Medium-resolution multispectral satellite imagery in precision agriculture: mapping precision canola (*Brassica napus* L.) yield using Sentinel-2 time series

Lan H. Nguyen[1,*], Samuel Robinson[1], and Paul Galpern[1]

[1] University of Calgary, Department of Biological Science, Calgary, AB T3B 4S3, Canada

* Correspondence: hoanglan.nguyen@ucalgary.ca

Abstract

Precision yield data is commonly recorded by modern combine harvesters and can be used to help growers optimize their operations. However, there have been very few attempts to predict variation in yield within a given field using multispectral satellite data. We used a precision yield dataset gathered in canola (*Brassica napus* L.) crops in central Alberta, Canada, and a time series of medium-resolution Sentinel-2 data collected over the growing season. Using two mapping methods, random forest regression and functional data analysis, we were able to predict crop yield to within 12-16% accuracy of actual yield, and to capture within-field variation. Our results demonstrate that time series of medium-resolution multispectral imagery is capable of mapping small-scale variation in crop yields, presenting new research and management applications for these techniques.

Keywords: crop yield; time series images; functional linear regression; random forest regression

1. Introduction

Precision agriculture (PA) uses technological innovations to increase production by conducting the right management practice at the right time and place. PA has been practiced commercially since the 1990's (*Mulla, 2013*) and is now deployed widely across the North American agricultural sector. For example, in the United States, guidance systems are used in about half of all planted acres, and GPS-based yield mapping and variable-rate technology are used in about 40% and 30% of operations, respectively (*Schimmelpfennig, 2016*). Adoption is even higher in Canada, as a survey by Agriculture and Agri-Food Canada indicated that 84% of Canadian farmers are currently using PA and have combine yield monitoring capability, and 75% of participants said they would use more PA in the future (*Steele, 2017*).

Remotely sensed imagery is a key data source for PA, summarizing soil properties (e.g., organic matter, moisture, pH) or plant conditions (e.g., crop nutrients, biomass, yield, water/heat stress, weed infestation, insects and plant diseases) (*Thorp & Tian, 2004; Liaghat & Balasundram, 2010; Ge et al., 2011; Mulla, 2013; Chlingaryan et al., 2018*). Crop yield maps are perhaps the most important data products for crop management, and medium-resolution multispectral satellite datasets (≥ 10-meter) have long aided in this objective, providing insights at coarse resolution over large areas (*Bauer et al., 1978; Rudorff & Batista, 1991; Doraiswamy et al., 2004; Bala & Islam, 2009; Liu et al., 2010; Salmon et al., 2015; Hunt et al., 2019; Sakamoto et al., 2020*). However, airborne multispectral, airborne hyperspectral, and high-resolution satellite imagery have typically been favoured for PA because of a closer match to the scale of variation typically observed within fields.

Medium-resolution multispectral satellite data has several advantages that make it attractive for use in PA. First, it is often free to the public, while the high-resolution or hyperspectral data can

be costly. Second, medium-resolution multispectral platforms (such as Landsat 8 or Sentinel-2) offer extra spectral bands, including Red Edge and Shortwave Infrared, which are useful in crop mapping and monitoring (*Liu et al., 2004; Herrmann et al., 2010; Delegido et al., 2013; Braga et al., 2020*). Lastly, data products from the medium-resolution multispectral satellites, especially those from space agencies such as NASA (National Aeronautics and Space Administration), JAXA (Japan Aerospace Exploration Agency), and ESA (European Space Agency), are already preprocessed to high-quality for time series analysis.

To our knowledge, only *Thenkabail (2003)* and more recently *Hunt et al. (2019)*, have attempted to use medium-resolution satellite imagery for monitoring within-field conditions. *Thenkabail (2003)* showed that Landsat Thematic Mapper images can be used to quantify between- and within-field variation in biophysical quantities (e.g., Leaf Area Index, biomass) of six crops, and classified quantities with an overall accuracy of 81%. They found a significant relationship between combine yield monitor data and Landsat-derived NDVI ($R^2 = 0.77$). *Hunt et al. (2019)* produced high resolution wheat yield maps at 10-meter resolution using Sentinel-2 imagery in random forest regression models (RMSE of 0.61 tonnes/ha). Incorporating environmental datasets further improved predictions (RMSE of 0.66 tonnes/ha). Although *Thenkabail (2003)* and *Hunt et al. (2019)* have successfully demonstrated that medium-resolution satellite imagery can be used in mapping precision crop yield, there remains potential to improve mapping method, and to better capture within-field variation. In addition, as climate, crop type and its spectral response can vary greatly due to geographic location, more studies are still needed to demonstrate the performance and limitations of medium-resolution imagery in precision agriculture across a broad range of field conditions.

Our objective in this study is two-fold. First, we aim to demonstrate the capability of medium-resolution multispectral satellite imagery, e.g. Sentinel-2, in mapping precision canola yield (dry seed mass in tonnes/ha). Second, we propose two yield mapping approaches that can be applied to any climate region and crop type. Because a single observation in time may not represent the best yield-spectra relationship, we propose two mapping methods that utilize the rich Sentinel-2 time series images. Accuracy of the proposed approaches is assessed by mean-absolute-errors (MAE) and a visual comparison between actual and predicted yield maps. Our results show how medium-resolution satellite imagery can be used to predict within-field yield variability, and they have potential applications in variable-rate fertilization and the evaluation of within-field and whole-field profitability. In addition, because our proposed mapping approach depends only on the presence of a remote sensing time series, and on no other geographically-determined variables, there is potential to extend it to other areas and crop types.

2. Data and methods

2.1. Data

2.1.1. Precision canola yield

We used precision yield data from eight canola fields (CF) located in the County of Vermilion River, Alberta, Canada (relative locations are shown in order to protect the data owner's privacy; Figure 1). Fields were seeded using three canola varieties (Liberty L234, L255, and Pioneer P501) in May of 2019, using standard seeding densities (~60 plants/m^2). Fields were treated with pre-emergence herbicide, and 1-2 passes of post-emergence herbicide, depending on weed and stand establishment. From late August to October, fields were first swathed and then harvested using a Case IH 8230 combine with a 10.7 m header bar. Precision yield data was recorded in segments by the combine's on-board yield monitor. Each segment is characterized by a starting

position of the combine (Figure 2b), width of the header bar (m), direction of travel (0-360° N), the length of a recorded segment (m), and the total dry mass of canola (tonnes). We used these attributes to construct harvested segments (polygons) within each field (Figure 2c), and rasterized the yield from all polygons using an area-weighted average yield (tonnes/ha) for each pixel of the Sentinel-2 images (Figure 2c,d). We only retained pixels that had at least 95% of their area covered by harvested polygons (Figure 2).

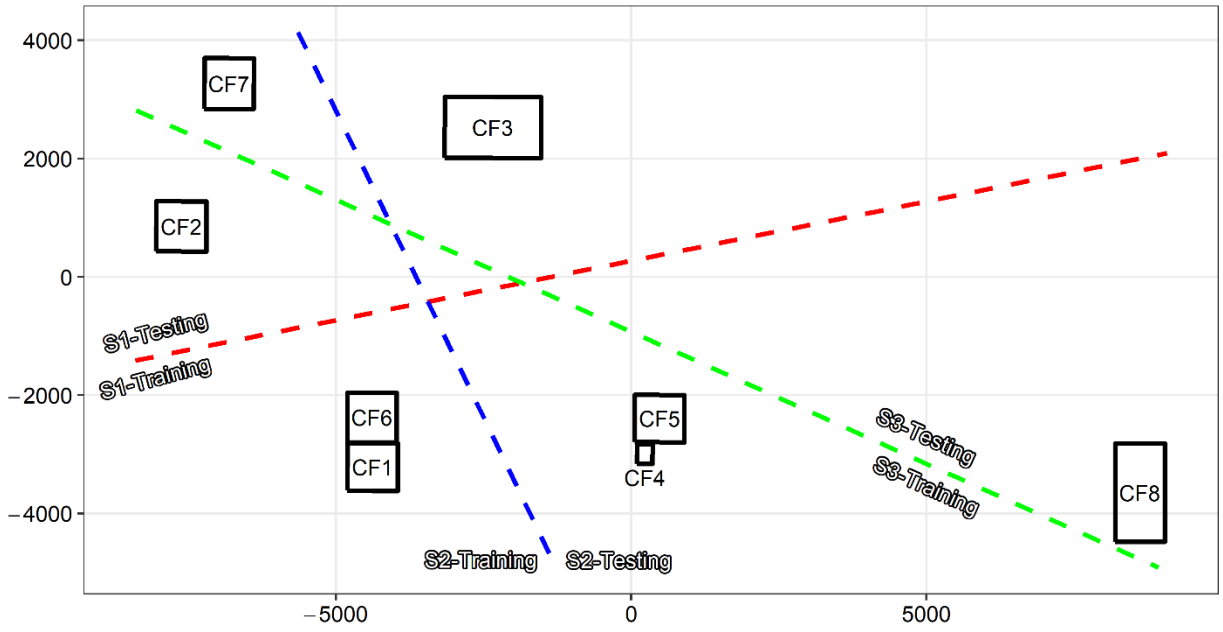


Figure 1. Relative locations of canola fields (CF1-CF8) with precision yield data shown using a UTM-based pseudo coordinate system (both axes are in meters). The fields for the three training scenarios (see Section 2.3.3) are separated by coloured lines.

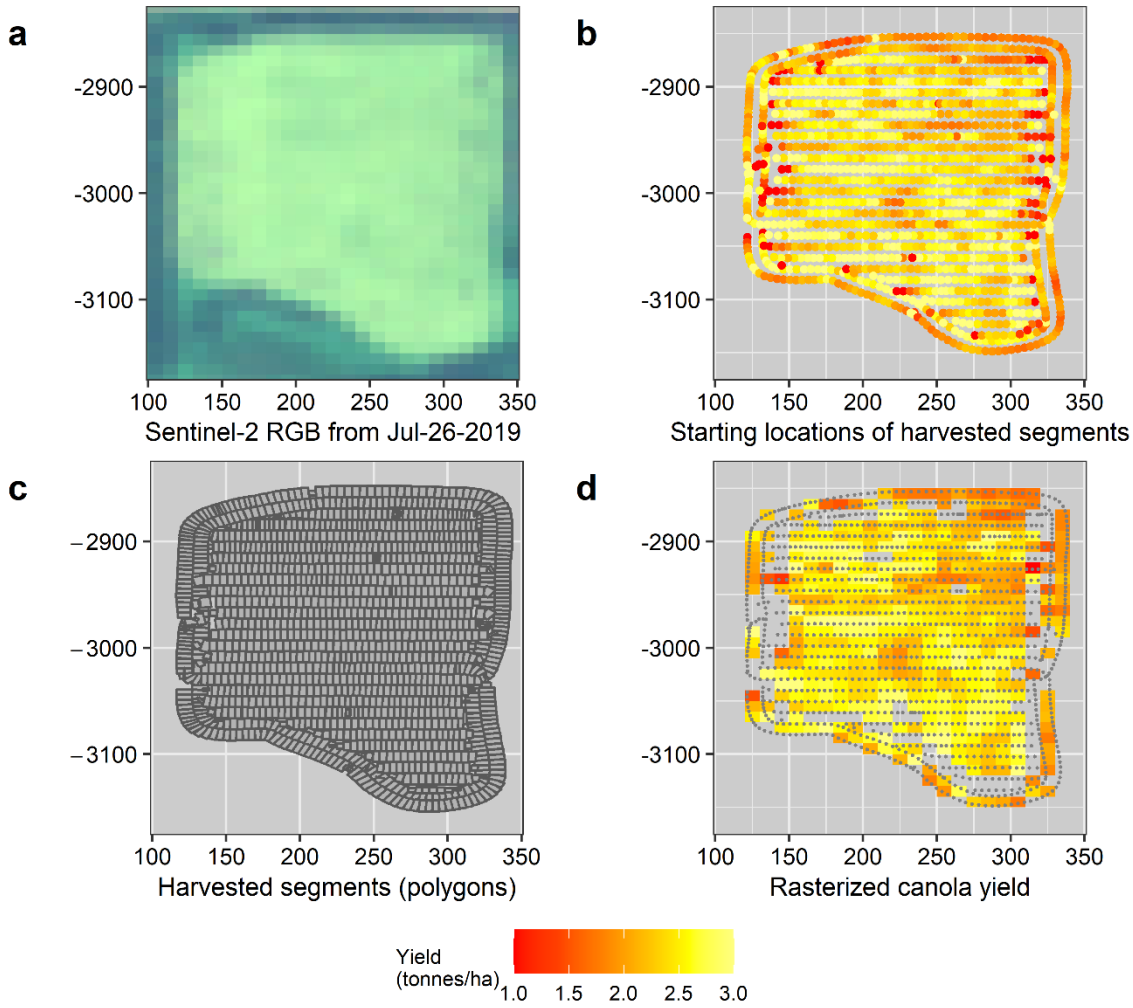


Figure 2. Preparation of precision canola yield from raw yield monitor data at CF4: (a) a sample Sentinel-2 RGB image of the field, (b) starting positions of the harvested segments, (c) harvested segments that were constructed from recorded attributes, and (d) precision canola yield in raster. Axes are in a UTM-based pseudo-coordinate system (meter)

2.1.2. Sentinel-2 images

Sentinel-2 is a European wide-swath, high-resolution, multi-spectral imaging mission. The mission specification of the twin satellites is designed to give a high revisit frequency of 5 days at the Equator. Each satellite carries an optical instrument payload that samples 13 spectral bands: four bands at 10 m (including Red, Green, Blue and NIR), six bands at 20 m, and three

bands at 60 m spatial resolution. There are two Sentinel-2 products available for users: Level-1C (top-of atmosphere reflectance) and Level-2A (bottom-of-atmosphere reflectance). Both products are available in 100x100 km^2 tiles and delivered free-of-charge through multiple sources, including the Google Earth Engine (GEE) platform. In this study, we used Level-1C product—the only product available for the study area through the GEE.

We downloaded all top-of-atmosphere Sentinel-2 tiles between Apr-01-2019 and Oct-31-2019 from GEE. A built-in cloud/snow mask was applied for each image to remove unreliable observations. In total, 67 Sentinel-2 images were available for the study area (Table 1). Among eleven spectral bands, only seven were used in the study, including three visible (B2-Red, B3-Green, B4-Blue), two NIR (near infrared, B8 and B8A), and two SWIR (shortwave infrared, B11 and B12) bands. After cloud/snow filtering, the remaining good observations in each band were stacked to create a time series dataset at each pixel.

Table 1. Sentinel-2 images used in the study.

Field Name	# of images	DOYs (days-of-year)	# of good obs.	# of pixels
CF1	67	96, 98, 106, 108, 111, 113, 118, 121, 123, 126, 131, 133, 136, 138, 141, 143, 146, 148, 151, 153, 158, 161, 163, 166, 168, 171, 173, 176, 178, 183, 186, 188, 191, 193, 196, 201, 203, 206, 208, 211, 213, 216, 218, 221, 226, 228, 231, 236, 238, 243, 246, 248, 258, 261, 263, 266, 273, 276, 278, 281, 283, 286, 288, 291, 293, 301	40	3358
CF2			43	4604
CF3			37	6933
CF4			42	440
CF5			40	4839
CF6			40	1868
CF7			41	3269
CF8			36	10826

Using the seven spectral bands, we also computed two spectral indices: normalized difference vegetation index (NDVI; *Tucker, 1979*; *Huete et al., 1997*) and normalized difference water index (NDWI; *Gao, 1996*). NDVI [(NIR - Red)/(NIR + Red)] is an indicator of the greenness of

the biomass (photosynthetic activity) while NDWI [(NIR - SWIR1)/(NIR + SWIR1)] is known to be strongly related to the plant water stress, making these two indices good proxies for plant health and productivity.

A current trend in remote sensing is to utilize a time series dataset to explore the spectral signatures of studied objects, as a single snapshot image may not capture the best relationship between an object of interest and its corresponding spectral signature. In addition, observations were often not uniform within and across fields due to a limited swath width and cloud cover, making a single-image approach impractical. There are two common approaches for using time series: (1) generating statistical composites of spectral bands and indices (statistical features) over a given time window, (e.g., growing season) and (2) reconstructing the time series to retrieve gap-free dataset at every pixel. In this study, we assessed both approaches of exploiting time series images in mapping precision yield.

Statistical features

From the time series of seven spectral bands and two indices, we computed six sets of annual statistical features, including the Min, Max, Mean, and 10th, 50th, and 90th percentiles of each pixel. For each Sentinel-2 tile, 54 composites (9 data layers x 6 statistical features) were generated.

Reconstructed NDVI and NDWI time series

Since the masking process on GEE is not able to completely remove unreliable observations from Sentinel-2 images (Figure 3), we applied an additional filter (Appendix A) to further reduce noise in the NDVI time series. Masked data points in the NDVI time series were also removed from the NDWI time series. We then retrieved daily NDVI and NDWI values between Apr-01

and Oct-30 (day-of-year or doy: from 91 to 304) from the corrected time series by a simple linear interpolation (Figure 3).

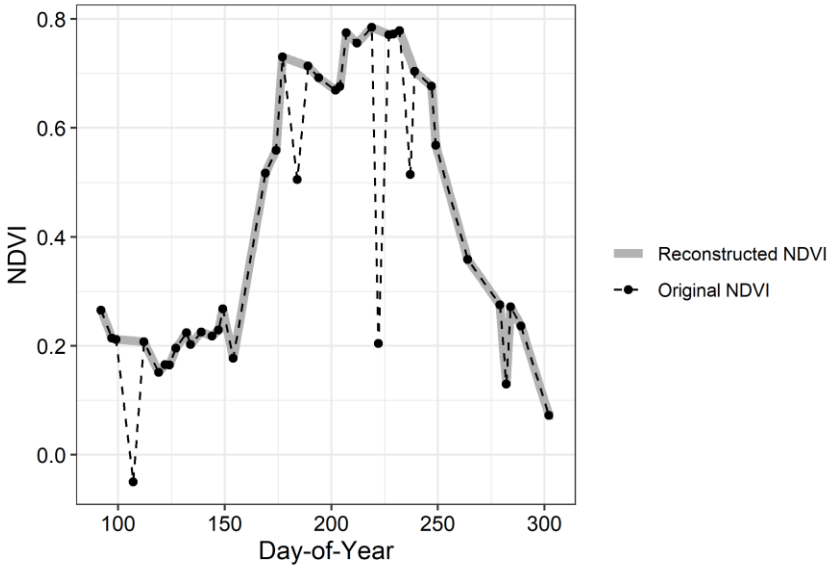


Figure 3. Reconstruction of NDVI time series at a sample location (pixel) of CF7. The additional filter was applied to further reduce noise in the GEE-retrieved (original) NDVI time series. Daily NDVI values (reconstructed NDVI) were obtained from the corrected time series by a simple linear interpolation. Note that additional filter cannot completely remove cloud-contaminated observations (e.g., at DOY 280) but majority of large noises were successfully removed.

2.2. Mapping precision canola yield

We used two statistical approaches for modelling canola yield given a time series of remotely sensed images. First, we used statistical features of each pixel as inputs into a random forest regression (RFR; *Breiman, 2001*). RFR has been used widely in agricultural remote sensing to predict crop health and development using spectral information (*Zhou et al., 2016*; *Liang et al., 2018; Hunt et al., 2019; Sakamoto, 2020*). Secondly, we used functional data analysis (FDA; *Ramsay & Silverman, 2005*) to predict the canola yield using the reconstructed NDVI/NDWI

time series, because it is more natural to think about plant growth as a continuous process (in a functional form of NDVI/NDWI time series) rather than sequence of individual observations. A specific type of FDA—functional linear regression (NDVI/NDWI time series) with scalar responses (canola yield)—was used in this study (*Kokoszka & Reimherr, 2017*).

2.2.1. Random Forest Regression

Random forest regression is an ensemble of randomized regression trees, each created with a random subset of training samples and features. The random forest predictor is then retrieved by averaging the results of all individual trees. Performance of RFR—prediction accuracy and computational time—can vary widely due to size of the sample dataset (*sample_size*) as well as the tuning of hyper-parameters (e.g., the number of regression trees: *n_estimators*, the number of features tested at each node: *max_features*, and the minimum number of samples required to split an internal node: *min_sample_split*). We experimented with different parameter settings, deciding on the following "optimal" settings (good balance between prediction accuracy and computation time) : *n_estimators* = 200, *max_features* = 54, *min_sample_split* = 20, *sample_size* = 50% of available pixels. We performed all random forest regression in Python using *scikit-learn* library (*Pedregosa et al., 2011*).

2.2.2. Functional Linear Regression

A functional linear regression (FLR) models crop yield, y, as:

$$y = f(X, \beta) + \varepsilon = \int X(t)\beta(t)dt + \varepsilon \qquad \text{(Model 1)}$$

where X is the value of a predictor variable at time t (NDVI and/or NDWI, in our case), while β is the instantaneous effect (slope) of that variable on y. One way of estimating β is to present the parameters (β) and the functional covariates (X_i) as a finite sum of pre-defined basis elements:

$\beta(t) = \sum_k \beta_k \theta_k(t) = \theta' b$; $X_i(t) = \sum_k c_{i,k} \psi_k(t) = C\Psi$. Replacing β and X of model 1 by their new forms results in model 2—a typical multiple linear regression.

$$y = f(X, \beta) + \varepsilon = C\Psi\theta' b + \varepsilon = Zb + \varepsilon \qquad \text{(Model 2)}$$

We performed the functional linear regression between precision canola yield and remote sensing time series in R using the *fda.usc* package (*Febrero-Bande & Oviedo de la Fuente, 2012*). We tested different types of basis functions for both β and X, and FLR using a B-spline basis (order of 4 and 4 control points) returned the best yield prediction.

2.2.3. Training-testing scenarios and accuracy assessment

We divided the eight canola fields into training fields and testing fields (Figure 1), in order to test how well an empirical model from a given area performs in another area with no reference data. Three training-testing scenarios were examined (Figure 1, Table 3). In all scenarios, we intentionally left CF5 in the testing set, because this field displayed obvious patches of higher/lower canola yield that would be helpful in evaluating performance of our models.

Table 2. Training-testing scenarios

Scenario	Training Fields	Training Pixels	Testing Fields	Testing Pixels
S1	CF2, CF3, CF7	14806	CF1, CF4, CF5, CF6, CF8	21331
S2	CF1, CF2, CF6, CF7	13099	CF3, CF4, CF5, CF8	23038
S3	CF3, CF7, CF8	21028	CF1, CF2, CF4, CF5, CF6	15109

Note: RFR models used only a half of "Training Pixels"

Performance of both RFR and FLR in mapping precision canola yield was compared using Mean-absolute-errors (MAE) at scenario-level and through a visual inspection of yield maps and regression residual maps.

3. Results

3.1. Descriptive statistics of precision canola yield

Figure 4 shows the descriptive statistics for precision canola yield (tonnes/ha) measured at the field-level and scenario-level. There was substantial variation between fields (Figure 4a). The least productive field (CF3; median value of 2.138 tonnes/ha) had approximately 18% lower yield than the most productive field (CF5; median value of 2.592 tonnes/ha). Across the three scenarios, S3 showed the most similar yield distribution between the training and testing fields (Figure 4b), while in other two scenarios, the testing fields were clearly more productive than the training fields. This is especially so in S1 where the difference in median yield was 13.7%.

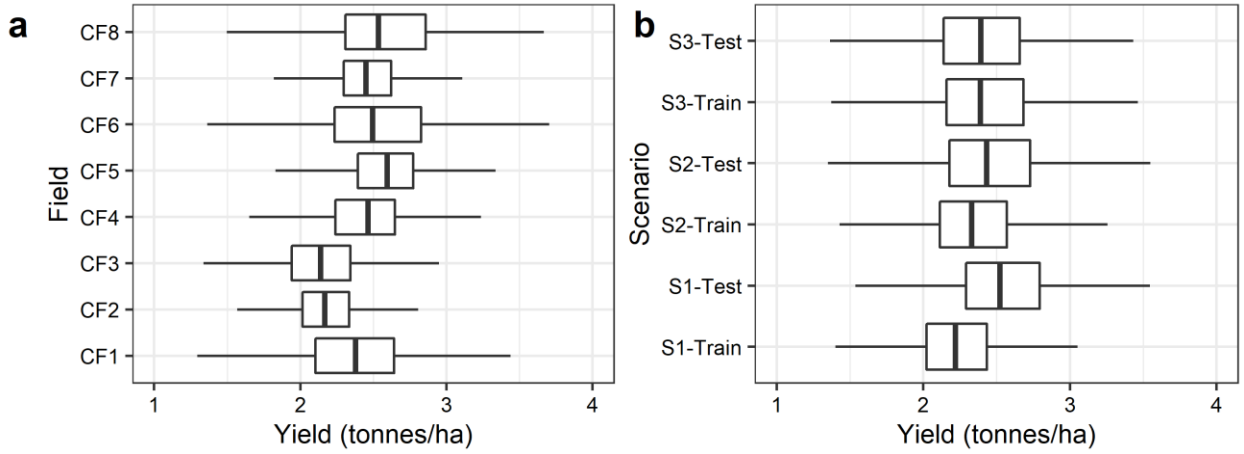


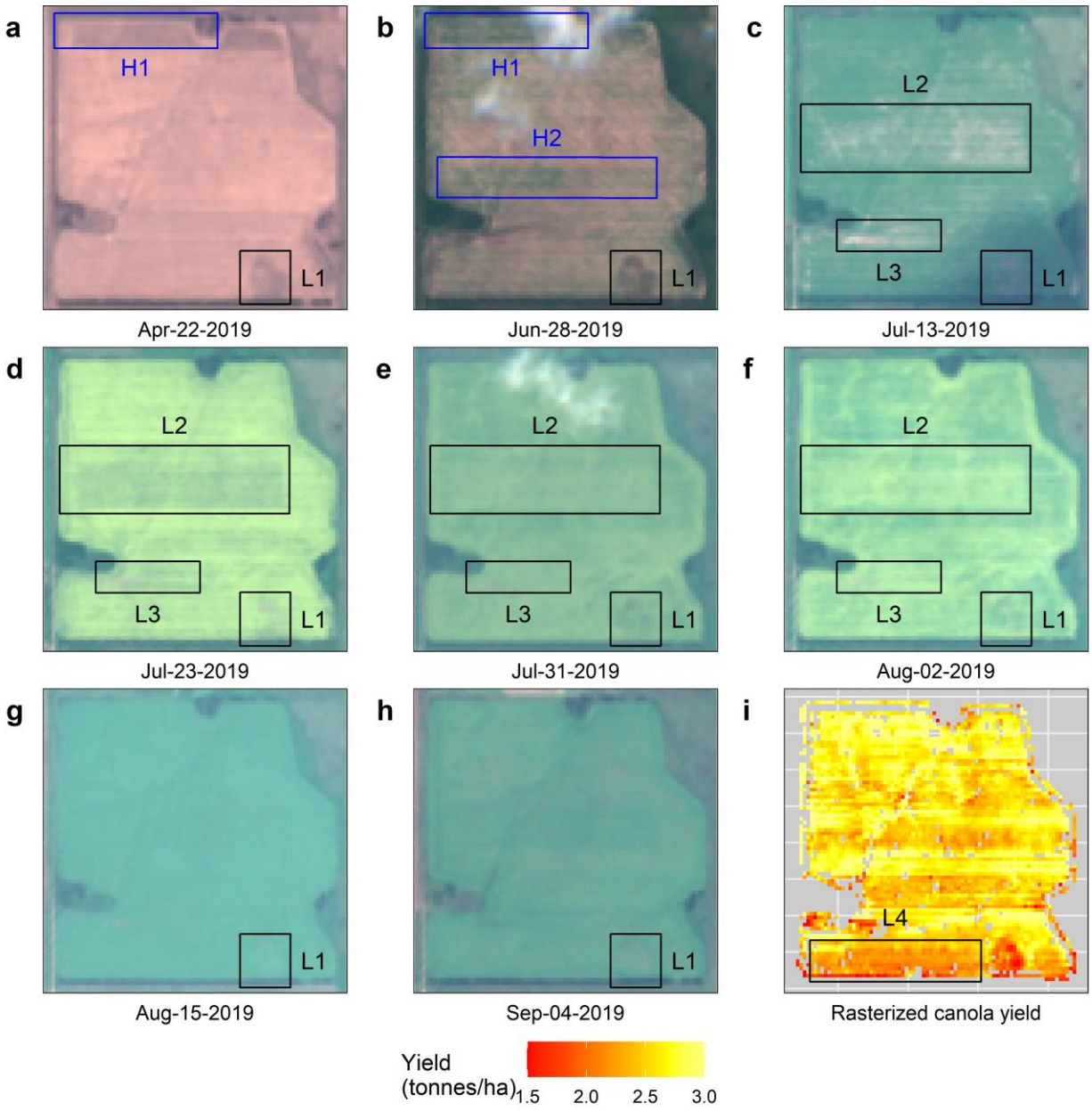
Figure 4. Distribution of precision canola yield at the (a) field level and (b) scenario level.

3.2. Qualitative visual analysis

Visual inspection of CF5 over the growing season (planted: May-15-2019 and harvested: Sep-24-2019; Figure 5a-h) revealed patches where canola productivity was lower (black boxes: L1, L2, L3) or higher (blue boxes: H1, H2) than other areas of the field. Soil conditions in the field are partly visible in the pre-planting image, where two patches appear darker than other parts (blue and black boxes; Figure 5a). Day 43 after the planting (Jun-28-2019), while most of CF5 was still soil-coloured, two patches—one at the top (H1) and another in the middle of field

(H2)—showed a strong green signal (Figure 5b), while the small patch at the bottom-right corner (L1) is still shaded in dark soil colour. Two weeks after that (Jul-13-2019), while the canola has greened up on most of CF5, a few patches still appear greyish (Figure 5c). Although L1 did not appear clearly in Figure 5c because of cloud cover, it is still easy to see that canola did not grow well there. Low or high growth patches of CF5 remained on RGB images until after flowering (flowering: Jul-23-2019, early ripening Aug-02-2019), especially at L1. RGB images at the end of season did not show differences between low or high growth patches (Figure 5g,h). The spatial distribution of canola yield (Figure 5i) is clearly reflected in the RGB images (Figure 5a-b), but we also observed a patch of lower yield at the bottom edge of the field (L4) that was not reflected in the visual inspection.

NDVI and NDWI images over the growing season were also related to canola yield (Figure 6). Similarities between yield and NDVI/NWDI were clearly visible in early season images (Jun-28-2019). While the pattern of low / high growth continued until very late in the canola season (Sep-04-2019) in the NDVI images, we did not see strong correlation between yield and NDWI after Jul-13-2019, except at L1 and L3. L1 and L3 showed consistently lower values in NDVI and NDWI over the entire season.



235

236 Figure 5. Sentinel-2 RGB images over the growing season (a to h) and precision canola yield (i) of CF5.

237 L1, L2, and L3 indicate low-yield patches, while H1 and H2 indicate high-yielding patches,

238 demonstrating that spatial patterns in yield are sometimes visually discernible across a time series of

239 true colour imagery.

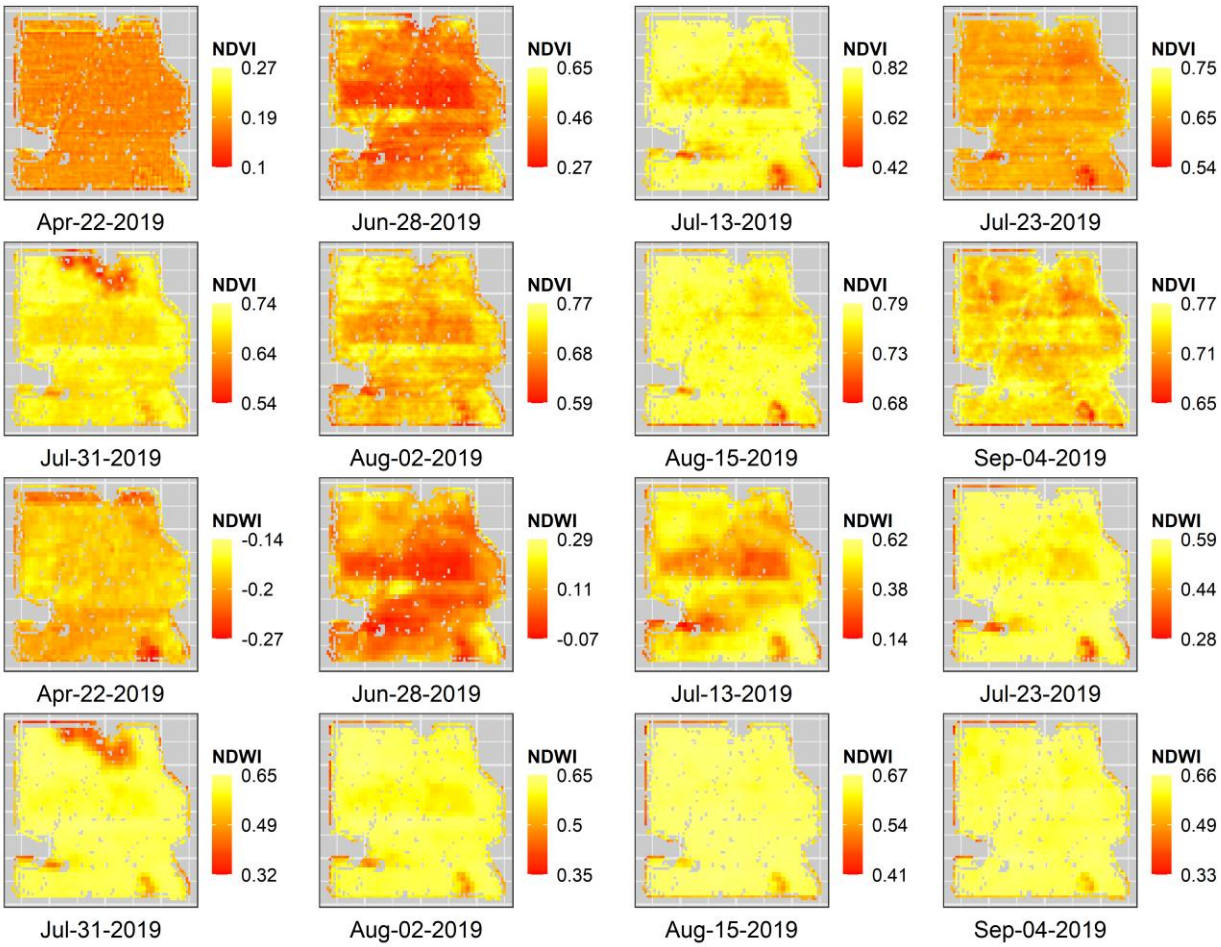


Figure 6. Snapshots of NDVI and NDWI over the growing season at CF5.

3.3. Predicted canola yield: Random Forest Regression vs. Functional Linear Regression

RFR had better prediction accuracy in the training datasets but performed similarly to FLR on testing datasets (i.e., out-of-field prediction; Table 3). S3 had the lowest prediction error, but differences between S3 and the other two scenarios were minor. Although RFR clearly performed better for the training fields, predictions of FLR were slightly more accurate for all three scenarios, indicating that RFR was likely overfitting. Overall, the testing MAEs on the testing dataset ranged from 300 to 390 kg/ha, approximately 12.6% to 15.5% of the median yield in each scenario. FLR models using solely NDVI or NDWI time series were similar, but the

NDVI time series provided more spatial detail due to the finer resolution of NDVI images (Red and NIR: 10 m, SWIR1: 20 m). Using both NDVI and NDWI yielded the best predictions in scenarios S1 and S3.

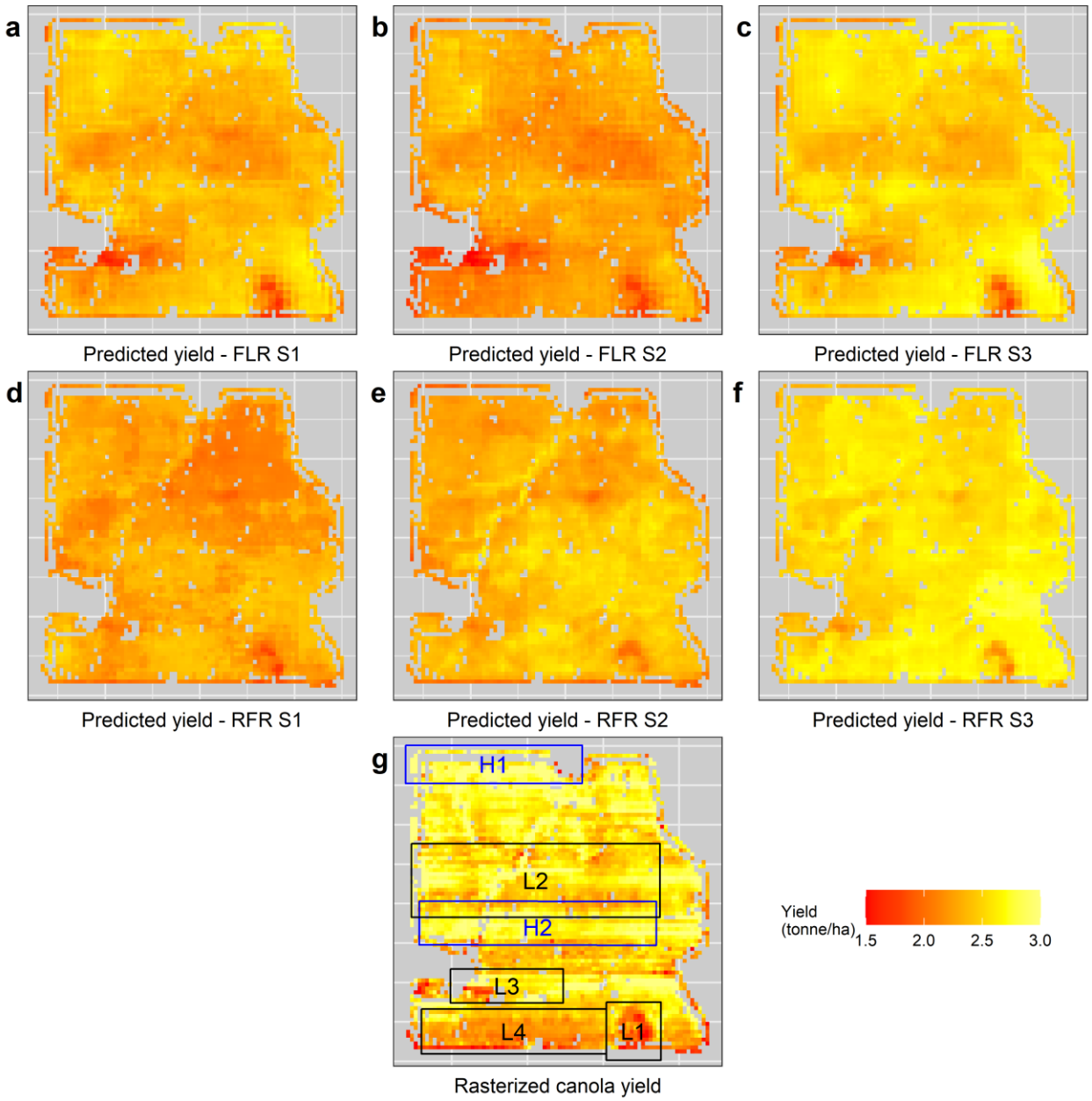
Table 3. Mean-absolute-errors of RFR and FLR. Both approaches performed similarly on the testing dataset with MAEs equivalent to 12%-16% of the median yields.

Scenario	Model	MAE Training (tonnes/ha)	MAE Testing (tonnes/ha)	MAE Testing (% of Median)
S1	FLR_NDVI	0.2152	0.3465	13.74
S1	FLR_NDWI	0.2281	0.3232	12.82
S1	FLR_Both	0.2082	0.3382	13.42
S1	RFR_Stats	0.1035	0.3912	15.52
S2	FLR_NDVI	0.2605	0.3207	13.20
S2	FLR_NDWI	0.2775	0.3212	13.22
S2	FLR_Both	0.2568	0.3486	14.35
S2	RFR_Stats	0.1299	0.3412	14.04
S3	FLR_NDVI	0.2732	0.3045	12.74
S3	FLR_NDWI	0.2650	0.3136	13.12
S3	FLR_Both	0.2624	0.3005	12.57
S3	RFR_Stats	0.1191	0.3014	12.61

FLR_NDVI: FLR using only NDVI time series; FLR_NDWI: FLR using only NDWI time series; FLR_Both: FLR using both NDVI and NDWI time series; RFR_Stats: RFR using statistical features

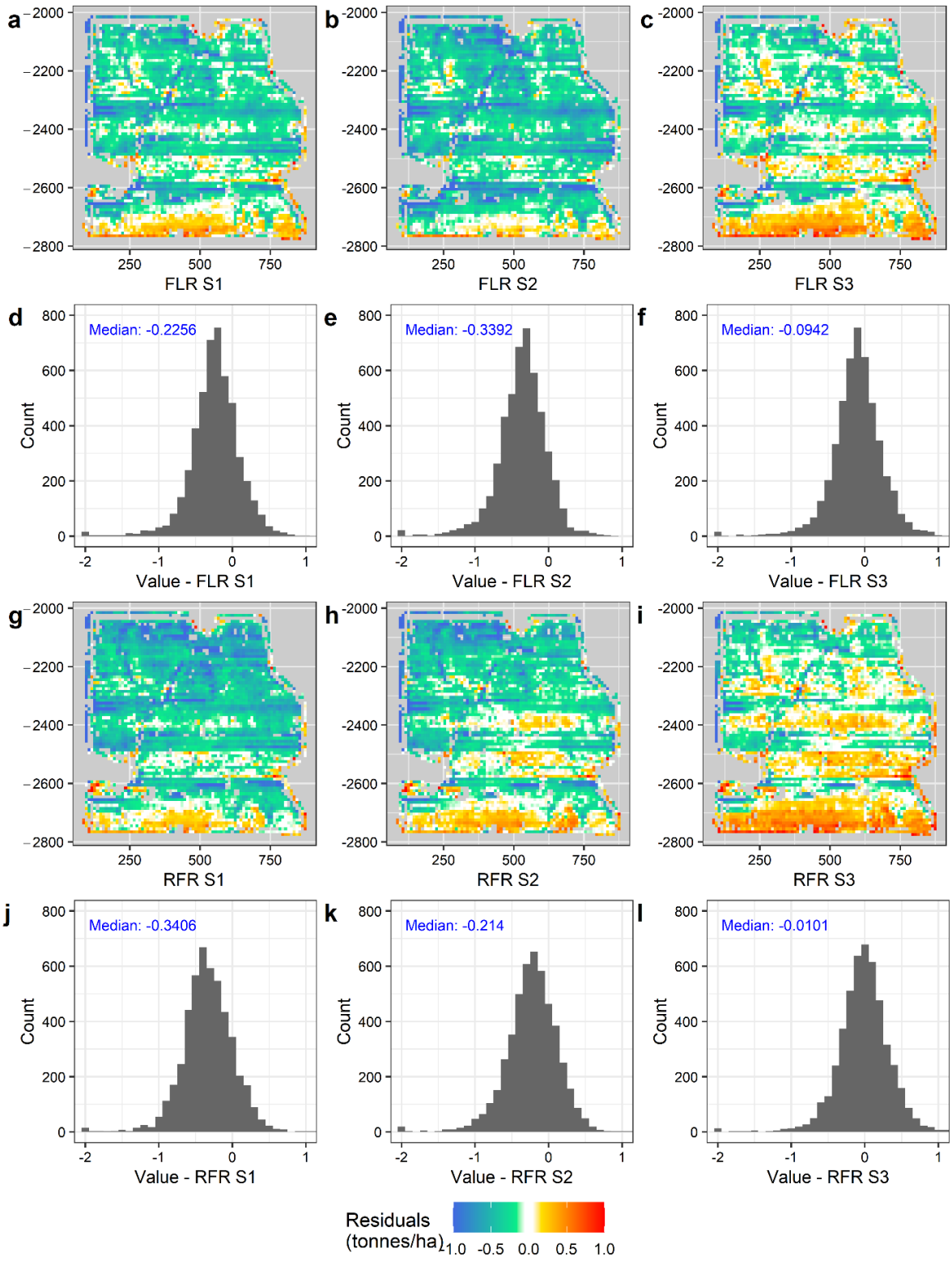
The results from field CF5 are helpful for explaining our overall results. Figure 7 shows predicted canola yield maps of CF5 from both RFR and FLR for three scenarios using a heat map (low yield: Red → high yield: Yellow). All yield maps were able to predict areas with the lowest yield (L1 and L3). Spatial yield patterns in FLR maps, especially in S3, appear to match the yield data better than those from RFR as they successfully predicted higher yields in H1 and H2, and lower yields in L1 to L3. However, none of models were able to predict a patch of lower yield at the bottom of CF5 (L4).

Figure 8 shows the spatial distribution of residuals (predicted - actual) in field CF5, showing distinct areas of negative and positive residuals. Both FLR and RFR predictions overestimated the actual canola yield in the northern part of the field. Predictions in scenario S3 are less biased as we see a good balance between the blue-green and the yellow-red colours (Figures 8f, 8i). In all three scenarios, both RFR and FLR consistently overestimated yield in areas with low actual values (especially in L4) and underestimated yield in areas with high actual values. Histograms of residuals (Figure 8d-f, 8j-i) confirm that the predicted yield maps underestimated the actual canola yield for field CF5. Median yield residuals in scenarios S1 and S2 (-0.2256, -0.3392 and -0.3406, -0.2140 tonnes/ha for FLR and RFR, respectively) were more positively biased than those in S3 (-0.0942 and -0.0101 tonnes/ha for FLR and RFR, respectively).



273

274 Figure 7. Spatial pattern of actual vs. predicted canola yield at CF5. L1-L4 (black boxes) indicate low-
275 yield patches, while H1 and H2 (blue boxes) indicate high-yield patches. Both NDVI and NDWI time
276 series were used in FLR models.



277

278 Figure 8. Spatial distribution and histogram of residuals at CF_5 (predicted - actual). Both NDVI and

279 NDWI time series were used in FLR models.

3.4. Contribution of remote sensing features in canola yield mapping.

We evaluated the contribution of remote sensing variables in both FLR (Table 4) and RFR (Table 5) for S3—the most accurate scenario. Those results help to explain findings in the previous section that patches of low or high growth were identifiable from early season NDVI images and were still distinguishable late in the season (around the end of the canola ripening stage). The early season NDWI image (Jul-13-2019) also showed a similar pattern with canola yield.

Among the 5 components of the NDVI b-spline basis, #3 and #4 were the more significant and larger (in terms of absolute values) than the other three (Table 4), suggesting that satellite observations around the peaks of those components were most influential for prediction—day 200 (flowering stage) for #3 and day 255 (early in the ripening stage) for #4 (Figure 9). Five components of NDWI were equally significant. In S3, the FLR model using NDVI performed better than the model using NDWI, meaning that FLR models picked up more information from the NDVI signal.

We aggregated scores of strongly correlated statistical features: “min” and “10^{th} percentile” into “low”, “mean” and “50^{th} percentile” into “medium”, “max” and “90^{th} percentile” into “high”. The RFR models showed that the peak NDVI value and the median NDWI value were the most important variables for predicting yield (Table 5). Peak NDVI is often observed in the later-season pod-ripening stage. The second most important feature was “medium” value of NDWI, which was observed around the end of the stem-elongation stage at approximately Jul-16-2019.

Table 4. Summary of FLR model using both NDVI and NDWI time series for S3 scenario.

Components	Estimate	Std. Error	t value	Pr(>|t|)	Significance
Intercept	2417.3871	2.2934	1054.071	< 2e-16	***
NDVI.bspl4.1	3.6266	0.8158	4.446	8.81e-06	***
NDVI.bspl4.2	0.7758	0.6073	1.277	0.201515	
NDVI.bspl4.3	-5.1700	0.7897	-6.547	6.00e-11	***
NDVI.bspl4.4	4.5442	0.6072	7.484	7.49e-14	***
NDVI.bspl4.5	-1.2803	0.3702	-3.458	0.000545	***
NDWI.bspl4.1	-4.8260	0.6014	-8.025	1.07e-15	***
NDWI.bspl4.2	3.8767	0.4213	9.202	< 2e-16	***
NDWI.bspl4.3	-5.0004	0.4918	-10.168	< 2e-16	***
NDWI.bspl4.4	4.8387	0.3630	13.329	< 2e-16	***
NDWI.bspl4.5	-3.2371	0.2474	-13.086	< 2e-16	***

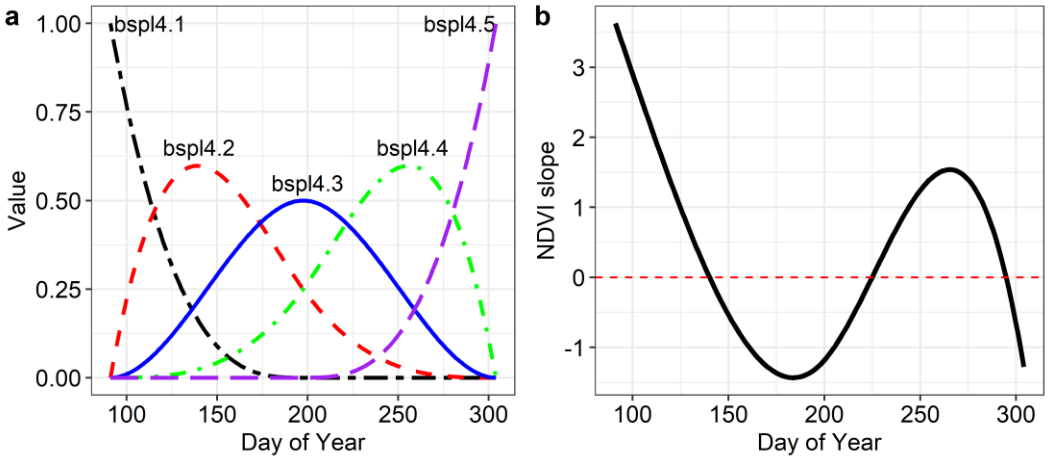


Figure 9. b-spline basis with order of 4 (a) and NDVI slope as a function of time — β(t) (b)

Table 5. Top 10 most important features of RFR model for S3.

Rank	Feature	Score	Rank	Feature	Score
1	NDVI_high	5.78	6	Band8(NIR1)_low	2.01
2	NDWI_medium	4.49	7	NDWI_low	1.98
3	Band2(Blue)_medium	2.36	8	Band11(SWIR1)_high	1.98
4	Band8(NIR1)_medium	2.25	9	NDWI_high	1.87
5	Band8A(NIR2)_low	2.17	10	Band12(SWIR2)_low	1.79

4. Discussion

It is worth noting that the eight canola fields cover a large area of approximately 360 hectares (36,137 pixels). However, canola growth dynamics and its corresponding spectral response still might not be fully captured in a few large and homogeneous training fields.

At the field-scale, our predicted yields had MAEs ranging from 0.30 to 0.39 tonnes/ha, equivalent to 12.6%-15.5% of the median yields (Table 3). This result is promising for several reasons. First, we predicted precision yield at a medium spatial resolution (10 m). Secondly, even though our reference dataset was relatively small (only 8 canola fields), our method was still able to predict well. Lastly, no ancillary data, such as soil moisture, climatic conditions, crop variety, or agricultural practices, were integrated into any of our models. Those variables are commonly used inputs of crop yield mapping (e.g., *Prasad et al., 2006*; *Jeong et al., 2016*; *Kern et al., 2018*). Nevertheless, our MAEs are in the same range as those reported in *Mkhabela et al., 2011* (0.08 – 0.4 tonnes/ha*),* which forecast regional canola yield in the Canadian Prairies based on a simple regression that used multiple-years of MODIS-derived NDVI to predict yield statistics.

Our prediction accuracy could probably be improved slightly if the training and testing pixels were selected randomly from the data pool (the training data would contain pixels from all studied fields). However, our purpose in selecting training or testing datasets was not only to achieve the highest prediction accuracy but also to create a geographic gradient to test how well an empirical model from one area performed in another area without reference data.

Our results showed that both RFR and FLR models performed the best in scenario S3 (Table 3). One possible reason for this was that the similarity in yield distribution between the training and testing datasets were highest for S3 (Figure 4b). A further examination of CF5 indicated that the models consistently underestimated canola yield of this field (Figure 8). This outcome is

reasonable because CF5 is the most productive field (Figure 4a) and the training dataset is likely not able to capture the complete yield dynamics of the field. In terms of spatial accuracy, we observed that both RFR and FLR maps successfully captured patches of low or high canola yields. However, residuals were not spatially homogeneous. Our predicted maps overestimated values in low-yield areas (L1-L4) and underestimated values in high-yield areas (H1, H2). This is likely caused by the noisy reference dataset. For example, operating the yield monitor without crop flowing through the combine will reduce the recorded yield. Additionally, yield monitors must be cleaned periodically, as debris can accumulate on the sensor surface, causing sensor drift over time (Producer-cooperator, *pers. comm.*). Our accuracy assessment focussed on CF5 because this field presented a distinctive pattern of high and low yield patches. Our intention was to demonstrate that using moderate resolution satellite images, we can capture accurately within-field variations of canola precision yield.

Another aim of this study was to examine two different approaches to using time series data in yield mapping: statistical features (input to a Random Forest Regression) and functional data (input to a Functional Linear Regression). While RFR is a robust machine learning algorithm that is a built-in option on several cloud-computing platforms including Google Earth Engine, Functional Data Analysis has received relatively little attention in the remote sensing community (but see *Zhao et al., 2009*; *Yu et al., 2010*; *Acar-Denizli et al., 2018*). Our results show that predictive performance of FLR is comparable to or better than RFR.

We note that accuracy of RFR could be improved if we had used all available pixels in the training datasets. Importantly, this would have come at the cost of computational time. Because of our smaller dataset, it is difficult to make broad statements about the relative performance of RFR versus FLR. However, we are cautiously optimistic about the potential of FLR in remote

sensing (and precision agriculture), and it deserves more consideration for a few reasons. First, FLR was computationally much faster than RFR and resulted in similar inferences to the widely used linear regression model. This could permit much more rapid modelling of yield using the large data sets that will be necessary to capture geographic variation in the yields across broad geographic areas. Second, all remote sensing time series are functional datasets—data in a form of function—as they provide spectral responses of an object over time (i.e., are functions of time). Thus, FLR (and more generally FDA with a scalar response) are from first principles an ideal tool for the analysis of remote sensing data, because these models can integrate the spectral state of an object over time and relate this to a single response value. In our study, this response was the crop yield measured at a pixel, which is likely to be the product of plant health conditions that vary continuously throughout the growing season, and therefore have the potential to be captured by a function of the spectral state of that pixel over time.

5. Conclusion

Here we presented a simple method to predict precision crop yield using a Sentinel-2 time series and a reference dataset recorded by a Case IH combine yield monitor. We examined two methods of using time series images: using statistical features in a Random Forest Regression and analyzing the spectral time series in a Functional Linear Regression. Results from both modeling approaches indicated that we can predict precision canola yield quite accurately from space. Mean-absolute-errors of all scenarios range from 300 to 390 kg/ha, equivalent to 12.6%-15.5% of the corresponding median values. The spatial distribution of crop yield within the field was also captured in predicted yield maps. Our study is among the first attempts to demonstrate the capability of the medium-resolution multispectral data in mapping precision crop yield. Predicted yield maps can likely be improved in future studies by using a larger reference dataset

or incorporating ancillary data such as soil moisture, climatic conditions, crop variety or agriculture practices, although this additional data collection burden might not be desirable when prediction across large geographic areas is the objective. Nonetheless, our findings underline that using medium-resolution multispectral satellite data in precision agriculture research and applications is promising. In this study, we also offer a way to leverage a rich dataset from combine harvester yield monitors to understand crop growth better at both field-level and regional-level. At the field-level, we can identify patches of low and high yields. This information is important for optimizing the productivity of farming operations, and could potentially be used in other applications, such as crop insurance adjustment. At a broader scale, precision yield maps can be created at the township or regional level to provide vital information for food security, agricultural management, and policymaking.

 # Acknowledgements

This research has been made possible by Alberta Canola Producers Commission, Manitoba Canola Growers Association, and Eyes High Postdoctoral Research Program at University of Calgary. We thank the producer who provided us with the precision canola yield dataset and valuable insight into the underlying field-level patterns of yield. We elect not to name them to maintain confidentiality. We also thank Laurel Thompson at Lakeland College in Vermillion, Alberta.

References

Acar-Denizli, N., Delicado, P., Başarır, G., & Caballero, I. (2018). Functional regression on remote sensing data in oceanography. *Environmental and Ecological Statistics*, 25(2), 277-304.

Bala, S. K., & Islam, A. S. (2009). Correlation between potato yield and MODIS-derived vegetation indices. *International Journal of Remote Sensing*, 30(10), 2491-2507.

Bauer, M. E., Hixson, M. M., Davis, B. J., & Etheridge, J. B. (1978). Area estimation of crops by digital analysis of Landsat data. *Photogrammetric Engineering and Remote Sensing*, 44(8), 1033-1043.

Braga, P., Crusiol, L. G. T., Nanni, M. R., Caranhato, A. L. H., Fuhrmann, M. B., Nepomuceno, A. L., ... & Mertz-Henning, L. M. (2020). Vegetation indices and NIR-SWIR spectral bands as a phenotyping tool for water status determination in soybean. *Precision Agriculture*, 1-18.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Delegido, J., Verrelst, J., Meza, C. M., Rivera, J. P., Alonso, L., & Moreno, J. (2013). A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. *European Journal of Agronomy*, 46, 42-52.

Doraiswamy, P. C., Hatfield, J. L., Jackson, T. J., Akhmedov, B., Prueger, J., & Stern, A. (2004). Crop condition and yield simulations using Landsat and MODIS. *Remote Sensing of Environment*, 92(4), 548-559.

Febrero Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4).

414 Gao, B. C. (1996). NDWI—A normalized difference water index for remote sensing of
415 vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), 257-266.

416 Ge, Y., Thomasson, J. A., & Sui, R. (2011). Remote sensing of soil properties in precision
417 agriculture: A review. *Frontiers of Earth Science*, 5(3), 229-238.

418 Herrmann, I., Karnieli, A., Bonfil, D. J., Cohen, Y., & Alchanatis, V. (2010). SWIR-based
419 spectral indices for assessing nitrogen content in potato fields. *International Journal of*
420 *Remote Sensing*, 31(19), 5127-5143.

421 Huete, A. R., Liu, H. Q., Batchily, K. V., & Van Leeuwen, W. J. D. A. (1997). A comparison of
422 vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of*
423 *Environment*, 59(3), 440-451.

424 Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High
425 resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment*, 233,
426 111410.

427 Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H.
428 (2016). Random forests for global and regional crop yield predictions. *PLoS One*, 11(6),
429 e0156571.

430 Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., ... & Lichtenberger, J.
431 (2018). Statistical modelling of crop yield in Central Europe using climate data and remote
432 sensing vegetation indices. *Agricultural and Forest Meteorology*, 260, 300-320.

433 Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. CRC Press.

Liaghat, S., & Balasundram, S. K. (2010). A review: The role of remote sensing in precision agriculture. *American Journal of Agricultural and Biological Sciences*, 5(1), 50-55.

Liang, L., Di, L., Huang, T., Wang, J., Lin, L., Wang, L., & Yang, M. (2018). Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. *Remote Sensing*, *10*(12), 1940.

Liu, J., Miller, J. R., Haboudane, D. & Pattey, E. (2004). Exploring the Relationship between Red Edge Parameters and Crop Variables for Precision Agriculture. In Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARRS), Fairbanks, Alaska, USA, 1276-1279.

Liu, J., Pattey, E., Miller, J. R., McNairn, H., Smith, A., & Hu, B. (2010). Estimating crop stresses, aboveground dry biomass and yield of corn using multi-temporal optical data combined with a radiation use efficiency model. *Remote Sensing of Environment*, 114(6), 1167-1177.

Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, 151(3), 385-393.

Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, 151(3), 385-393.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1), 26-33.

Ramsay JO, Silverman BW (2005). Functional Data Analysis. 2nd edition. Springer-Verlag, New York.

Rudorff, B. F. T., & Batista, G. T. (1991) Wheat yield estimation at the farm level using TM Landsat and agrometeorological data. *International Journal of Remote Sensing*, 12, 2477–2484.

Rudorff, B. F. T., & Batista, G. T. (1991) Wheat yield estimation at the farm level using TM Landsat and agrometeorological data. *International Journal of Remote Sensing*, 12, 2477–2484.

Sakamoto, T. (2020). Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160, 208-228.

Sakamoto, T. (2020). Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160, 208-228.

Salmon, J. M., Friedl, M. A., Frolking, S., Wisser, D., & Douglas, E. M. (2015). Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data. *International Journal of Applied Earth Observation and Geoinformation*, 38, 321-334.

Salmon, J. M., Friedl, M. A., Frolking, S., Wisser, D., & Douglas, E. M. (2015). Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data. *International Journal of Applied Earth Observation and Geoinformation*, 38, 321-334.

Schimmelpfennig, D., (2016). *Farm profits and adoption of precision agriculture*. United States Department of Agriculture, Economic Research Service. ERR-217.

Steele, D. (2017). Analysis of precision agriculture adoption & barriers in Western Canada. URL. (Accessed: 08-15-2020).

Thorp, K. R., & Tian, L. F. (2004). A review on remote sensing of weeds in agriculture. *Precision Agriculture*, 5(5), 477-508.

Yu, Q., Tian, Y. Q., Chen, R. F., Liu, A., Gardner, G. B., & Zhu, W. (2010). Functional linear analysis of in situ hyperspectral data for assessing CDOM in rivers. *Photogrammetric Engineering & Remote Sensing*, 76(10), 1147-1158.

Zhao, K., Popescu, S., & Nelson, R. (2009). Lidar remote sensing of forest biomass: A scale-invariant estimation approach using airborne lasers. *Remote Sensing of Environment*, 113(1), 182-196.

Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, 4(3), 212-219.

Appendix

A – NDVI time series filtering

At any particular location (pixel), let X be a vector containing *n* values of days-of-year (DOYs) and Y be a vector containing *n* values of the GEE-retrieved NDVI observed at DOYs in X.

Rule 1: If a large change in NDVI is detected within a searching window of 14 days, a pixel is considered as cloud contaminated.

We define ΔY—an *n x n* matrix containing distances between a particular Y value and all values in the Y vector ($\Delta Y_i = y_i - Y$ with i indicates a value position in X, Y vectors and a row position in ΔY). At a random position *j*, we split ΔY_j into two part: ΔY_{j1} — distances between y_j and Y values to the left of *j* (start → *j*), and ΔY_{j2} — distances between y_j and Y values to the right of j position in Y (*j* → end).

If $\min(\Delta Y_{j1}) < -0.1$ & $\min(\Delta Y_{j2}) < -0.1$ & $(\min(\Delta Y_{j1}) + \min(\Delta Y_{j2})) < -0.3$, we conclude that the NDVI observation at *j* is contaminated by cloud.

Rule 2: If a very large change in NDVI is detected between three consecutive valid observations (no matter how far they are from each other in terms of DOY), a pixel is considered as cloud contaminated.

Let x_1, x_2, x_3 be three consecutive DOYs with GEE-filtered NDVI observations—y_1, y_2, y_3.

If $(y_1-y_2) \geq 0.15$ & $(y_2-y_3) \leq -0.15$ & $(y_1+y_3) \geq 0.35$, we conclude that NDVI at x_2 is contaminated by cloud.