The International Society of Precision Agriculture presents the
15th International Conference on Precision Agriculture
26–29 JUNE 2022
Minneapolis Marriott City Center | Minneapolis, Minnesota USA

Best use of modern data for field-specific decision support

Paul B. Hegedus[a], Bruce D. Maxwell[a]

[a] Department of Land Resources and Environmental Sciences, Montana State University, United States

A paper from the Proceedings of the
15th International Conference on Precision Agriculture
June 26-29, 2022
Minneapolis, Minnesota, United States

Abstract.

Many agronomic input decisions are based on data from previous production years. Data from the gap in time between a previous harvest and the point that an input (e.g. fertilizer) decision is made may improve crop response predictions. This research was designed to study the potential value of using the extra data in decision making, by comparing the uncertainty in predictions from agronomic models derived solely from data from previous production years (PY) versus data collected until immediately before the need for top-dress nitrogen fertilizer application decisions (DP). As there is additional uncertainty in the appropriate functional form of agronomic models related to crop responses to fertilizer, this research utilized two models: generalized additive models (GAMs) and random forest regression (RF). The ultimate objective was to evaluate differences in management recommendations and ability to accurately predict rain-fed winter wheat responses to nitrogen fertilizer rates under the two data constraints (PY vs. DP). To evaluate differences in management recommendations, a wet and dry year were selected from the historical record in which site-specific (precision) rate recommendations for nitrogen fertilizer were simulated. A difference in the accuracy of predictions between the PY and DP data constraint was observed in 5 out of 28 cases across both model types. Functional form of the model used to predict responses to variable N rate fertilizer played a more important role in characterizing uncertainty surrounding recommendations than did the PY or DP data constraint. Where inclusion of more recent data only occasionally influenced model accuracy, accuracy was never sacrificed with its inclusion. Thus, the potential value of more recent data in decision making and its inclusion in decision support systems should consider using models fit with all available data up to the point a crop manager must make decisions.

Keywords.
List both specific and general terms that will aid in searches.

Introduction

Decision support for site-specific inputs is typically based on empirical models that predict crop responses such as crop yield, or quality, and covariate data collected from sensors and satellites. These models are commonly parameterized based on fits to data across experiments and/or farms that are pooled to generate a generalized model for informing field-specific management. However, crop responses vary across space and time, within and between fields, even those fields that share a border (Hegedus & Maxwell, 2022a). This variation also influences differences in the most appropriate functional form of crop response models for a given crop (Hegedus & Maxwell, 2022b). Thus, field-specific management should logically be informed by the now available field specific data (Cook, Cock, Oberthür, & Fisher, 2004; Hegedus & Maxwell, 2022a; Lacoste et al., 2022).

Field-specific experiments have potential to locally optimize input rates, which overcomes the uncertainty introduced by potentially non-representative data from off-site experimentation, and to reveal the role of field-specific variables influencing crop production and farmer net-returns (Luschei et al., 2001; Hegedus & Maxwell, 2022a). Uncertainty associated with crop response variability over time becomes the primary concern for using empirical models to convert site-specific information into decision recommendations. A common approach for developing empirical predictive models is to calibrate a model with data from previous years, quantify the quality of that calibrated model by comparing its predictions with observations gathered in subsequent years, and to retrain and amend the model as necessary (Hair, 2007). This process is then repeated over time, with each passing year providing a new dataset to test and improve a model. For example, a range of crop responses in each field over different weather and economic conditions is required to predict the most profitable input level before the next year, assuming the next year is average for these variables and the past years sampled represent the range of possible conditions. The backlog of conditions and responses can be used to simulate the probability of an outcome given a particular treatment if the past responses and economic variability (i.e., training data) are applied to making a current year decision (i.e., validation data). Simulation models have determined that high degrees of certainty in probabilities of net return can be achieved by conducting experiments on a field with a given crop for about 6-8 years (Lawrence, Rew, & Maxwell, 2015). Future anomalous weather conditions due to climate change may introduce further uncertainty in decision support recommendations based on any assumptions of stationarity in this empirical approach. However, incorporating strategic sampling of past years in which to simulate projected climate changes increases the power of decision support systems by allowing farmers to examine outcomes in years of the past that may represent anomalous future conditions of an upcoming year.

Technology has given producers data sources not conventionally used in agronomic decision making (given the recency of the agricultural data revolution) that allow the application of data gathered immediately before the crop manager decides on the amount of input to buy and apply (decision point). Simulating management outcomes in a decision support system requires empirical predictive models that utilize weather data, however one element of uncertainty for constructing field-specific models is determining what data to include to make the best prediction of crop response. More specifically, uncertainty surrounds the question of what portion of past weather and crop reflectance data to include in analysis, e.g., whether covariate data is constrained to only past years, includes data from past years plus weather data into the growing season up to the decision point, or whether covariate data from further into the growing season is included. Many studies have been conducted to predict crop responses, such as yield, using in-season covariate data. However, estimates of future yield from within-season measurements after a decision point are not appropriate for use in decision support systems (Croft et al., 2020; Gaso et al., 2019; Gonzalez-Sanchez et al., 2014; Stepanov et al., 2020).

In a rain-fed winter wheat (*Triticum aestivum*) production system, response models of grain yield and protein concentration to nitrogen (N) fertilizer are hypothesized to predict crop responses more accurately when the cumulative influence of weather and early-season growth prior to a decision point in mid-growing season are included in the model, compared to just using information from previous years (Fig. 1). A rainfed crop would logically be influenced by weather conditions during the growing season as well as previous seasons, thereby providing increased predictive power if included in the crop response models. Mid-growing season application of N fertilizer is thought to boost yields, but especially increase grain protein concentrations that can receive a price premium (Jaynes & Colvin, 2006; Jones, 2013; Qianqian et al., 2021; Zeng et al., 2012).

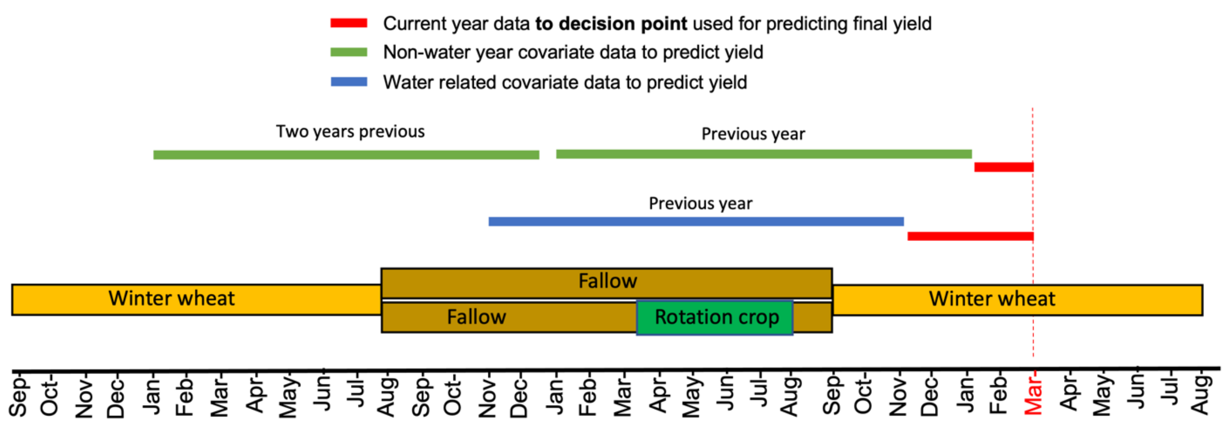

Fig 1. Timeline of the data constraints for non-water and water related covariates. Water related covariates are based on water years, while non-water related covariates follow calendar years. The two situations for collecting data correspond to when in relation to harvest the data are collected, where the difference between the approaches is indicated by the red bars.

This study attempts to address a critical component of the agronomic data revolution by addressing how constraints on data used to fit models of crop production and quality influence the ability of a model to support effective application decisions. Multiple models for characterizing crop responses to N fertilizer and covariate data were used in recognition of the uncertainty surrounding the appropriate functional form of crop response models between fields (Hegedus & Maxwell, 2022b). Two models were selected, based on prior investigations into the adequacy of modeling crop responses: a generalized additive model (GAM) and a random forest regression (RF) model. Both models were fit to characterize winter wheat yield and grain protein concentration response to variable N fertilizer rates with climate, environmental, and vegetation index covariates to assess the influence of data constraints on predictions of crop responses. Model performance when data were constrained to previous calendar years was compared to performance with data collected up to a March 30th (current growing season) decision point for top-dress N fertilizer (Fig.1). Assessing performance between the same type of model fit under the two data constraints was the main objective of this paper, where performance was defined as the ability of the model to predict responses in test datasets and measured by root mean square error (RMSE). In addition, assessments of the crop model performance were repeated under simulated wet and dry weather conditions to assess how nitrogen application recommendations are influenced by interannual uncertainty in weather.

Methods

The objectives were evaluated using seven dryland fields from three farms distributed across Montana that had at least three years of on-field N fertilizer experimentation (Fig. 2, Table 1). The response variables of interest were crop productivity (yield in kg ha^{-1}) and quality (grain protein concentration), both of which are gathered from monitors mounted on farmers' combine harvesters. Data from yield monitors were gathered on average every three seconds. All yield monitors were calibrated every spring by the farmers, according to their respective manufacturing instructions. Grain protein concentration (%) was measured with a CropScan 3000H near infrared monitor (Clancy, 2019). Beyond data collected from the machines on the field, remotely sensed covariate data from open sources were gathered (Table 2). These data were obtained or derived from Google Earth Engine (Gorelick et al., 2017). Note that while elevation, aspect, slope, TPI, and OpenLandMap data were collected every year, they did not vary from year to year. Temporal covariate data included Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), precipitation, and growing degree days (GDD). Yield and protein datasets were created by aggregating the response variable data, as-applied N, and remote sensing data to the centroids of a 10m grid laid across each field.

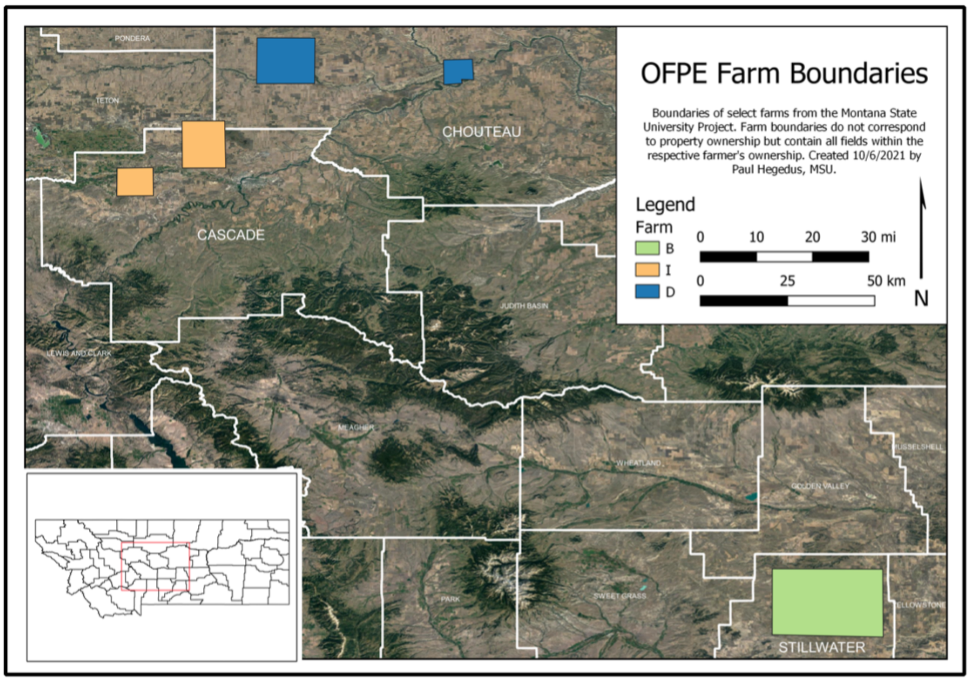

Fig 2. Map of OFPE farm boundaries for the three selected farmer collaborators with Montana State University. Colors represent different farmers, while shapes represent general areas in which their respective fields are located.

Table 1. Crop histories, field sizes, and years in VRA treatment for each field for given farmers. The farm identifier corresponds to the map above and is used instead of the farmer's name for privacy.

Farm	Field	Field size (ha)	**Crop History[1]:** *2014 / 2015 / 2016 / 2017 / 2018 / 2019 / 2020 / 2021*	Years N rate treatment
B	B1	79	SF / WW / CF / WW / CF / WW / CF / WW	2017, 2019, 2021
	B2	94	WW / CF / WW / CF / SF / WW / CF / WW	2016, 2019, 2021
	B3	64	SW / CF / WW / CF / WW / CF / WW / CF	2016, 2018, 2020
D	D1	46	CF / WW / CF / WW / CF / WW / CF / WW	2017, 2019, 2021
	D2	48	WW / SW / WW / CF / WW / CF / WW / CF	2016, 2018, 2020
	D3	20	SW / SW / WW / CF / WW / CF / WW / CF	2016, 2018, 2020
I	I1	94	SW / CF / WW / CF / WW / CF / WW / CF	2016, 2018, 2020

[1] WW = winter wheat, CF = chemical fallow, SW = spring wheat, SF = safflower

Table 2. Table of covariates gathered from Google Earth Engine to enrich the crop yield and protein datasets gathered from farms. In some cases, multiple sources are used, however only one data source is used when aggregating to yield and protein.

Data Type	Data Sources	Resolution	Years Collected	Description
Normalized Difference Vegetation Index (NDVI)	Sentinel 2, Landsat 5/7/8	10m, 30m	S2: 2016-present L5: 1999-2011 L7: 2012-2013 L8: 2014 - present	Sentinel 2 is from the European Space Agency as part of the Copernicus program. Landsat is an ongoing USGS and NASA collaboration. Bands (NIR, red) S2: B8 and B4 L5/L7: B4 and B3 L8: B5 and B4
Normalized Difference Water Index (NDWI)	Sentinel 2, Landsat 5/7/8	10m, 30m	S2: 2016-present L5: 1999-2011 L7: 2012-2013 L8: 2014 - present	Sentinel 2 is from the European Space Agency as part of the Copernicus program. Landsat is an ongoing USGS and NASA collaboration. Bands (NIR, red) S2: B3 and B5 L5/L7: B2 and B4 L8: B2 and B5
Normalized Difference Red Edge (NDRE)	Sentinel 2	20m	S2: 2016-present	Bands B5 and B6
Red Edge Chlorophyll Index (CIRE)	Sentinel 2	20m	S2: 2016-present	Bands B7 and B5
Elevation	USGS NED	~10m (1/3 arc second), ~23m (3/4 arc second)	1999-present	USGS National Elevation Dataset. Measured in meters.
Aspect	USGS NED	~10m (1/3 arc second), 30m	1999-present	Direction the surface faces, function of neighboring elevations, in radians. Also calculated for each E/W and N/S direction as cosine and sine.
Slope	USGS NED	~10m (1/3 arc second), 30m	1999-present	Rate of change of height from neighboring cells, in degrees. Measured in degrees.
Topographic Position Index (TPI)	USGS NED	~10m (1/3 arc second), 30m	1999-present	Measure of divots and low spots as a function of neighboring elevation.
Precipitation	DaymetV3	1km	1999-present	Estimates from the NASA Oak Ridge National Laboratory (ORNL). Measured in mm.
Growing Degree Days (GDD)	DaymetV3	1km	1999-present	Estimates from the NASA Oak Ridge National Laboratory (ORNL).
SMAP	susm	10km	2016-present	Surface (0-5cm) and sub-surface (5-100cm) soil moisture content.
OpenLandMap	grtgroup	250m	1999-present	Predicted USDA soil taxonomy great group probabilities.
OpenLandMap	texture	250m	1999-present	Soil texture classes (USDA system) averaged over 6 soil depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	bulkdensity	250m	1999-present	Soil bulk density (fine earth) 10 x kg / m3 averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	claycontent	250m	1999-present	Clay content in % (kg / kg) averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	sandcontent	250m	1999-present	Sand content in % (kg / kg) averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	pH (phw)	250m	1999-present	Soil pH in H_2O averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	watercontent	250m	1999-present	Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted and averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).
OpenLandMap	carboncontent	250m	1999-present	Soil organic carbon content in x 5 g / kg averaged over 6 standard depths (0, 10, 30, 60, 100 and 200 cm).

All temporal covariates from Google Earth Engine were collected under two constraints, 1) the first, hereafter labeled PY, included all previous years of information, and 2) hereafter labeled DP, included all previous years plus the current year to the mid growing season when the decision point for top-dress fertilizer rates is determined for winter wheat. The PY data were collected up to December 31st of the year prior to upcoming harvest for vegetation index and GDD data and October 31st of the year prior to upcoming harvest for precipitation to follow the water year. Data collected to the decision point included the extra three months from January 1st to March 30th of the harvest year for vegetation index and GDD data and the extra five months from November 1st of the previous year to March 30th of the harvest year for precipitation (Fig. 1). A rainfed crop would logically be influenced by weather conditions (precipitation and temperature) during the winter portion of the growing season and therefore provide increased predictive power if included in the crop response models

The first model type fit to the yield and protein datasets was a GAM. Generalized additive models have been used in the literature for estimating crop responses to variable rates of N fertilizer and were chosen in our case because of their flexibility and ability to characterize the response of observed data (Chen et al., 2019; Estes et al., 2013; Joshi et al., 2021; Liew & Forkman, 2015). A gamma family distribution was used due to the realistic constraint that yield and protein of a crop cannot be negative. To account for non-constant variance found in initial model fits, a log link function was used. Thin plate shrinkage splines were used for all variables to allow the estimated degrees of freedom of parameters to shrink to zero, combining the process of model fitting and selection. To account for spatial autocorrelation, coordinates were included as smoothing terms with a Gaussian process basis function (Gotway & Stroup, 1997; Guisan et al., 2002; Holland et al., 2000; Zuur & Camphuysen, Kees, 2012). The exception to using thin plate shrinkage splines was the use of a Gaussian process for the interaction of coordinates accounting for spatial autocorrelation between observations within the field. While time is not explicitly included in the model, the singularities within years for variables that vary across time, such as precipitation and growing degree days, mean these variables serve as proxies for the effect of time. Models were fit using the *mgcv* package in R (Wood, 2003; Wood et al., 2016).

The second model used was a RF where the number of trees and the number of covariates sampled at each node were optimized during the fitting process. To account for spatial autocorrelation, the coordinates were included as covariates (Janatian et al., 2017; Langella et al., 2010; Walsh et al., 2017; Y. Wang et al., 2017). Random forest regression has also been used to fit crop responses to agricultural inputs (Everingham et al., 2016; Jeong et al., 2016; Lawes et al., 2019; Mariano & Mónica, 2021; Marques Ramos et al., 2020; Paccioretti et al., 2021; L. Wang et al., 2016), and initial testing of various model forms indicated that the predictive ability of the random forests outcompeted parametric and Bayesian empirical models (Hegedus & Maxwell, 2022b). The *ranger* package in R was used for fitting and generating predictions (Wright & Ziegler, 2017).

Predictors that did not vary across time were centered by taking the difference between each value and the mean value for the predictor to reduce discrepancies in scale between covariates. Covariates that varied across time (vegetation indices and weather data) were left uncentered because the distribution and subsequent mean of these predictors could vary across time. The covariates used for both models for both crop responses (yield and protein) under the two data constraints are shown in Table 3.

Table 3. Covariates used under the two data constraints. Units and sources are in Table 2. CY indicates data from the harvest year, PY indicates data from the year prior to the harvest year, 2PY indicates data from two years prior to the harvest year.

PY	DP
As-Applied N	As-Applied N
UTM Coordinates	UTM Coordinates
Aspect	Aspect
Slope	Slope
Elevation	Elevation
Topographic Position Index	Topographic Position Index
Precip. from November 1st (2PY) to October 31st (PY)	Precip. from November 1st (2PY) to October 31st (PY)
	Precip. from November 1st (PY) to March 30th (CY)

GDD from January 1st (PY) to December 31st (PY)	GDD from January 1st (PY) to December 31st (PY)
	GDD from January 1st (CY) to March 30th (CY)
NDVI from January 1st (PY) to December 31st (PY)	NDVI from January 1st (PY) to December 31st (PY)
NDVI from January 1st (2PY) to December 31st (2PY)	NDVI from January 1st (2PY) to December 31st (2PY)
	NDVI from January 1st (CY) to March 30th (CY)
NDWI from January 1st (PY) to December 31st (PY)	NDWI from January 1st (PY) to December 31st (PY)
NDWI from January 1st (2PY) to December 31st (2PY)	NDWI from January 1st (2PY) to December 31st (2PY)
	NDWI from January 1st (CY) to March 30th (CY)
Bulk Density averaged over 0cm – 200cm	Bulk Density averaged over 0cm – 200cm
Clay Content averaged over 0cm – 200cm	Clay Content averaged over 0cm – 200cm
pH of water averaged over 0cm – 200cm	pH of water averaged over 0cm – 200cm
Water Content averaged over 0cm – 200cm	Water Content averaged over 0cm – 200cm
Carbon Content averaged over 0cm – 200cm	Carbon Content averaged over 0cm – 200cm

Data analyses was performed using R (R Core Team, 2021) where data management, aggregation, cleaning, and analysis with the two model types used the *OFPE* package (Hegedus, 2020). Comparison of the performance of a given model using the PY or DP data was assessed by 5x2 cross validation (CV) developed by Dietterich (1998) for each field, crop response (yield and protein), and for both model types (GAM and RF). The dataset was split 50/50 into a training and test set, and the model using both the PY and DP constraints were fit on the training set and RMSE was calculated on the test set. Then the training and test sets were swapped and both models were fit on the new training set (previously the test set) with RMSE calculated on the test set (previously the training set). This was repeated five times, and a 5x2 CV t statistic was calculated (Dietterich, 1998). Thus, for each field, response, and model type, a t statistic and corresponding two tailed p-value were computed to compare the predictive accuracy (RMSE) of the model fit using data solely from previous years (PY) and data using covariates up to the decision point of the harvest year (DP).

After fitting each model type for both grain yield and protein concentration for each field under both data constraints, crop responses at every location in the field under varying weather conditions were simulated to derive a site-specific N fertilizer recommendation optimized to maximize net-return. For each field, two weather conditions were simulated based on data for a given field: the wettest year from 1999-2021, and the driest year from 1999-2021.This resulted in eight simulations per field, where the yield and grain protein models of each type (GAM or RF) for each data constraint (DP or PY) were used to find optimized N fertilizer rates in the two selected types of years (wet or dry). Selecting the year defined the data fed into the models. For simulations in the given year under the DP data constraints, data were collected up until March 30th of the wet or dry year and used in the DP models to make predictions. For simulations with the PY data constraints, the models fit with PY data were used to predict crop responses and find optimized N fertilizer rates using data collected to January 1st of the year selected. Thus, temporal covariate data based on the years selected were supplanted into each dataset to perform the simulation under the different weather conditions for each field. Crop yield and grain protein concentration were predicted at the centroid of a 10m grid across the field for N rates ranging from 0 to 168 kg ha^{-1}. For every point and N rate, the predicted yield and protein concentration were used to calculate a net-return using the economic conditions from the simulated weather year, beginning with a calculation of the amended price received based on base price and the protein concentration of a given observation;

$$P = Bp + (B0pd + B1pd * protein + B2pd * protein^2) \quad (1)$$

where *P* is the final price received (\$ kg^{-1}), *Bp* is the base price received (\$ kg^{-1}), *B0pd* is the intercept of the protein premium/dockage function set by the grain elevator, *B1pd* is the coefficient on the grain protein concentration (protein, %) and *B2pd* is the coefficient on the squared protein term. Using this price, net-return was calculated as;

$$NR = yield * P - CA * AA - FC - ssAC \quad (2)$$

where *NR* is the net-return (\$ ha^{-1}) received and a function of the product of the *yield* (kg ha^{-1}) and *P*, minus the cost of the applied input (*CA*) multiplied by the as-applied rate of the input (*AA*), the fixed costs (*FC*) associated with production (\$ ha^{-1}) that do not include the input, and the cost per hectare of the site-specific application (*ssAC*). The net-return at each point was used to identify the N rates across the field that maximized farmer net-returns and generate a N fertilizer recommendation on a 10m scale.

Results

Model Performance Comparison

The 5x2 CV was performed first using the GAM to simulate yield responses for each field to compare the error estimates relative to observed data between the GAM fit with PY rather than the GAM fit with the DP, measured by RMSE. In two (B1, D2) out of seven fields there was strong evidence against the null hypothesis that there was no difference in the predictive ability between a GAM fit with PY or DP data ($\alpha < 0.05$), with the GAM for yield responses resulting in a higher predictive accuracy under the DP constraint than the PY constraint (Table 4). In two (B2, I1) out of seven fields, there was moderate evidence ($\alpha < 0.15$) against the null hypothesis, and weak to no evidence in three (B3, D1, D3) out of seven fields against the null hypothesis ($\alpha \geq 0.15$; Table 4). The 5x2 CV was then performed for each field using the RF to simulate yield responses in the same manner as the GAM to compare RMSE between a RF fit with the PY and DP. In contrast to the results of the GAM, there was no evidence against the null hypothesis for any field where either of the data constraints resulted in a difference in performance of the RF for yield responses (Table 5).

Table 4. Results from 5x2 CV for the GAM fit to yield responses. The mean RMSE, in kg ha^{-1}, were calculated across each split and the 5 folds. An asterisk indicates significance at an alpha level of 0.05 and bolded RMSE values indicate the data constraint with a lower RMSE.

Field	5x2 t statistic	p-value	Mean RMSE PY	Mean RMSE DP
B1	4.3416	0.0074*	858.8653	**848.3403**
B2	2.1672	0.0824	788.9887	783.4343
B3	0.4362	0.6809	876.8696	899.8229
D1	-0.8908	0.4139	538.0681	536.7981
D2	4.8024	0.0049*	719.2763	**714.4325**
D3	0.1869	0.8591	762.5487	763.6518
I1	1.9276	0.1118	812.1172	805.8488

Table 5. Results from 5x2 CV for the RF fit to yield responses. The mean RMSE, in units of kg ha^{-1}, were calculated across each split and the 5 folds.

Field	5x2 t statistic	p-value	Mean RMSE PY	Mean RMSE DP
B1	-1.5857	0.1737	572.5701	578.1144
B2	-0.5464	0.6083	533.3079	534.0513
B3	-0.5831	0.5852	590.6175	592.1910
D1	1.2008	0.2836	387.3872	385.2688
D2	0.7394	0.4929	398.1208	397.2827
D3	0.1423	0.8924	548.1305	546.6204
I1	-0.1685	0.8728	556.0405	561.4883

Assessing the comparative model performance of protein response models, only one (D1) out of seven fields showed a difference in the predictive ability between a GAM fit under the PY or DP ($\alpha < 0.05$), where a higher predictive accuracy under the DP was observed (Table 6). In three (B2, B3, I1) out of seven fields, there was moderate evidence ($\alpha < 0.15$) against the null hypothesis of no difference between data constraints, and weak to no evidence ($\alpha > 0.15$) in three (B1, D2, D3) out of seven fields (Table 6). While there was no evidence in any field that data constraints played a role in model performance of the RF for yield responses, two (D1, D3) out of seven fields were identified as having a difference in the predictive ability between a RF fit with PY or DP data ($\alpha < 0.05$), with the RF for protein responses under the DP constraint resulting in greater predictive accuracy compared to the RF for protein responses under the PY constraint (Table 7). In one (B3) out of seven fields, there was moderate evidence ($\alpha = 0.15$) against the null hypothesis that there was no difference in the data constraints, and weak to no evidence in four (B1, B2, D2, I1) out of seven fields (Table 7).

Table 6. ***Results from 5x2 CV for the GAM fit to grain protein concentration responses. The mean RMSE, in % protein, were***

calculated across each split and the 5 folds. An asterisk indicates significance at an alpha level of 0.05 and bolded RMSE values indicate the data constraint with a lower RMSE.

Field	5x2 t statistic	p-value	Mean RMSE PY	Mean RMSE DP
B1	1.5650	0.1784	1.6953	1.6963
B2	-2.3742	0.0636	1.3179	1.3212
B3	1.9562	0.1078	1.3187	1.3116
D1	4.5167	0.0063*	1.1303	**1.1026**
D2	1.3191	0.2443	1.2219	1.2160
D3	0.2611	0.8045	1.2785	1.2815
I1	-2.4841	0.0556	1.5910	1.6007

Field	5x2 t statistic	p-value	Mean RMSE PY	Mean RMSE DP
B1	1.5946	0.1717	1.6697	1.6683
B2	0.1305	0.9013	1.2754	1.2780
B3	1.7632	0.1382	1.2871	1.2830
D1	2.7191	0.0418*	1.0866	**1.0786**
D2	-1.3101	0.2471	1.1660	1.1684
D3	2.9727	0.0311*	1.2420	**1.2394**
I1	-0.4395	0.6786	1.4567	1.4566

Table 7. Results from 5x2 CV for the RF fit to grain protein concentration responses. The mean RMSE, in % protein, were calculated across each split and the 5 folds. An asterisk indicates significance at an alpha level of 0.05 and bolded RMSE values indicate the data constraint with a lower RMSE.

Data Constraint Impact on N Fertilizer Recommendations

For each field, crop yield and protein responses were simulated in the driest and wettest year from 1999-2021 for the field's geographic area, using both model types with both data constraints to compare how the data constraint influenced the pattern of site-specific N fertilizer recommendations and the total N fertilizer applied when fertilizer rates were optimized on maximizing net-returns.

Across all of Montana, 2021 was the driest year on record from 1999-present and used across farms as the dry weather year. For all other farms, the wettest year varied based on the geographic location. The wettest year was 2018 for farm B, 2016 for farm D, and 2006 for farm I. Due to constraints on data availability, the protein premium dockage scheme used for all simulated years was from 2021 and equaled a \$0.02 increase in the base price for every half percent protein above 11.5% up to 14% and an \$0.08 dockage to the base price for every half percent protein below 11.5% down to 9.5%. All the farmers that manage the fields used in the study were able to apply variable rate fertilizer with their own equipment, so the *ssAC* parameter in equation 2 was zero. The fixed costs (*FC*), base price (*Bp*), and cost of N (*CA*) from equation 1 were derived from a survey of our farmers and the USDA Economic Research Service (ERS) using data from the USDA Agricultural Resource Management Survey (Table 8).

Table 8. Economic parameters across the years used in the dry (2021) and wet (2018, 2016, 2006) simulations that were used to calculate net-return in equation 1.

Economic Parameter	2021	2018	2016	2006
Fixed Costs (FC)	\$181.87 ha^{-1}	\$181.94 ha^{-1}	\$169.61 ha^{-1}	\$136.08 ha^{-1}
Base Price (Bp)	\$0.27 kg^{-1}	\$0.18 kg^{-1}	\$0.13 kg^{-1}	\$0.15 kg^{-1}
Cost of N (CA)	\$0.89 kg^{-1}	\$0.71 kg^{-1}	\$0.75 kg^{-1}	\$0.95 kg^{-1}

All fields used economic conditions from 2021 to calculate net-return in the simulation of a dry year, while economic conditions from 2018 was used to calculate the net-return in the simulations of a wet year in farm B, 2016 was used in the simulation of a wet year for Farm D, and 2006 was used in the simulation of a wet year for farm I.

While there was evidence of a difference in the predictive ability of the GAM between the PY and DP data constraint for yield responses in field B1 (Table 4), this did not appear to play a role in generating a difference in the pattern of site-specific profit maximizing N fertilizer rates and total N applied across the field in either weather year (Fig. 3). In general, across all fields of farm B and both weather conditions, there were minimal differences between the pattern of profit maximizing rates and total N applied between the data constraints for simulations using the GAM or RF (Fig. 3-5).

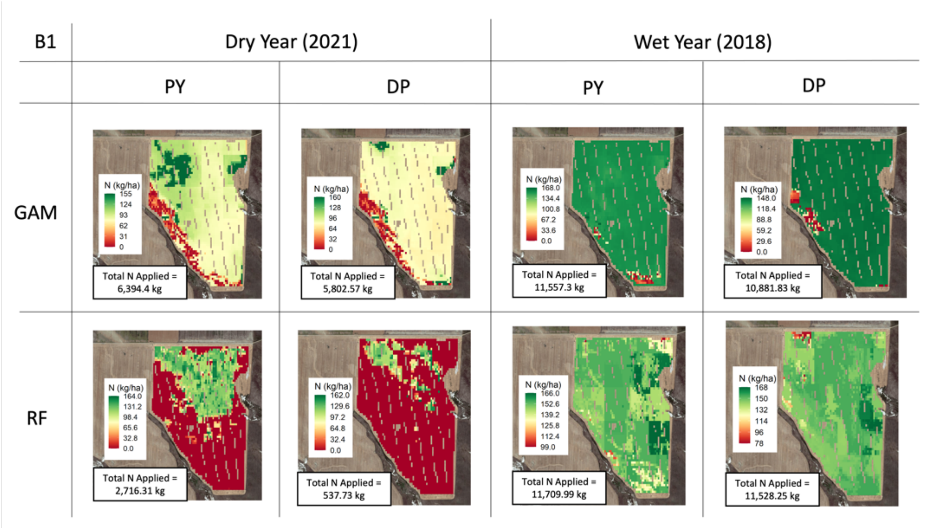

Fig 3. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field B1.

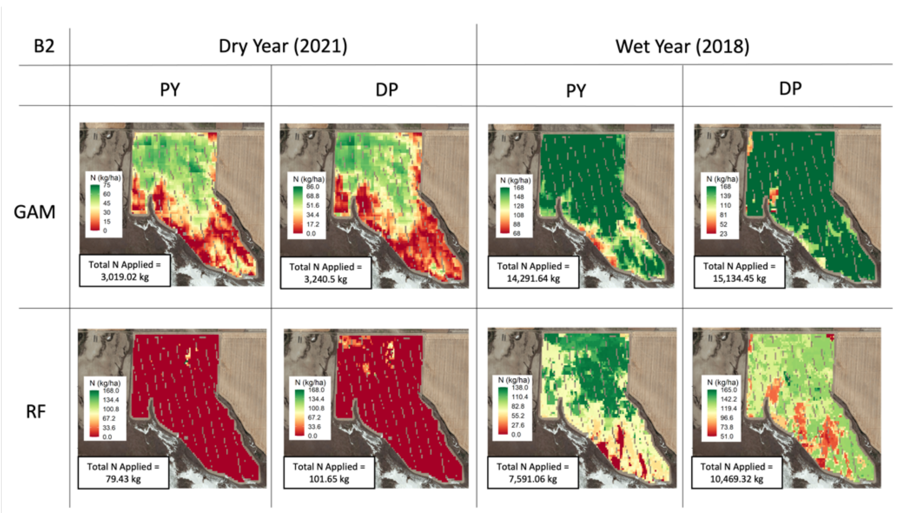

Fig 4. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field B2.

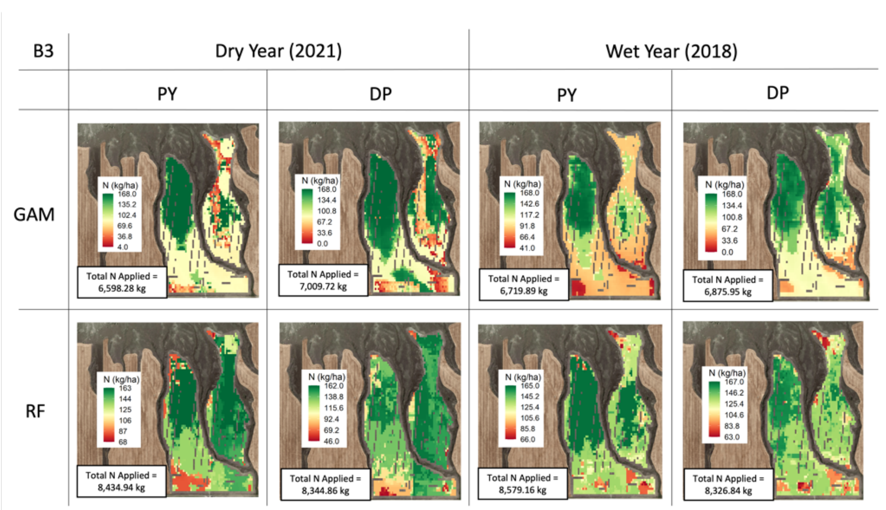

Fig 5. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field B3.

In field D1, there was a pronounced difference in the pattern of N fertilizer rates produced from the GAM between the DP and PY in the wet year for D1 yet no apparent differences between management outcomes with the GAM in the dry year between DP and PY (Fig. 6). While there were minimal differences in the pattern of N rates between the DP and PY for the RF model, there were slight differences between DP and PY in the amount of total N recommended for both the wet and dry years (Fig. 6). For both models fit under the DP constraint there was evidence that it produced more accurate predictions of protein (Tables 6 & 7), however no conclusive evidence is shown that a greater predictive ability for protein responses with the DP constraint contributed to the differences in management recommendations for either model in the wet or dry year. While the scaling is deceptive for field D2, there were no drastic differences in the pattern of N recommendations or total amount of N between DP and PY for either model in the wet or dry year, with the largest difference in recommendations between DP and PY occurring with the GAM in the dry year between PY and DP (Fig. 7). However, in field D3, we observed stark differences in the pattern of N fertilizer recommendations and total amount of N applied between DP and PY in both the wet and dry year with the GAM (Fig. 8). This occurred despite any difference in the ability of the GAM to predict yield or protein responses between DP and PY, furthering the conclusion that the ability of a model to predict responses is not conclusively correlated to differences in management recommendations between data constraints.

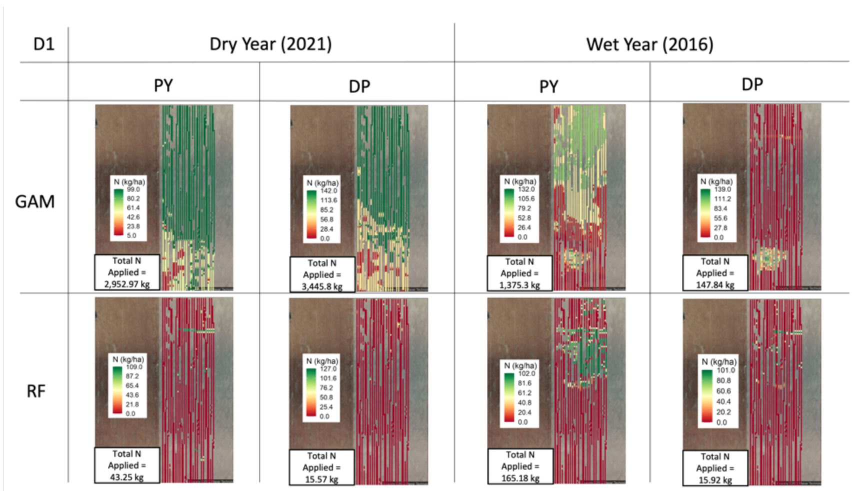

Fig 6. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field D1.

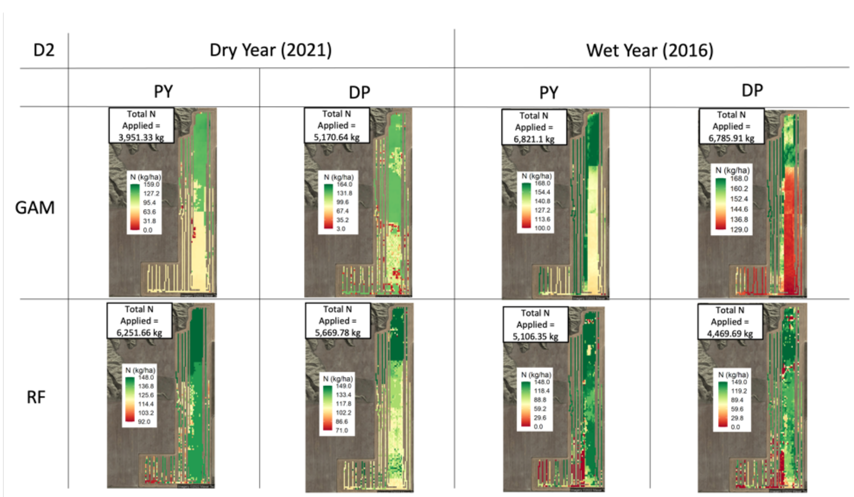

Fig 7. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field D2.

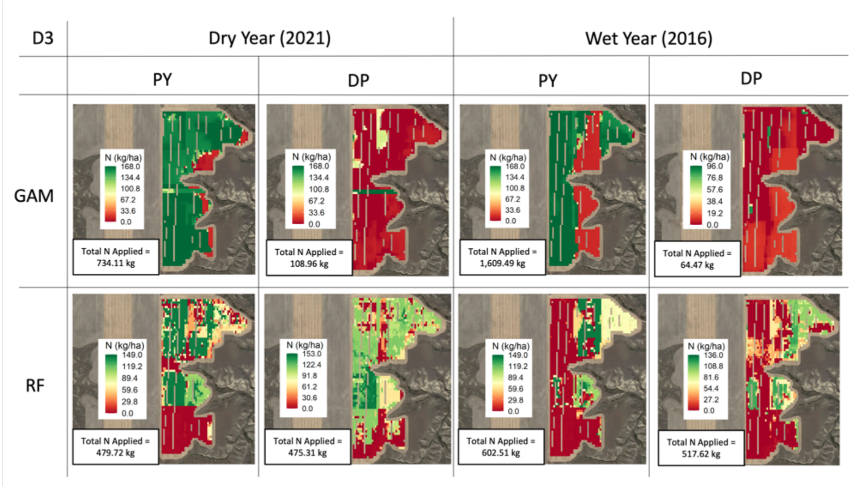

Fig 8. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field D3.

In field I1, comparisons between DP and PY of recommendations in the dry year with either model or the wet year with the RF model showed little to no variation in recommendations (Fig. 9). There was a slight difference in the pattern of profit maximizing N rates and total N applied in the simulated wet year between the GAM fit with PY data compared to the GAM fit with DP data (Fig. 9). However, there was no difference in the ability of either model to predict crop responses between data constraints so any difference in recommendations cannot be contributed to differences in the predictive ability of a model between DP and PY.

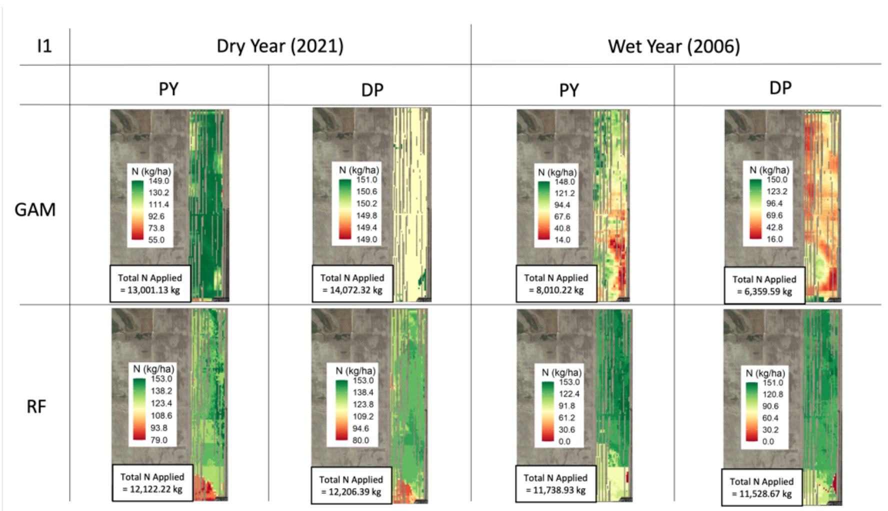

Fig 9. Site-specific profit maximizing nitrogen fertilizer rates for each model type and data constraint in a simulated dry and wet year for field I1.

In most fields, there were only slight differences in the pattern of profit maximizing N fertilizer rates and total N applied between simulations generated from the two data constraints for a given model in a given weather year, albeit with a few exceptions. Field D3 showed the most significant differences in the pattern of N fertilizer rates and total fertilizer applied between data constraints, however this was only apparent between a GAM fit with PY compared to a GAM fit with DP data, but not with the RF model (Fig. 8). Across most fields, simulations using GAMs appeared to be most sensitive to the different data constraints, resulting in slightly different patterns of N fertilizer rates and total N fertilizer applied, though mostly in wet years, while dry years tended to show less of a difference. The RF seemed to be more robust to differences in the data constraints with minimal variation in either wet or dry years between management recommendations from the PY or DP data constraint. While not the intended comparison, differences in the pattern of profit maximizing N fertilizer rates and total N applied were most prevalent between simulated outcomes from the different model forms (GAM or RF) compared to between DP and PY data constraints for a given model.

Discussion

Grain yield and protein response models to variable top-dress N fertilizer rates were hypothesized to benefit from utilizing covariate information into the crop growing season (DP) rather than relying only on data from the previous years (PY) because the DP contains more information on the current growing season. Using a glimpse of the conditions in the current growing season to predict crop responses at harvest was expected to increase the predictive ability of a model because uncertainty in what the growing conditions were at the beginning of the season would be reduced. It may seem obvious that training models with data up to a decision point would result in better predictions of crop responses at harvest compared to models fit with data from past years. Our analysis explicitly tested the value of utilizing previous year and well as data closer to the top-dress fertilizer application decision point for the prediction of crop responses and generation of management recommendations, filling a gap in the literature. Rainfed winter wheat is a good system to test this difference in data use as the crop has gown for 6 months (with at least 3 months dormant) by the time fertilizer is applied and rates of application selected.

Contrary to our hypothesis, there were only a few cases where using data up to a decision point when using either a GAM or RF model improved model predictive performance. In field B1, one of the few cases where utilizing data up to the mid-season reduced uncertainty in yield predictions, the field sharing a border (B2) did not share the same result, indicating a difference in data constraints in adjacent fields (Table 4). These results highlight the site-specificity of crop responses and justifies the use of on-field experimentation and field specific modeling to make decisions for a field (Hegedus & Maxwell, 2022a).

While in two of seven fields using data up to the decision point improved yield predictions from a GAM compared to a GAM fit with data only from past years, there was no evidence across any fields that yield predictions from RF models were improved by including data up to the decision point (Table 4). The RF model resulted in an average 251 kg ha^{-1} reduction in RMSE of yield predictions compared to the GAM, and no observed differences in the predictive ability of the RF between data constraints (Table 5). This indicates that the greater general ability to predict crop responses of the RF resulted in less sensitivity of predictive accuracy when the RF was fit under different data constraints.

Further evidence of the insensitivity of the RF to data constraints was found when simulating management outcomes in dry and wet years (Fig. 3-9). Even when there was no difference in the RMSE of a GAM (for either crop response) between the PY or DP data constraint, there were more differences between simulated management outcomes of a GAM fit with the PY and DP compared to between RF models fit with either the PY or DP constraint. Similar to how the RF was more robust to differences in raw predictive ability between data constraints than the GAM, the low degree of differences in N fertilizer recommendations from the RF between DP and PY further indicates how the RF is robust to differences in data constraints, even when simulating management outcomes in extreme weather conditions (Fig. 3-9).

Despite the sensitivity of the GAM, there did not seem to be a clear pattern between observing a difference in the predictive ability of either model fit to the two data constraints from the 5x2 CV and observing a difference in predicted outcomes in simulated weather conditions. For example, in field B1, there was strong evidence that the predictive ability of a GAM was greater using data up to the decision point compared to data from past years (Table 4), yet in both the wet and dry year simulations, there was no discernable difference in the pattern of profit maximizing N rates or the recommended total N applied to the field between a GAM fit with DP or PY data (Fig. 3). On the other hand, the 5x2 CV analysis of field I1 showed no indication of a difference in predictive ability for the GAM between DP and PY (Table 4), yet in the wet year there were some differences in the recommended pattern of profit maximizing N rates and total N applied between a GAM fit with data up to the decision point and a GAM fit with data from past years (Fig. 9).

Differences in the predictive ability of a given model between PY and DP constraints did not tend to translate into differences in management recommendations. Across all fields, the greatest difference in the pattern of recommended profit maximizing N rates and total N applied for either a wet or dry year was between the model forms, rather than between data constraints for a given model.

Management decisions for N fertilizer management require high quality data that is appropriate for usage under realistic data constraints. Using models under the DP constraint compared to the PY constraint resulted in greater predictive ability of the model in 5/28 cases. Thus, because of the ease of obtaining weather information, crop response models used to make N fertilizer management recommendations should be fit with data constrained to the point in time that decisions need to be made. Using models with better predictive abilities results in less uncertainty in predictions of yield and protein, an important objective for increasing the resiliency of farmer livelihoods, as their profits are predominately dictated by yield and protein (Hegedus & Maxwell, 2022a). In the cases where there was no difference in model prediction accuracy, such as for RF models fit to yield data, there are only potential benefits to fitting models with all the data that the farmer will have available, and models should still be fit with data up to the point where the farmer must make management decisions.

Open source remotely sensed satellite data was used for this study (Table 2) but using data up to a decision point rather than past years may have a greater benefit when collecting more accurate on-farm weather station data, likely to come available in the future. Data collected on soil moisture or weeds influencing soil moisture, for example, should be gathered up to the point that producers need to make decisions, so that information can be obtained in the year where decisions in that year affect outcomes. Only covariate data where measurements from the remote sensing sources varied over time (precipitation, GDD, NDVI, and NDWI) differed between the data constraints, while intransient edaphic and topographic variables that did not vary over time were held constant (Table 3). However, some of these edaphic variables realistically change over time, for example soil water content, which is sensitive not only to precipitation up until the decision point but throughout the growing season and has a large impact on crop yield and grain protein. This represents a limitation in the nature of the open-source data used in this study. Further limitations of the open-source data beyond not capturing realistic temporal changes are accuracy and scale. This study used the best open-source data available on hand and trusted that the developers of these datasets did due diligence in verifying the accuracy of measurements. Downscaling of remote sensing datasets from satellites constitutes a career of research itself, and was beyond the scope of this paper, though it must be recognized that imperfections and averaging across spatial scales could have contributed to the inability in discerning differences between the predictive ability of models between data constraints. However, despite these limitations, these tools still need to be applied towards the goal of improving efficiency of agricultural inputs. Collaboration between scientific disciplines is needed for improving the spatial scale and accuracy of measurements by ground truthing open-source satellite data to aid in informing farmer decision making at a low-cost.

There will still be uncertainty in predicted yield and protein responses when making management recommendations no matter the data constraint applied to training crop response models. The functional form of the model, rather than the data used to constrain a model, has a greater effect on uncertainty in predicted outcomes, highlighting the need for evaluating the most appropriate crop response model form for a given field in a given year. However, no matter the functional form of crop response models used, accumulating data over time increases producers' understanding of their fields and generates a rich backlog of data to use in precision agriculture decision support systems. These decisions support systems will be critical to sustaining the resources that agriculture relies on for production by harnessing the data from modern farms to increase producer net-returns and reduce pollution from inputs such as N fertilizer.

Conclusion

With increasing uncertainty faced by producers, decision support systems will need to be developed that reduce as much uncertainty in crop response predictions as possible for farmers to make the most informed management decisions. In all cases where there was evidence of a difference between data constraints, models using data up to the decision point that a farmer needs to make management decisions resulted in higher accuracies of crop response predictions. The increased ease of scraping site-specific data from the internet within analysis code relaxes much of the concern of analysis efficiency and thus data inclusions is of little concern, so there is no reason not to include data up to a decision point. Regardless of the data constraints used to fit a model, the greatest uncertainty in predicted management recommendations resulted from the functional form of the model itself, highlighting the need for decision support systems to assess various model types when providing field specific farmers management recommendations.

Acknowledgements

The authors wish to thank the team members of the On-Field Precision Experiment (OFPE) project and the Agroecology Lab at Montana State University. This research was supported by a USDA-NIFA-AFRI Food Security Program Coordinated Agricultural Project, titled "Using Precision Technology in On-farm Field Trials to Enable Data-Intensive Fertilizer Management," (Accession Number 2016-68004-24769), the USDA-NRCS Conservation Innovation Grant from the On-farm Trials Program, titled "Improving the Economic and Ecological Sustainability of US Crop Production through On-Farm Precision Experimentation" (Award Number NR213A7500013G021), and the Montana Fertilizer Advisory Council from 2016 - 2021.

References

Chen, K., O'Leary, R. A., & Evans, F. H. (2019). A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool. *Agricultural Systems*, *173*(December 2018), 140–150. https://doi.org/10.1016/j.agsy.2019.02.009

Clancy, M. (2019). Real-time protein monitoring. *Next Instruments*, *June*.

Cook, S., Cock, J., Oberthür, T., & Fisher, M. (2004). On-Farm Experimentation. *Better Crops*, *97*, 17–20.

Croft, H., Arabian, J., Chen, J. M., Shang, J., & Liu, J. (2020). Mapping within-field leaf chlorophyll content in agricultural crops for nitrogen management using Landsat-8 imagery. *Precision Agriculture*, *21*(4), 856–880. https://doi.org/10.1007/s11119-019-09698-y

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Estes, L. D., Bradley, B. A., Beukes, H., Hole, D. G., Lau, M., Oppenheimer, M. G., Schulze, R., Tadross, M. A., & Turner, W. R. (2013). Comparing mechanistic and empirical model projections of crop suitability and productivity: Implications for ecological forecasting. *Global Ecology and Biogeography*, *22*(8), 1007–1018. https://doi.org/10.1111/geb.12034

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, *36*(2). https://doi.org/10.1007/s13593-016-0364-z

Gaso, D. V., Berger, A. G., & Ciganda, V. S. (2019). Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images. *Computers and Electronics in Agriculture*, *159*(February 2018), 75–83. https://doi.org/10.1016/j.compag.2019.02.026

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328. https://doi.org/10.5424/sjar/2014122-4439

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Gotway, C. A., & Stroup, W. W. (1997). A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(2), 157–178.

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, *157*(2–3), 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

Hair, J. F. (2007). Knowledge creation in marketing: The role of predictive analytics. *European Business Review*, *19*(4), 303–315. https://doi.org/10.1108/09555340710760134

Hegedus, P.B. (2020). OFPE: On-Farm Precision Experiments (OFPE) Data Management and Analysis Tools. R package version 1.8.11. https://github.com/paulhegedus/OFPE.git

Hegedus, P.B & Maxwell, B.D. (2022a). Rationale for field specific on-farm precision experimentation. *Agriculture, Ecosystems & Environment.* Manuscript submitted for publication, in review.

Hegedus, P.B & Maxwell, B.D. (2022b). Assessing performance of empirical models for forecasting crop responses to variable fertilizer rates using on-farm precision experimentation. Manuscript in preparation.

Holland, D. M., De Oliveira, V., Cox, L. H., & Smith, R. L. (2000). Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics*, *11*(4), 373–393. https://doi.org/10.1002/1099-095X(200007/08)11:4<373::AID-ENV419>3.0.CO;2-2

Janatian, N., Sadeghi, M., Sanaeinejad, S. H., Bakhshian, E., Farid, A., Hasheminia, S. M., & Ghazanfari, S. (2017). A statistical framework for estimating air temperature using MODIS land surface temperature data. *International Journal of Climatology*, *37*(3), 1181–1194. https://doi.org/10.1002/joc.4766

Jaynes, D. B., & Colvin, T. S. (2006). Corn yield and nitrate loss in subsurface drainage from midseason nitrogen fertilizer application. *Agronomy Journal*, *98*(6), 1479–1487. https://doi.org/10.2134/agronj2006.0046

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, *11*(6), 1–15. https://doi.org/10.1371/journal.pone.0156571

Jones, C. (2013). *Mid- to late-season nitrogen fertilizer: a viable option to increase wheat yield and grain protein*. MSU News Service. https://www.montana.edu/news/11934/msu-extension-mid-to-late-season-nitrogen-fertilizing-can-boost-yield-protein-in-wheat-crops

Joshi, V. R., Kazula, M. J., Coulter, J. A., Naeve, S. L., & Garcia y Garcia, A. (2021). In-season weather data provide reliable yield estimates of maize and soybean in the US central Corn Belt. *International Journal of Biometeorology*, *65*(4), 489–502. https://doi.org/10.1007/s00484-020-02039-z

Lacoste, M., Cook, S., Mcnee, M., Gale, D., Ingram, J., Bellon-maurel, V., Macmillan, T., Sylvester-bradley, R., Kindred, D., Bramley, R., Tremblay, N., Longchamps, L., Thompson, L., Ruiz, J., Oscar Garcia, F., Maxwell, B., Griffin, T., Oberthür, T., Huyghe, C., … Hall, A. (2022). On-Farm Experimentation to transform global agriculture. *Nature Food*. https://doi.org/https://doi.org/10.1038/s43016-021-00424-4 On-Farm

Langella, G., Basile, A., Bonfante, A., & Terribile, F. (2010). High-resolution space-time rainfall analysis using integrated ANN inference systems. *Journal of Hydrology*, *387*(3–4), 328–342. https://doi.org/10.1016/j.jhydrol.2010.04.027

Lawes, R. A., Oliver, Y. M., & Huth, N. I. (2019). Optimal nitrogen rate can be predicted using average yield and estimates of soil water and leaf nitrogen with infield experimentation. *Agronomy Journal*, *111*(3), 1155–1164. https://doi.org/10.2134/agronj2018.09.0607

Lawrence, P. G., Rew, L. J., & Maxwell, B. D. (2015). A probabilistic Bayesian framework for progressively updating site-specific recommendations. *Precision Agriculture*, *16*(3), 275–296. https://doi.org/10.1007/s11119-014-9375-4

Liew, J., & Forkman, J. (2015). A guide to generalized additive models in crop science using SAS and R. *Communications in Biometry and Crop Science*, *10*(1), 41–57.

Luschei, E. C., Van Wychen, L. R., Maxwell, B. D., Bussan, A. J., Buschena, D., & Goodman, D. (2001). Implementing and conducting on-farm weed research with the use of GPS. *Weed Science*, *49*(4), 536–542. https://doi.org/10.1614/0043-1745(2001)049[0536:iacofw]2.0.co;2

Mariano, C., & Mónica, B. (2021). A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. *Computers and Electronics in Agriculture*, *184*(March). https://doi.org/10.1016/j.compag.2021.106094

Marques Ramos, A. P., Prado Osco, L., Elis Garcia Furuya, D., Nunes Gonçalves, W., Cordeiro Santana, D., Pereira Ribeiro Teodoro, L., Antonio da Silva Junior, C., Fernando Capristo-Silva, G., Li, J., Henrique Rojo Baio, F., Marcato Junior, J., Eduardo Teodoro, P., & Pistori, H. (2020). A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. *Computers and Electronics in Agriculture*, *178*(July). https://doi.org/10.1016/j.compag.2020.105791

Paccioretti, P., Bruno, C., Gianinni Kurina, F., Córdoba, M., Bullock, D. S., & Balzarini, M. (2021). Statistical models of yield in on-farm precision experimentation. *Agronomy Journal*, *113*(6), 4916–4929. https://doi.org/10.1002/agj2.20833

Qianqian, D., Mieno, T., Bullock, D. S., & Edge, B. (2021). Economically Optimal Nitrogen Side-dressing Based on Vegetation Indices from Satellite Images Through On-farm Experiments. In Harper Adams University (Ed.), *Proceedings of the 4th Symposium on Agri-Tech Economics for Sustainable Futures* (Issue September, pp. 51–62). HAU Publications.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stepanov, A., Dubrovin, K., Sorokin, A., & Aseeva, T. (2020). Predicting soybean yield at the regional scale using remote sensing and climatic data. *Remote Sensing*, *12*(12). https://doi.org/10.3390/rs12121936

Walsh, E. S., Kreakie, B. J., Cantwell, M. G., & Nacci, D. (2017). A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. *PLoS ONE*, *12*(7), 1–18. https://doi.org/10.1371/journal.pone.0179473

Wang, L., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop Journal*, *4*(3), 212–219. https://doi.org/10.1016/j.cj.2016.01.008

Wang, Y., Wu, G., Deng, L., Tang, Z., Wang, K., Sun, W., & Shangguan, Z. (2017). Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-07197-6

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Wright, M.N. & A. Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17. doi:10.18637/jss.v077.i01

Zeng, X., Han, B., Xu, F., Huang, J., Cai, H., & Shi, L. (2012). Effects of modified fertilization technology on the grain yield and nitrogen use efficiency of midseason rice. *Field Crops Research*, *137*, 203–212. https://doi.org/10.1016/j.fcr.2012.08.012

Zuur, A. F., & Camphuysen, Kees, C. (2012). 6 Generalized Additive Models applied on northern gannets. In *A Beginners Guide to Generalized Additive Models with R* (Issue January, p. 15). Highland Statistics Ltd.