Using hyperspectral remote sensing and machine learning for potato yield forecasting in irrigated sandy soils

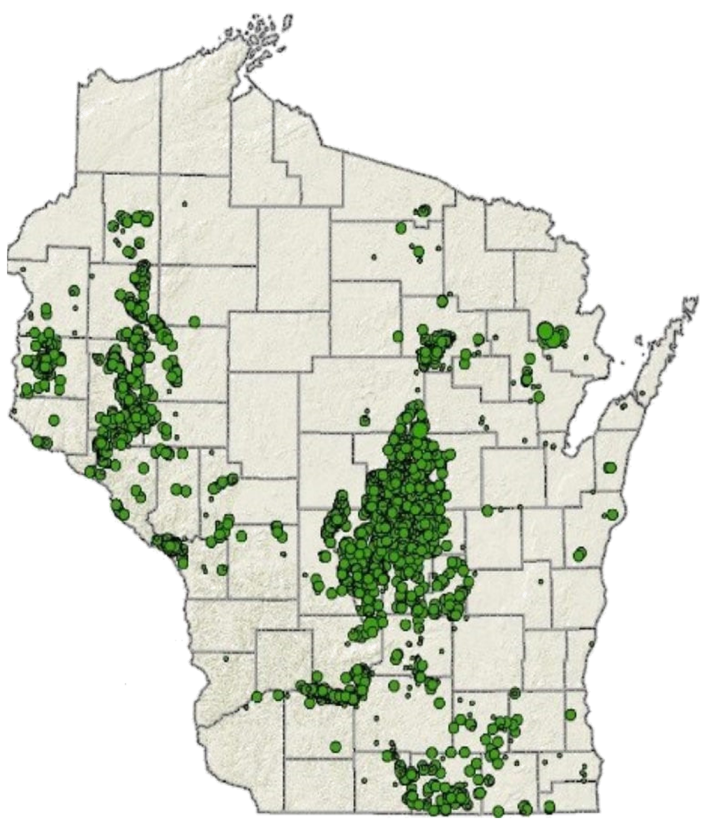
Alfadhl Alkhaled[1], Philip A. Townsend[2], Yi Wang[1]

1: Department of Plant and Agroecosystem Sciences, UW-Madison;

2: Department of Forest and Wildlife Ecology, UW-Madison

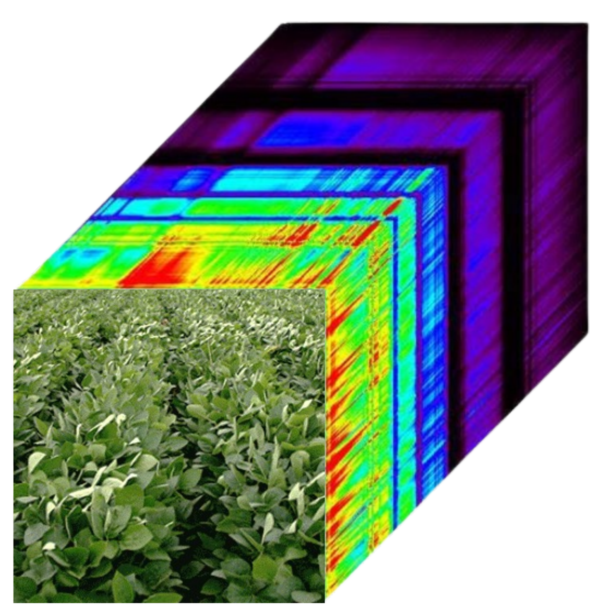
INTRODUCTION

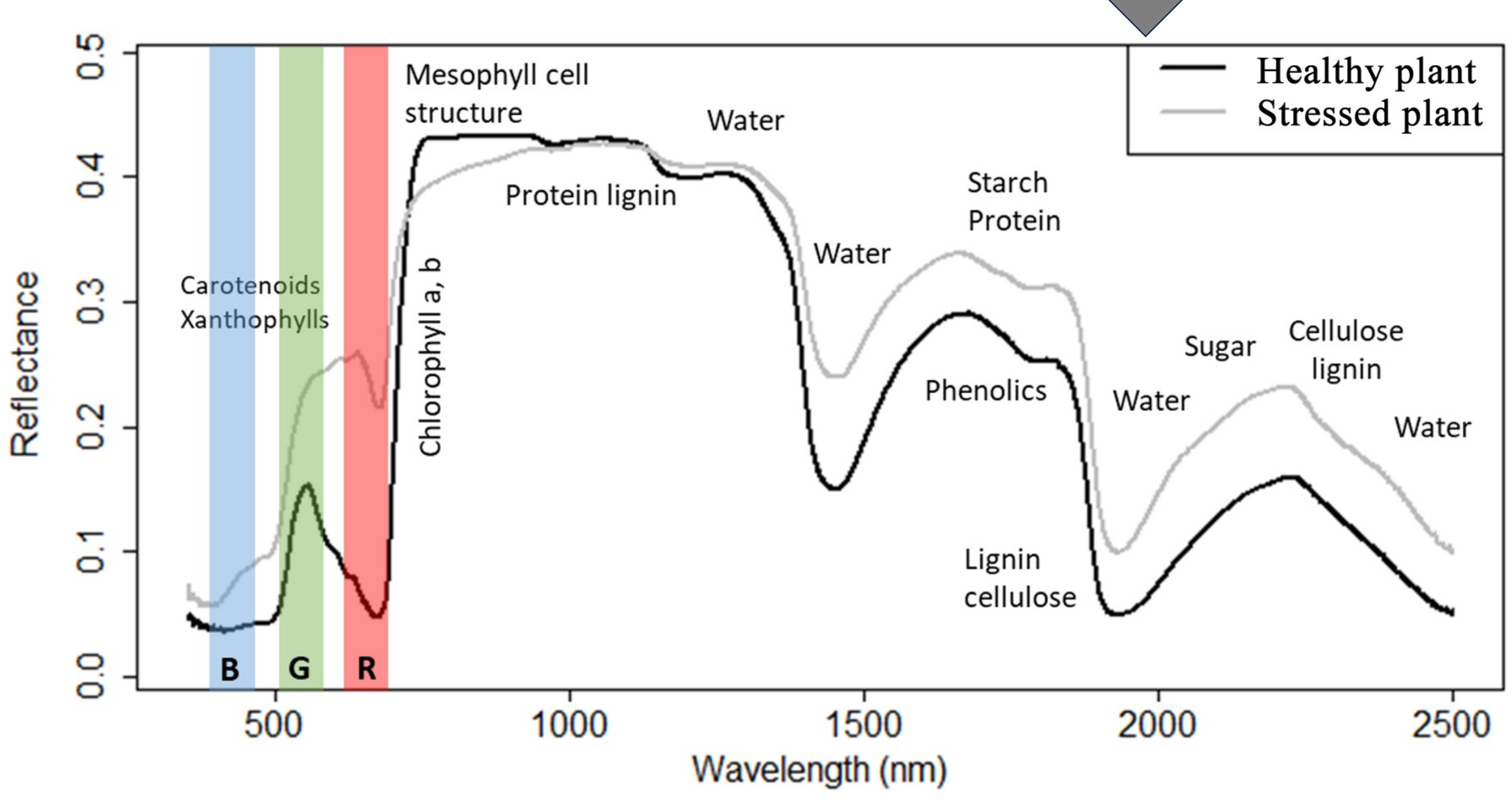
- Potato production's significance in global food security highlights the need for precise yield prediction.
- Wisconsin, a leading state in U.S. for potato production, grows over **63,000 acres annually**, yielding more than **2.3 billion pounds of potatoes**. However, sandy soils and frequent rainfall in the summer contribute to nitrate-N leaching and groundwater contamination.
- This study addresses optimizing potato nitrogen management using the innovative hyperspectral imaging and machine learning techniques.



REMOTE SENSING & HYPERSPECTRAL IMAGING

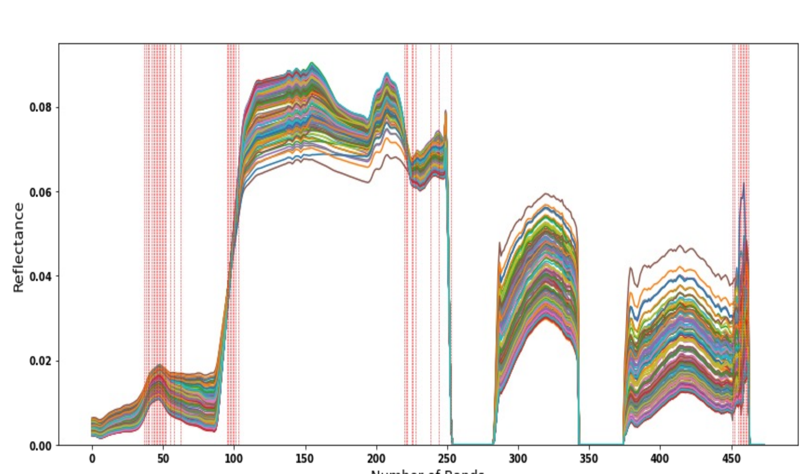
- Remote sensing technologies cover spatio-temporal variability and are non-destructive.
- Hyperspectral imaging, capturing extensive reflectance data from plant canopies, facilitates precise crop growth modeling.
- With a Cessna-180 airplane, reflectance data from **474 spectral bands** were collected at visible, NIR, and SWIR ranges (**400 - 2500 nm**) across multiple years (**2018-2022**) and sites in Central Wisconsin, offering insights into ***precise potato yield prediction*** using advanced machine learning algorithms.





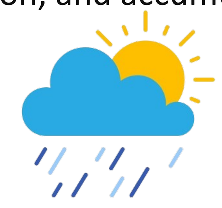
DATA COLLECTION AND G × E × M

- Over five years, we collected diverse datasets from seven research sites and eight potato varieties.
- Feature selection using Random Forest generates the top 20 bands that are the most highly related to final tuber yield.
- Besides the spectral data, we have also collected the following information (G, E, M):



Genotype: potato varieties

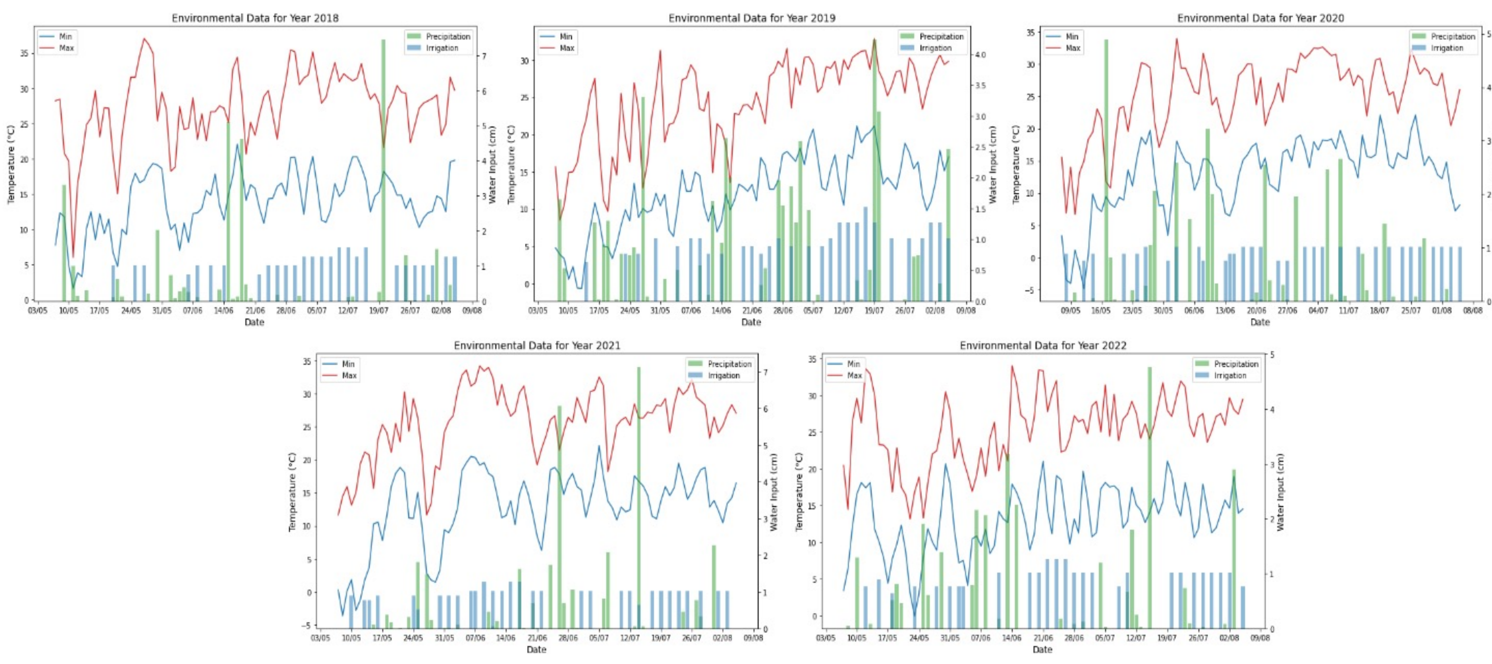
Environment: Soil temperature, precipitation, and accumulated GDD



Management: Site, Year, Growth stage, N rate

MACHINE LEARNING AND FINDINGS

Year-to-year weather variations:



Comparison between feature selection and all bands:

	All bands		Selected top 45 bands	
	R^2	RMSE	R^2	RMSE
2018	0.341	52.1	**0.376**	50.7
2019	0.465	52.1	**0.482**	51.3
2020	0.546	45.9	**0.575**	44.4
2021	**0.160**	73.2	0.135	72.4
2022	0.032	56.2	**0.099**	54.2
All years	0.343	84.5	**0.387**	81.6

Single factor effects:

	Genotype		Environment		Management		Selected bands	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
2018	0.369	50.9	0.044	65.5	0.129	59.9	**0.376**	50.7
2019	0.142	66.0	0.057	73.3	**0.483**	51.2	0.482	51.3
2020	**0.728**	35.5	0.014	68.6	0.009	67.8	0.575	44.4
2021	**0.564**	44.9	0.355	79.1	0.473	82.5	0.135	72.4
2022	0.037	56.1	0.017	56.6	**0.160**	52.4	0.099	54.2
All years	0.481	75.1	0.496	74.0	**0.616**	64.6	0.387	81.6

Two-factor effects:

	Genotype + Environment		Environment + Management		Genotype + Management	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
2018	0.333	52.4	0.129	59.9	**0.478**	46.3
2019	0.100	67.6	0.483	51.2	**0.640**	42.7
2020	0.684	38.3	0.011	67.8	**0.751**	34.0
2021	0.459	50.0	**0.473**	82.5	0.301	56.8
2022	0.037	58.2	0.160	52.4	**0.387**	44.7
All years	0.627	63.7	0.615	65.6	**0.766**	50.4

Multiple-factor effects:

	Genotype + Environment + Management		Genotype + Environment + Management + Selected bands	
	R^2	RMSE	R^2	RMSE
2018	0.478	46.3	**0.667**	37.0
2019	0.640	42.7	**0.722**	37.5
2020	0.751	34.0	**0.834**	27.7
2021	0.301	56.8	**0.394**	52.9
2022	**0.387**	44.7	0.361	45.7
All years	**0.762**	50.8	0.716	55.5

DISCUSSIONS AND CONCLUSION

- When working with hyperspectral dataset, using **random forest** with **feature selection** showed consistent prediction accuracy, emphasizing the significance of selecting the most relevant information out of the large amount of collected data. Big data mining is a concern that future research needs to address, as efficient modeling should focus on retrieving relevant and demanded patterns and extracting value hidden in data of an immense volume.
- Using information about ***genotype***, ***management***, ***environment***, together with the ***spectral signatures*** as input variables into machine learning algorithms yields good prediction accuracy (**R^2 up to 0.87**), highlighting their importance and non-linear relationship. Future research should take all factors that can synergistically affect crop productivity into consideration.
- Utilizing **multi-year-site** data yields significantly **higher accuracy** compared to only using individual years-sites. Aggregating data across years and sites captures broader trends and diverse influences, and can fortify model reliability, whereas models developed from single year-site focus on specific nuances but lack comprehensive understanding.

Better Potatoes for a Better World
PAA
THE POTATO ASSOCIATION OF AMERICA

POTATO EXPO
AUSTIN
TEXAS 2024