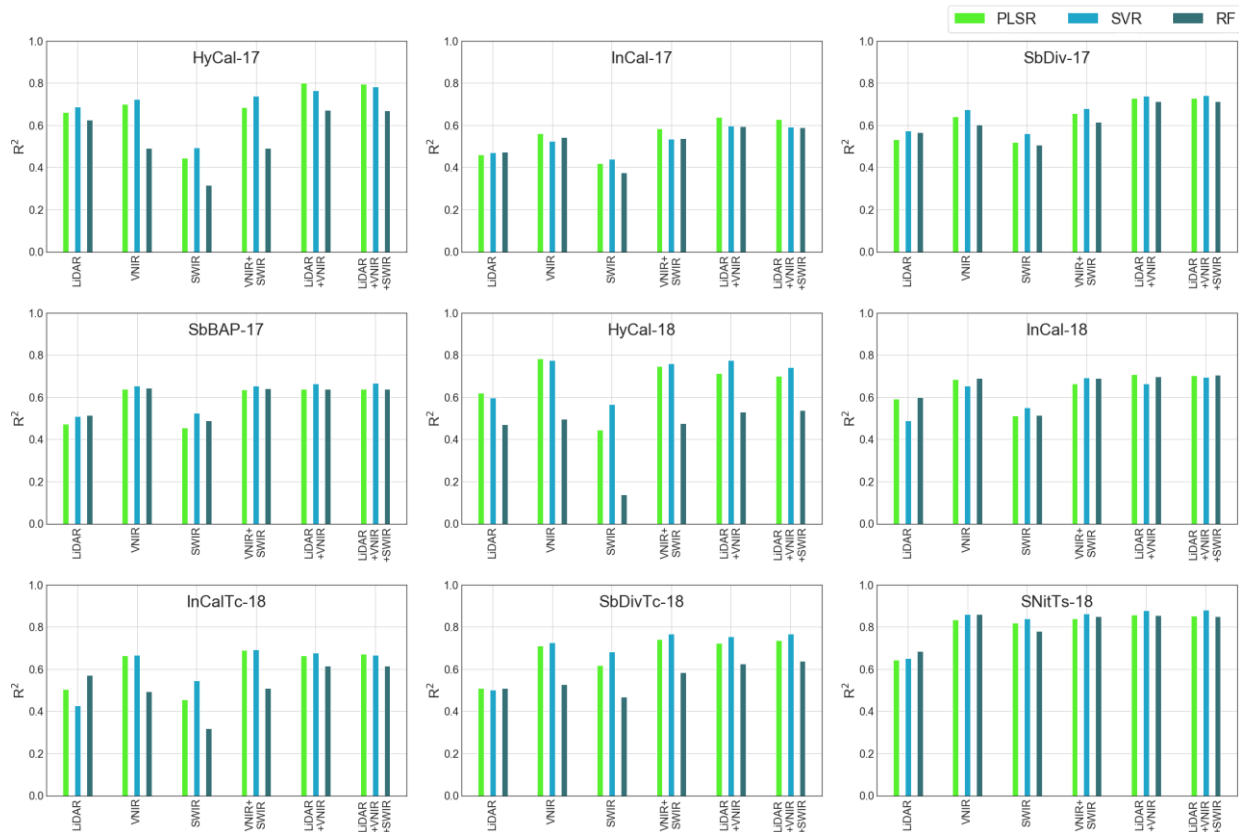


Figure 2.19.  $R^2$  values of the end of season fresh biomass predictions, using six data sources, and the three PLSR, SVR, and RF methods for all the experiments conducted in the 2017 and 2018 growing seasons.

Figure 2.19 shows that the highest accuracy of end of season biomass prediction using all combinations of data sources was achieved for the SNitTs-18 experiment. There were two nitrogen treatments in this experiment; half of the plots in this experiment were fertilized with 250 kg/ha nitrogen while the other half were not fertilized, causing high and low biomass values for the plots, high diversity in the reflectance data from the hyperspectral images, as well as high diversity in geometric-based features extracted from LiDAR point cloud (reason i). As was noted, samples were assigned to the training and test sets for this experiment differently from the other experiments, causing multiple samples of each variety to be assigned to both the training and test sets, resulting in increased similarity in the two sets (reason iii).

The highest accuracy of prediction using LiDAR features as the sole input was obtained for the HyCal-17 and HyCal-18 experiments, which include hybrid cultivars that are more diverse in structural characteristics compared to the inbred cultivars; thus, the regression model can distinguish and relate the LiDAR-based features to the biomass data (reason i), which is consistent

with the results in Figure 2.15. In general, the predictions are more accurate for the experiments that include hybrid cultivars. As was shown in Figure 2.9, the InCal-17, InCal-18, and SbDiv-17 have the smallest standard deviation in the LiDAR-based height, indicating that the associated varieties have similar structural characteristics.

The predictions for the SbDiv-17 are more accurate than the predictions for the InCal-17 as more samples were available for the training set, 200 for SbDiv-17, and 80 for InCal-17 (reason ii). For the SbBAP-17, the prediction accuracies are lower than most of the other experiments. This experiment included varieties that were highly diverse in terms of structural characteristics (Figure 2.9), the distribution of the biomass values for this experiment (see Figure 2.1), also had a much larger range compared to the other experiments. This resulted in dissimilar samples in the training and test sets (reason iii). Figure 2.20 shows a box plot for the fresh biomass data for all the experiments, and the SbBAP-17 experiment had the greatest range of biomass values among the experiments, the lowest accuracies for the predictions, while SNitTs-18 experiment has the smallest range and the highest  $R^2$  values for the predictions. The predictions results with respect to the diversity, number of samples, and similarity between test and training sets for all the experiments are summarized in Table 2.7.

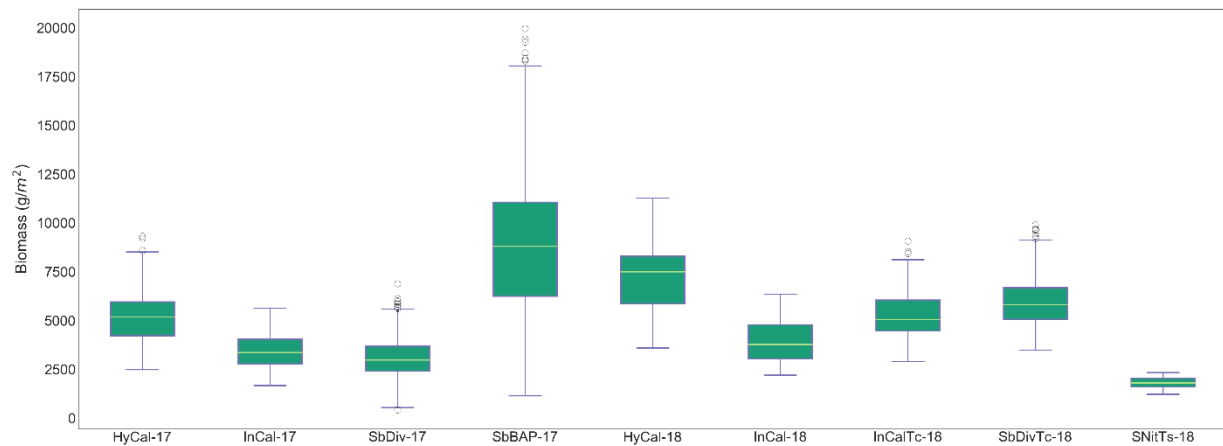


Figure 2.20. The box plot for the fresh biomass data for all the experiments conducted in the 2017 and 2018 growing seasons.

Table 2.7. Summary of the prediction results of the experimental trials

	HyCal-17 and 18	InCal-17 and 18	SbBAP-17	SbDiv-17	InCalTc-18	SbDivTc- 18	SNitTs
Diversity	Very High	Low	Very High	Low	High	High	High
Number of samples	Low	Low	High	High	Low	High	Low
Test-training set Dissimilarity	High	Low	Very High	Low	High	High	Low
Prediction accuracy (maximum $R^2$ )	High (0.80)	Medium (0.71)	Medium (0.67)	Medium (0.74)	Medium (0.69)	High (0.77)	Very high (0.88)

In general, for the experiments with hybrid cultivars, RF models had lowest prediction accuracies among the three methods, which is related to the fact that there was more dissimilarity between the training and test sets in both RS and biomass data among the hybrid cultivars compared to those that were inbred, and RF models can be overfitted to the training data set; thus, they may not provide as accurate predictions as SVR and PLSR. For the InCal-18, however, RF yielded the highest accuracies for most of the data sources. For the experiments with a sample size of 200 (SbDiv-17, SbBAP-17, and SbDivTc-18), the SVR models provided the most accurate results, while for the experiments with a lower number of data samples, PLSR provided the highest prediction accuracies.

A summary of the prediction results for various data sources and regression methods is provided in Figure 2.21, where the  $R^2$  values of the nine experiments are shown in a box plot. For the LiDAR data, the RF method provided slightly higher median accuracies than PLSR and SVR, with lower variability in  $R^2$  values (more reliability). When VNIR data was the only input, PLSR yielded more accurate results, while for all other data sources, SVR yielded a higher median  $R^2$ . For SWIR and VNIR combined with SWIR sources, SVR provided more reliable results, while for VNIR, VNIR combined with LiDAR, as well as a combination of all data sources, PLSR provided more reliable results. Also, the SVR models provided the maximum  $R^2$  for all the data sources.

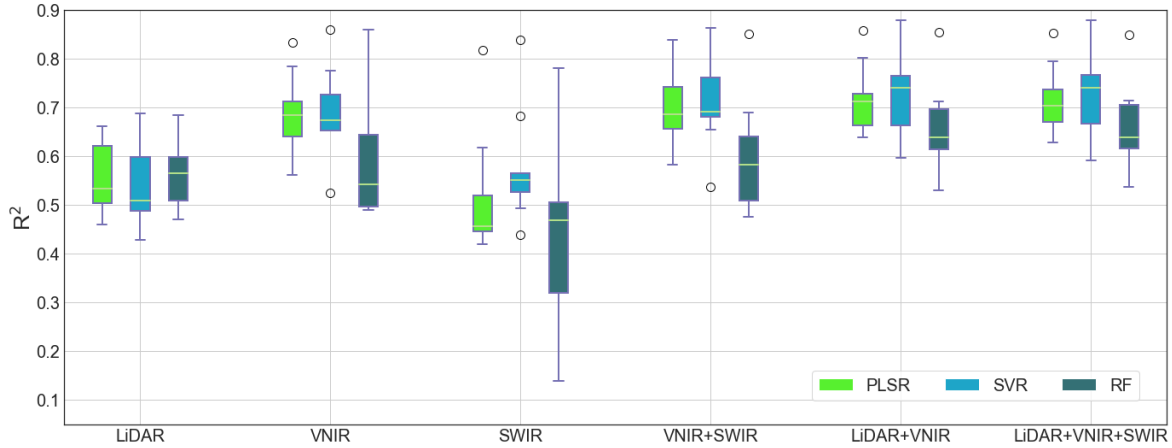


Figure 2.21. Box plot of the prediction results for various data sources and regression methods.

Similar to (de Almeida et al., 2019), an ANOVA test was performed on the prediction results to determine the impact of the method and data source (e.g. sensor-based features) on the prediction results. The ANOVA results provided in Table 2.8 show that the data source is the cause of 69% of the variation in the prediction results, 28% for the regression method, and 3% for the interaction their interaction. This indicates that the data source should also be considered to determine the regression method in the design of similar experiments. A similar ANOVA test was performed on the  $R^2$  values for six experiments, including HyCal-17, HyCal-18, InCal-17, InCal-18, SbDiv-17, and SbDiv-18, of which three include hybrid cultivars, and three include inbred cultivars. On this test, the cultivar type (hybrid or inbred) was considered as the third factor. Recall that the sorghum hybrid cultivars are more variable in their characteristics than inbreds in terms of biomass and structural characteristics. The results of this test in Table 2.9 indicate that the data source also has the highest contribution to the variation in the predictions (44%). The regression method is the cause of 26% of the variation in the prediction results, and less than 1% is attributed to the cultivar type. However, the interaction between the regression method and cultivar is responsible for 24% of the variation, suggesting that the cultivar type is another important factor to consider when determining the regression method for developing predictive models.

Table 2.8. Analysis of variance of the  $R^2$  respective to the data source, regression method, and their interaction

Factor	Sum of squares	Degree of freedom	F value	p-value	$\eta^2$ (%)
Data source	103.70	5	564.09	< 2e-014	68.59
Method	17.15	2	233.23	< 2e-014	28.36
Data source:method	9.24	10	25.14	< 2e-014	3.06
Residuals	594.95	16182			

Table 2.9. Analysis of variance of the  $R^2$  respective to the data source, regression method, cultivar type, and their interactions

Factor	Sum of squares	Degree of freedom	F value	p-value	$\eta^2$ (%)
Data source	80.95	5	412.58	< 2.e-06	44.18
Method	19.29	2	245.79	< 2.e-06	26.32
Cultivar type	0.35	1	8.86	0.003	0.95
Data source:method	6.84	10	17.44	< 2.e-06	1.87
Data source:cultivar type	4.03	5	20.53	< 2.e-06	2.20
Method:cultivar type	17.59	2	224.12	< 2.e-06	24.00
Data source:method:cultivar type	1.82	10	4.63	0.093	0.50
Residuals	422.37	10764			

For HyCal-17 and HyCal-18 experiments, PLSR, SVR, and RF models were developed using all three data sources and leave-one-out cross validation strategy, where in each fold, one variety was assigned as test and the other 17 varieties were included in the training set. The results are shown in Figure 2.22. For both experiments, the SVR method provided the highest  $R^2$  values for the predictions. For HyCal-17, all three regression methods underestimated the value of the biomass for one of the photoperiod varieties (which also had the maximum biomass, as noted previously); however, the RF model had the lowest accuracy. RF for both years resulted the lowest  $R^2$  as a result of overfitting as was discussed earlier. All three methods resulted in less accurate predictions for the experiment in 2018 compared to 2017, which could be because the end of season biomass data were measured at an earlier date in 2018, when all the plants had not reached full maturity.

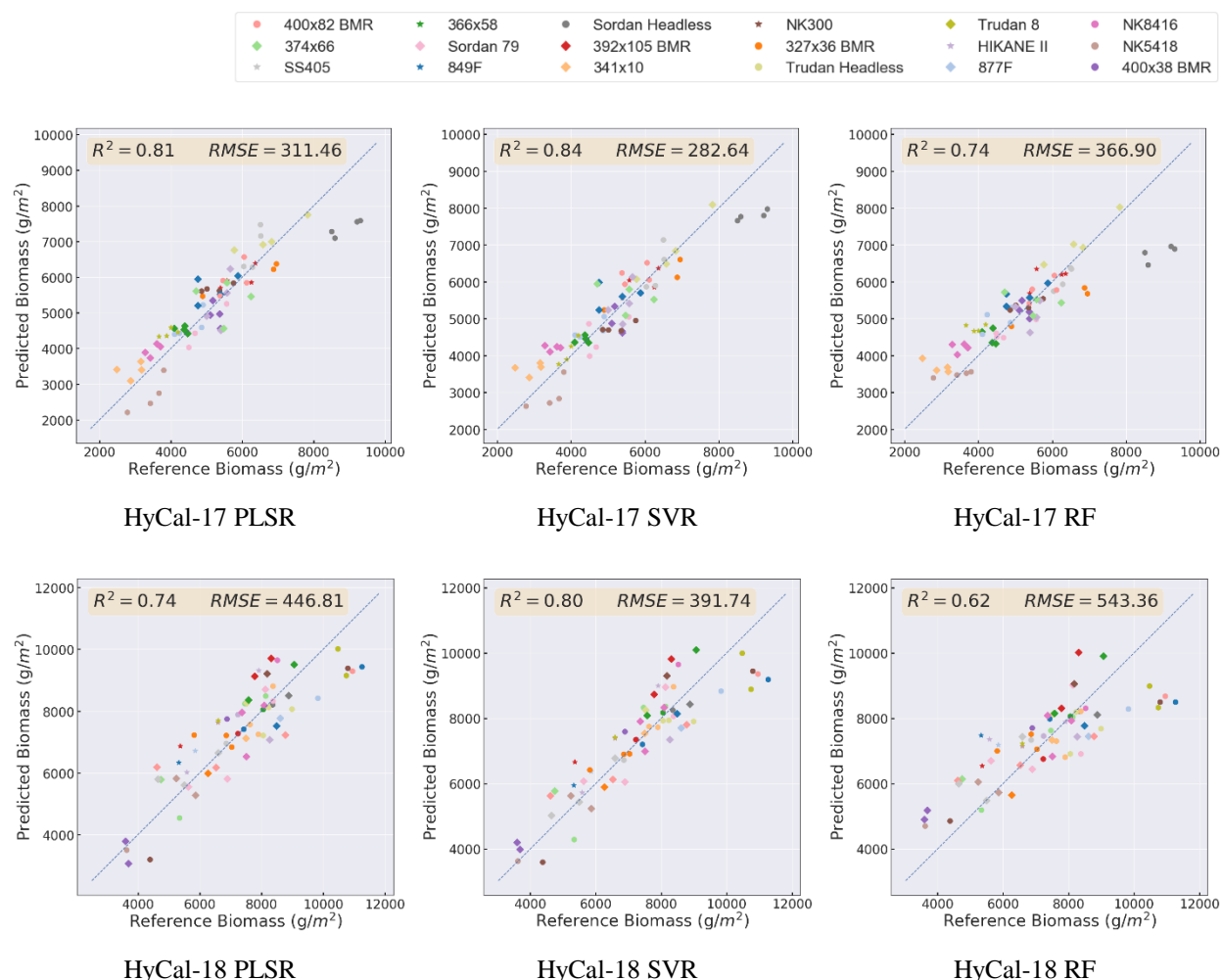


Figure 2.22. Prediction results of PLSR, SVR, and RF models developed for the HyCal-17 and HyCal-18 experiments, using all the data sources and leave-one-out cross validation strategy.

### *Predictions in time*

To evaluate the capability of remotely sensed data for predicting biomass through the growing season, SVR models were developed for six dates in the 2017 and 2018 growing season for the HyCal-17 and HyCal-18 experiments. The  $R^2$  values of the predictions relative to the reference data are shown in Figure 2.23. For each prediction, all the VNIR hyperspectral and LiDAR data collected prior to the date of biomass measurement were used in the SVR models. The  $R^2$  values of the predictions at the beginning of the season were lower compared to the end of the season, especially when using only VNIR features. Early season growth is focused on the production of biomass from stalks and leaves, while mid-season development is related to flowering and early development of panicles. Plant structural characteristics do not change

significantly after flowering in the mid-season, while spectral characteristics change significantly especially during flowering with the emergence of the panicles.

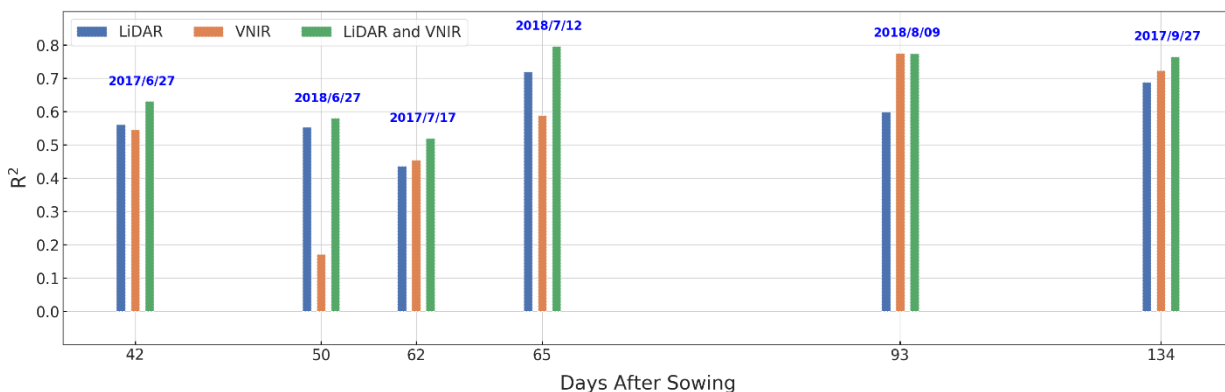


Figure 2.23.  $R^2$  values for biomass predictions during the 2017 and 2018 growing seasons using the SVR models developed based on VNIR hyperspectral and LiDAR data for the hybrid calibration panels.

### ***Multi-temporal predictions of end-of-season biomass***

It is highly desirable to predict the end-of-season biomass as early as possible during the growing season to avoid unnecessary investment of phenotyping resources in non-productive varieties. The SbBAP-17, SbDiv-17, and SbDiv-18 were chosen to conduct the evaluations in this section as they had an adequate number of samples as well as RS data points, and included both hybrid and inbred cultivars. Figure 2.24 shows the accuracy of end-of-season biomass predictions for these experiments using hyperspectral VNIR, SWIR, and LiDAR data from each date individually, and in combination with the earlier dates. For both SbBAP-17 and SbDiv-17 experiments, the earliest data set yielded very low prediction accuracies. For SbBAP-17, the best results when using features from individual sensors for VNIR and SWIR data sets were achieved from Sep. 10<sup>th</sup> with  $R^2=0.60$  and  $R^2=0.54$ , respectively. Based on LiDAR data, Aug. 23<sup>th</sup> resulted in the highest values, with  $R^2=0.46$ . For the SbDiv-17 experiment, the combined VNIR and LiDAR data sets of July 25<sup>th</sup> and Aug. 2<sup>nd</sup> provided the highest accuracy of using individual data sets. For SbDivTc-18, using the data inputs from July 11<sup>th</sup> resulted an  $R^2$  of 0.75, which shows the July data sets have good potential for biomass predictions. For all three experiments, the best results were obtained when features from all the hyperspectral and LiDAR data sets from the whole season were used, resulting in  $R^2$  of 0.63, 0.75, and 0.78 for the SbBAP-17, SbDiv-17, and SbDiv-18 experiments, respectively. Although the best results were obtained using the whole season RS data,

the models developed using middle season data (DAS of ~60 to 80) were also able to provide comparable accuracies.

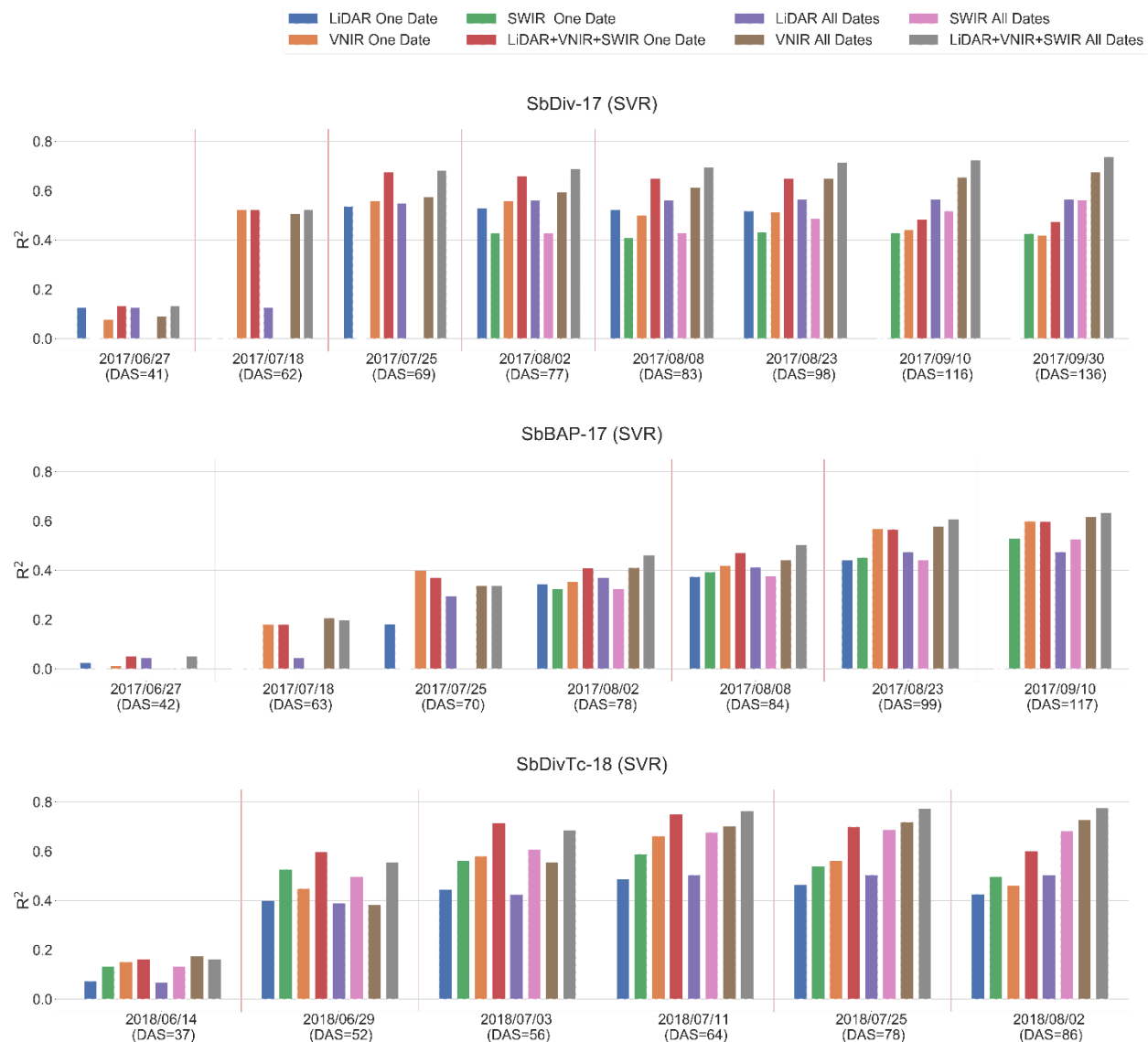


Figure 2.24.  $R^2$  for end of season predictions using hyperspectral and LiDAR data collected on different dates for the SbBAP-17, SbDiv-17, and SbDivTc-18 experiment.

### ***Impact of the number of samples on biomass prediction***

As noted earlier, measuring biomass in the field is time-consuming and expensive; however, it is still required for training the regression models. Generally, the greater the number of samples for training, the more accurate the predictions. To evaluate the impact of the number of samples

in training set, the SVR and PLSR models are developed for end-of-season biomass prediction using all the data sources and various number of samples in training set. Each model was trained with a specific number of samples, and the process was repeated 100 times, each time with a different, but same sized set of randomly selected samples. The rest of the available samples were assigned to the testing set. Figure 2.25 shows the median and standard deviation of the  $R^2$  values of developed models for some of the experimental trials. The  $R^2$  of predictions for both SVR and PLSR models increase as the number of training samples increases. However, the accuracy of PLSR models is higher than SVR models when a smaller number of training samples is used. The rate of increase of  $R^2$  with the respective increases in the number of training samples is higher for SVR compared to PLSR; thus, when the maximum of available samples is used in training for experiments (e.g. for SbDivTc-18), the SVR models had higher  $R^2$  values. For all the experiments, the standard deviation of the  $R^2$  values decreases as the number of training samples increases, showing more reliable (repeatable) prediction models are developed when more samples are available for training, as expected. However, for some experiments such as HyCal-17, the standard deviation of  $R^2$  decreases –initially, reaches a minimum, then increases. This is attributed to the small total number of samples: using more samples in the training set implies a smaller number of samples is available in the test set.

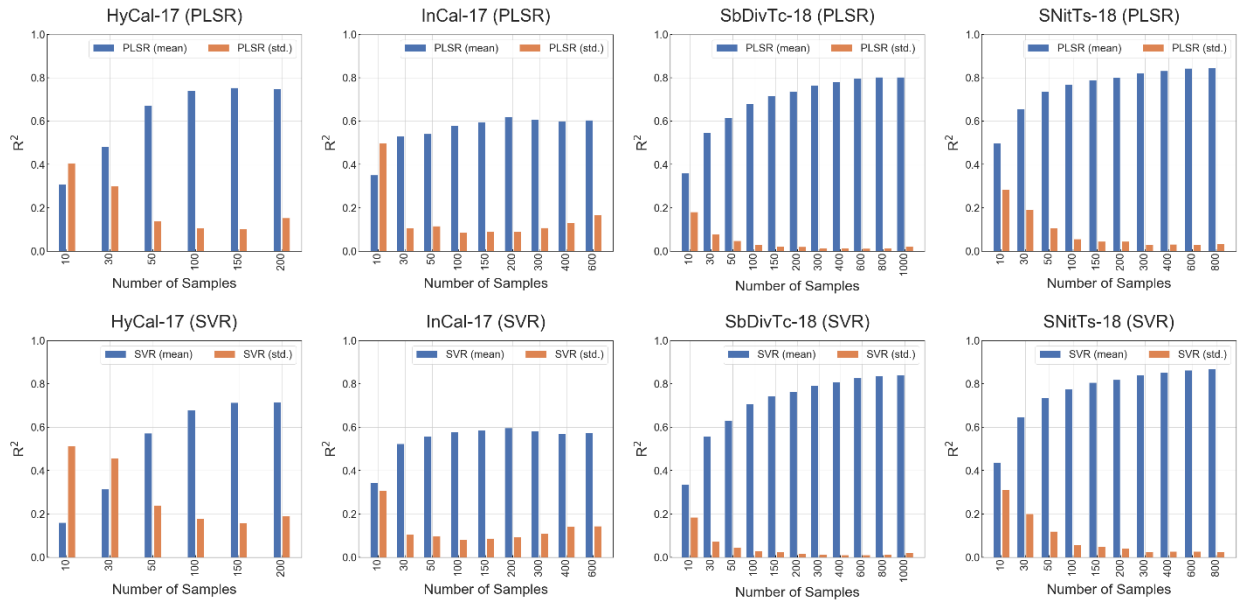


Figure 2.25. Impact of the number of training samples on  $R^2$  of the predictive models using SVR and PLSR models.

## 2.5 Conclusions

In this paper, we explored the potential for reliable prediction of sorghum biomass using multi-temporal hyperspectral and LiDAR data acquired by sensors mounted on UAV platforms. We developed prediction models using three nonlinear regression models for nine experiments conducted in the 2017 and 2018 growing seasons at the Agronomy Center for Research and Education (ACRE) at Purdue University. Experiments included multiple sorghum varieties with different sample sizes, providing an opportunity for multiple statistical tests and models. Based on the experiments conducted in this study, nitrogen and photosynthesis related features extracted from hyperspectral data and geometric based features derived from the LiDAR data provided reliable and accurate prediction of biomass. The 750-1100 nm range of the spectrum provided the most relevant information for biomass prediction.

Both data source and regression method are important factors for a reliable prediction; however, the ANOVA results show that the data source was more important with 69% significance, versus 28% significance for the regression method. The number of samples in training set for the prediction is an important factor for determining the accuracy of the predictions. Generally, the PLSR method provided more accurate prediction models when the number of samples in training was limited. With increasing samples, the rate of increase in the accuracy of the SVR models was higher than PLSR.

We also evaluated the prediction models with respect to the time of the RS data acquisition and the time of harvest. The end-of-season biomass predictions were more reliable and accurate than the mid-season predictions, as more varieties in the field were at the same stage of growth. With respect to the remote sensing data, the best results were obtained using the RS data from the whole season; however, the models developed using mid-season data (DAS of ~60 to 80) were also able to provide comparably accurate results, which were useful for early screening of varieties.

### **3. ACCURATE PHENOTYPING USING UAV-BASED REMOTE SENSING DATA, GENOMIC MARKERS, AND WEATHER DATA**

#### **3.1 Abstract**

High-throughput phenotyping using remote sensing (RS) platforms as well as genomic selection have been widely investigated in recent years as potential approaches for reducing the need for the expensive, time-consuming traditional field phenotyping for breeding programs. In this paper, the potential for accurate and reliable sorghum biomass prediction using visible/near infrared (VNIR) hyperspectral data and light detection and ranging (LiDAR) data acquired by sensors mounted on UAV platforms combined with the genomic information and weather data are investigated for prediction of the end of season sorghum biomass yield. Predictive models are developed using four regression methods based on experiments conducted in the 2017, 2018, and 2019 growing seasons at the Agronomy Center for Research and Education (ACRE) at Purdue University. Models based on one year and two years of data are used to predict the biomass yield for the future years. The results show the high potential of the models for biomass and biomass rank predictions. While models developed using one year of data are able to predict biomass rank using two years of data resulted in more accurate models, especially when RS data were included. Also, the possibility of developing predictive models using the RS data collected until mid-season, rather than the full season, is investigated. The results show that using the RS data until 60 DAS in the models can predict the rank of biomass with  $R^2$  values of around 0.65-0.70. This not only reduces the time required for phenotyping by avoiding the manual sampling process, but also decreases the time and the cost of the RS data collections and the associated challenges of time-consuming processing and analysis of large data sets, and particularly for hyperspectral imaging data.

#### **3.2 Introduction**

Plant breeding pipelines typically start with generating inbred and hybrid lines, planting them in multiple locations, phenotyping various traits of each line, and finally, selecting the “best” genotypes with respect to the trait of interest such as yield, resistance to pest pressures and climate-related abiotic stresses (Bhat, 2019). Researchers have widely investigated using new

methods/technologies to increase breeding efficiency by reducing the time and cost of commercializing new products (Cobb et al., 2019). This mainly includes enhancements in phenotyping, as it provides the critical information required for the accurate genotype selection (Desta & Ortiz, 2014) while it is also currently the most expensive component of a plant breeding operations (Cobb et al., 2019). Thus, reducing the need for traditional field phenotyping, which can limit the ability to analyze the genetics of quantitative traits, particularly those related to yield and stress tolerance (Araus & Cairns, 2014), has become an objective of many studies. This can be achieved by estimating the phenotypes using predictive models developed from various sources of data, mainly remote sensing (RS) and genomic information (i.e. marker data).

Recent advances in high throughput genotyping platforms (3rd generation platforms) have made it possible to obtain genomic information for thousands of lines with limited cost in both time and expense, that are useful for a variety of purposes, including genome-wide association studies, quantitative trait loci (QTL) mapping, and genomic selection (GS) (Nakaya & Isobe, 2012). In GS, a training population is used to estimate the effects of markers based on prior phenotypic and marker data, and marker effects are then used to predict the genomics based breeding value of individuals in the prediction population which has not been phenotyped (Meuwissen et al., 2001). GS can greatly reduce the length of the breeding cycle, and avoid the loss of genetic variability in comparison to traditional selection based on only phenotypic data, without sacrificing selection gains (Shamshad, Mohd and Sharma, 2018). However, there are challenges to the use of GS that must be addressed, such as its low accuracy for low heritability traits, including yield (Nakaya & Isobe, 2012). Traditionally, prediction accuracy is defined as the correlation between the predicted and the measured phenotype for the testing data set (Rabier et al., 2016). Several methods, including multi-trait and trait assisted GS (Fernandes et al., 2018) have been developed in recent years to improve the accuracy of predictions. Despite the improvements observed in the prediction accuracy, significant advances are still required for plant breeding programs to fully exploit the potential of GS for being included in breeding pipelines (Heffner et al., 2009; Shamshad, Mohd and Sharma, 2018).

Some researchers have adopted remote sensing based measurements to reduce the need for traditional phenotyping by developing high-throughput phenotyping platforms that phenotype large numbers of lines rapidly and relatively accurately (Maes & Steppe, 2019; Tattaris et al., 2016). Unmanned Aerial Vehicle (UAV) based RS has been widely utilized in recent years due to

its capability to acquire both high temporal and high spatial resolution data required for high throughput phenotyping over agricultural fields. Researchers have demonstrated that predictive models based on RS-derived features and field data can be used to estimate phenotypes such as yield (Foster et al., 2017; Thorp et al., 2017), LAI (X. Li et al., 2014; Liang et al., 2015; Nguy-Robertson et al., 2014), and biomass (Jiating Li et al., 2018; Masjedi et al., 2019; Z. Zhang et al., 2017).

If the predictive models are acceptable, the estimates can be injected into the breeding pipeline instead of manual phenotypic data, which decreases the time and cost of the breeding programs. The accuracy of the predictive models depends on the quantity of samples (population size) (Combs & Bernardo, 2013), a key limitation of the current phenotyping processes. Also, developing predictive models for end-of-season traits such as biomass and yield is not possible during the growing season, except retrospectively, as the sample data are typically only available after the harvest is completed, although, for some breeding programs, it is important to predict the final biomass or yield of the planted varieties as soon as possible, so that genotype selection can be accomplished in a shorter period of time, and thus, less expensively.

To overcome this drawback, we explore the development of multi-year predictive models for which the data from earlier years of the breeding cycle are used for training models, and then predict traits for experiments in subsequent year(s). Further, more observations are available as the sample data from multiple years are included in a model. In this study, the potential of biomass prediction of varieties of sorghum using hyperspectral and LiDAR data acquired by sensors mounted on UAV platforms, as well as genomic markers, is investigated, with the goal of achieving a faster and cheaper breeding pipeline. Sorghum is a multipurpose crop that has attracted attention in recent years, both for its broad-based potential usage for human consumption and animal feed, as well as biofuel production (Stanton et al., 2017) and its drought and heat tolerance (Spindel et al., 2018). Having a short growth cycle, low input requirements, mechanized management, and the ability to produce reasonable quantities of ethanol make sorghum an efficient biofuel crop (Fernandes et al., 2018). Many researchers are developing enhanced genotypes that can produce more energy-rich plant material (biomass) (Ogbaga et al., 2019).

In this study, we focus on two main objectives; 1) developing predictive models using the RS data collected multiple times during a season in previous years, and 2) investigating the possibility of developing the same models using the RS data collected until mid-season rather than

the full season. This will not only reduce the time of the phenotyping by avoiding the manual sampling process, but also decrease the time and the cost of the RS data collection, as well as the associated challenges of processing, managing, and analysis of large data sets, especially for hyperspectral imaging. Multiple challenges must be considered for developing multi-year predictive models: 1) different management practices (e.g. weather dependent planting dates), 2) changes in protocols for the acquisition of biomass data throughout the multi-year experiments for which data are available, 3) changes in protocols for acquiring remote sensing data over the three years of experimental period, 4) diverse sources and characteristics of input data.

As the phenotype is the result of genotype by environment (GxE) interactions, environmental data are incorporated in our multi-year prediction models. The experiments in this study were conducted at the same geographic location; thus, we can assume some environmental conditions such as soil information are essentially the same for the experiments conducted in different years. However, weather information can be highly variable from one year to another. Local solar radiation, temperature, and precipitation are included in the models as they can impact the plant growth strongly.

### **3.3 Materials and Methods**

#### **3.3.1 Field Design**

The field experiments were conducted in the 2017, 2018, and 2019 growing seasons, each year in 2.8 ha sorghum breeding trials located at the Purdue University Agronomy Center for Research and Education (ACRE), West Lafayette, Indiana. All trials were arranged in a randomized complete block design and planted at 220,000 plants per ha. In 2017, the field experiments consisted of two distinct trials in two panels; the inbred calibration (InCal) panel and the sorghum biodiversity (SbDiv) panel. The SbDiv panel consisted of 840 inbred lines replicated twice in four row plots. The inbred calibration panel included 60 inbred lines, a subset of the SbDiv panel, which encompasses the genetic diversity of the SbDiv panel, and were planted in four row plots. The InCal experiment was also repeated in the 2018 growing season in a different field at ACRE, with 54 inbred lines. In 2018, two other experiments were also conducted in the same field as 2017 experiments, sorghum biodiversity test cross (SbDivTc) and inbred calibration test cross (InCalTc), both planted in four row plots. The SbDivTc and InCalTc are hybrids derived from the

Sbdiv and InCal experiments with a common maternal parent. The InCalTc included 54 hybrid lines, also a subset of the SbDivTc panel. The SbDivTc included 630 hybrids replicated twice in four row plots in the 2018 growing season. In 2019, the SbDivTc was repeated with the same number of genotypes. An additional panel was added in 2019: sorghum biodiversity test cross calibration (SbDivTcCal), which included 72 hybrid lines, a subset of the SbDivTc panel that is representative its genetic diversity was planted in 10 row plots. Details of experiment trials are provided in Table 3.1. Figure 3.1 shows the RGB bands of the hyperspectral data and LiDAR point cloud data of two panels of the experimental field in the 2019 growing season

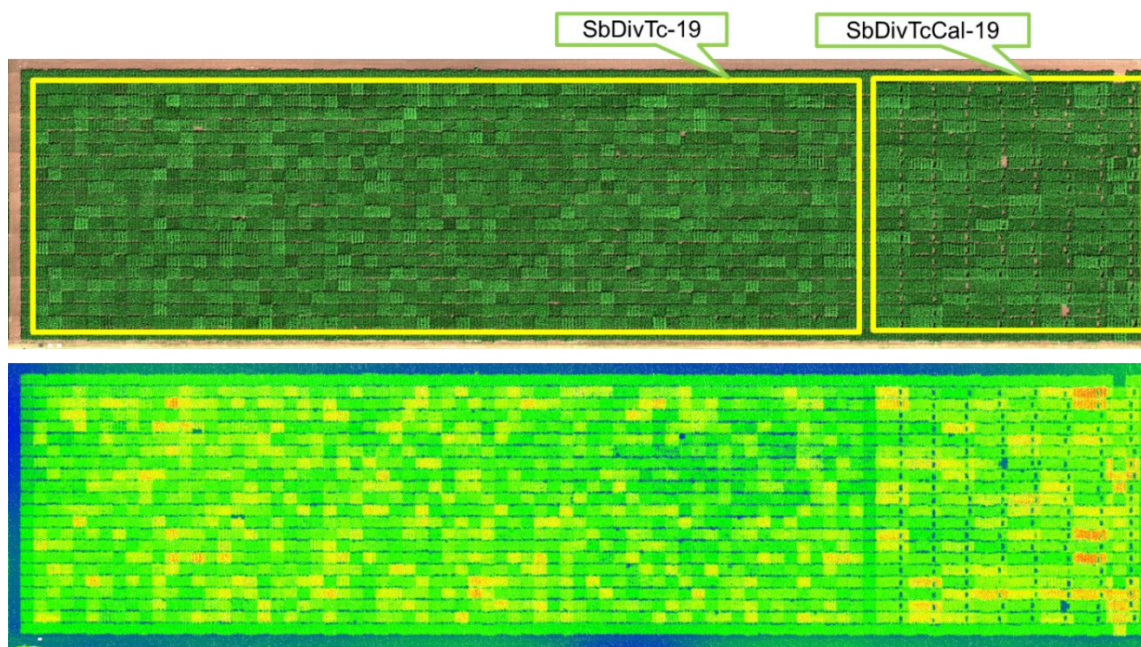


Figure 3.1. RGB bands of the hyperspectral data and the LiDAR point cloud of the SbDivTcCal and SbDivTc panels on August 2<sup>nd</sup>, 2019.

### 3.3.2 Data Description

#### *Ground Sampling of Biomass*

For all the experiments, the biomass data were collected by harvesting rows two and three of the four-row or 10-row plots at the end of each growing season. The biomass per square meter was obtained by dividing the measured fresh weight of the plants by the area of the row segment within the plot of two harvested rows. Figure 3.2 shows an area of the InCal-17 experiment that included two plots and rows two and three of each plot extracted based on RGB imagery using the

methods explained in (Ribera, J., Chen, Y., Boomsma, C., & Delp, 2017). Note that the end of each row segment is trimmed by 40 cm to eliminate the border effect as it was shown that it can improve the heritability of the remote sensing measurements {Neal's paper}. Figure 3.3 is a boxplot of the fresh biomass data measured for all the experiments in the three growing seasons.

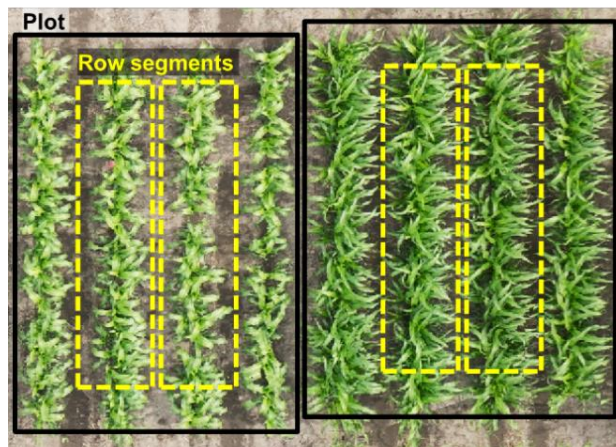


Figure 3.2. An example area of the InCal-17 experiment. Two plots (black rectangles) and rows two and three (yellow rectangles) are highlighted.

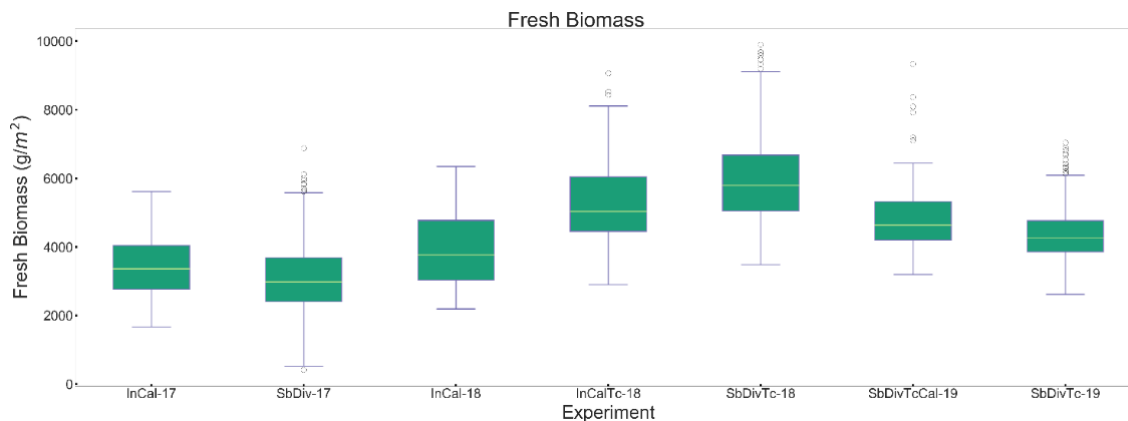


Figure 3.3. Box plot for the fresh biomass data for all the experiments conducted in the 2017, 2018, and 2019 growing seasons.

### ***Remote Sensing Data***

This study includes RGB, hyperspectral, and LiDAR remote sensing data collected by custom designed UAV platforms. All remote sensing data acquisition platforms were flown with Global Navigation Satellite System/Inertial Navigation System (GNSS/INS) units for direct georeferencing. RGB data for this study were collected using a Sony Alpha ILCE-7R RGB camera

with a Sony 35mm lens. The Sony Alpha 7R features a full-frame 36.4MP sensor delivering high-resolution UAV-based aerial imagery. LiDAR data were collected with a Velodyne VLP-16 3D LiDAR sensor operating in the strongest return mode and 600 rotation per minute (RPM). The VLP-16 is a 16 beam LiDAR unit with a 360 degree horizontal FOV, and a maximum range of 100 m. Both the RGB camera and the VLP-16 sensor are mounted on a DJI Matrice 600 Pro (M600P) platform. Spatial and temporal system calibration for the datasets used in this study were conducted using the approaches described in (Ravi et al., 2018) and (LaForest et al., 2019), respectively. Also, the georeferenced orthomosaics were generated using the structure from motion strategies introduced in (Hasheminasab et al., 2020; He et al., 2018).

Visible Near Infrared (VNIR) hyperspectral data were collected with two Headwall Photonics push-broom scanners. The VNIR data were acquired by a Nano-Hyperspec imaging sensor in 272 spectral bands at 2.2 nm/band from 400 nm to 1000 nm with 640 spatial channels at 7.4  $\mu\text{m}/\text{pixel}$ . In 2017, the sensor was flown at an altitude of 60 m with a 12 mm Schneider lens, resulting in a Ground Sampling Distance (GSD) of  $\sim 4$  cm. An 8 mm lens was used in 2018 and 2019, and the flying height was  $\sim 40$  m to maintain the GSD at  $\sim 4$  cm. In 2019, the RGB, LiDAR, and VNIR sensors were integrated and flown together on a single UAV platform (see Figure 3.4). A rigorous boresight calibration process was performed (Habib et al., 2018), yielding aligned RGB, LiDAR, and VNIR data. All the hyperspectral data were converted to reflectance using the empirical line method to relate the spectra collected from the UAV to data acquired by an SVC 1024i field spectrometer over the calibration targets placed in the field for each acquisition. Table 3.1 lists the data acquired by the sensors in 2017, 2018, and 2019 growing seasons.

Table 3.1. Experimental design for the 2017, 2018, and 2019 growing seasons.

Year	Trial	Genotype	# of plots	# of varieties	Sowing Date	Harvest Date	LiDAR Dates	VNIR Dates
2017	InCal	inbred	120	60	May 15	Sep 27	June 16, 21, and 27; July 5, 7, 11, 14, 17, and 25; Aug. 2, 8, 23, and 30	June 21, 27 <sup>*1</sup> , and 28 <sup>*2</sup> ; July 4, 12 <sup>*2</sup> , 18, 25, and 30 <sup>*1</sup> ; Aug. 8, 14 <sup>*1</sup> , 23, and 30 <sup>*2</sup> ; Sep. 10, 15 <sup>*2</sup> , 24 <sup>*1</sup> , and 30 <sup>*1</sup>
	SbDiv	inbred	1800	840	May 25	Oct 15		
2018	InCal	inbred	108	54	May 1	Aug 9	May 22 and 29; June 4, 11, 20, and 27; July 2, 11, 18, and 23; Aug. 1 and 6	June 4, 8, 14 and 29; July 3, 6, 10 <sup>*3</sup> , 11, 18, and 25; Aug. 2 and 9 <sup>*4</sup>
	InCalTc	hybrid	108	54	May 8	Aug 6		
	SbDTc	hybrid	1200	630	May 8	Aug 2		
2019	SbDTcCal	hybrid	144	72	June 4	Sep 5	June 14, 18, and 26; July 1, 12, and 23; Aug. 2, 10, and 24; Sep. 5	July 12, 19, and 23; Aug. 2, 10, and 24; Sep. 5
	SbDTc	hybrid	1200	630	June 4	Sep 5		

<sup>\*1</sup> only SbDiv-17, <sup>\*2</sup> only InCal-17, <sup>\*3</sup> only SbDTc-18, <sup>\*4</sup> only InCal-18

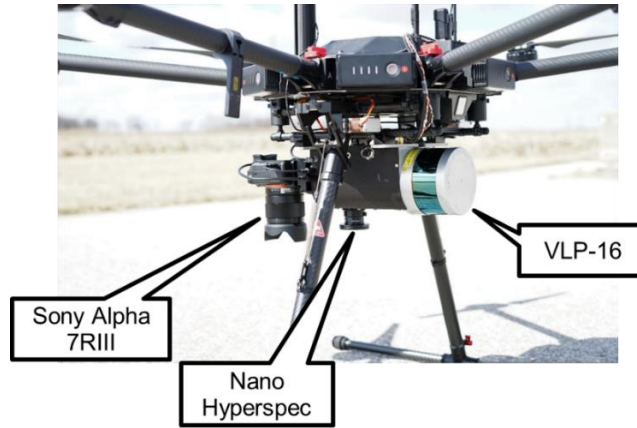


Figure 3.4. M600 platform with the RGB, LiDAR, and VNIR sensors flown in the 2019 growing season. A GNSS/INS unit (APX-15) was also mounted on the platform providing accurate position and orientation information required for geo-referencing of the captured data.

### *RS Feature Extraction*

From each LiDAR point cloud and for each plot, 16 features including 30<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup> percentiles of height, canopy cover, and other statistical features were extracted. From hyperspectral data, first and second derivative features, vegetation indices, and the area under the average of the spectral curves for each variety on a given date were extracted and used as input features for further analysis. Additional information about feature extraction of the data is available in our previous studies (Masjedi et al., 2019, 2018; Z. Zhang et al., 2017).

### *Genotype Data*

More than 80,000 markers are available for the inbred and hybrid varieties planted in the three growing seasons.

### *Genomic Feature Extraction*

The goal is to use the genetic information individually and in conjunction with other features extracted from remote sensing data sources in the prediction models. High dimensional data can be problematic for regression algorithms; thus, the number of features need to be reduced either with feature selection or extraction methods. In this study, both PCA and wavelet transformations are investigated for feature extraction from the genomic markers. PCA

decomposes variance into orthogonal components, while the wavelet transform decomposes data via time-scale analysis (Barclay et al., 1997). Transforming the signal using wavelets results in two decomposed signals, one focused on low frequency information and the other the high frequency information. The resulting low frequency signal can be considered as a compressed version of the original signal. The wavelet transformation can be applied in multiple levels, each of which reduces the length by a factor of two. Figure 3.5 shows the wavelet transformation of the genetic markers for the inbred and hybrid genotypes planted in the SbDiv and SbDivTc panels, respectively. Each row of the images in Figure 3.5 is associated with the genetic markers after the wavelet transformation in 6, 10, and 10 levels which provided 1252, 313, and 72 features, respectively.

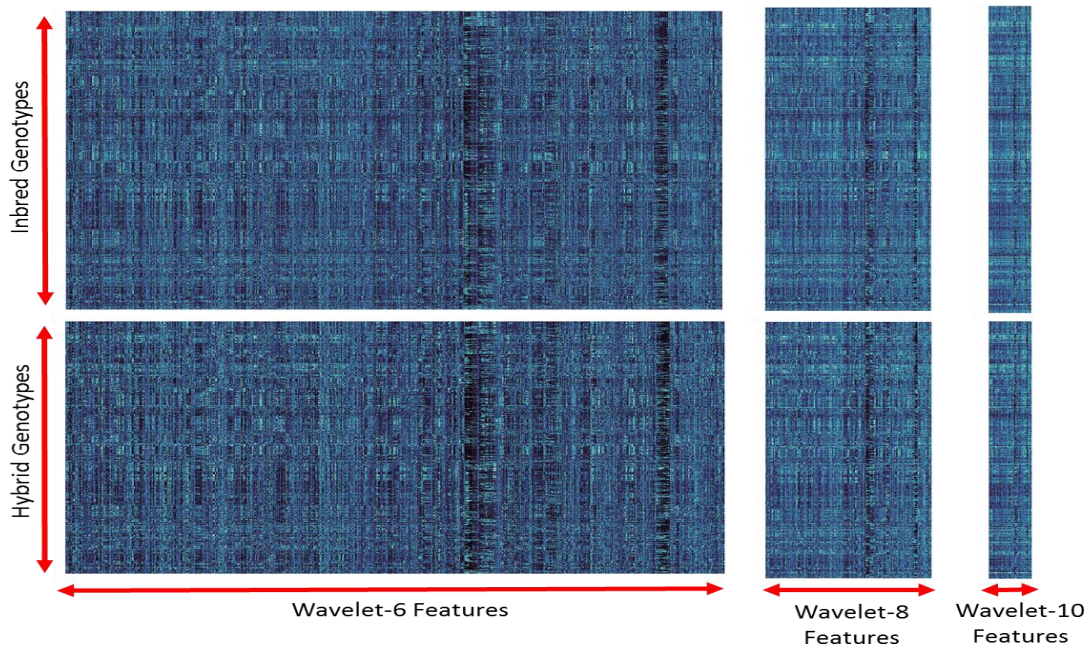


Figure 3.5. Feature extraction using wavelet transform in 6, 8, and 10 levels. The X and Y axes are associated with the number of features and number of genotypes, respectively.

Principal component analysis (PCA) has been widely utilized for analysis, visualization, and interpretation of the high dimensional genomic data as it can provide a concise picture of the whole data set in a few components (Lee et al., 2010; Scholz & Selbig, 2007). Principal components for the hybrid and inbred genotypes are shown in Figure 3.6, which illustrates how the first three principal components can be used for visualization of high dimensional genomic

markers. Figure 3.6 also shows where the genotypes planted in the calibration experiments (InCal, InCalTc, and SbDivTcCal) were located in the PC1-PC2 plot with respect to the hybrid and inbred genotypes. The plot demonstrates that the calibration panels were included the genotypes that encompass the genetic diversity of all the hybrid or inbred genotypes.

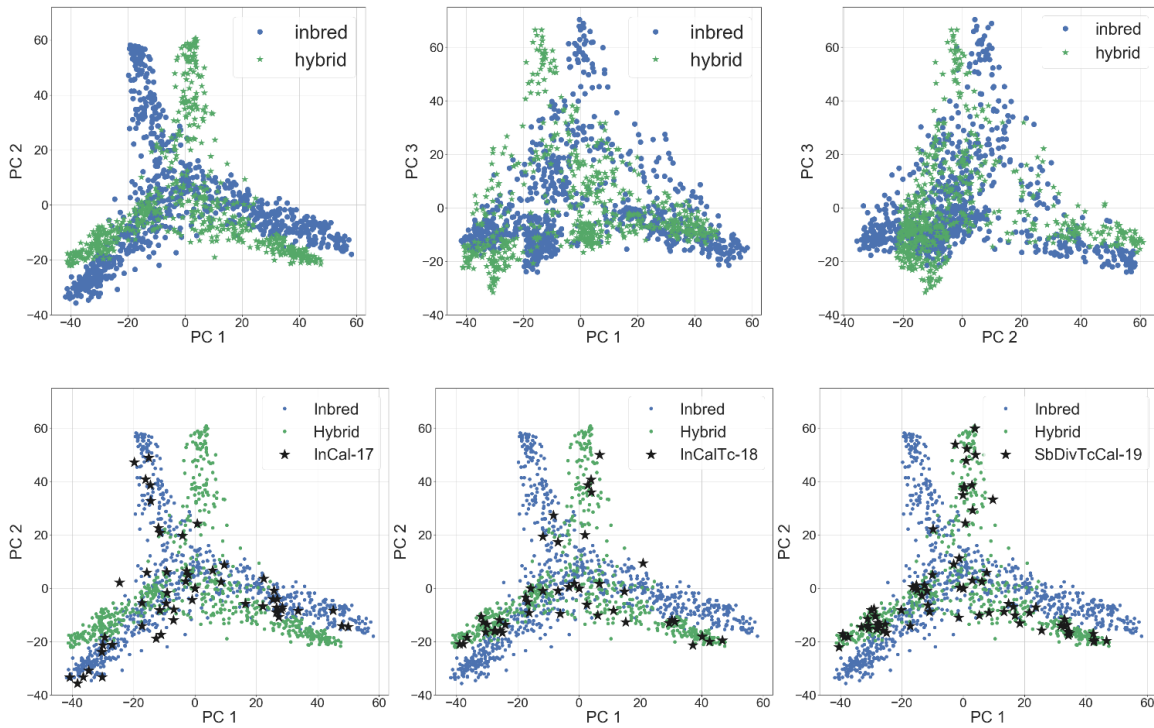


Figure 3.6. Feature extraction using PCA. The top row contains plots of combinations of the first 3 PCs for the combined hybrid and combined inbred genotypes. The plots on the bottom row show the first two principal components of the InCal-17, InCalTc-18, and SbDtcCal-19 genotypes are located with respect to those of the combined hybrid and inbred genotypes.

### *Features-Biomass Correlation*

To evaluate the prediction potential of the genomic markers as well as wavelet-based features for biomass estimation, linear regression-based models were developed using each marker/feature for three experiments conducted in the 2017-2019 growing seasons. Figure 3.7 shows the  $R^2$  values of the resulting models for each marker as well as features extracted by wavelet transforms. The markers in the range of 60,000 for both SbdivTc-18 and SbdivTc-19 provided the highest  $R^2$ , with the values 0.14 and 0.12, respectively. The corresponding wavelet-based features (features in the range of 930-960) provided the highest  $R^2$  using wavelet features with the values

0.10 and 0.08 for SbdivTc-18 and SbdivTc-19, respectively. Similar  $R^2$  values obtained with wavelets suggest that the wavelet features can be as informative as the original markers; however, as the number of features was much lower, they might provide more accurate predictions than using all the markers. We further evaluate these features for genomic-based biomass prediction in the subsequent sections.

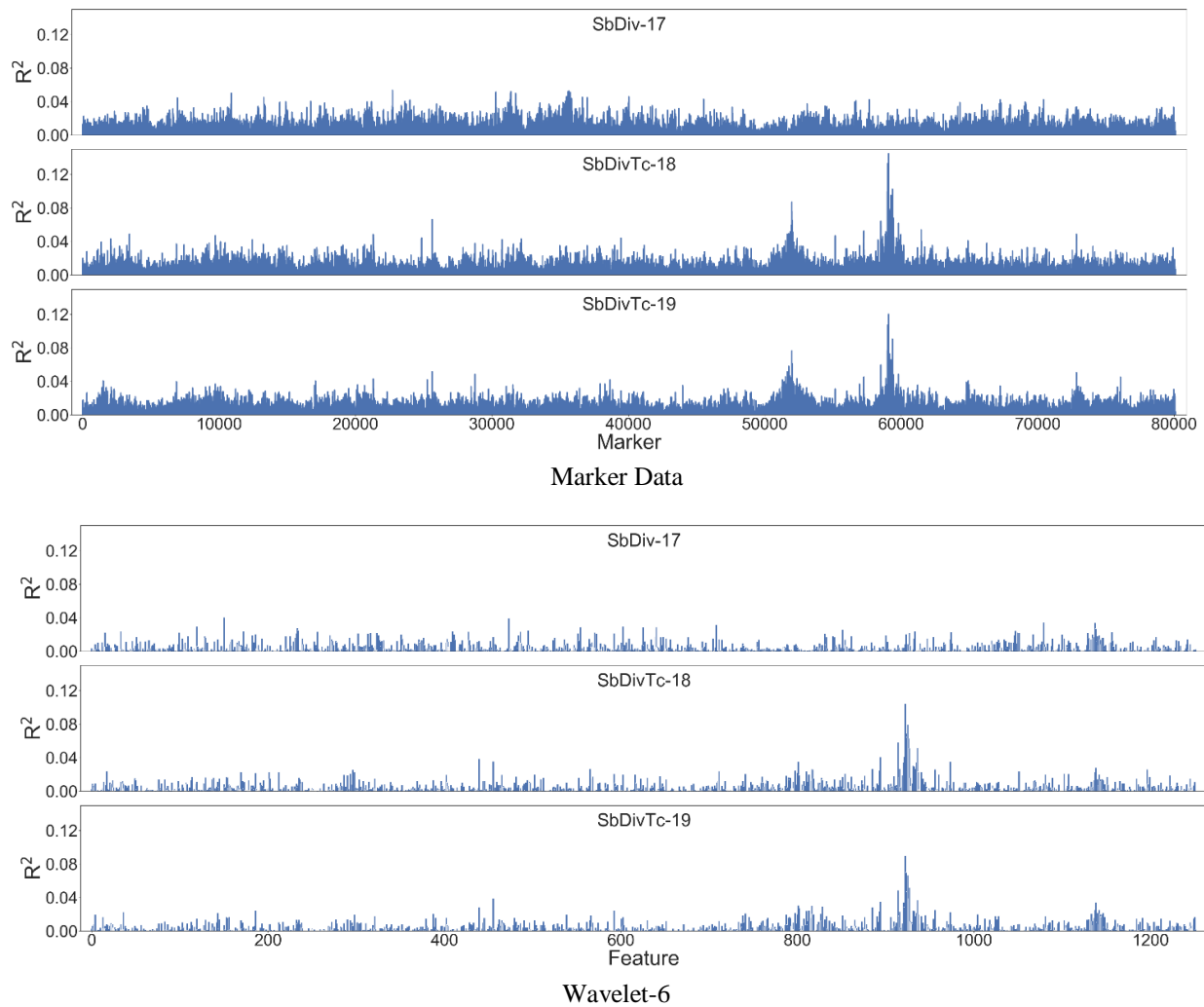


Figure 3.7.  $R^2$  values of the linear regression-based models developed for fresh biomass using genomic markers and wavelet-6 features for three experiments, SbDiv17, SbDtc18, and SbDtc19 conducted in 2017-2019.

### ***Weather Data***

As noted, weather strongly impacts crop growth and development, and should be considered in developing multi-year prediction models. For the experiments conducted at the same

location but different years, precipitation, temperature, and radiation, as well as planting date can be significantly different for different years. Figure 3.8 shows precipitation, GDDs, and solar radiation for the three years that experiments were conducted at ACRE from May 1<sup>st</sup> until October 1<sup>st</sup>. In 2017, the annual cumulative precipitation was higher, although early season precipitation was higher in 2019. Cumulative radiation received compared to the other years was also higher, while the average temperature was higher in 2018.

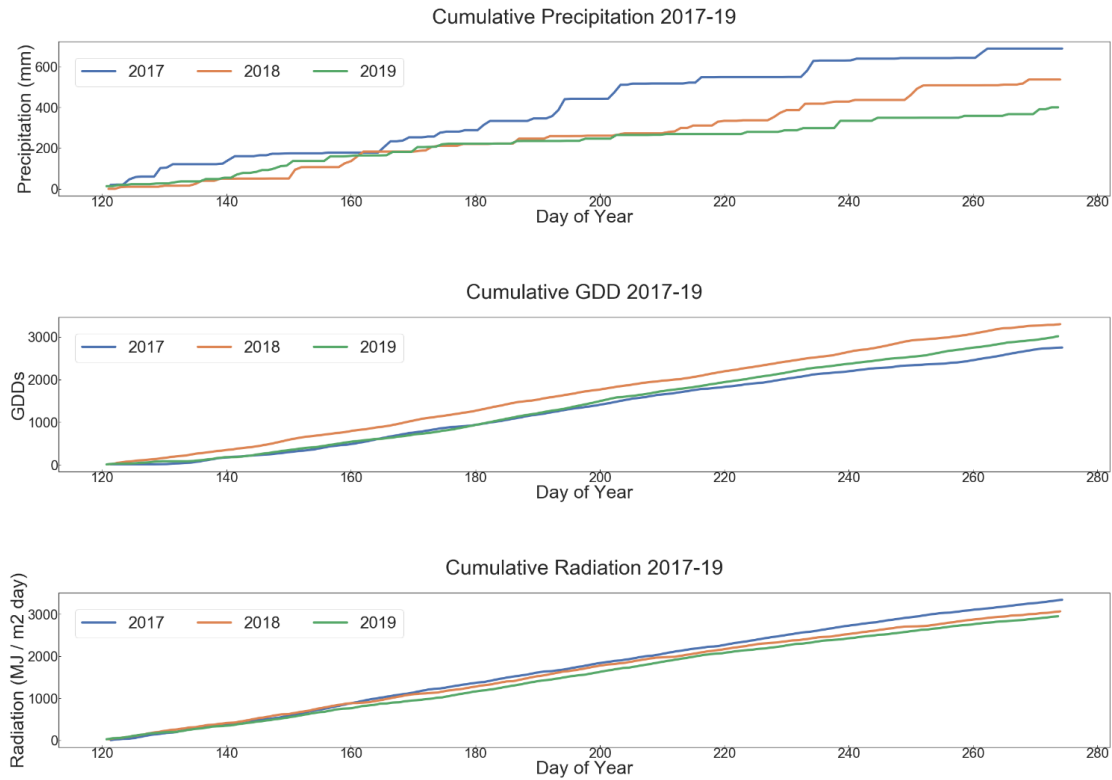


Figure 3.8. Weather conditions in the 2017-2019 growing seasons at ACRE.

From the weather data shown in Figure 3.8, nine features were extracted for each experiment as follows. The weather data for each season were split into two groups, prior and post flowering, by considering 60 days after sowing (DAS) as the average flowering time for sorghum at the location experiments were conducted. The average values of daily precipitation, growing degree days (GDD), and solar radiation for each group were calculated and considered as the input weather features for the predictive models. Also, as the planting and harvest dates varied in the three years, the cumulative precipitation, GDDs, and solar radiation for the whole season were considered as the input features. Figure 3.9 shows the relationship between the cumulative weather

data and the biomass data for the InCal experiments in the 2017 and 2018, SbDivTc experiments in the 2018 and 2019 growing seasons. Generally, the fresh biomass observed in 2018 was higher than in 2017, while the ranges of biomass in 2018 and 2019 were similar. Note that Figure 3.9 only shows the end of season biomass values with respect to cumulative weather data; however, there are many other weather related parameters that can impact the biomass yield such as when the precipitation occurs, amount of radiation received at different growth stages, as well as the interaction between the radiation and temperature.

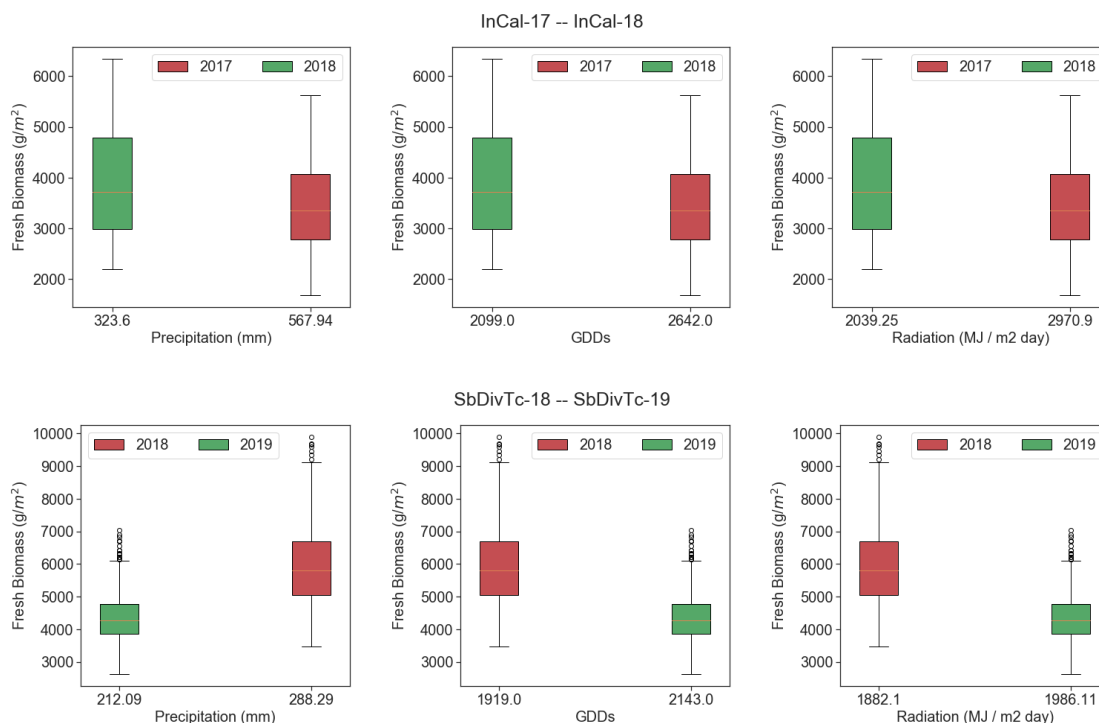


Figure 3.9. Weather-biomass relationship in the four experiments, and two pairs of similar experiments conducted in two years: InCal-17 vs. InCal-18 and SbDTc-18 vs. SbDTc-19.

### 3.4 Results and Discussion

#### 3.4.1 Genomic-based Predictive Models

Genomic-based features, including wavelet based features in 2, 4, 6, 8, 10, 12, and 14 levels, as well as 50, 100, and 200 principal components, and original markers, were used as the input features for biomass prediction using four regression-based models. Results are presented for Ridge Regression (Ridge), Partial Least Squares Regression (PLSR), Support Vector Regression

(SVR) with a Radial Basis Function (RBF) kernel, and Random Forests (RF), which are commonly utilized to build predictive models in genomic selection (Azodi et al., 2019) or phenotype prediction via remote sensing (Masjedi et al., 2019). A grid search method was used to select the best hyperparameters for each model. Table 3.2 lists the candidate parameters that were tested for each method in the grid search process. The coefficient of determination ( $R^2$ ) in this study is calculated using the following equation as suggested in (Kvålseth, 1985):

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

where  $y_i$  is the observed value,  $\hat{y}_i$  is predicted value,  $\bar{y}$  is mean value, and  $i$  an integer from 1 to  $n$  (number of samples in test data set).

Table 3.2. Grid search parameters for regression methods

Algorithm	Hyperparameter	Values Tested
Ridge Regression	alfa	0.5, 1, 10, 100
PLSR	Number of components	2, 3, 5, 10, 15, 20
SVR (rbf)	$C$	10, 100, 1000, $n^*$
	gamma	1/ $n$ , 0.0001, 0.001, 0.01, 0.1
Random Forest (RF)	Max tree depth	5, 10, 100
	Min sample split	2, 10
	Number of trees	50, 100, 500

\* $n$ : number of features

Figure 3.10 shows the average accuracy of the predictive models for three experiments; SbDiv-17, SbDivTc-18, and SvDivTc-19 where 250 genotypes were used for training and the rest (380 to 510) for evaluation. This was repeated five times, and each time the training samples were selected randomly. From Figure 3.10, the various regression methods provided similar results in terms of accuracy, with correlations of 0.35 to 0.4 between the predicted biomass and the reference data. The predictions for SbDivTc-18 and 19 were slightly more accurate than the predictions for the SbDiv-17 for all the methods. The models based on wavelet features up to 10 levels generally provided predictions with similar accuracies as using all the markers, while the number of features was much smaller. RF resulted in less accurate predictions using all the markers compared to using the wavelet based features, especially for the SbDivTc-18 and 19 experiments, where a smaller

quantity of sample data was available. This is mainly because RF requires optimization of many parameters in training, and as the number of features increases, the number of parameters also increases; thus, more sample data would be required to achieve the same accuracy as the other methods. In the next sections, we evaluate the impact of using these features in multi-year predictive models.

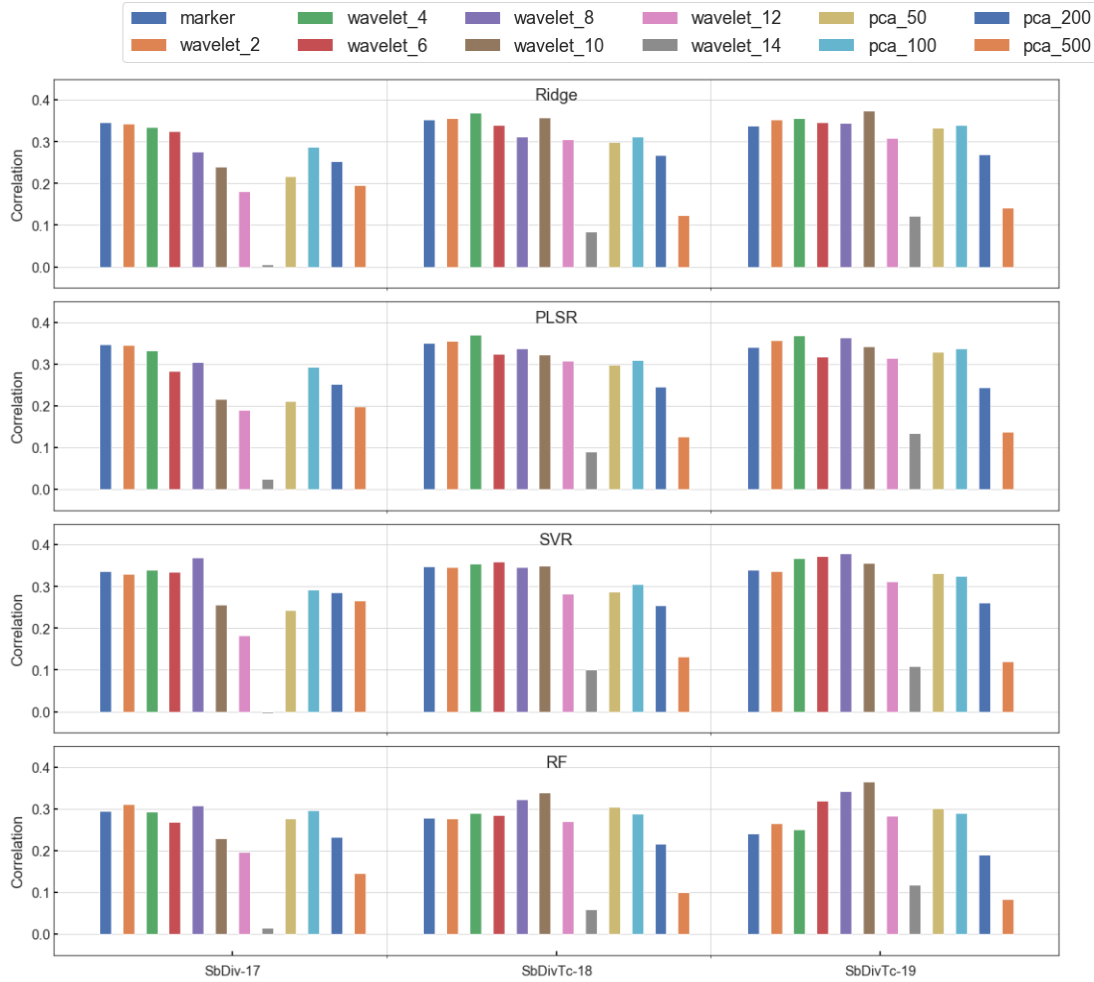


Figure 3.10.  $R^2$  of predictions using derived genomic features in Ridge, PLSR, SVR, and RF regression models.

### 3.4.2 Comprehensive Predictive Models

#### *Models trained on one year of data*

Different management practices, such as different dates for planting and harvesting the crops, in various years is a significant challenge for multi-year biomass prediction, including

different ranges for the biomass data in each year (see Figure 3.3). Also, the RS data were collected each year on different dates during the growing season. The regression models we use in this study require the same number of features for all the samples from various years, so the same number of RS datasets from each year should be included for the predictive models. Also, the dates the data were collected should be as similar as possible. Thus, for each experiment, four data sets collected on approximately 40, 60, 80, and 95 DAS were selected from the available RS data sets, and features were extracted from LiDAR and hyperspectral data. Also, nine features from the weather data were extracted as noted in Section 2.2.4 for each experiment. Based on the results obtained in the last section, the features extracted by the wavelet transform in 6 levels were chosen as the genomic-based input features in the multi-year predictive models in both this and the following sections.

Figure 3.11 shows the biomass prediction results for three experiments in the 2018 and 2019 growing seasons using the SVR models trained on the 2017 experimental data (both InCal-17 and SbDiv-17 experiments). The biomass in 2017 was harvested on ~120-150 DAS, which was ~60 days later than the harvest in 2018 (93 DAS) and 2019 (95 DAS). As plant water content is lower by the end of the season, it is lower if the harvest is later, and as a result, fresh biomass is lower. This is one of the reasons why the end-of-season biomass values were lower in 2017 compared to 2018 and 2019. Also, the genotypes planted in 2017 were all the inbred genotypes, while in the 2018 and 2019 hybrid genotypes, which usually provide higher biomass values, were planted. As expected, models trained on the 2017 experiments underestimated the biomass for the experiments in 2018 and 2019. The estimated biomass in this figure shows a bias with the reference biomass, but the relative relationship between the genotypes in terms of their biomass seems to be accurate. As the goal of breeding programs is to select the genotypes with the highest return with respect to the trait of interest (biomass in our study), we calculated the rank of the genotypes in terms of biomass production. The higher the rank, the higher the biomass in our models. Figure 3.11 also shows the rank of the genotypes estimated from the biomass predictive models with respect to the rank calculated from the reference biomass data. It shows that even though the  $R^2$  values for biomass predictions are negative relative to a 1:1 prediction, the  $R^2$  values for rank predictions are 0.43 to 0.84 for the 2018 and 2019 experiments without using any reference data from those years for training the predictive models. Also, the RMSEs of the biomass rank predictions are 5% - 11%.

Figure 3.12 shows the predictions of the SVR models trained on the 2018 experiments (all three experiments conducted in 2018) and used to predict biomass for the 2019 experiments. The range of biomass values in both experiments in 2019 is smaller than the range of predicted biomass by the predictive models trained on the 2018 experimental data, indicating that more variability existed in the 2018 data. The predictions for SbDivTc-19 are however more accurate than the prediction from models trained on the 2017 experiments shown in Figure 3.11 which is expected as the harvest in both 2018 and 2019 occurred at ~95 DAS. Also, the experiments in both years included the same genotypes, which is important, especially when the genomic-based features are used in the predictions. Figure 3.12 also shows the rank of the genotypes calculated from the predicted biomass and the reference biomass. The  $R^2$  values of the predicted rank, similar to the results presented in Figure 3.11, were higher than the  $R^2$  values of the predicted biomass. This indicates that the same genotypes performed similarly in both years; most of the genotypes that provided high biomass were still able to provide higher biomass in 2019. The RMSEs of biomass rank for both experiments were ~7%.

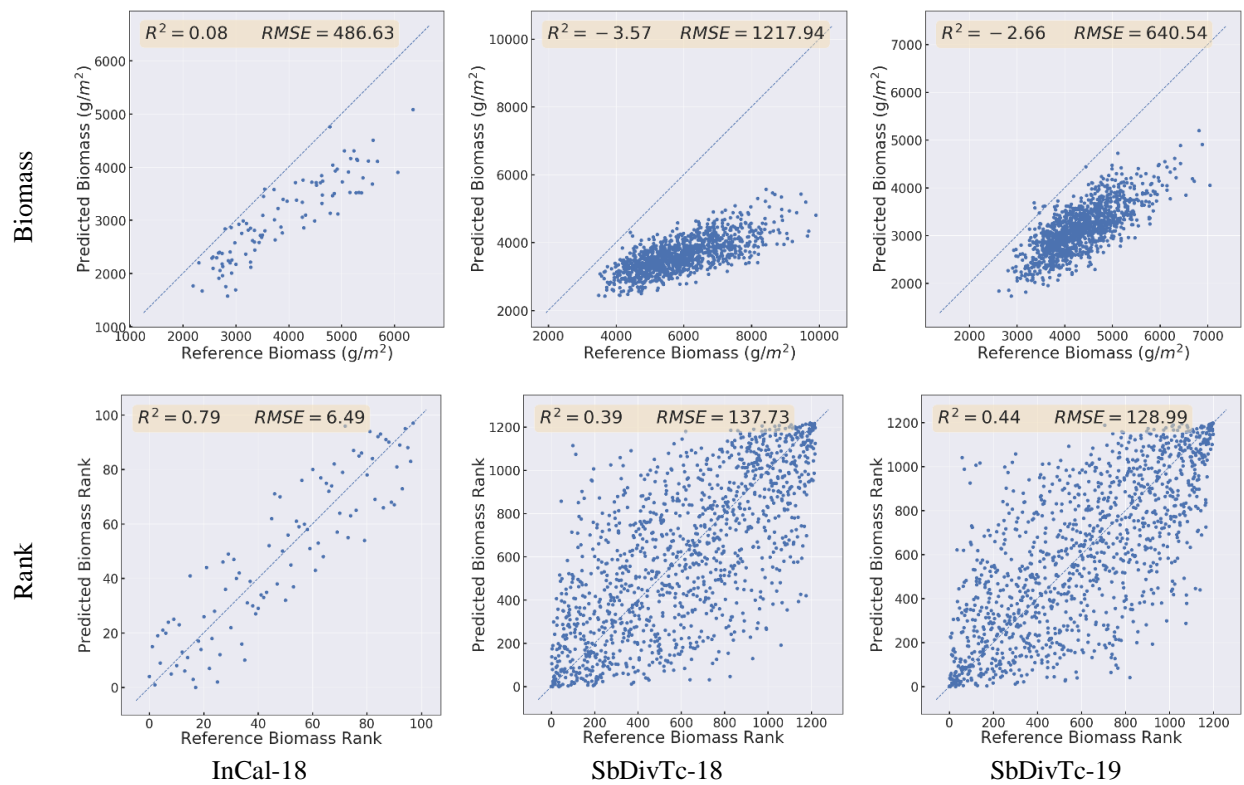


Figure 3.11. Results of biomass and rank prediction for example experiments in 2018 and 2019 from SVR models trained on the 2017 experiments using RS, genotype, and weather data.

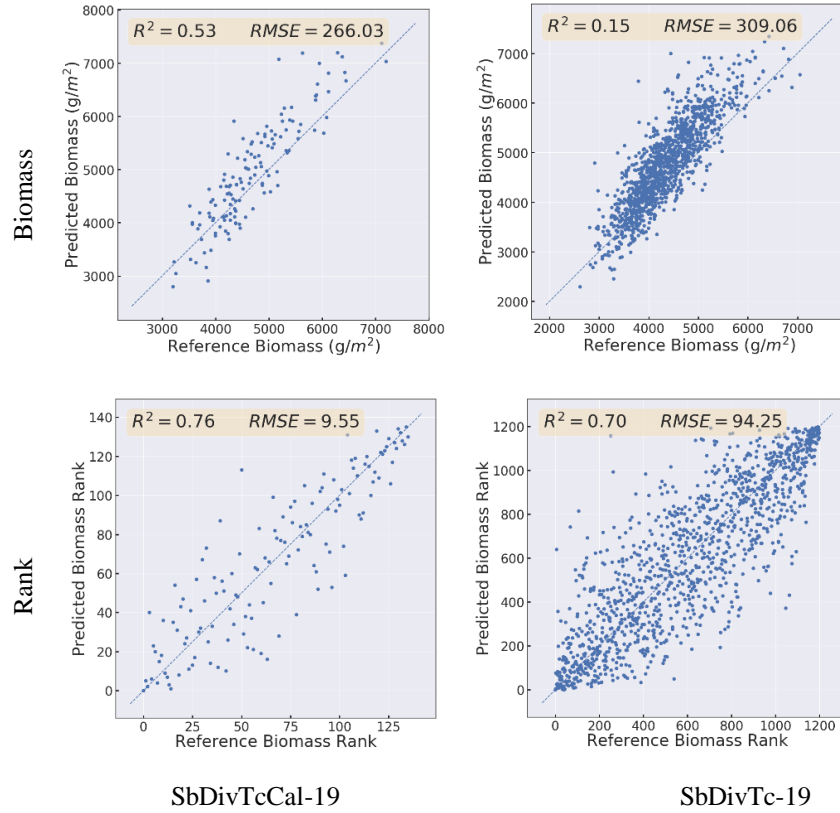


Figure 3.12. Results of biomass and rank prediction for example experiments in 2019 from SVR models trained on the 2018 experiments using RS, genotype, and weather data.

The results of using 2017 experimental data to train the models and predict the rank of genotypes in the following years are provided in Figure 3.13. As the same genotypes were planted in InCal-17 and InCal-18, using the models trained on the 2017 experiment provided the most accurate results for InCal-18. Although the models were trained on data from 2017, when all the inbred genotypes were planted, they were still able to provide useful information for the TcCal-18 and SbDivTcCal-19, which included the hybrid genotypes. However, using the genomic-based information did not improve the prediction accuracies for the SbDivTc-18 and 19 models. Based on Figure 3.6, the TcCal-18 and SbDivTcCal-19 included many genotypes that were located close to the inbred genotypes, meaning there were similarities with the inbred genotypes, while the SbDivTc-18 and 19 experiments included many varieties that were very different from the inbred genotypes.



Figure 3.13. Results of biomass and rank prediction for example experiments in 2018 and 2019 from predictive models trained on the 2017 experiments using genomic (G), genomic and weather data (G+W), RS (RS), RS and weather data (RS+W), RS and genomic data (RS+G), and RS, genomic, and weather data (RS+G+W).

Figure 3.14 shows the  $R^2$  for biomass rank of the two experiments in 2019 predicted using the models trained on the 2018 experimental data. In most of the experiments, the models developed using genomic-based features (wavelet-6 feature) as well as RS based features provided the most accurate prediction results. As the same genotypes are planted in 2018 and 2019, the predictive models developed using only genomic-based features were able to provide  $R^2$  of ~0.6 and higher. Adding the weather-driven features did not significantly change the  $R^2$  values. The Ridge Regression and SVR models provided more accurate models when using the RS data, while the RF models had higher accuracies based only on the genomic-based features. The prediction

accuracies of SVR, PLSR, and Ridge Regression were very similar and slightly higher than the  $R^2$  of the RF model when using all the input data sources.

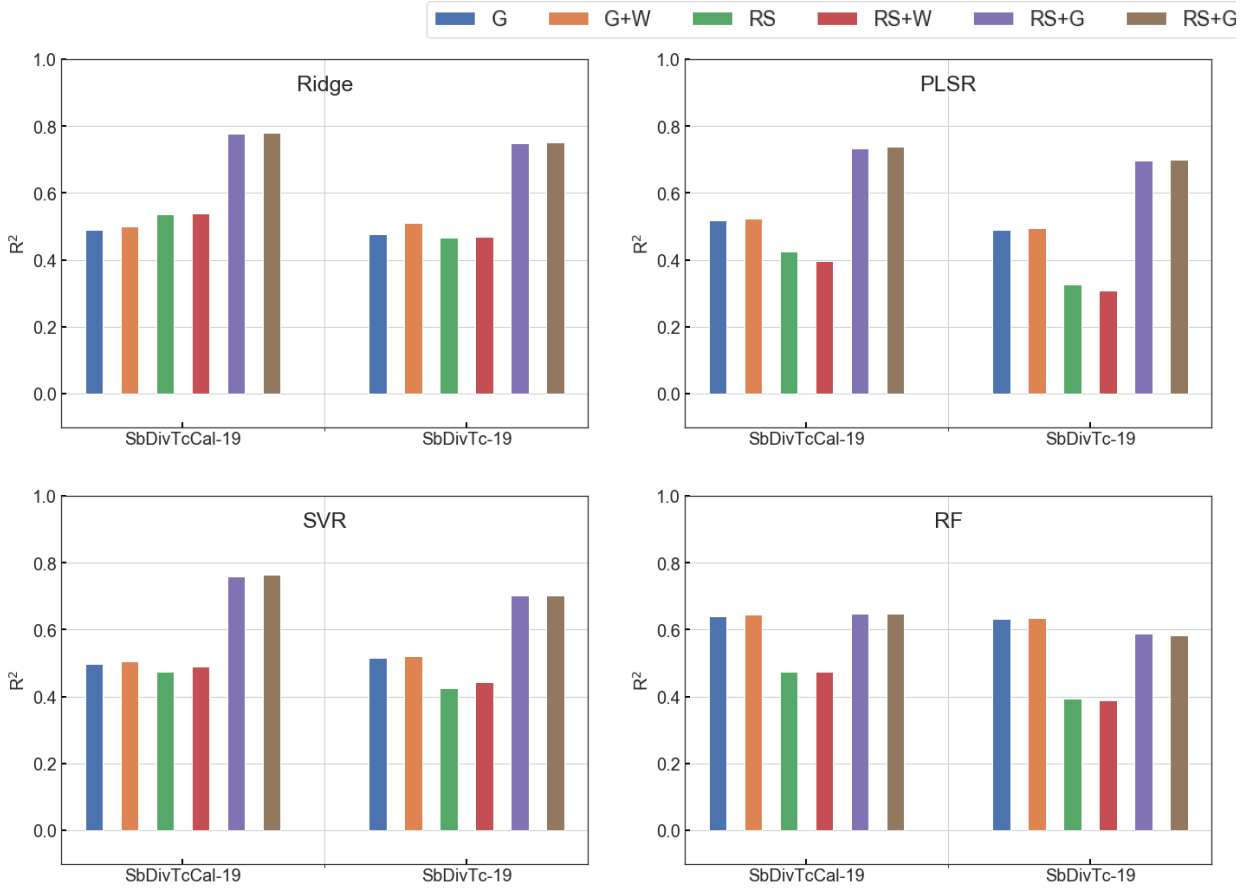


Figure 3.14. Biomass and rank prediction for example experiments in 2019 from predictive models trained on the 2018 experiments using various data sources.

### *Models trained on two previous years of data*

Figure 3.15 shows that the models trained on the combined 2017 and 2018 experiments are more accurate than the models trained on only the 2017 or 2018 data that were presented in the last section. This was expected as the models were able to train better when multiple years of data were used in for developing the model, especially when weather-driven features were used. There is still a bias between the predicted and reference biomass data as the planting and harvest dates were different in the three years of experiments. Figure 3.15 also shows the results of the biomass rank from the SVR models trained on data from the 2017 and 2018 experiments. The  $R^2$  values of the biomass rank prediction for the two experiments conducted in the 2019 growing season were

slightly higher than the results of the models trained only on the 2018 data, which was again expected as more samples are used in training (from both 2017 and 2018 experiments), and also, the models were trained on two years of weather related features. The RMSE of rank-based models for both SbDivTcCal-19 and SbDivTc-197 experiments was  $\sim 7\%$ .

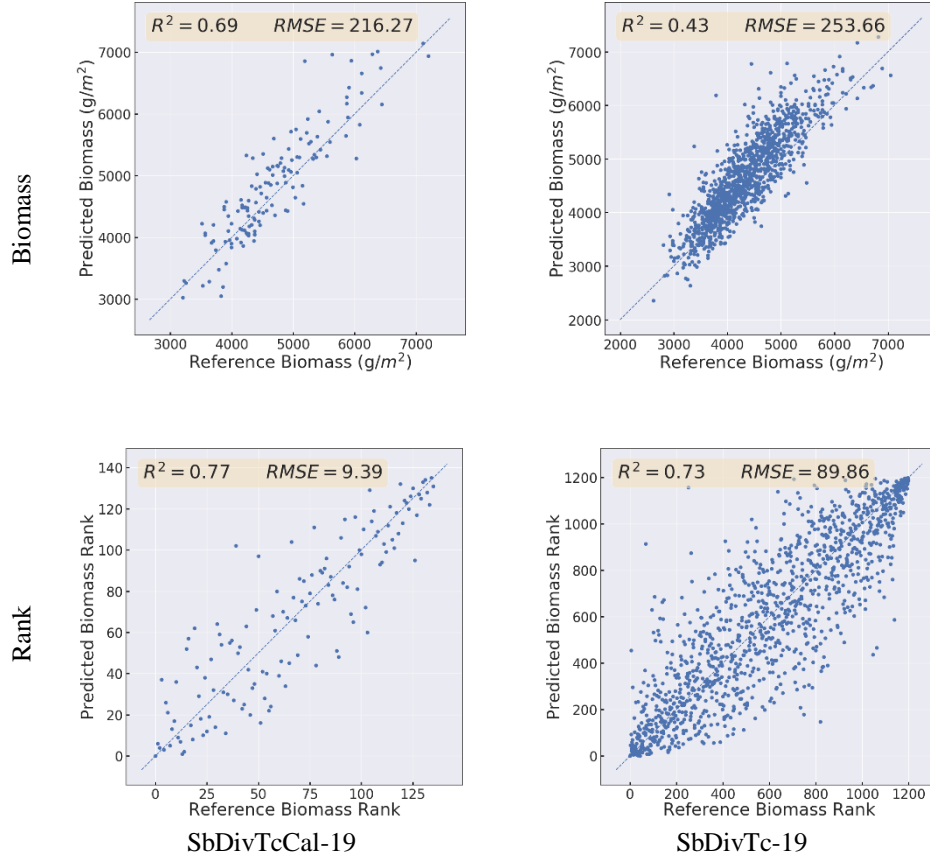


Figure 3.15. Biomass and rank prediction for example experiments in 2019 from SVR models train on the 2017 and 2018 experiments using RS, genotype, and weather data

Figure 3.16 shows the results of predicting the biomass rank for the 2019 experiments using the models trained on the 2017 and 2018 experiments data using Ridge Regression, PLSR, RF, and SVR methods. RF provided more accurate results when only genomic-based features or genomic and weather based features were used in the prediction. The most accurate predictions were obtained when all the data sources were used in the predictive models, except RF models. The models developed with both weather and genomic based features were more accurate than the models developed only with the genomic based features. However, as for other experiments adding the weather based features to the RS models did not typically improve the  $R^2$  values much. One

explanation could be that while the genomic based features represent the functions encoded in the plant's DNA that controls the growth and other behavior of the plants, this can change with environments. The weather-based features represent an environmental factor in the plant's behavior and contribute to variation in biomass estimates. However, models that include RS data also include the impact of the interaction of the plants (and their genomics) with the environment. Adding the weather related information represented by temperature, radiation, and precipitation to the models that included RS features did not provide significant additional information for the models, although the most accurate results were obtained when all the data sources are used in the models for both experiments in 2019.

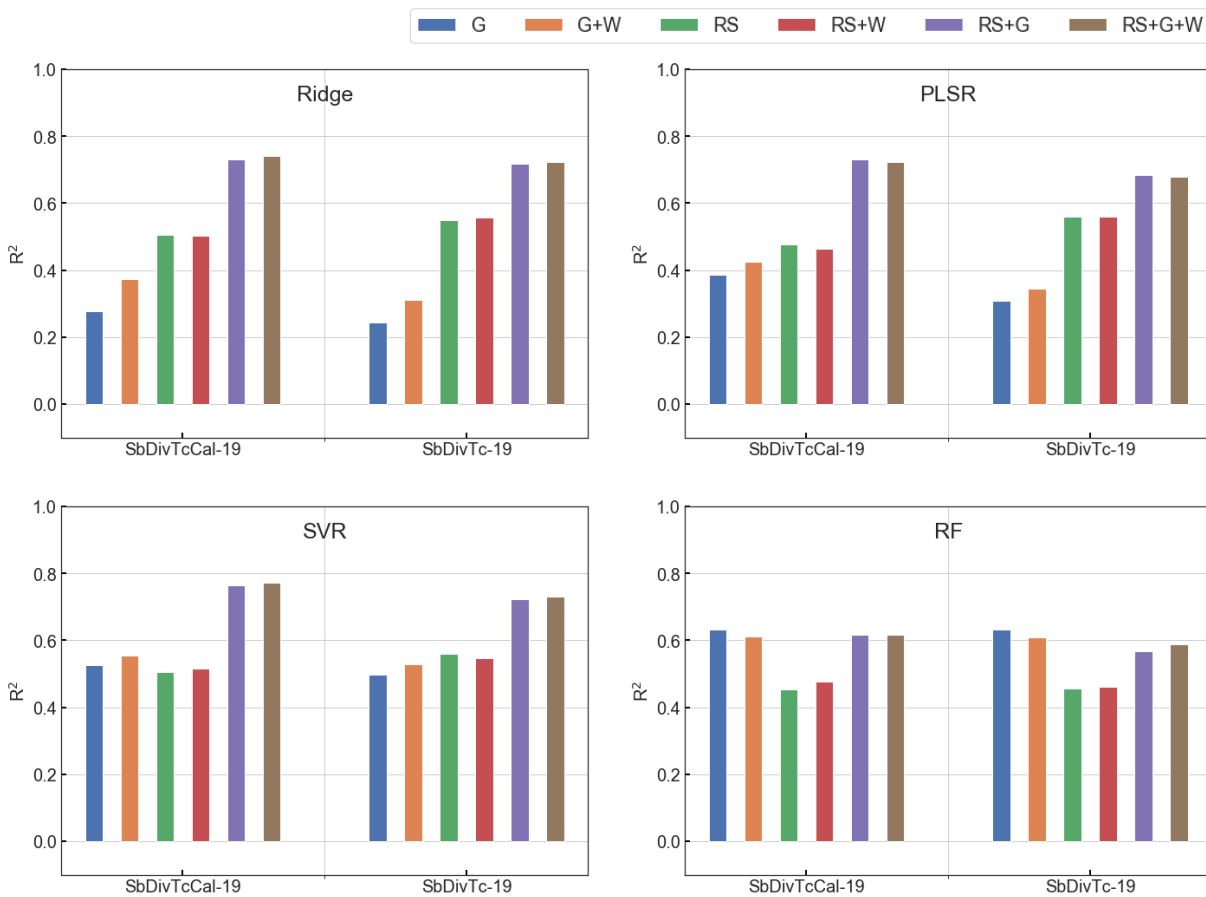


Figure 3.16. Results of biomass and rank prediction for experiments in 2019 from predictive models train on the 2017 and 2018 experiments using various data sources.

### Genomic-driven feature evaluation

In Section 3.4.1, the impact of using genomic-based features in predictive models was investigated. In this section, we repeat the evaluation, but using those features combined with the RS and weather-related data. Models are trained using the data of the 2017 and 2018 experiments, and evaluation is performed on the two experiments in the 2019 growing season. The results are presented in Figure 3.17, which shows PLSR, Ridge Regression, and RF yielded the highest  $R^2$  values when using wavelet-4, although the  $R^2$  values for models based on wavelets-2 were only slightly lower. With the PCA based inputs, the SVR models provided higher accuracies than the other methods, but the accuracies were generally lower than using wavelet-based features. The maximum  $R^2$  values achieved by SVR models were based on the wavelet-6 genomic features.

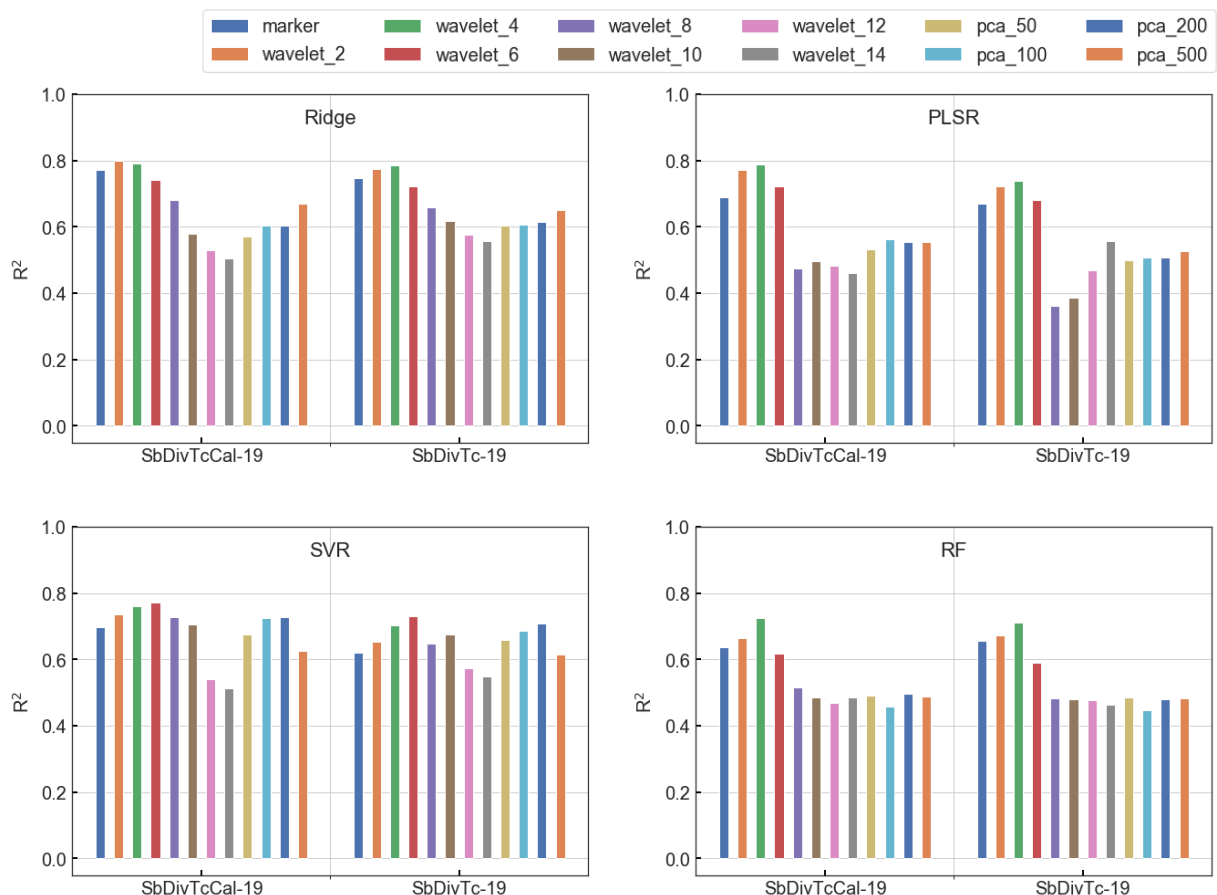


Figure 3.17.  $R^2$  of predictions based principal components and wavelet-based genomic features in combination with RS and weather inputs in Ridge Regression, PLSR, SVR, and RF models

### Multi-temporal Predictions

In the previous sections, we showed that the predictive models developed using the weather, genomic, and RS data as inputs are able to predict the rank of genotypes with respect to their biomass values for the genotypes in the subsequent years without any reference data from these years. This implies that it may not be necessary to perform expensive, time-consuming plot level sampling of biomass at the end of the season. Additionally, the models that were developed used the RS data collected from the beginning of the season to the end of the season, which implies that RS data acquisitions were also required until the end of the season. In this section, we investigate whether the biomass rank can be predicted earlier in the season to reduce both field sampling and acquisition of RS data.

Figure 3.18 shows the impact of using RS data in G+W+RS models based on RS data sets collected on DAS of 40, 60, 80, and 95. Models based on the 2017 and 2018 experiments, using all four data sets provided the highest accuracy for the Ridge Regression, PLSR, and SVR models. RF models, whose  $R^2$  values were lower, had the highest  $R^2$  using RS samples from only the earliest date. Generally, the models developed with data from three or four dates provided similar  $R^2$  values. These models had  $R^2$  values of 0.6 or better using only data sets at 40 and 60 DAS, which implies that genotype selection for biomass production can be initiated earlier, and models with DAS=60 can be considered in the selection of genotypes for biomass production.

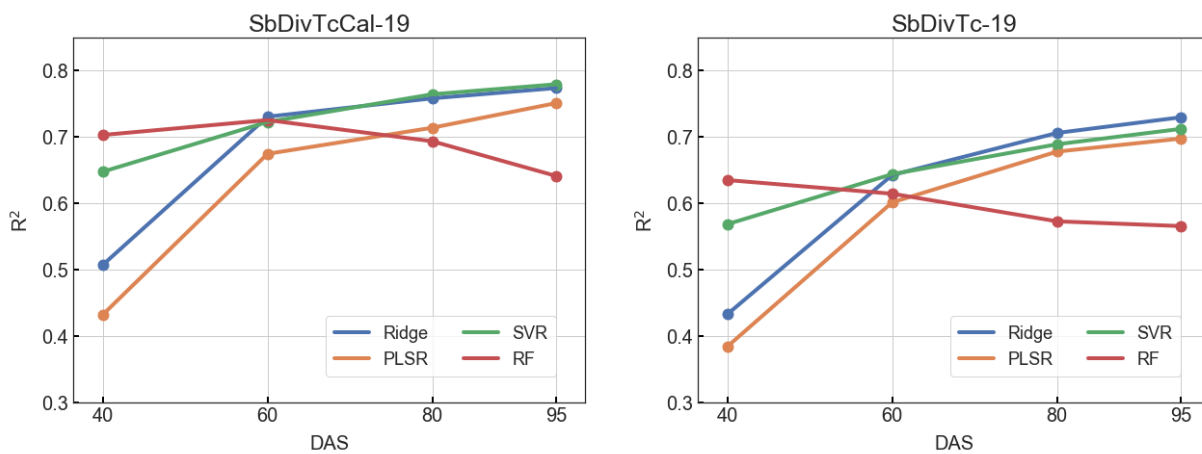


Figure 3.18. Impact of number of RS datasets on  $R^2$  of predictions using Ridge, PLSR, SVR, and RF regression models

### 3.5 Conclusions

In this study, we focused on developing regression based models using the RS data collected multiple times during a growing season combined with the genomic information as well as weather data to predict the end of season sorghum biomass yield. The models based on one and two years of data were used to predict the biomass yield for the future years based on using sample data from previous years. The results showed the high potential of the models for biomass and biomass rank predictions, with models based on two years of data having the higher  $R^2$  values. We also investigated the use of wavelet decomposition and PCA to extract a smaller number of features from the markers. The results indicated that level 4 and 6 wavelets generally provided models with higher accuracies than using all markers. We also investigated the possibility of developing predictive models using the RS data collected until mid-season, rather than the full season. The results showed that the rank of biomass can be predicted with  $R^2$  values of around 0.65-0.70 using the RS data until 60 DAS. This not only reduces the time required for phenotyping by reducing the manual sampling process, but also decreases the time and the cost of the RS data collection, processing, and analysis, and particularly for hyperspectral sensing.

## **4. PREDICTION OF CROP YIELD USING UAV TIME SERIES DATA AND RECURRENT NEURAL NETWORKS**

### **4.1 Abstract**

Phenotyping via Unmanned Aerial Vehicles (UAVs) is of increasing interest for many applications because of their capability to carry advanced sensors and achieve accurate positioning required to collect both high temporal and high spatial resolution data required over relatively limited areas. This study focuses on the development of a data analytics based predictive modeling strategy that incorporates multi-sensor data acquisition systems and accommodates environmental inputs. Supervised and unsupervised feature learning based on fully connected, convolutional, recurrent neural networks are investigated. For hyperspectral data, supervised feature extraction provides more accurate predictions, while the features extracted from LiDAR data in an unsupervised training yield more accurate prediction. Predictive models based on Recurrent Neural Networks (RNNs) are designed and implemented to accommodate high dimensional, multi-modal, multi-temporal data. Remote sensing data, including Light Detection and Ranging (LiDAR) and hyperspectral inputs, as well as weather data, are incorporated in RNN models. Results from multiple experiments focused on high throughput phenotyping of sorghum for biomass predictions are provided and evaluated for agricultural test fields at the Agronomy Center for Research and Education (ACRE) at Purdue University. The results of using proposed RNNs for training on one experiment and predicting biomass for other experiments with different types of sorghum varieties illustrate the potential of the network for biomass prediction, and the challenges relative to small sample sizes, including weather and sensitivity to the associated ground reference information.

### **4.2 Introduction**

With the advances in science and technology, it is now possible to sequence the genome of plants rapidly and at low cost (Davey et al., 2011). In order to evaluate the performance of the plant varieties with given genetic characteristics, numerous physical and agronomic traits are measured in controlled facilities or in the field during the growing season (phenotyping). Planting the varieties of a crop and measuring dozens of phenotypic traits for each is time consuming,

laborious, and expensive. Recently, researchers have begun to explore the use of remotely sensed data to augment and even replace some traditional in-field phenotyping measurements.

Biomass is an important plant characteristic for sorghum-based biofuels, as it is indicative of both the crop condition and the quantity of ethanol that can be produced. Many researchers have studied predictive models for complex phenotypes such as biomass, based on data derived from Remote Sensing (RS) technologies, via empirical approaches. For example, the importance of sample size, prediction method, and sensor type for biomass prediction was investigated in (Fassnacht et al., 2014) using support vector machines, random forests, and Gaussian process models. Although traditional empirical methods are powerful tools to model the complex relationships between the RS data and biomass, they are unable to effectively exploit the temporal aspects of the data, fully exploit multiple sensing modalities, or include environmental inputs.

In recent years, deep neural networks have been investigated by many researchers in the area of pattern recognition and machine learning (ML). Deep Learning (DL) adds more “depth” (complexity) into the model compared to classical ML and represents the data in a hierarchical way in multiple layers using various functions to transform the data (Schmidhuber, 2015). If properly trained with adequate labeled data, DL can provide very accurate classification and prediction outputs because of its ability to use more complex models, which also increases flexibility and potential adaptability needed for many scenarios.

The extraction of appropriate features, along with using a powerful predictive model, is required for accurate prediction of complex phenotypes such as biomass. Hyperspectral data consist of hundreds of channels that provide detailed spectral information from the image data from which useful predictive features can be extracted. From LiDAR point clouds, multiple features describing the geometric characteristics of the canopy including height, canopy coverage, and vertical distribution of the plant material can be extracted. Although having all these features provides more information for predictive models, it can degrade the accuracy of prediction as many features are highly correlated. Generally, this is not a concern for DL-based models, as they have the advantage of being able to learn the features and automatically extract appropriate representations from input data (Lecun et al., 2015) if adequate labeled data are available. For biomass prediction, however, the number of reference data samples is limited, as destructive sampling and processing of biomass is time-consuming and expensive. In this situation, it is

important to separate feature learning from the prediction by first extracting appropriate features, and then using them in the predictive models.

DL networks, including stacked autoencoder (SAE), convolutional NNs (CNNs), and recurrent NNs (RNNs) have been investigated in different RS applications. In general, DL is beneficial for RS problems where physical models are complex, e.g., nonlinear, or even not yet well understood and/or cannot be generalized (Zhu et al., 2017). Using DL to solve RS problems also has multiple challenges, including:

1. Multi-modality; in contrast with most fields, the data in RS problems are often acquired in multi-modality scenarios.
2. Geotagged; the spatial relationship between different data units (usually pixels) is important in RS.
3. Time-related; in many applications, multi-temporal RS data are acquired to represent temporally varying processes.
4. High dimensional, big data; RS data collected with HSI sensors have hundreds of channels, and for some problems, the experiments cover an extended area.
5. Limited reference samples; in many applications of RS, collecting ground reference data is very expensive and time-consuming which results in a small number of samples.

To date, most deep learning-related research in the area of remote sensing has focused on classification using benchmark data sets. Chen et al. (Chen et al., 2014) used three fully connected networks for the classification of HSI data. Zhao et al. (W. Zhao & Du, 2016) proposed a novel, spectral-spatial classification framework for HSI which first uses a feature extraction method to extract low-dimensionality spectral features, and then employs a CNN to extract high-level spatial features. Spectral and spatial features are stacked as the inputs for a classifier. The method was compared to other well-known classifiers; it was reported that better classification performance was achieved because of combining the spectral and spatial discriminant features (it should be noted that this is also a common conclusion of traditional classification methods when applied to image data). In (Jiaojiao Li et al., 2018), a fully CNN proposed for HSI classification consisted of convolution, deconvolution, and pooling layers. In these studies, the DL network was used to extract features from the HSI data, and the extracted features were used in a traditional ML classification method.

In our study, we use deep learning networks in two steps, first feature learning, and then prediction (using the learned features). The features extracted by this method can also be the input of traditional methods such as SVR. The advantage of this feature extraction method is that the feature learning network can be trained in an unsupervised manner, which means that the data from all available experiments can be used for training. This also means that for the prediction we can use a shallower network with a few parameters as the “appropriate features” are already extracted from the data; thus, the network has much fewer parameters to optimize, and a smaller number of samples are required.

For feature learning, we investigated two strategies; feature learning from the many already extracted features (feature reduction), and feature extraction from the original data based on auto-encoder decoder networks. Two networks are designed for feature reduction that can be used for features extracted from either hyperspectral or LiDAR data. One consists of several fully connected layers, and the second consists of convolutional, pooling, and de-convolutional layers. For feature extraction from the original data, a 2D CNN is used for hyperspectral data, and two RNN based networks are designed for the LiDAR point cloud data.

For biomass prediction, two RNNs are designed to include both hyperspectral and LiDAR data and weather-related data (e.g. solar radiation, temperature, precipitation). For biomass prediction, having a method that can incorporate the data from multi-year experiments and different sources is important, both for increasing the quantity of data samples and representing the growth scenarios under different weather-related growing conditions. The results and analysis provide useful insights on feature importance for prediction and demonstrate the ability to estimate the biomass from remotely sensed data, thereby greatly reducing the required human labor.

### **4.3 Feature Learning**

Feature learning has been investigated by many researchers in different fields (Farabet et al., 2013; Shin et al., 2013; Tang & Eliasmith, 2010). In hyperspectral image classification, Tao et al. (Tao et al., 2015) showed that feature learning based on a stacked sparse auto-encoder model provides robust results. In (F. Li et al., 2018), a network of fully convolutional auto-encoders for image feature learning was developed for image clustering. In (Wu & Prasad, 2017), a convolutional recurrent neural network (CRNN) consisted of several convolutional layers and a

few recurrent layers to extract locally-invariant and spectrally-contextual information from the hyperspectral data.

Inspired by these studies, in this paper, we develop deep neural networks for unsupervised learning of features from both hyperspectral and LiDAR data. We investigate the networks that can reduce the number of extracted features by learning a smaller number of features from extracted features. Toward this end, we designed a fully connected auto-encoder network (FAEN) and a convolutional auto-encoder network (CAEN). The FAEN and CAEN learn fewer features from the original features extracted from the RS data for each plot. These networks are described in Section 4.3.1.

The auto-encoder decoder networks can be used to extract latent features from hyperspectral data and LiDAR point cloud data. For hyperspectral data, we used a 3D convolutional auto-encoder network (CAEN-2d), including six hidden convolutional layers, three layers for encoding the data, and three layers for decoding the encoded features to reconstruct the hyperspectral data. For LiDAR data, as different numbers of points exist in each plot, we designed multiple recurrent auto-encoder networks (RAEN) that use LSTM units for encoding and decoding the point cloud data. These networks are described in Section 4.3.2.

The design of each network architecture is based on the evaluation of many different hyperparameter settings, including learning rate, number of layers, size of each layer, number of encoded features, size of the kernels in CNN-based networks, dropout rate. The architecture that provides the best results in terms of the reconstruction error, computational complexity, and best encoded features (with regard to prediction accuracy) is selected.

#### **4.3.1 Feature Reduction by Auto-encoders**

The FAEN network consists of three fully connected hidden layers as well as input and output layers as shown in Figure 4.1. The first hidden layer consists of more neurons than the input layer, allowing the network to expand the information of the input features to a higher dimensional space. The second layer then has a few neurons (four in this paper, based on evaluations of different hyperparameters) that are considered as the learned features. This part of the network is referred to as the encoder, which abstracts (encodes) the information of the input features into a few features. Then, the second part of the network (decoder), which is symmetric with the first part, reconstructs the input features from the learned features.

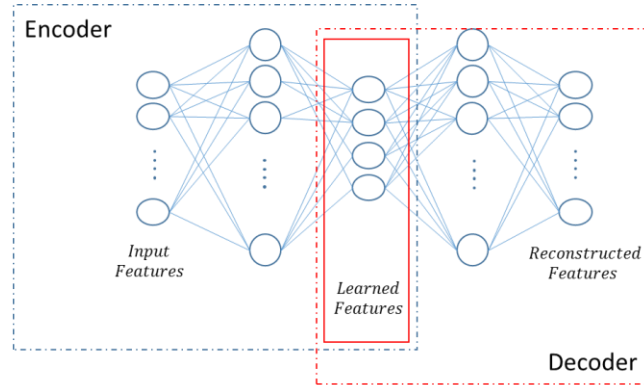


Figure 4.1. Unsupervised feature learning using a fully connected auto-encoder network (FAEN).

The second network, CAEN, is similar to FAEN, but uses convolutional layers instead of fully connected layers, as shown in Figure 4.2. Using convolutional layers for feature learning allows the network to take advantage of potential relationships between the input features. Also, there are fewer parameters to optimize during the training of the network. The size of the output layer in both networks is the same as the input layer as they reconstruct the input features. The cost function is the average sum-of-squared error between the input and reconstructed features over the entire sample data.

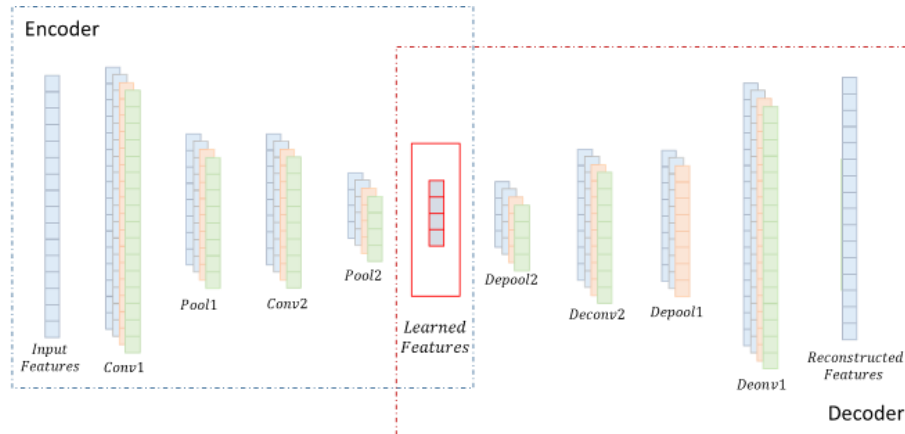


Figure 4.2. Unsupervised feature learning using a convolutional auto-encoder network (CAEN).

### 4.3.2 Feature Extraction by Auto-encoders

DL has the advantage of being able to learn and automatically extract features from the original data, in multiple levels of the hierarchy (Lecun et al., 2015). Similar to CAEN-1d, the goal is to use the auto-encoders to encode a few features from each plot, but using the image of the plot including all the bands, rather than the extracted features. This exploits the spatial information from the image as well as the variation in the pixel's values within a plot. Compared with CAEN-1d, there are more parameters to optimize, and thus a more time-consuming training process. The proposed 2D convolutional auto-encoder network (CAEN-2d) is shown in Figure 4.3. The network consists of three convolutional layers that reduce the size of the image at each layer until a desired number of features is obtained. Then, in the decoding part of the network, three symmetric layers increase the size of the image at each layer until the original image data are reconstructed. As this network is used for training and extracting features from all the experiments, the average size of the plots in all experiments,  $64 \times 32$  pixels is considered for the network design. If a plot covers a smaller or larger area, padding or trimming is used to resize all the plots. Also, the number of hyperspectral bands differs slightly in the three years (271 to 273), but we have binned the data by four so that a spectral resolution of 10 nm is obtained, and by ignoring the first and last bands, so all the plots had 64 spectral bands. Details on each layer of the designed network are presented in Table 4.1.

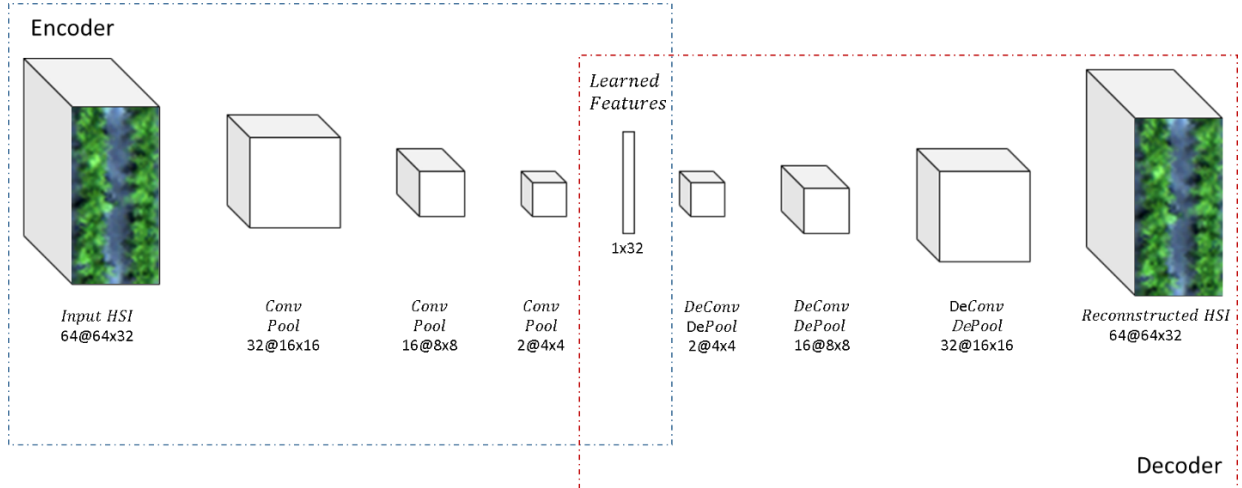


Figure 4.3. Unsupervised feature learning using 2d convolutional auto-encoder network (CAEN-2d).

Table 4.1. CAEN-2d details

Input Image	Size: (64,32,64)
First layer	Kernel: size (11,7,16), stride (2,1) Activation: Sigmoid, Max polling: Yes Output size: (16,16,16)
First layer	Kernel: size (5,5,8), stride (1,1) Activation: Sigmoid, Max polling: Yes Output size: (8,8,8)
First layer	Kernel: size (3,3,2), stride (1,1) Activation: Sigmoid, Max polling: Yes Output size: (4,4,2)
Encoded Features	Size (32)

Typically, convolutional or fully connected networks are fed through a regular input data format; however, the point cloud data usually represent a spatially irregular data format, requiring a conversion to extract regular 3D voxel grids or collections of images. This unnecessarily increases the size of the data set, which is not efficient, and also can reduce natural invariance of the data by introducing quantization artifacts (Su et al., 2018). In recent years, several network architectures have been developed based on using point clouds directly. Qi et al. (Qi et al., 2017) proposed using the point cloud directly in a network that relies on max pooling as a single symmetric function to learn the global descriptor for the entire shape (shape classification) or predicting per point labels (shape segmentation) from informative points of the point cloud. In this study, we investigate the possibility of using recurrent auto-encoder networks for learning features from the point cloud data. Two architectures are investigated in this section based on LSTM units.

RNNs are powerful models for representing sequential data and are particularly effective in learning long-term dependencies. However, it is demonstrated that conventional RNNs suffer from the vanishing gradient and long-distance dependency problem. Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) architectures were designed to address the issue. LSTMs and gated recurrent units (GRU) (Cho et al., 2014) have proven to be successful in accommodating the difficulty of training simple recurrent networks (Karpathy et al., 2015). An LSTM unit is capable of learning long-term dependencies due to its structure, which incorporates gates that regulate the learning process. The operations are accomplished by the input, forget, and output gates. Based on the information provided by the input and forget gates, the cell state, which represents the memory of the unit, is updated. The GRU architecture is similar to LSTM, but

simpler to compute and implement as it consists of only two gates: reset and update gates (Fu et al., 2017).

The RNN based encoder-decoders consist of a many-to-one network that encodes the input data to a few encoded features, and a one-to-many network that decodes the encoded features to reconstruct the original input data (Figure 4.4).

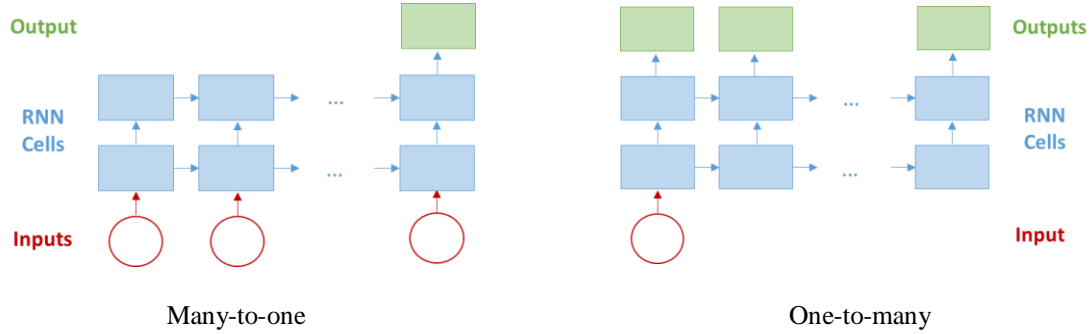


Figure 4.4. One-to-many vs. many-to-one RNN network.

Inspired by the LSTM encoder-decoder model used for text simplification in (Wang et al., 2016), we designed two networks shown in Figure 4.5 and Figure 4.6. In RAEN-1, we first select a group of points and assign them to a bag, then we have a number of bags which is considered an input to the RNN first layer. The number of points in each bag and the total number of points in each plot determine the number of bags. The larger the number of bags, the longer the training time. From Figure 4.5, the bags are fed into two layers of RNN cells (LSTM in our study) and the output of the two layers is considered as the encoded features. The encoded features are then copied as the same number of bags, and fed to the decoder part of the network. The output of the RNN cells in this part of the network is the reconstructed bags of the points which were originally fed to the decoder.

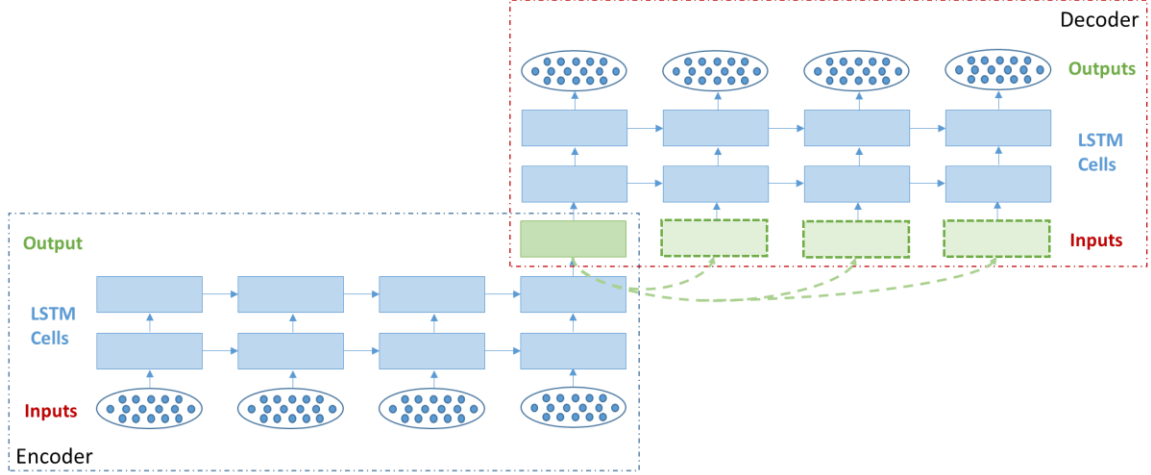


Figure 4.5. RNN auto-encoder network (RAEN-1) proposed for feature extraction from LiDAR data.

One possible issue with the RAEN-1 is that all the layers have the same number of inputs (equal to the number of bags) which can create a long sequence of data, which not only increases the training time, but can also negatively impact the training itself. To avoid this issue, we propose a modified architecture where the hidden RNN layers have a smaller number of inputs by concatenating the outputs of several outputs of the previous layers (see Figure 4.6). Also, we investigate a non-symmetric decoder to reduce the number of hidden layers and ultimately the gradient vanishing problem that RNNs often encounter.

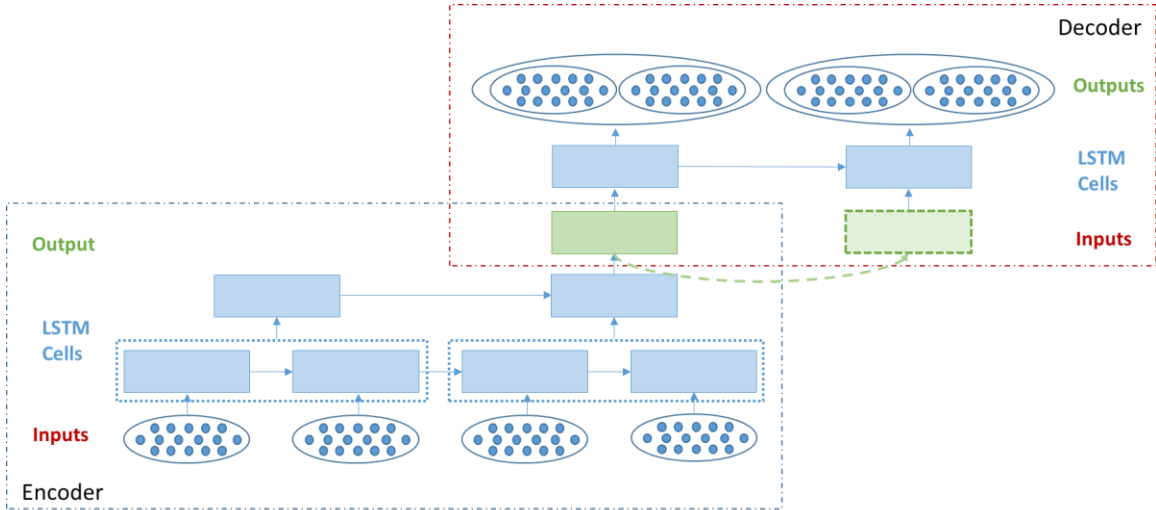


Figure 4.6. RNN auto-encoder network (RAEN-2) proposed for feature extraction from LiDAR data.

### 4.3.3 Supervised Feature Learning by Auto-encoders

Feature extraction can be challenging for large data sets, as the AE networks might encode features that are not especially useful for a given application. To avoid this issue, it is possible to force the network to learn features for a specific goal. For example, here we force the network to minimize the error of biomass prediction using encoded features in a linear regression model. We modify the cost function so that it reduces the reconstruction error, as well as the error of the biomass prediction using the encoded features in each iteration:

$$\begin{aligned} Cost_{reconst} &= \sum (reconstruction - original)^2 / n \\ Biomass_{pred} &= W * Learned_{Features} + B \\ Cost_{biomass} &= \sum (Biomass_{pred} - Biomass_{references})^2 / n \\ Cost &= Cost_{reconst} + \alpha Cost_{biomass} \end{aligned}$$

where  $n$ ,  $W$ ,  $B$  and  $\alpha$  are the number of batches, a random coefficients tensor, a random bias tensor, and regularization factor, respectively.

## 4.4 Predictive Models

Two models are developed for biomass prediction, one based on Support Vector Regression (SVR) and the other based on RNN.

### 4.4.1 Support Vector Regression

SVR is a supervised non-parametric regression technique, and therefore, no assumptions regarding the underlying data model are required. SVR transforms the original feature space into a higher dimensional space (T. Zhang, 2001), with the goal of finding a hyperplane to predict the training data set. More details for SVR are provided in (Smola & Schölkopf, 2004). The optimal values of the kernel function parameters are found by a general k-fold cross-validation in a grid search.

### 4.4.2 Recurrent Neural Network for Prediction

In this study, GRU and LSTM models are developed and compared for multi-temporal prediction of sorghum biomass. Mou et al. (Mou et al., 2017) proposed an RNN model to classify

hyperspectral pixels by considering the sequence-based data structure. In (Kong et al., 2018), a novel RNN network was proposed for online disturbance detection from a satellite image time series. RS-based prediction was also investigated in (You et al., 2017) for soybean yield prediction using an RNN. The proposed RNN architecture with two hidden layers (either GRU or LSTM based cells) is shown in Figure 4.7. The input vector for the network includes hyperspectral features, LiDAR features, and weather-related features as shown in Figure 4.7. In this study, both LSTM and GRU models are explored.

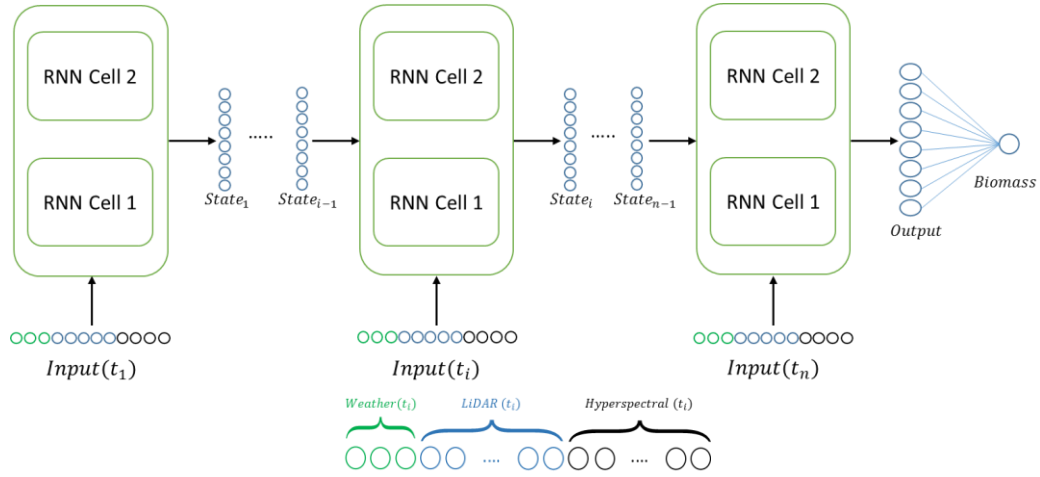


Figure 4.7. Recurrent neural network (RNN-1) developed in this study for biomass prediction, and its input vector at time  $t_i$ .

One issue with RNN-1 is that if the number of hyperspectral and LiDAR data sets differ, or they are collected on different dates, it not possible to simply use the input vector directly (also an issue for traditional regression models such as SVR). To overcome this issue, we propose a second architecture shown in Figure 4.8. The proposed network includes two sets of RNN layers: one set for features derived from hyperspectral data and one for the LiDAR-based features. This network can accommodate the differing number of datasets from each source as input. The final outputs of the two RNN cells are concatenated and considered as the input for the fully connected layer which outputs the prediction results of the network.

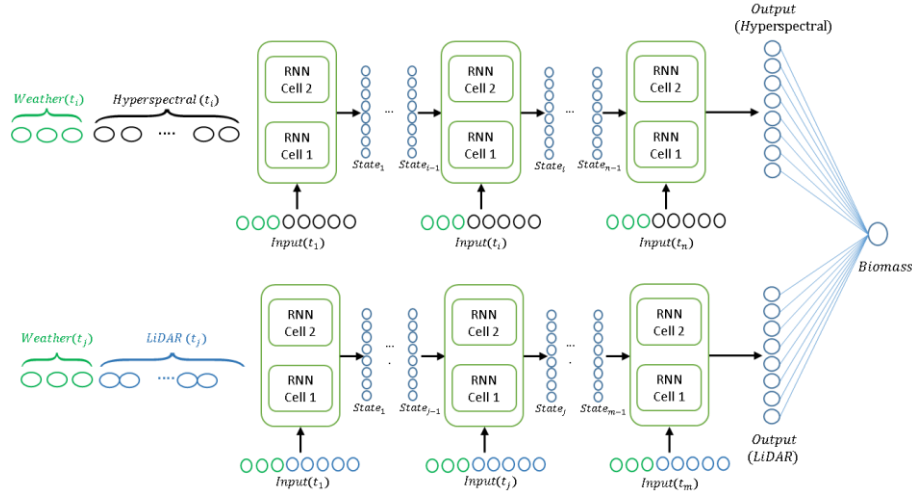


Figure 4.8. Recurrent neural network (RNN-2) developed in this study for multi-temporal biomass prediction, and its input vectors at time  $t_i$  and  $t_j$ .

## 4.5 Data and Experimental Setup

### 4.5.1 Field Ground Reference Data

The field experiments were conducted in the 2017, 2018, and 2019 growing seasons, each year in 2.8 ha sorghum breeding trials located at the Purdue University Agronomy Center for Research and Education (ACRE), West Lafayette, Indiana. All trials were arranged in a randomized complete block design and planted at 220,000 plants per ha. In 2017, the field experiments consisted of two distinct trials in two panels; the inbred calibration (InCal) panel and the sorghum biodiversity (SbDiv) panel. The SbDiv panel consisted of 840 inbred lines replicated twice in four row plots. The inbred calibration panel included 60 inbred lines, a subset of the SbDiv panel, which encompasses the genetic diversity of the SbDiv panel, and were planted in four row plots. The InCal experiment was also repeated in the 2018 growing season in a different field at ACRE, with 54 inbred lines. In 2018, two other experiments were also conducted: sorghum biodiversity test cross (SbDivTc) and inbred calibration test cross (InCalTc), both planted in four row plots. The SbDivTc and InCalTc are hybrids derived from the Sbdiv and InCal experiments with a common maternal parent. The InCalTc included 54 hybrid lines, also a subset of the SbDivTc panel. The SbDivTc included 630 hybrids replicated twice in four row plots in the 2018 growing season. In 2019, the SbDivTc was repeated with the same number of genotypes. An additional panel was added in 2019: sorghum biodiversity test cross calibration (SbDivTcCal),

which included 72 hybrid lines, a subset of the SbDivTc panel that is representative its genetic diversity planted in 10 row plots. Details of experiment trials are provided in Table 4.2.

Table 4.2. Experiment designs for the 2017-2019 growing seasons.

Year	Trial	Genotype	# of plots	# of genotypes	Sowing Date	Harvest Date	Comment	Available Biomass Data
2017	HyCal-17	Hybrid	72	18	May 16	Sep 27	--	6/27, 7/17, 7/31, 8/08, 9/27
	InCal-17	Inbred	120	60	May 16	Sep 27	60 replicated genotypes from SbDiv-17	9/27
	SbBAP-17	Inbred	760	350	May 16	Sep 28	--	9/28
	SbDiv-17	Inbred	1800	840	May 17	Nov 9	--	11/09
2018	HyCal-18	Hybrid	72	18	May 8	Aug 9	Same genotypes as HyCal-17	6/27, 7/12, 8/09
	InCal-18	Inbred	108	54	May 8	Aug 9	54 replicated genotypes from InCal-17	8/09
	InCalTc-18	Hybrid	108	54	May 8	Aug 6	54 replicated genotypes from SbDivTc-18	8/06
	SbDivTc-18	Hybrid	1260	630	May 8	Aug 2	--	8/2 and 8/14
	SNitTs-18	Inbred	112	4	June 4	Oct 2	--	10/02
2019	SbDivTcCal-19	Hybrid	144	72	June 4	Aug 2	72 replicated genotypes from SbDivTc-19	9/5
	SbDivTc-19	Hybrid	1260	630	June 4	Aug 2	Same experiment as SbDivTc-18	9/5

## 4.5.2 Remote Sensing Data

This study includes RGB, hyperspectral, and LiDAR remote sensing data collected by custom designed UAV platforms. All remote sensing data acquisition platforms were flown with Global Navigation Satellite System/Inertial Navigation System (GNSS/INS) units for direct georeferencing. RGB data for this study were collected using a Sony Alpha ILCE-7R RGB camera with a Sony 35mm lens. The Sony Alpha 7R features a full-frame 36.4MP sensor delivering high-resolution UAV-based aerial imagery. LiDAR data were collected with a Velodyne VLP-16 Puck Hi-Res in 2017 and 2018 at an altitude of 25 m, and a Velodyne VLP-16 Puck LITE in 2019 in an altitude of 45 m. The LiDAR sensors in all three years were operated in the strongest return mode and 600 rotations per minute (RPM). The VLP-16 is a 16 beam LiDAR unit with a 360-degree horizontal FOV, and a maximum range of 100 m. Both the RGB camera and the VLP-16 sensor are mounted on a DJI Matrice 600 Pro (M600P) platform. Spatial and temporal system calibration

for the datasets used in this study were conducted using the approaches described in (Ravi et al., 2018) and (LaForest et al., 2019), respectively. Also, the georeferenced orthomosaics were generated using the structure from motion strategies introduced in (Hasheminasab et al., 2020; He et al., 2018).

Visible Near Infrared (VNIR) hyperspectral data were collected with two Headwall Photonics push-broom scanners. The VNIR data were acquired by a Nano-Hyperspec imaging sensor in 272 spectral bands at 2.2 nm/band from 400 nm to 1000 nm with 640 spatial channels at 7.4  $\mu\text{m}/\text{pixel}$ . In 2017, the sensor was flown at an altitude of 60 m with a 12 mm Schneider lens, resulting in a Ground Sampling Distance (GSD) of  $\sim 4$  cm. An 8 mm lens was used in 2018 and 2019, and the flying height was  $\sim 40$  m to maintain the GSD at  $\sim 4$  cm. In 2019, the RGB, LiDAR, and VNIR sensors were integrated and flown together on a single UAV platform (see Figure 3.4). A rigorous boresight calibration process was performed (Habib et al., 2018), yielding aligned RGB, LiDAR, and VNIR data. All the hyperspectral data were converted to reflectance using the empirical line method to relate the spectra collected from the UAV to data acquired by an SVC 1024i field spectrometer over the calibration targets placed in the field for each acquisition. Table 4.3 lists the data acquired by the sensors in 2017, 2018, and 2019 growing seasons.

Table 4.3. Remote Sensing Data Sets Details

Year	Data Type	Dates
2017	LiDAR	6/28, 7/18, 8/08, 8/23
	Hyperspectral	6/27, 7/14, 8/08, 8/23
2018	LiDAR	6/14, 7/06, 7/25, 8/09
	Hyperspectral	6/11, 7/02, 7/23, 8/06
2019	LiDAR	7/12, 8/02, 8/24, 9/05
	Hyperspectral	7/13, 8/02, 8/24, 9/05

### 4.5.3 Experimental Setup

Results of three sets of experiments are presented: the proposed method for feature learning is investigated initially. For each experiment, the data from all other experiments are used to train the proposed the FAEN, CAEN, CAEN-2d, and RAEN models, then the learned features are evaluated as the input features in SVR models to predict the end of season biomass for that experiment. In the second set, the robustness of the performance of the SVR models and the proposed RNN-1 is evaluated by training these models on one experiment and using the models to predict biomass of all other experiments using the features learned via FAEN. In the third set of

experiments, predictive models based on RNN-2 are developed for biomass estimating at multiple times during the growing season.

## **4.6 Results**

### **4.6.1 Feature Learning Results**

The results of feature learning include both the proposed FAEN and CAEN architectures. For training, all the features from all the multi-variety plots and all five available dates are stacked in an input vector. For example, for the HyCal 2017, there are 72 plots (from Table 1) and five dates, so  $4 \times 72 = 360$  sample data points are available for training. Once training is completed, the trained network can be used to extract features for other data sets (any experiment at any date).

Figure 4.9(a) shows the original, reconstructed, and learned features from hyperspectral data for one of the varieties in the HyCal-17 experiment using FAEN trained on the same experiment. Figure 4.9 (b) shows the features using the same network architecture, but training was performed on the SbDiv-17 experiment and the data for the HyCal experiment were reconstructed. The Root Mean Square Error (RMSE) of the reconstructions is also provided in Figure 4.9. The reconstructed signatures are very similar to the original signatures in both training scenarios based on the small RMSE values observed for each reconstruction. This implies that it may be possible to train the network on some experiments, and then use/transfer the trained network for other similar experiments. This is important for two reasons: 1) if there are experiments where the number of samples for training is not adequate, pre-trained networks can be used; 2) this provides the opportunity to use a trained network to extract features for all the experiments in both 2017 and 2018, which represent a broader range of environmental conditions. This can potentially facilitate transfer learning required for multi-year predictions.

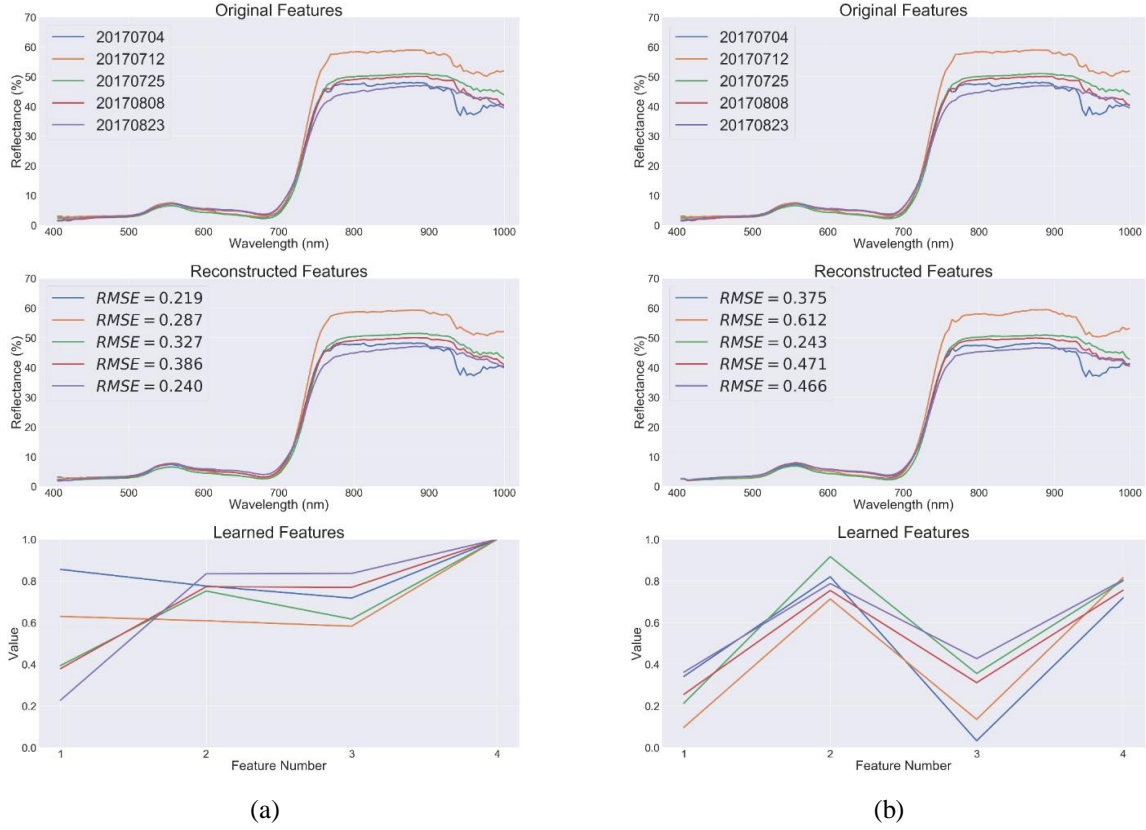


Figure 4.9. Original, reconstructed, and learned features for one variety of the HyCal-17 experiment using FAEN trained on the (a) HyCal-17 and (b) SbDiv-17 hyperspectral data sets.

To further investigate these issues, the FAEN and CAEN are trained on data from each experiment first, and the trained networks are then used to learn the features for all the other experiments. The learned features for each experiment are then used in an SVR model to predict the end-of-season biomass for that experiment using 3-fold cross validation. The  $R^2$ s of the predictions compared to ground reference data are provided in Table 4.4 and Table 4.5. In most cases, the highest accuracies were obtained when the training and testing were both associated with the same experiment (panel). However, for all experiments using the networks trained on data sets from other experiments provided comparable results. In some cases, training using the data from the combined experiments (last column of Table 4.4 and Table 4.5) provided the most accurate prediction. These results indicate that it may not be necessary to train the proposed networks on the data sets of the same experiment for biomass prediction. It should be noted however, that it is critical to use high quality data for training and evaluating networks over multiple panels and time periods.

Table 4.4.  $R^2$  values of SVR-based biomass prediction (of field-based results) using latent features extracted by the proposed **FAEN** trained on different experimental data sets and tested on all experiments. The maximum  $R^2$  of the predictions for each experiment is in bold.

Train \ Test	Data Type	2017				2018				2019		All
		HyCal	InCal	SbBAP	SbDiv	HyCal	InCal	TcCal	SbDivTc	SbDivTcCa <sub>1</sub>	SbDivTc	
SbBAP-17	H	0.54	0.53	0.49	0.49	0.47	0.47	0.37	0.53	0.37	0.50	<b>0.55</b>
	L	0.52	0.50	<b>0.54</b>	0.52	0.52	0.53	0.51	0.53	0.51	0.52	0.51
	HL	0.60	0.61	0.58	0.49	0.59	0.58	0.53	0.58	0.53	0.57	<b>0.62</b>
SbDiv-17	H	0.64	0.66	0.67	<b>0.68</b>	0.67	<b>0.68</b>	0.59	<b>0.68</b>	0.59	0.66	<b>0.68</b>
	L	0.63	0.62	0.62	0.64	0.58	0.61	0.60	0.61	0.60	0.63	<b>0.65</b>
	HL	0.74	0.75	0.75	0.76	0.74	0.75	0.73	0.74	0.73	0.74	<b>0.76</b>
SbDivTc-18	H	0.58	0.60	0.56	0.58	0.59	0.57	0.43	0.63	0.43	0.60	<b>0.64</b>
	L	<b>0.51</b>	0.50	<b>0.51</b>	<b>0.51</b>	0.50	<b>0.51</b>	0.50	0.48	0.50	0.50	0.49
	HL	0.65	0.65	0.62	0.64	0.65	0.64	0.61	0.66	0.61	0.66	<b>0.68</b>
SbDivTc-19	H	0.61	<b>0.63</b>	<b>0.63</b>	<b>0.63</b>	<b>0.63</b>	0.62	0.36	0.60	0.36	0.61	<b>0.63</b>
	L	<b>0.53</b>	0.52	0.51	0.52	0.48	0.52	0.50	0.49	0.50	0.52	<b>0.53</b>
	HL	0.67	0.67	<b>0.69</b>	<b>0.69</b>	0.68	<b>0.69</b>	0.62	0.66	0.62	0.67	<b>0.69</b>

\*\* H refers to hyperspectral, L refers to LiDAR, and HL is combined hyperspectral and LiDAR data sets

For most experiments in Table 4.4 and Table 4.5, the combined set of learned features from LiDAR and hyperspectral data provided more accurate results than using features from only hyperspectral or LiDAR. This indicates the importance of using both geometric and spectral related features for biomass predictions. For HyCal experiments in both 2017 and 2018, LiDAR-based features provided more accurate results than hyperspectral-based features, while for all other experiments, hyperspectral-based features provided more accurate results. This is likely due to the diversity of genotypes in the HyCal experiments. The HyCal experiments consisted of 18 commercial genotypes with a wide range of structural characteristics that can be reflected in LiDAR-based geometric features, while the InCal-17, InCal-18, and SvDiv-17 experiments had more similar genotypes, and thereby more similar plant structure.

Table 4.5.  $R^2$  values of SVR-based biomass prediction (of field-based results) using latent features extracted by the proposed **CAEN** trained on different experimental data sets and tested on all experiments. The maximum  $R^2$  of the predictions for each experiment is bold.

Train \ Test	Data Type	2017				2018				2019		All
		HyCal	InCal	SbBAP	SbDiv	HyCal	InCal	TcCal	SbDivTc	SbDivTcCa <sub>1</sub>	SbDivTc	
SbBAP-17	H	0.50	0.52	0.52	0.50	0.49	0.52	0.51	0.50	<b>0.53</b>	<b>0.53</b>	0.52
	L	0.50	0.45	0.47	<b>0.51</b>	0.42	<b>0.51</b>	0.38	0.46	0.45	<b>0.51</b>	0.49
	HL	0.59	0.55	0.59	<b>0.61</b>	0.54	0.60	0.53	0.57	0.57	0.58	0.59
SbDiv-17	H	0.63	0.68	0.68	0.69	0.67	0.68	0.67	<b>0.70</b>	0.67	0.69	<b>0.70</b>
	L	0.55	0.55	0.59	<b>0.61</b>	0.56	0.59	0.47	0.48	0.47	0.60	0.59
	HL	0.73	0.74	0.74	<b>0.76</b>	0.73	0.75	0.71	0.74	0.72	<b>0.76</b>	<b>0.76</b>
SbDivTc-18	H	0.62	0.59	0.62	<b>0.64</b>	0.61	0.62	0.59	0.63	0.61	0.61	0.62
	L	0.44	0.45	0.50	0.51	0.43	0.51	0.43	0.42	0.38	<b>0.52</b>	0.50
	HL	0.65	0.64	0.66	<b>0.68</b>	0.65	0.66	0.64	0.67	0.66	0.65	0.66
SbDivTc-19	H	0.61	0.62	0.63	0.62	0.63	0.63	0.62	0.63	0.63	<b>0.64</b>	0.61
	L	0.41	0.41	0.48	<b>0.49</b>	0.43	0.46	0.32	0.40	0.34	<b>0.49</b>	0.47
	HL	0.66	0.68	<b>0.69</b>	0.67	0.68	0.67	0.64	0.66	0.66	<b>0.69</b>	0.68

\*\* H refers to hyperspectral, L refers to LiDAR, and HL is combined hyperspectral and LiDAR data sets

#### 4.6.2 Feature Extraction Results

##### *Feature Extraction from Hyperspectral Data Using CAEN-2d*

Figure 4.10 shows the reconstruction results of CAEN-2d trained on the SbDivTc-18 experiment and used for feature extraction for all other experiments in both supervised and unsupervised mode. From Figure 4.10, the unsupervised reconstructed results appear to be more similar to the original data which is expected as the optimization goal was only to reduce the reconstruction error, while in supervised training, the optimization not only minimizes the reconstruction error, but also maximizes the relation between the extracted features and biomass as was noted earlier.

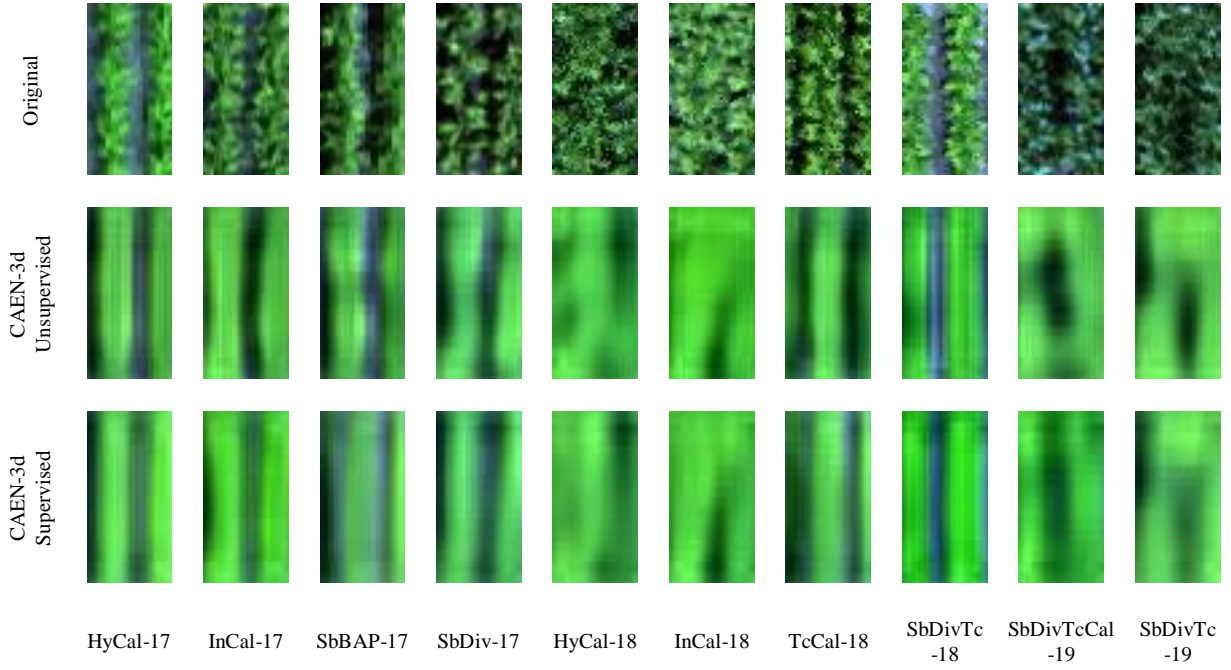


Figure 4.10. CAEN-2d reconstruction results trained on SbDivTc-18.

Figure 4.10 also indicates that even when the training is performed with data from another experiment, the reconstruction results are similar for both for supervised and unsupervised approaches. To further investigate this, the average spectra of the original and reconstructed images, as well as extracted features for an example plot from HyCal-17 experiment using the CAEN-2d trained on HyCal-17 and SbDivTc-18 are shown in Figure 4.11. The low RMSE of the reconstructed average spectra indicated that it is possible to train the network on some experiments, and then use the trained network for other experiments, as was observed for FAEN and CAEN in previous sections.

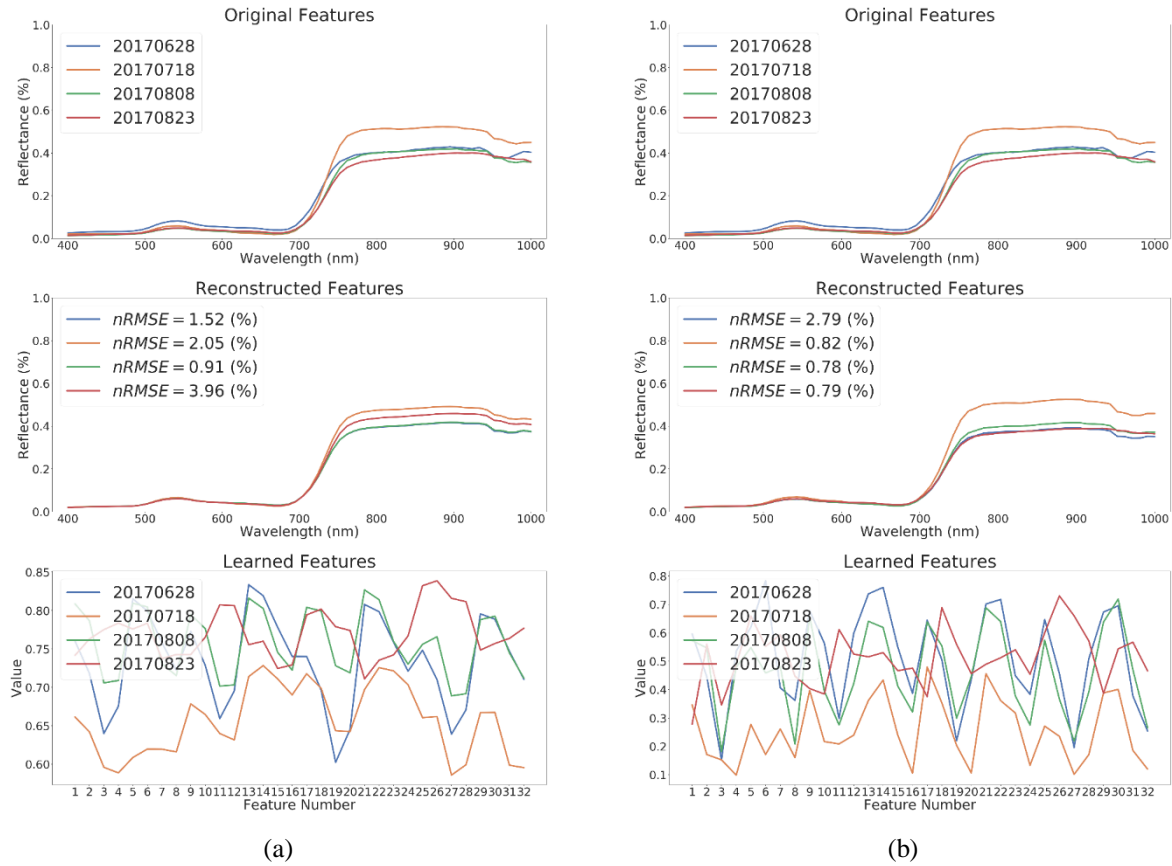


Figure 4.11. Original, reconstructed, and learned features for one variety of the HyCal-17 experiment using CAEN-2d trained on the (a) HyCal-17 and (b) SbDiv-17 experiments.

The CAEN-2d was trained on each of the experiments individually and then used for feature extraction for all the experiments. Table 4.6 shows the results of using those extracted features for biomass prediction using SVR models in both supervised and unsupervised training. The maximum  $R^2$  value for each experiment indicates that the supervised training provided more informative features for biomass prediction, even when the feature extraction network was trained on other experiments. Comparing the results in Table 4.6 with those in Table 4.4 and Table 4.5 shows that the features from FAEN and CAEN provided slightly more accurate predictions. A potential reason for this is that here we used all the pixels in each plot where there are many non-vegetative pixels, which can negatively impact the extracted features, especially for the early season data. Also, comparing Figure 4.9 and Figure 4.11 indicates that the reconstructed signal using FAEN preserved the details and resulted in a lower RMSE compared to CAEN-2d. The number of extracted features is also different from these networks, which can impact the prediction results using SVR models.

Table 4.6.  $R^2$  values of SVR-based biomass prediction (of field-based results) using latent LiDAR-based features extracted by the proposed **CAEN-2d** trained on different experimental data sets and tested on all experiments. Blue and green colors show unsupervised and supervised, respectively. The maximum  $R^2$  of each row is highlighted.

Train \ Test		2017								2018								2019			
		HyCal-17		InCal-17		SbBAP-17		SbDiv-17		HyCal-18		InCal-18		TcCal-18		SbDivTc-18		SbDivTcCal-19		SbDivTc-19	
		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
2017	HyCal-17	0.08		0.18	0.46	0.29	0.02	0.30	0.21	0.11	0.44	0.0	0.32	-0.3	0.41	0.10	0.2	0.11	0.5	0.11	0.0
	InCal-17	0.42	0.46	0.41		0.53	0.51	0.52	0.52	0.45	0.11	0.50	0.63	0.46	0.31	0.34	0.46	0.39	0.40	0.44	0.49
	SbBAP-17	0.28	0.44	0.26	0.44	0.35		0.34	0.36	0.31	0.36	0.32	0.45	0.30	0.41	0.26	0.34	0.25	0.33	0.27	34
	SbDiv-17	0.46	0.49	0.47	0.62	0.57	0.49	0.56		0.47	0.52	0.53	0.61	0.46	0.39	0.44	0.52	0.47	0.50	0.49	0.49
2018	HyCal-18	0.40	0.57	0.22	0.52	0.43	0.42	0.46	0.36	0.34		0.45	0.50	0.33	0.48	0.42	0.35	0.24	0.23	0.27	0.43
	InCal-18	0.49	0.43	0.39	0.69	0.49	0.43	0.51	0.57	0.46	0.5	0.46		0.45	0.49	0.34	0.52	0.37	0.55	0.48	0.52
	TcCal-18	0.25	0.62	0.27	0.51	0.27	0.37	0.29	0.43	0.28	0.52	0.27	0.54	0.33		0.26	0.52	0.38	0.60	0.33	0.47
	SbDivTc-18	0.42	0.48	0.44	0.55	0.54	0.50	0.51	0.52	0.45	0.53	0.45	0.62	0.43	0.65	0.46		42	0.46	0.43	0.49
2019	SbDivTcCal-19	0.29	0.60	0.41	0.50	0.34	0.48	0.39	0.40	0.37	0.40	0.32	0.45	0.28	0.62	0.35	0.46	0.37		0.37	0.36
	SbDivTc-19	0.31	0.46	0.30	0.53	0.37	0.37	0.41	0.39	0.28	0.38	0.34	0.55	0.30	0.53	0.26	0.38	0.39	0.42	0.28	

### Feature Extraction from LiDAR Data Using RAEN

Using AE for feature extraction from LiDAR data is more challenging, as was noted earlier. Using RAEN-1 and 2, we conducted seven experiments to explore the impact of the number of points in a bag, number and the size of the hidden layers, and supervised vs. unsupervised training in terms of the reconstruction error, as well as the accuracy of prediction using the encoded features. Table 4.7 lists the details of all the experiments, including training time. RAEN-2 is much faster than RAEN-1, and increasing the number of points in a bag reduces the time needed for training. All experiments were carried out on a cluster node equipped with an Intel 16-core CPU with 192 GB memory and two NVIDIA GeForce P100 GPUs.

Table 4.7. List of experiments for feature extraction using RAEN-1 and RAEN-2

Experiment	Network	alfa	Encode layer	Encode layer	# of points in a bag	# of points in a plot	Training time (hh:mm:ss)
Ex.1	RAEN-1	0 (unsupervised)	Bi-LSTM (64, 8)	Bi-LSTM (8, 48)	(32, 1)	1024	5:26:19
Ex.2	RAEN-1	0 (unsupervised)	Bi-LSTM (32, 8)	Bi-LSTM (32, 8)	(16, 1)	1024	7:24:02
Ex.3	RAEN-1	10 (supervised)	Bi-LSTM (32, 8)	Bi-LSTM (32, 24)	(16, 1)	1024	7:54:19
Ex.4	RAEN-2	0 (unsupervised)	Bi-LSTM (16, 8)	Bi-LSTM (96)	(16, 4)	1024	5:33:04
Ex.5	RAEN-2	10 (supervised)	Bi-LSTM (16, 8)	Bi-LSTM (96)	(16, 4)	1024	5:42:12
Ex.6	RAEN-2	0 (unsupervised)	Bi-LSTM (16, 8)	Bi-LSTM (24)	(8, 1)	1024	11:03:10
Ex.7	RAEN-2	10 (supervised)	Bi-LSTM (16, 8)	Bi-LSTM (24)	(8, 1)	1024	10:05:46

The differences in X, Y, and Z values of the reconstructed points from the original point cloud for an example plot from SbDivTc-18 experiment using RAEN-1 and 2 are shown in Figure 4.12 and Figure 4.13, respectively. Based on these figures, the reconstruction error using RAEN-2 is slightly lower than those of RAEN-1, especially in Z. The figures also show that the reconstruction error in X and Y is much higher for the points in the middle of the sequence. Note that the RNN networks tend to eventually forget the first inputs during the training, especially when the length of the sequence is large. To avoid this issue, we used bidirectional LSTM cells for the proposed RAEN-1 and 2. This is the reason that the error for the points at the beginning and the end of the sequence are smaller than those in the middle. This issue is also more severe in the RAEN-1 as its hidden layers have more inputs compared with RAEN-2. The extracted features are also shown in the figures, which indicates some of them have the same value for all the dates, but comparing the features from different plots showed that they can be very different for different plots.

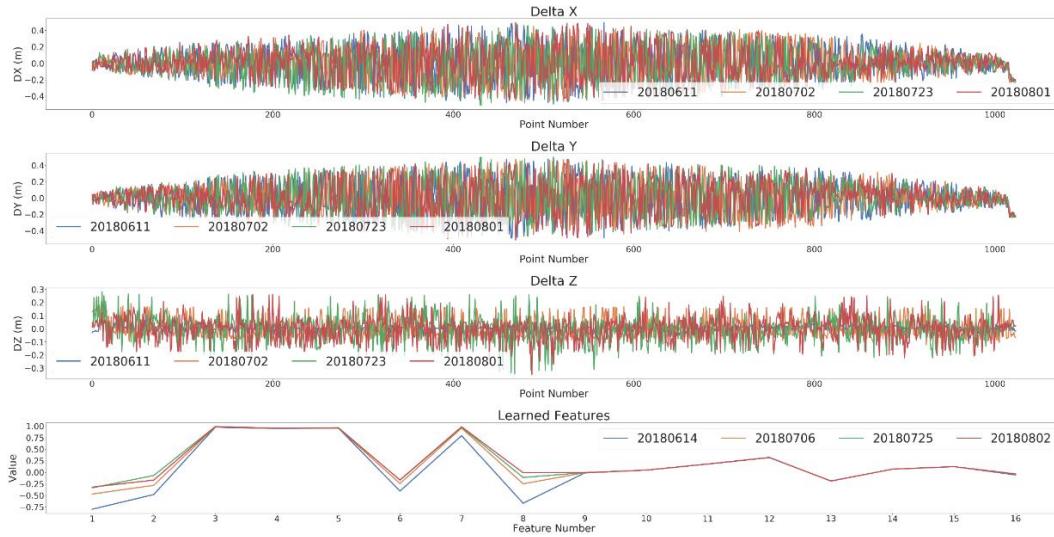


Figure 4.12. Error of reconstruction in X, Y, and Z for the points in an example plot with 1024 point as well as extracted features for one variety of the SbDivTv-18 experiment using RAEN-1 trained on the same experiment data

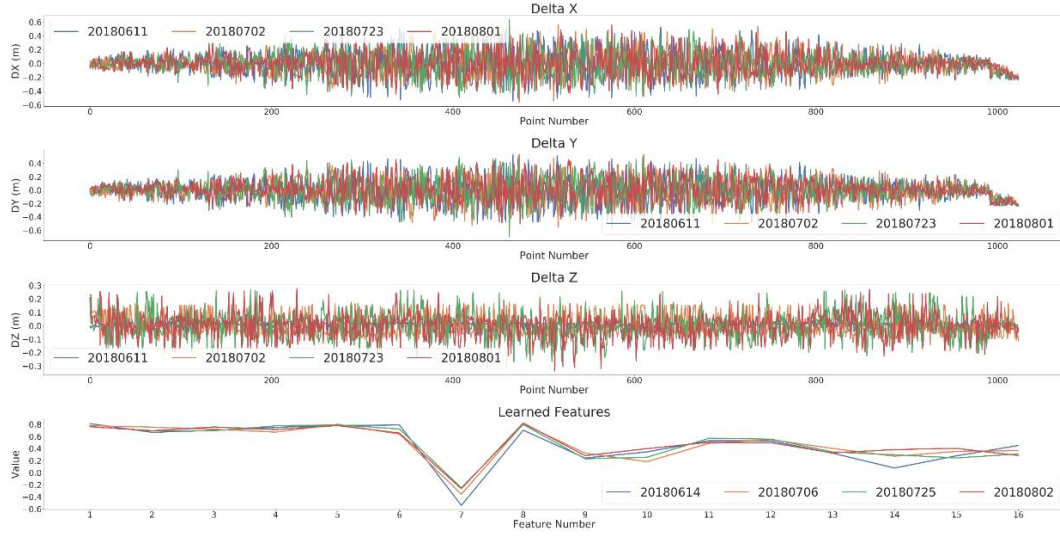


Figure 4.13. Error of reconstruction in X, Y, and Z for the points in an example plot with 1024 points as well as extracted features for one variety of the SbDivTv-18 experiment using RAEN-1 and RAEN-2 trained on the same experiment data.

Figure 4.14 and Figure 4.15 show the RMSE in X, Y, and Z for the reconstructed point clouds for all the plots in the SbDivTc-18 experiment using RAEN-1 (Ex. 2) and RAEN-2 (Ex.4). Based on these figures, the RMSE in X, Y, and Z were about 2-3 cm smaller using RAEN-2.

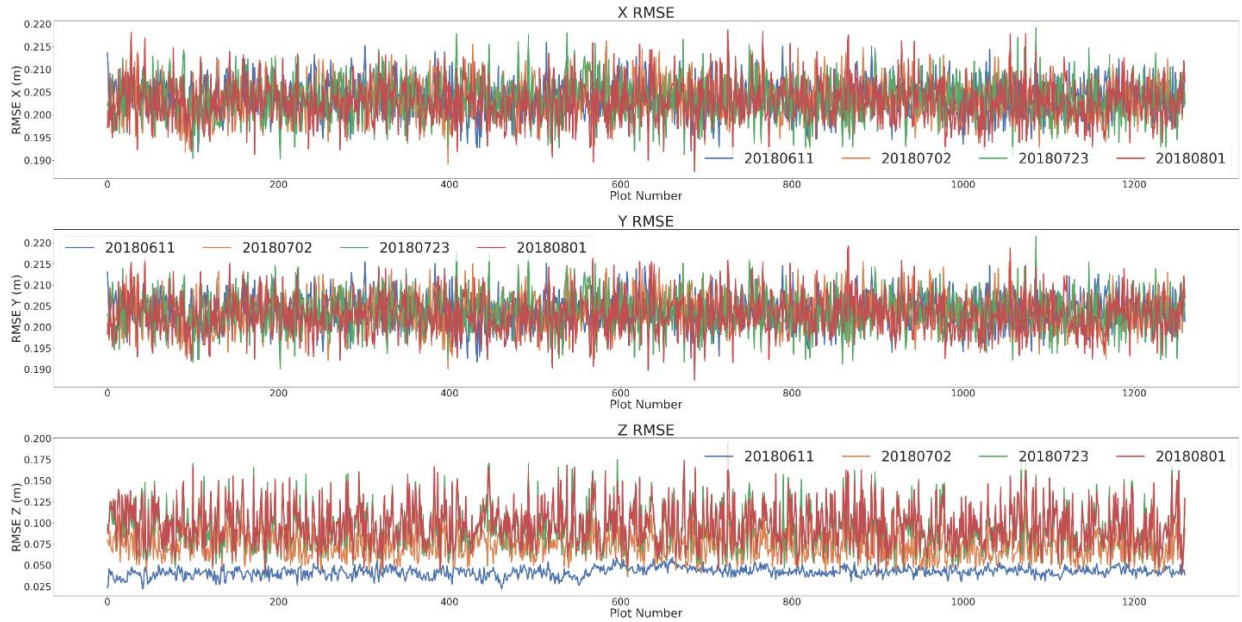


Figure 4.14. Reconstruction X, Y, and Z RMSE for all the plots of SbDivTv-18 experiment using RAEN-1 trained on the same experiment data.

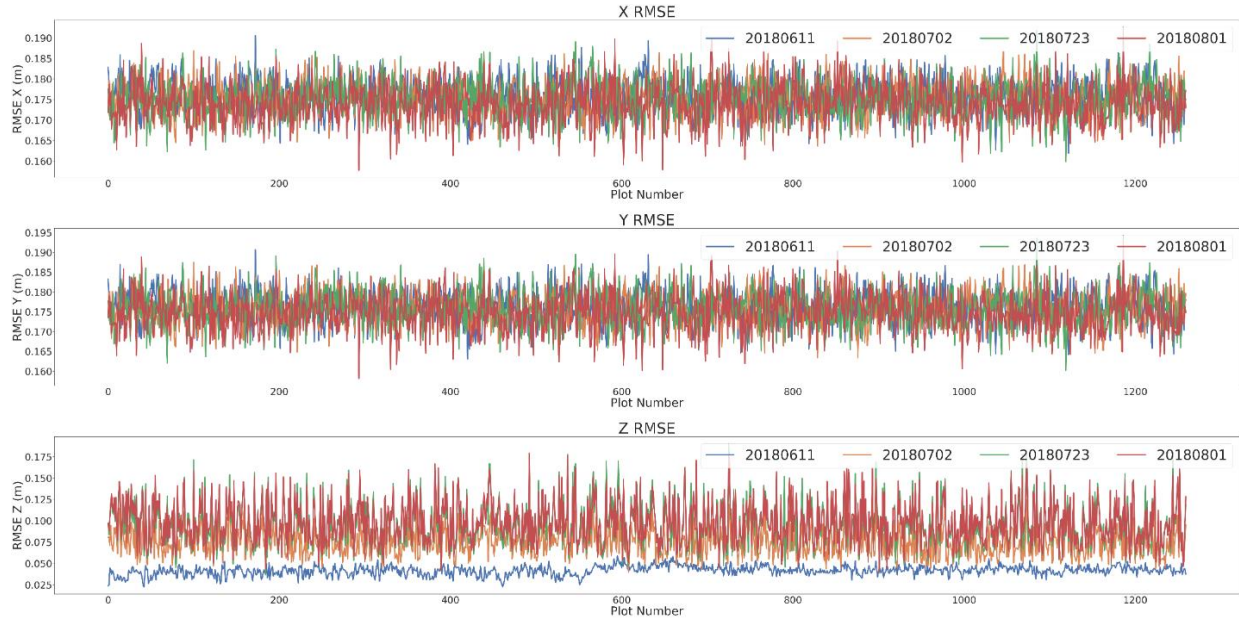


Figure 4.15. Reconstruction X, Y, and Z RMSE for all the plots of SbDivTv-18 experiment using RAEN-2 trained on the same experiment data.

Table 4.8 shows the  $R^2$  of the biomass predictions using the SVR models with the features extracted from the seven experiments in Table 4.7 using RAEN-1 and 2. Here, the most accurate predictions were obtained using the unsupervised training. Also, in most cases, the predictions were slightly more accurate than using RAEN-1 compared with RAEN-2. Also, comparing these results with those of CAEN and FAEN (Table 4.4 and Table 4.5) indicates that the new results are slightly more accurate. A potential reason could be the fact that here we used all the points' information including X, Y, and Z coordinates, while the input features for the FAEN and CAEN were calculated from Z only.

Table 4.8.  $R^2$  values of SVR-based biomass prediction (of field-based results) using latent LiDAR-based features extracted by the proposed **RAEN-1** and **RAEN-2** trained on different experimental data sets and tested on all experiments. Blue and green colors show unsupervised and supervised, respectively. The maximum  $R^2$  value for each row is highlighted.

Train \ Test		SbDiv-17							SbDivTc-18							SbDivTc-19						
		RAEN-1				RAEN-2			RAEN-1				RAEN-2			RAEN-1				RAEN-2		
		Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
2017	HyCal-17	0.65	0.55	0.35	0.64	0.58	0.59	-0.07	0.55	0.64	0.37	0.72	0.54	0.73	0.58	0.46	0.62	0.53	0.65	0.55	0.65	0.71
	InCal-17	0.27	0.29	0.19	0.26	0.34	0.21	0.32	0.31	0.32	0.10	0.19	0.0	0.22	0.11	0.36	0.32	0.10	0.28	0.12	0.19	0.16
	SbBAP-17	0.45	0.53	0.34	0.41	0.39	0.44	0.31	0.43	0.49	0.35	0.35	0.36	0.44	0.46	0.41	0.49	0.44	0.42	0.39	0.45	0.46
	SbDiv-17	0.52	0.51		0.52		0.50		0.49	0.53	53	49	45	45	0.45	0.52	0.52	0.52	0.51	0.48	0.48	0.50
2018	HyCal-18	0.50	0.65	0.25	0.62	0.42	0.66	0.47	0.71	0.70	0.61	0.67	0.34	0.67	0.66	0.49	0.48	0.30	0.61	0.56	0.64	0.71
	InCal-18	0.20	0.23	-0.02	0.15	0.29	0.02	0.14	0.20	0.02	-0.36	0.10	-0.08	-0.08	-0.13	0.19	0.29	0.08	0.20	-0.06	0.13	-0.04
	TcCal-18	0.52	0.48	0.09	0.50	0.44	0.44	0.33	0.53	0.55	0.48	0.45	0.86	0.39	0.53	0.40	0.42	0.36	0.40	0.27	0.47	0.43
	SbDivTc-18	0.41	0.42	0.41	0.40	0.39	0.40	0.38	0.41	0.43		0.38		0.47		0.37	0.41	0.42	0.39	0.40	0.39	0.39
2019	SbDCal-19	0.50	0.57	0.23	0.47	0.42	0.53	0.34	0.49	0.60	0.49	0.49	0.51	0.53	0.60	0.24	0.48	0.45	0.50	0.48	0.56	0.58
	SbDivTc-19	0.36	0.36	0.34	0.35	0.37	0.37	0.34	0.36	0.36	0.35	0.33	0.34	0.35	0.33	0.35	0.36		0.34		0.37	

The highest  $R^2$  values for each experiment using the proposed auto-encoder networks are provided in Table 4.9, which shows for hyperspectral data, the FAEN provided slightly more accurate results for the SbBAP-17 experiment while CAEN resulted higher  $R^2$  for the SbDiv-17 and SbDivTc-19 experiments. For SbDivTc-18, the CAEN-2d provided slightly most accurate results. For LiDAR data, the FAEN provided the maximum  $R^2$  values for the prediction for most of the experiments which indicates the original features, including height-based, canopy cover, and volume features were more informative for biomass predictions.

Table 4.9. Maximum  $R^2$  values of SVR-based biomass prediction using latent features extracted from hyperspectral and LiDAR data by the proposed networks in this study. The maximum  $R^2$  value for each row and from each data type is highlighted.

Experiment	Hyperspectral			LiDAR			
	FAEN	CAEN-1d	CAEN-2d	FAEN	CAEN-1d	RAEN-1	RAEN-2
SbBAP-17	0.55	0.53	0.45	0.54	0.51	0.53	0.45
SbDiv-17	0.68	0.70	0.62	0.65	0.61	0.53	0.52
SbDivTc-18	0.64	0.64	0.65	0.51	0.52	0.43	0.47
SbDivTc-19	0.63	0.64	0.55	0.53	0.49	0.36	0.37

### 4.6.3 Multi-year Biomass Prediction

In the previous section, the training and validation of the SVR models were performed on sample data of the same experiment. In this section, the SVR and RNN models were trained for each experiment, and the models are then used for biomass prediction of all the experiments. For all predictions in this section, the features learned from the proposed FAEN (trained on all the data sets of all experiments) were used as the input features for the predictive models as we believe they were more robust based on the results provided in previous sections.

Table 4.10 shows the results of biomass prediction using SVR, RNN with GRUs (RNN-GRU) and LSTMs (RNN-LSTM) models trained on each experiment and validated on all other experiments. The highest  $R^2$  for each experiment was achieved when the training and validation were performed on the same experiment (splitting the sample data into training and validation categories and using 3-fold cross validation). The number of iterations for RNN-1 was set to 100,000. Figure 4.16 shows the predictions of an LSTM model that was trained and tested on three experiments from 2017 to 2019. It shows that although the  $R^2$  is negative in some cases, a relationship between the predicted and ground reference biomass data is obvious.

Table 4.10. The  $R^2$  value (relative to field based measurements) of biomass predictions using SVR and RNN-1 with GRU and LSTM units trained on three experimental data sets and tested on all experiments.

	Train \ Test	SbDiv-17			SbDivTc-18			SbDivTc-19		
		SVR	GRU	LSTM	SVR	GRU	LSTM	SVR	GRU	LSTM
2017	HyCal	0.72	0.46	0.52	0.38	0.37	0.70	-0.08	-0.41	-0.14
	InCal	0.20	-0.73	-0.63	-6.30	-0.49	0.23	-4.10	-4.52	-3.45
	SbBAP	-0.58	-0.91	-0.96	-3.43	-0.72	-0.98	0.01	-0.35	-0.35
	SbDiv	0.77	0.73	0.75	-20.79	0.49	0.43	-1.81	-1.03	-0.86
2018	HyCal	0.37	-0.09	-0.40	0.12	0.52	0.38	-0.13	0.56	0.53
	InCal	0.42	0.68	0.57	-0.92	0.47	0.19	-9.05	0.08	0.18
	TcCal	0.12	0.41	0.42	0.69	0.69	0.70	-6.60	-0.90	-0.80
	SbDivTc	0.29	0.41	0.38	0.68	0.66	0.67	-6.62	-1.86	-1.62
2019	SbDivTcCal	-2.94	-0.05	-0.76	-53.03	-1.87	-2.99	0.65	0.56	0.58
	SbDivTc	-6.33	0.03	-1.21	-26.02	-2.74	-5.02	0.69	0.65	0.66

Table 4.10 indicates that when the training and validation are performed with different experiments of the same year, predictions are more accurate than training and validation with different experiments of different years. Training on the SbDiv-17 and validating on 2018

experiments yielded more accurate predictions than validating on 2019 experiments. In general, the  $R^2$ s of predictions with RNN-1 were higher than those with SVR. The possible explanations for this are that RNNs are more powerful than SVRs in modeling time series data, and that with RNN, the training can be controlled to avoid the overfitting by early termination of the training. RNN-GRU provided marginally better results than RNN-LSTM which could be because LSTM is more complex than GRU and requires learning of more parameters. Also, predictions for the HyCal-17 and HyCal-18, the models trained on SbDiv-17 and SbDivTc-18 experiment provided more accurate results, respectively.

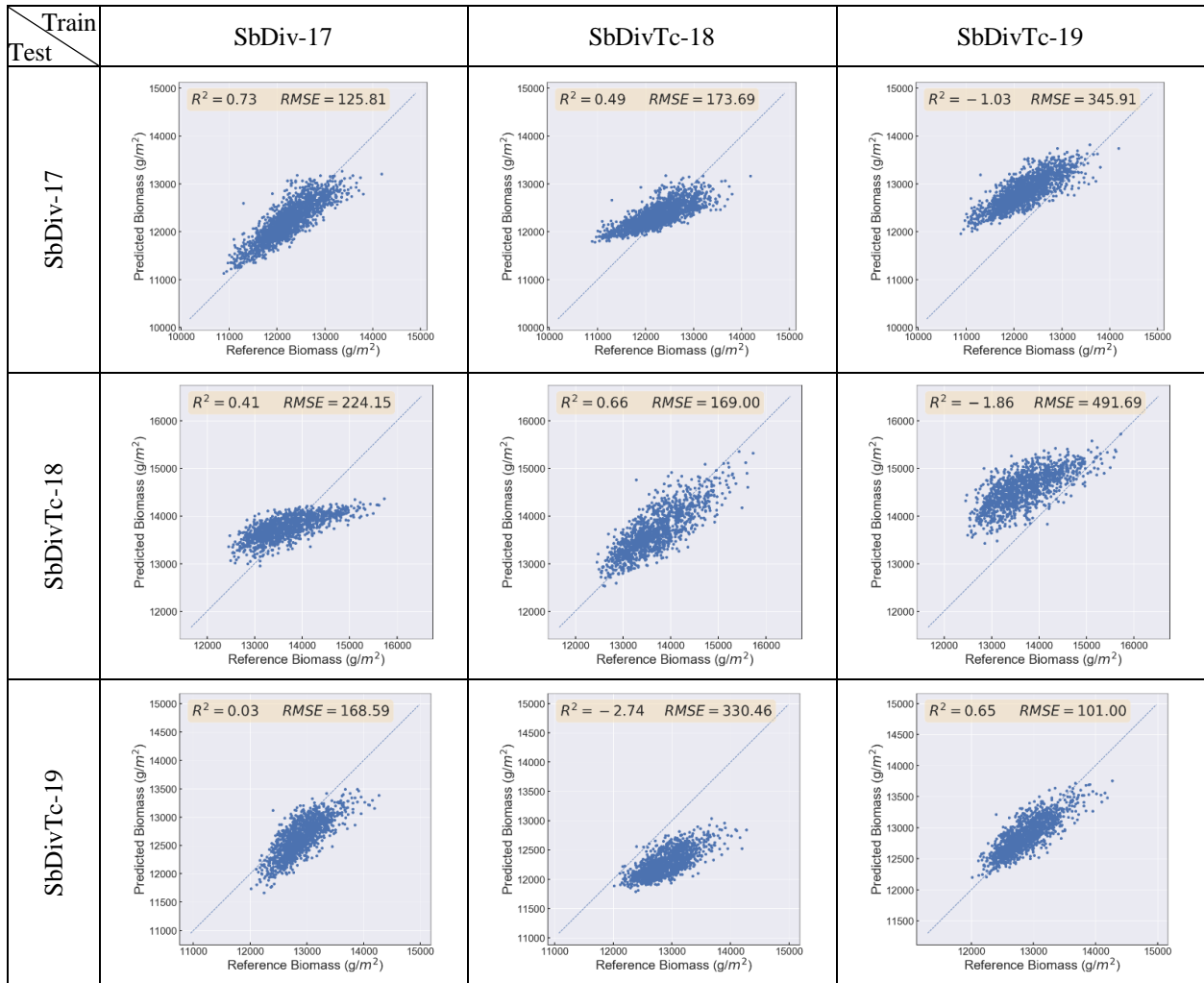


Figure 4.16. Biomass prediction results of RNN-1 with LSTM units trained and tested on different experimental data sets.

#### 4.6.4 Multi-time Predictive Models

The hyperspectral and LiDAR data in this study were not always acquired simultaneously. Generally, fewer hyperspectral data sets are available as hyperspectral data should be acquired during cloud-free days near solar noon. Our goal in this paper is to develop a predictive model for each field that can estimate biomass at any time during the season using the available RS data including hyperspectral and LiDAR. Such a model is useful for early selection of varieties and in applications such as radiation use efficiency estimation in agriculture where multiple measurements of yield/biomass in a season are required.

In this section, both LSTM and GRU models are developed for multi-temporal prediction of biomass using the RNN-2 network shown in Figure 4.8. The proposed network includes two sets of RNN cells (either GRU or LSTM based cells in this study): one set for features derived from hyperspectral data and one for the LiDAR-based features. This network can accommodate the differing number of data sets from each source as input. The final outputs of the two RNN cells are concatenated and considered as the input for the fully connected layer which outputs the prediction results of the network.

Figure 4.17 shows the results of biomass prediction for the three trials using the proposed RNN with LSTM and GRU cells. The training and evaluation for each network are accomplished using 3-fold cross validation. For each fold, one third of the plots are considered as the test set, and the remaining samples are used for training (90%) and validation (10%). For each trial, the network is only trained once using the destructively sampled biomass data collected on each date as a predictive variable, and the RS data collected prior to that date as the predictor variable. As noted in Table 4.2, there are 5, 4, and 2 biomass data sets available for HyCal-2017, HyCal-2018, and TcCal-2019 trials, which means that each plot for these experiments provides 5, 4, and 2 sample data sets for the network. Once the network is trained, it can be used to estimate biomass at any given time during the season using the remotely sensed hyperspectral and LiDAR data, as well as concurrent weather information.

From Figure 4.17, the LSTM-based network provided more accurate results than the GRU-based network, especially for the TcCal-2019 experiment where fewer biomass data sets were available compared to the other experiments. To evaluate the performance of the networks on each biomass data set individually, the normalized Root Mean Square Error (nRMSE) of the biomass predictions for each date in the HyCal-2017 and HyCal-2018 are shown in Figure 4.18. From

Figure 4.18, the network for each experiment is trained well as the nRMSE values for most of the dates are in the same range.

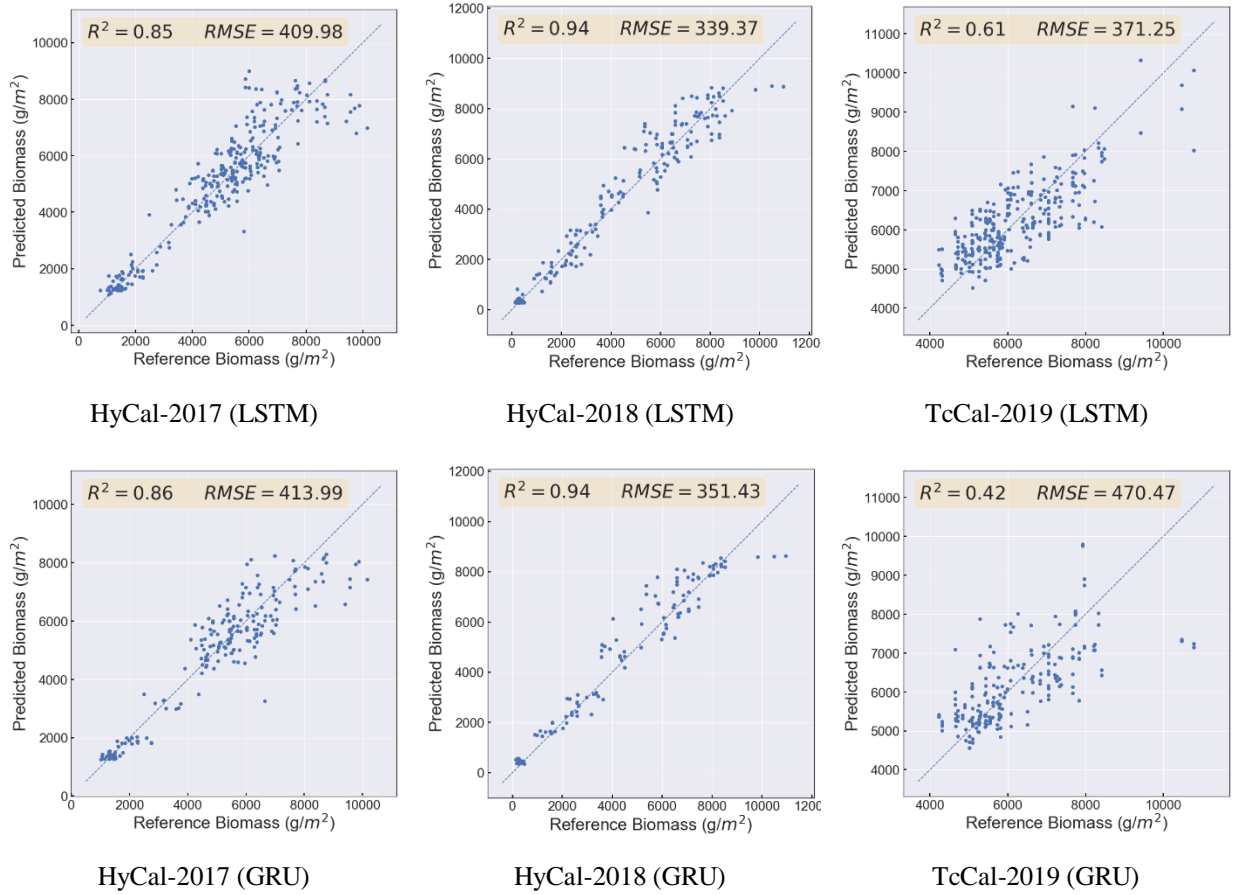


Figure 4.17. Biomass prediction results for the three trials using the proposed RNN with LSTM and GRU cells.

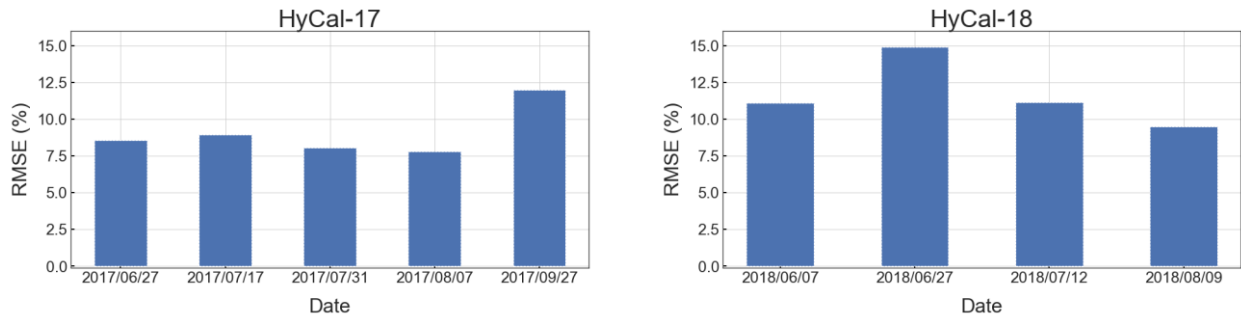


Figure 4.18. nRMSE of biomass estimation for individual dates for HyCal-2017 and HyCal-2018 experiments using the proposed network with LSTM cells.

## 4.7 Conclusions

Five deep learning-based networks were investigated in this study for unsupervised and supervised feature learning from hyperspectral and LiDAR data. The reduction of extracted features from hyperspectral and LiDAR features using FAEN and CAEN showed that it is possible to train the network on an experiment and use it for feature learning on other experiments. Both FAEN and CAEN provided features with similar prediction accuracy when they were used in SVR models. We also investigate the potential of using auto-encoders for feature extraction from the original data. We investigated using a 2d-CNN based network for hyperspectral data in both supervised and unsupervised training procedures. The results of using the extracted features from this network in SVR models were slightly lower than those of FAEN and CAEN. For LiDAR data, we proposed using RNN based auto-encoders and evaluated them in terms of the accuracy of the prediction results using the extracted features, reconstruction error, and training time. For hyperspectral data, supervised feature extraction provided more accurate predictions, while the features extracted from LiDAR data in an unsupervised training yielded more accurate predictions.

Two recurrent neural networks (RNN-1 and RNN-2) were then developed with hyperspectral, LiDAR, and weather data inputs to predict sorghum biomass. The results of using RNN-1 for training on one experiment and predicting biomass for other experiments with different types of sorghum varieties illustrate the potential of the network for biomass prediction, and the challenges relative to small sample sizes, including weather and sensitivity to the associated ground reference information. RNN-2 was designed in a way that can accommodate different numbers of hyperspectral and LiDAR data sets, and thus, it was capable of estimating biomass at multiple times during the growing season.

## **5. CONCLUSIONS AND RECOMMENDATIONS**

### **5.1 Summary**

Currently, phenotyping provides the critical information for selecting the “best” genotypes with respect to the trait(s) of interest, but is considered as the bottleneck for the plant breeding chain as traditional methods are time-consuming and expensive. Fast and accurate phenotyping can be achieved by using predictive models developed using multiple data sources, such as RS data and genomic information. This dissertation focused on developing models for\_high-throughput phenotyping based on high spatial, spectral, and temporal resolution RS data, genomic information, and weather data. A key to good predictive models is extracting the appropriate features that are predictors of the trait of interest so that they can support the machine learning algorithms. Feature extraction was investigated in this thesis in that context, where various novel features, as well as traditional features, were extracted from the hyperspectral imagery and LiDAR point cloud data for regression-based models. Then, the impact of the regression method, data source, timing of RS and field-based biomass reference data acquisition, and the number of samples on the prediction results were investigated. To accelerate the breeding cycle and minimize the need for sample data, the possibility of developing multi-year predictive models for which the data from earlier years were used for training models, and then predicting traits for experiments in future year(s) was explored. Having robust features for experiments conducted in multiple years, particularly as the data were captured by different sensors/platforms, was critical in multi-year prediction methods. Thus, more advanced supervised and unsupervised feature learning tools through deep neural networks were developed. Finally, advanced predictive models to accommodate high dimensional, multi-modal, multi-temporal data.

### **5.2 Key Objectives and Findings**

The objectives and contributions of the dissertation are summarized for each chapter as follows:

- Chapter 2: In this chapter, the objective was to provide a robust framework for predicting sorghum biomass which is suitable for plant breeding research and industrial applications. To accomplish this, we: 1) evaluated multiple prediction models, 2) investigated the impact

of the timing of the RS data acquisition on prediction relative to days since sowing, and 3) evaluated the importance of the features extracted from the data. We developed prediction models using three nonlinear regression models for nine experiments conducted in the 2017 and 2018 growing seasons. Based on the experiments conducted in this chapter on trials, which included multiple sorghum varieties with different sample sizes, nitrogen and photosynthesis related features extracted from hyperspectral data and geometric based features derived from the LiDAR data provided reliable and accurate prediction of biomass. Our ANOVA results showed that the data source was the most important factor impacting the prediction accuracy with 69% significance, versus 28% significance for the regression method. The number of training samples for the prediction was another important factor for determining the accuracy of the predictions. Generally, the PLSR method provided more accurate prediction models when the number of training samples was limited. With increasing numbers of samples in training, the rate of increase in the accuracy of the SVR models was higher than PLSR. We also evaluated the prediction models with respect to the time of the RS data acquisition and the time of harvest. The end-of-season biomass predictions for the experiments with diverse varieties were more reliable and accurate than the mid-season predictions, as more varieties in the field were at the same stage of growth. With respect to the remote sensing data, the best results were obtained using the RS data from the whole season; however, the models developed using mid-season data (DAS of ~60 to 80) were also able to provide useful results, particularly for screening varieties. To our knowledge, this work is the most thorough, rigorous evaluation of remotely sensed features for traditional regression models, based on current literature.

- Chapter 3: In this chapter, we focused on two main objectives; 1) developing predictive models using the RS data collected multiple times during a season in previous years, and 2) investigating the possibility of developing the same models using the RS data collected until mid-season rather than the full season. We focused on developing regression based models using the RS data combined with the genomic information as well as weather data to predict the end of season sorghum biomass yield. The models based on one and two years of data were used to predict the biomass yield for the future years based on using sample data from previous years. The results showed the high potential of the models for biomass and biomass rank predictions, with models based on two years of data having

higher  $R^2$  values. We investigated the use of wavelet decomposition and PCA to extract a smaller number of features from the markers. The results indicated that level 4 and 6 wavelets generally provided models with higher accuracies than using all markers. Predictive models were developed using the RS data collected until mid-season, rather than the full season. The results showed that the rank of biomass can be predicted with  $R^2$  values of around 0.65-0.70 using the RS data until 60 DAS. The role of genetics in data driven predictive modeling is a topic of significant research from multiple points of view. Extraction of meaningful features is a critical problem. The investigation of wavelets is not new for modeling markers, but the integration of wavelet-based features with comprehensive remote sensing based measurements for predictive models using this unique data set is a contribution.

- Chapter 4: Extraction of appropriate features, along with using a powerful predictive model, is required for accurate prediction of complex phenotypes such as biomass. Thus, in this chapter, the first objective was to develop deep neural networks for robust feature extraction. The second goal was to develop predictive models for representing sequential data, which could effectively exploit long-term dependencies. For the first objective, five deep learning-based networks were investigated for unsupervised and supervised feature learning from hyperspectral and LiDAR data. The reduction of extracted features from hyperspectral and LiDAR features using a fully connected auto-encoder network (FAEN) and a convolutional auto-encoder network (CAEN) showed that it is possible to train the network on an experiment and use it for feature learning on other experiments. Both the FAEN and CAEN architectures provided features with similar prediction accuracy when they were used in SVR models. We also investigated the potential of using autoencoders for feature extraction from the original data, as they require no prior definition of features (e.g. indices, derivatives, integral features) and may provide new insights. We investigated using a 2d-CNN based network for hyperspectral data in both supervised and unsupervised training procedures. The  $R^2$  values of models developed using the extracted features from this network in SVR models were slightly lower than those of FAEN and CAEN. For LiDAR data, we proposed using RNN based autoencoders and evaluated them in terms of the accuracy of the prediction results using the extracted features, reconstruction error, and training time.

- Two recurrent neural networks (RNN-1 and RNN-2) were then developed with hyperspectral, LiDAR, and weather data inputs to predict sorghum biomass. RNN-2 was designed to accommodate different numbers of hyperspectral and LiDAR data sets, and thus, it was capable of full utilization of the remote sensing data and estimating biomass at multiple times during the growing season. The results of using these models for training on one experiment and predicting biomass for other experiments with different types of sorghum varieties illustrated the potential of the network for biomass prediction, and the challenges relative to small sample sizes, including weather and sensitivity to the associated ground reference information. To our knowledge, this is the first time these deep learning architectures which incorporate multi-modality data and weather have been utilized for predicting biomass yield.

### 5.3 Future Work

The current work may be extended in the following directions:

- Transfer learning:  
Transfer learning approaches can be utilized to achieve an effective strategy for improving multi-year predictions; a possible way would be using domain adaptation techniques so that the models trained on an experiment can be adjusted and used for another experiment on the same or future years. Few shot learning, which updates the proposed networks using a few samples from the target experiments, is a simple strategy that should be investigated.
- Incorporate crop models:  
Developing crop models, APSIM for example, is another way of predicting (simulating) crop traits such as biomass. If they are well parameterized, they can be used to estimate traits for the plots/dates that sampling is not conducted. This can generate “augmented” data that is useful for training the models if more sample data is needed. Crop models also can be used for extracting latent features from some of data sources such as weather data. Extracted features from weather data can be inputs to the predictive models; however, it would be difficult for the models to learn from the weather features as their impact on the traits is complex, and we typically only have consistent local sample data from a few years. Alternatively, the crop models are developed based on many years of weather data and are able to convert those data into latent features that are more related to the biomass.

- Multi-temporal feature extraction via RNNs

In our proposed feature learning networks, we extracted encoded features for each plot and each date regardless of other dates. Instead, it would be possible to use RNN networks, for example RCNN, to extract only a few latent features from all dates at the same time. This would reduce the number of extracted features, which is important for some predictive models, especially when the quantity of sample data is limited. Also, many of the features extracted from successive weeks are highly correlated, which can negatively impact the predictive models.

- Application to diverse datasets:

It would be appropriate to evaluate our models on the data from multiple locations, or use the same strategies for other traits of sorghum such as LAI, or APSIM parameters. Also, we believe that our models (with minor modifications) can be applied for yield-related traits similar to biomass in other crops.

- Improve parameter selection strategies:

In this research, we determined the hyperparameters for our neural networks in a simple trial-and-error procedure. It would be beneficial to use more formal optimization approaches to search a greater range of the possible options for each parameter and/or models the behavior of each parameter.

- Investigate other data sources:

Lastly, as was noted, having appropriate input features is the key to an accurate prediction. Considering all the factors that affect the crops and RS data in the multi-year predictive models can increase the prediction accuracy. Thermal data, available from UAV-based thermal imagery, and soil related information (i.e. soil moisture and composition) are example data sources that could potentially be included in the multi-year predictive models to increase robustness across different sites. Thermal data would be especially beneficial for drought/heat stress environments.

## REFERENCES

- Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, 19(1), 52–61.
- Asner, G. P., & Martin, R. E. (2008). Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. *Remote Sensing of Environment*, 112(10), 3958–3970.
- Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., & Shiu, S. H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, 9(11), 3691–3702.
- Barclay, V. J., Bonner, R. F., & Hamilton, I. P. (1997). Application of Wavelet Transforms to Experimental Spectra: Smoothing, Denoising, and Data Set Compression. *Analytical Chemistry*, 69(1), 78–90.
- Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 114, pp. 24–31).
- Bhat, B. V. (2019). Breeding Forage Sorghum. In *Breeding Sorghum for Diverse End Uses* (pp. 175–191). Woodhead Publishing.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Carter, G. A. (1994). Ratios of leaf reflectances in narrow wavebands as indicators of plant stress. *International Journal of Remote Sensing*, 15(3), 517–520.
- Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6), 2094–2107.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734.
- Chu, T., Starek, M. J., Brewer, M. J., Murray, S. C., & Pruter, L. S. (2018). Characterizing canopy height with UAS structure-from-motion photogrammetry—results analysis of a maize field trial with respect to multiple factors. *Remote Sensing Letters*, 9(8), 753–762.

- Clevers, J. G. P. W. (1994). *Imaging Spectrometry in Agriculture - Plant Vitality And Yield Indicators BT - Imaging Spectrometry — a Tool for Environmental Observations* (J. Hill & J. M  gier (eds.); pp. 193–219). Springer Netherlands.
- Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., & Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder’s equation. *Theoretical and Applied Genetics*, 132(3), 627–645.
- Combs, E., & Bernardo, R. (2013). Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *The Plant Genome*, 6(1), plantgenome2012.11.0030.
- Daughtry, C. S. T., Walthall, C. L., Kim, M. S., Colstoun, E. B. de, & McMurtrey, J. E. (2000). Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sensing of Environment*, 35, 229–239.
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), 499–510.
- de Almeida, C. T., Galv  o, L. S., Arag  o, L. E. de O. C. e, Ometto, J. P. H. B., Jacon, A. D., Pereira, F. R. de S., Sato, L. Y., Lopes, A. P., Gra  a, P. M. L. de A., Silva, C. V. de J., Ferreira-Ferreira, J., & Longo, M. (2019). Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms. *Remote Sensing of Environment*, 232, 111323.
- Demetriades-Shah, T. H., Steven, M. D., & Clark, J. A. (1990). High resolution derivative spectra in remote sensing. *Remote Sensing of Environment*, 33(1), 55–64.
- Desta, Z. A., & Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9), 592–601.
- Duan, S. B., Li, Z. L., Wu, H., Tang, B. H., Ma, L., Zhao, E., & Li, C. (2014). Inversion of the PROSAIL model to estimate leaf area index of maize, potato, and sunflower fields from unmanned aerial vehicle hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, 26(1), 12–20.

- Duan, T., Chapman, S. C., Guo, Y., & Zheng, B. (2017). Dynamic monitoring of NDVI in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Research*, 210, 71–80.
- Ecartot, M., Compan, F., & Roumet, P. (2013). Assessing leaf nitrogen content and leaf mass per unit area of wheat in the field throughout plant cycle with a portable spectrometer. *Field Crops Research*, 140(1), 44–50.
- Eitel, J. U. H., Magney, T. S., Vierling, L. A., Greaves, H. E., & Zheng, G. (2016). An automated method to quantify crop height and calibrate satellite-derived biomass using hypertemporal lidar. *Remote Sensing of Environment*, 187, 414–422.
- Farabet, C., Couprie, C., Najman, L., & Lecun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1915–1929.
- Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, 154(1), 102–114.
- Feng, W., Guo, B. Bin, Wang, Z. J., He, L., Song, X., Wang, Y. H., & Guo, T. C. (2014). Measuring leaf nitrogen concentration in winter wheat using double-peak spectral reflection remote sensing data. *Field Crops Research*, 159, 43–52.
- Féret, J. B., François, C., Gitelson, A., Asner, G. P., Barry, K. M., Panigada, C., Richardson, A. D., & Jacquemoud, S. (2011). Optimizing spectral indices and chemometric analysis of leaf chemical properties using radiative transfer modeling. *Remote Sensing of Environment*, 115(10), 2742–2750.
- Fernandes, S. B., Dias, K. O. G., Ferreira, D. F., & Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, 131(3), 747–755.
- Foster, A. J., Kakani, V. G., & Mosali, J. (2017). Estimation of bioenergy crop yield and N status by hyperspectral canopy reflectance and partial least square regression. *Precision Agriculture*, 18(2), 192–209.

- Fu, R., Zhang, Z., & Li, L. (2017). Using LSTM and GRU neural network methods for traffic flow prediction. *Proceedings - 2016 31st Youth Academic Annual Conference of Chinese Association of Automation, YAC 2016*, 324–328.
- Gamon, J. A., Peñuelas, J., & Field, C. B. (192 C.E.). A Narrow-Waveband Spectral Index That Tracks Diurnal Changes in Photosynthetic Efficiency. *Remote Sensing of Environment*, 44, 35–44.
- Gao, B.-C. (1996). NDWI A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space. *Remote Sensing of Environment*, 58, 257–266.
- Gerik, T., Bean, B., & Vanderlip, R. (2003). Sorghum Growth and Development. *Texas FARMER Collection*.
- Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58, 289–298.
- Gitelson, A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology, B: Biology*, 22(3), 247–252.
- Gouache, D., Beauchêne, K., Mini, A., Fournier, A., de Solan, B., Baret, F., & Comar, A. (2016). Applying remote sensing expertise to crop improvement: progress and challenges to scale up high throughput field phenotyping from research to industry. *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping*, 9866, 986604.
- Habib, A., Zhou, T., Masjedi, A., Zhang, Z., Evan Flatt, J., & Crawford, M. (2018). Boresight Calibration of GNSS/INS-Assisted Push-Broom Hyperspectral Scanners on UAV Platforms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5), 1734–1749.
- Hasheminasab, S. M., Zhou, T., & Habib, A. (2020). GNSS/INS-Assisted structure from motion strategies for UAV-Based imagery over mechanized agricultural fields. *Remote Sensing*, 12(3), 1–35.
- He, F., Zhou, T., Xiong, W., Hasheminnasab, S. M., & Habib, A. (2018). Automated aerial triangulation for UAV-based mapping. *Remote Sensing*, 10(12).
- Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, 49(1), 1–12.

- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
- Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. *ArXiv:1506.02078*. <http://arxiv.org/abs/1506.02078>
- Kong, Y. L., Huang, Q., Wang, C., Chen, J., Chen, J., & He, D. (2018). Long short-term memory neural networks for online disturbance detection in satellite image time series. *Remote Sensing*, 10(3), 1–13.
- Kvålseth, T. O. (1985). Cautionary note about  $r^2$ . *The American Statistician*, 39(4), 279–285.
- LaForest, L., Hasheminasab, S. M., Zhou, T., Flatt, J. E., & Habib, A. (2019). New strategies for time delay estimation during system calibration for UAV-Based GNSS/INS-Assisted imaging systems. *Remote Sensing*, 11(15).
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. In *Nature* (Vol. 521, Issue 7553, pp. 436–444).
- Lee, D., Lee, W., Lee, Y., & Pawitan, Y. (2010). Super-sparse principal component analyses for high-throughput genomic data. *BMC Bioinformatics*, 11(2).
- Lewis, B., Smith, I., Fowler, M., & Licato, J. (2017). The robot mafia: A test environment for deceptive robots. *28th Modern Artificial Intelligence and Cognitive Science Conference, MAICS 2017*, 189–190.
- Li, F., Qiao, H., & Zhang, B. (2018). Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, 83, 161–173.
- Li, Jiaojiao, Zhao, X., Li, Y., Du, Q., Xi, B., & Hu, J. (2018). Classification of Hyperspectral Imagery Using a New Fully Convolutional Neural Network. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 292–296.
- Li, Jiating, Shi, Y., Veeranampalayam-Sivakumar, A. N., & Schachtman, D. P. (2018). Elucidating sorghum biomass, nitrogen and chlorophyll contents with spectral and morphological traits derived from unmanned aircraft system. *Frontiers in Plant Science*, 9(1406), 1–12.
- Li, X., Zhang, Y., Bao, Y., Luo, J., Jin, X., Xu, X., Song, X., & Yang, G. (2014). Exploring the best hyperspectral features for LAI estimation using partial least squares regression. *Remote Sensing*, 6(7), 6221–6241.

- Liang, L., Di, L., Zhang, L., Deng, M., Qin, Z., Zhao, S., & Lin, H. (2015). Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment*, 165, 123–134.
- Maes, W. H., & Steppe, K. (2019). Perspectives for Remote Sensing with Unmanned Aerial Vehicles in Precision Agriculture. *Trends in Plant Science*, 24(2), 152–164.
- Maimaitijiang, M., Ghulam, A., Sidike, P., Hartling, S., Maimaitiyiming, M., Peterson, K., Shavers, E., Fishman, J., Peterson, J., Kadam, S., Burken, J., & Fritschi, F. (2017). Unmanned Aerial System (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134, 43–58.
- Marshak, A., Knyazikhin, Y., Davis, A. B., Wiscombe, W. J., & Pilewskie, P. (2000). Cloud-vegetation interaction: use of normalized difference cloud index for estimation of cloud optical thickness. *Geophysical Research Letters*, 27(12), 1695–1698.
- Masjedi, A., Carpenter, N., Crawford, B., & Tuinstra, M. (2019). Prediction of Sorghum Biomass Using UAV Time Series Data and Recurrent Neural Networks. *IEEE Computer Society CVPR Computer Vision Problems in Plant Phenotyping (CVPPP)*, June.
- Masjedi, A., Zhao, J., Thompson, A. M., Yang, K. W., Flatt, J. E., Crawford, M. M., Ebert, D. S., Tuinstra, M. R., Hammer, G., & Chapman, S. (2018). Sorghum biomass prediction using uav-based remote sensing data and crop model simulation. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 7719–7722.
- McGlinchy, J., Van Aardt, J. A. N., Erasmus, B., Asner, G. P., Mathieu, R., Wessels, K., Knapp, D., Kennedy-Bowdoin, T., Rhody, H., Kerekes, J. P., Ientilucci, E. J., Wu, J., Sarrazin, D., & Cawse-Nicholson, K. (2014). Extracting structural vegetation components from small-footprint waveform lidar for biomass estimation in savanna ecosystems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(2), 480–490.
- McMurtrey, J. E., Chappelle, E. W., Kim, M. S., Meisinger, J. J., & Corp, L. A. (1994). Distinguishing nitrogen fertilization levels in field corn (*Zea mays* L.) with actively induced fluorescence and passive reflectance measurements. *Remote Sensing of Environment*, 47, 36–44.
- Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819–1829.

- Mou, L., Ghamisi, P., & Zhu, X. X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3639–3655.
- Nakaya, A., & Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Annals of Botany*, 110(6), 1303–1316.
- Nguy-Robertson, A. L., Peng, Y., Gitelson, A. A., Arkebauer, T. J., Pimstein, A., Herrmann, I., Karnieli, A., Rundquist, D. C., & Bonfil, D. J. (2014). Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. *Agricultural and Forest Meteorology*, 192–193, 140–148.
- Ogbaga, C. C., Bajhaiya, A. K., & Gupta, S. K. (2019). Improvements in biomass production: Learning lessons from the bioenergy plants maize and sorghum. *Journal of Environmental Biology*, 40(3), 400–406.
- Phua, M. H., Johari, S. A., Wong, O. C., Ioki, K., Mahali, M., Nilus, R., Coomes, D. A., Maycock, C. R., & Hashim, M. (2017). Synergistic use of Landsat 8 OLI image and airborne LiDAR data for above-ground biomass estimation in tropical lowland rainforests. *Forest Ecology and Management*, 406, 163–171.
- Potgieter, A. B., George-Jaeggli, B., Chapman, S. C., Laws, K., Suárez Cadavid, L. A., Wixted, J., Watson, J., Eldridge, M., Jordan, D. R., & Hammer, G. L. (2017). Multi-Spectral Imaging from an Unmanned Aerial Vehicle Enables the Assessment of Seasonal Leaf Area Dynamics of Sorghum Breeding Lines. *Frontiers in Plant Science*, 8(1), 1532–1546.
- Pugh, N. A., Horne, D. W., Murray, S. C., Carvalho, G., Malambo, L., Jung, J., Chang, A., Maeda, M., Popescu, S., Chu, T., Starek, M. J., Brewer, M. J., Richardson, G., & Rooney, W. L. (2018). Temporal Estimates of Crop Growth in Sorghum and Maize Breeding Enabled by Unmanned Aerial Systems. *The Plant Phenome Journal*, 1(1), 1–10.
- Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua*, 77–85.
- Rabier, C. E., Barre, P., Asp, T., Charmet, G., & Mangin, B. (2016). On the accuracy of genomic selection. *PLoS ONE*, 11(6), 1–23.

- Ravi, R., Lin, Y. J., Elbahnasawy, M., Shamseldin, T., & Habib, A. (2018). Simultaneous System Calibration of a Multi-LiDAR Multicamera Mobile Mapping Platform. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5), 1694–1714.
- Ribera, J., Chen, Y., Boomsma, C., & Delp, E. J. (2017). Counting plants using deep learning. *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 1344–1348.
- Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, 55, 95–107.
- Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8), 1627–1639.
- Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. In *Neural Networks* (Vol. 61, pp. 85–117).
- Scholz, M., & Selbig, J. (2007). Visualization and analysis of molecular data. *Methods in Molecular Biology (Clifton, N.J.)*, 358, 87–104.
- Serrano, L., Peñuelas, J., & Ustin, S. L. (2002). Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals. *Remote Sensing of Environment*, 81, 355–364.
- Shamshad, Mohd and Sharma, A. (2018). The Usage of Genomic Selection Strategy in Plant Breeding. In *Next Generation Plant Breeding* (pp. 93–108). Intechopen.
- Shao, G., Shao, G., Gallion, J., Saunders, M. R., Frankenberger, J. R., & Fei, S. (2018). Improving Lidar-based aboveground biomass estimation of temperate hardwood forests with varying site productivity. *Remote Sensing of Environment*, 204, 872–882.
- Shin, H. C., Orton, M. R., Collins, D. J., Doran, S. J., & Leach, M. O. (2013). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1930–1943.
- Sims, D. A., & Gamon, J. A. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *International Journal of Remote Sensing*, 81, 337–354.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. In *Statistics and Computing* (Vol. 14, Issue 3, pp. 199–222). Springer.

- Spindel, J. E., Dahlberg, J., Colgan, M., Hollingsworth, J., Sievert, J., Staggenborg, S. H., Hutmacher, R., Jansson, C., & Vogel, J. P. (2018). Association mapping by aerial drone reveals 213 genetic associations for *Sorghum bicolor* biomass traits under drought. *BMC Genomics*, *19*(1), 1–18.
- Stanton, C., Starek, M. J., Elliott, N., Brewer, M., Maeda, M. M., & Chu, T. (2017). Unmanned aircraft system-derived crop height and normalized difference vegetation index metrics for sorghum yield and aphid stress assessment. *Journal of Applied Remote Sensing*, *11*(2), 026035.
- Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M. H., & Kautz, J. (2018). SPLATNet: Sparse Lattice Networks for Point Cloud Processing. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2530–2539.
- Tang, Y., & Eliasmith, C. (2010). Deep networks for robust visual recognition. *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 1055–1062.
- Tao, C., Pan, H., Li, Y., & Zou, Z. (2015). Unsupervised Spectral-Spatial Feature Learning With Stacked Sparse Autoencoder for Hyperspectral Imagery Classification. *IEEE Geoscience and Remote Sensing Letters*, *12*(12), 2438–2442.
- Tattaris, M., Reynolds, M. P., & Chapman, S. C. (2016). A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. *Frontiers in Plant Science*, *7*(1131), 1–9.
- Thorp, K. R., Wang, G., Bronson, K. F., Badaruddin, M., & Mon, J. (2017). Hyperspectral data mining to identify relevant canopy spectral features for estimating durum wheat growth, nitrogen status, and grain yield. *Computers and Electronics in Agriculture*, *136*, 1–12.
- Thulin, S., Hill, M. J., Held, A., Jones, S., & Woodgate, P. (2014). Predicting Levels of Crude Protein, Digestibility, Lignin and Cellulose in Temperate Pastures Using Hyperspectral Image Data. *American Journal of Plant Sciences*, *05*(07), 997–1019.
- Ullah, S., Skidmore, A. K., Ramoelo, A., Groen, T. A., Naeem, M., & Ali, A. (2014). Retrieval of leaf water content spanning the visible to thermal infrared spectra. *ISPRS Journal of Photogrammetry and Remote Sensing*, *93*, 56–64.

- Vaglio Laurin, G., Chen, Q., Lindsell, J. A., Coomes, D. A., Frate, F. Del, Guerriero, L., Pirotti, F., & Valentini, R. (2014). Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 89, 49–58.
- Vaglio Laurin, G., Puletti, N., Chen, Q., Corona, P., Papale, D., & Valentini, R. (2016). Above ground biomass and tree species richness estimation with airborne lidar in tropical Ghana forests. *International Journal of Applied Earth Observation and Geoinformation*, 52, 371–379.
- Vastaranta, M., Holopainen, M., Karjalainen, M., Kankare, V., Hyypä, J., & Kaasalainen, S. (2014). TerraSAR-X stereo radargrammetry and airborne scanning LiDAR height metrics in imputation of forest aboveground biomass and stem volume. *IEEE Transactions on Geoscience and Remote Sensing*, 52(2), 1197–1204.
- Wang, T., Chen, P., Amaral, K., & Qiang, J. (2016). An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification. *ArXiv Preprint ArXiv:1609.03663*. <http://arxiv.org/abs/1609.03663>
- Wu, H., & Prasad, S. (2017). Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sensing*, 9(3), 298.
- You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 4559–4565.
- Yue, J., Feng, H., Yang, G., & Li, Z. (2018). A comparison of regression techniques for estimation of above-ground winter wheat biomass using near-surface spectroscopy. *Remote Sensing*, 10(1), 1–23.
- Zhang, T. (2001). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods A Review. In *AI Magazine* (Vol. 22, Issue 2). Cambridge university press.
- Zhang, Z., Masjedi, A., Zhao, J., & Crawford, M. M. (2017). Prediction of sorghum biomass based on image based features derived from time series of UAV images. *International Geoscience and Remote Sensing Symposium (IGARSS), 2017-July*, 6154–6157.
- Zhao, K., Suarez, J. C., Garcia, M., Hu, T., Wang, C., & Londo, A. (2018). Utility of multitemporal lidar for forest and carbon monitoring: Tree growth, biomass dynamics, and carbon flux. *Remote Sensing of Environment*, 204, 883–897.

- Zhao, W., & Du, S. (2016). Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4544–4554.
- Zhao, Y. R., Li, X., Yu, K. Q., Cheng, F., & He, Y. (2016). Hyperspectral Imaging for Determining Pigment Contents in Cucumber Leaves in Response to Angular Leaf Spot Disease. *Scientific Reports*, 6(27790), 1–9.
- Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. In *IEEE Geoscience and Remote Sensing Magazine* (Vol. 5, Issue 4, pp. 8–36).