Mohamed Jabloun,
Emmanuel Udugbezi,
Mike Rivington
01/06/2023

Developing Crop Yield Forecasting Capabilities in Scotland.

A SEFARI Fellowship Report

This report details research and technical developments made to enable forecasting of harvest time crop yield as the growing season develops using satellite imagery, data integration and modelling. The aim was to develop crop yield prediction capabilities to enhance the RESAS Scottish Crop Map, and to contribute to improvement and the future direction of this product. The objective was to develop an annual yield prediction model for high profile crops in Scotland.

The James Hutton Institute

SEFARI

RESAS
Rural & Environmental Science and Analytical Services

@SGRESAS

Contents

Citation:

This report should be cited as:

Jabloun M, Udugbezi E, Rivington M (2023) Developing Crop Yield Forecasting Capabilities in Scotland. SEFARI Fellowship Report. The James Hutton Institute, Aberdeen. Scotland. DOI 10.5281/zenodo.8146670

Contact:

Mike Rivington: mike.rivington@hutton.ac.uk

Acknowledgements

The project was funded by the Scottish Government's Rural and Environment Science and Analytical Services Division through SEFARI Gateway. The James Hutton Institute is supported by the Strategic Research Programme (2022-2027) funded by the Scottish Government's Rural and Environment Science and Analytical Services Division

Summary

This report details research and technical developments made to enable forecasting of harvest time crop yield as the growing season develops using satellite imagery, data integration and modelling. The aim was to develop crop yield prediction capabilities to enhance the Scottish Crop Map,, and to contribute to improvement and the future direction of this product. The objective was to develop an annual yield prediction model for high profile crops in Scotland. This report provides brief details on the method and data processing developed and the results gained. We have focused the development, testing and application of the method to Spring Barley, as this is a key economic crop and one that has best available observed yield data for calibration and testing purposes.

Key Messages

- It is feasible to use optical imagery and radar satellite data as input into a model to estimate crop yield.
- We have successfully developed a data processing, analysis and modelling pipeline that enables yield forecasting for any field in Scotland.
- In the best instances during calibration using training data, the model makes yield estimates that have a Root Mean Square Error (RMSE) of 0.38 t/ha for field-specific observed yields.
- When the model is tested for predictive skill using data not used for calibration, the accuracy declines with an RMSE of 0.71 t/ha.
- The model performs best for yields +/- 1 ton of the mean (c. 6.0 t/ha) but the size of error increases with lower (under-estimates) and higher (over-estimates) yields.
- The over- and under-estimation of lower and higher yields will likely compensate when yield estimates are aggregated spatially.
- There are two key issues that limit the current potential of the approach:
    - The lack of field-specific observed yield data with a Scotland-wide geographic spread that would enable more comprehensive model calibration and validation. There is good potential to improve the forecasting skill if more observed yield becomes available.
    - The amount of cloud cover over Scotland and ability to utilise time series of satellite optical imagery-based indices to observe crop development.
- Use of radar data can help overcome limitations of optical based indices.

There is very good potential to develop approaches for farmer participation to facilitate the collection of key data at the field and farm scale that will facilitate crop growth simulation and improved forecasting skill. OurSmartFarm is an example of an on-going online application under development.

1 Crop Yield Forecasting in Scotland

1.1 Introduction

The purpose of this report is to present the development of a satellite imagery data integration, analysis and modelling approach to forecast harvest time crop yield at the field scale. The context is to build the technical and expertise capabilities to enable crop yield forecasting at the field scale which can be aggregated to regional and national scales.

The aim of this research has been to develop this crop yield prediction capability to enhance the Scottish Crop Map, and to contribute to improvement and the future direction of this product. This requires novel ways of estimating crop yields to support or replace existing National Statistics, providing insight into Scottish crop production capability by producing predicted yield and production values for key crops. Using a modelling approach that supports industry intelligence and survey information can improve the timeliness of the statistical publication Cereal and oilseed rape harvest estimates whilst reducing the burden on farmers to provide yield values.

The objective was to develop an annual yield prediction model for high profile crops in Scotland. Developing a yield prediction capability adds value to the Scottish Crop Map, which uses sentinel-1 radar images and a machine learning methodology to estimate areas and locations of Scotland's high-profile crops (barley, wheat and oats) and agricultural grassland.

Hence this report demonstrates the increasing capabilities within the Scottish Government and Environment, natural resources and agriculture strategic research programme to better predict crop yields.

1.1.1 Defining the challenge

The challenge this research addresses is how to provide estimates of crop yields at the field scale using remotely sensed data from satellites of observe biomass growth and soil conditions prior to harvest time and use these to estimate a final yield.

This challenge includes addressing issues of sparse spatially and temporally representative observed yield data and the fact that there is often cloud cover in Scotland, that reduces the opportunity to utilise satellite optical imagery.

1.2 Data for yield prediction

1.2.1 Available yield data

The development of the yield forecasting model has utilised observed barley yield data supplied to Scottish Government by farmers using the annual Scottish Cereal Production and Disposal Survey. This is a survey of c.600 randomly selected farm businesses across Scotland. The farms selected vary between years, meaning data for the same businesses can occur in multiple years.

A limitation of this data is that it is provided at the business holding level, rather than for individual fields, making direct alignment between an observed yield and the satellite data problematic. For some holdings there may be data for 1-2 contiguous fields, whereas for others the fields may be several kilometres apart. See Appendix 5.1 for further details.

The period coverage of data is 2017, 2018 and 2019 with 309 yield records that can be assigned to individual fields.

1.2.2 Topography

Field topography is considered one of the driving variables behind within-field variation in which it can affect the crop growth and yield directly from its effect on microclimate condition such as solar radiation and air temperature, or indirectly from its effect on soil properties such as soil nutrients and soil temperature that can affect crop growth and development. The average field elevation, slope and aspect were retrieved for each field using the R package ‘elevatr’ (Hollister et al., 2021).

1.2.3 Remote sensed data

In this study, a time-series of Sentinel-2 images are utilized. Four bands in the visible and near-infrared regions with a 10m spatial resolution and two bands in the Red Edge and near-infrared regions with a 20m spatial resolution of Sentinel-2 A and B are used. The number of images for each field is different due to the various cloudiness conditions during growing season of barley (March to September) in each location and year.

In addition, we explored the possibility of the use of Sentinel-1 Synthetic Aperture Radar (SAR) data to supplement the use of optical remotely sensed imagery. In the context of agricultural crop monitoring in Scotland, there is significant benefit in the use of SAR data since radar, as an active sensor operates under all weather conditions and can penetrate clouds. The microwave signal is sensitive to the dielectric and geometrical properties of crops (Ulaby, 1975). Sentinel-1 provides a unique opportunity to monitor plant growth progressively at a temporal resolution of 2-6 days intervals.

1.3 Technical and Methods Development

1.3.1 Image processing and analysis

Atmospherically corrected Sentinel-2 reflectance products were used in this study. Sentinel-2 level-2A images covering the fields were obtained from Google Earth Engine (Gorelick et al., 2017). Only cloud free images were selected. Four vegetation indices (VI) which describe different aspects of crop growth and status were derived based on the Sentinel-2 Level-2A imagery:

- The Normalized Difference Vegetation Index (NDVI): it has been recognised as the most popular VI for biomass and crop productivity assessment. It describes the vigour level of the crop. However, under high-biomass conditions, reflectance in the red region becomes saturated, and further increases in chlorophyll content do not affect reflectance.

- The Normalized Difference Red Edge index (NDRE): The red-edge (700-740 nm) region does not suffer saturation effect as the one observed for the red region of the spectrum, and thus has been found to be a better predictor of chlorophyll content and canopy N status. It is calculated as: NDRE = (NIR-RED EDGE)/(NIR+RED EDGE) where NIR and RED EDGE refer to near-infrared bands (842 nm) and red-edge band (705 nm), respectively.
- The Weighted Difference Vegetation Index (WDVI): it offers a good correction for soil background in estimating the Leaf Area Index (LAI) of green vegetation, e.g. cereals at the vegetative stage (Clevers, 1991).
- The Normalized Difference Moisture Index (NDMI): The NDMI describes a crop's water stress level (Gao, 1996) and is calculated as the ratio between the difference and the sum of the refracted radiations in the near infrared and SWIR (1610 nm), that is: (NIR-SWIR)/(NIR + SWIR).

In this study, the seasonal maximum VI was chosen since it enabled timely forecasting of yield approximately a month before harvest. The seasonal maximum VI was also used in several other studies to forecast cereal yields (Becker-Reshef et al., 2010; Franch et al., 2015; Johnson et al., 2021).

1.3.2 Yield Model Development

Prior to developing the yield forecasting model, a two-step analysis was carried out.

Step 1: data exploratory analysis to detect and remove outliers. Yield observation beyond the 25% and 75% quartiles are considered as outliers.

Step 2: A correlation analysis was carried out between the observed barley yield and each of the covariates (topography and VIs). The Pearson correlation coefficient (R) was used to assess the role of the different variables. The results were used to discard the explanatory variables having low R coefficient with yield and those presenting strong collinearity.

We used Random Forests (RF) to develop the yield forecasting model. RF performs nonlinear regression by model averaging of many regression trees where each tree uses a random number of predictors sampled with replacement according to a uniform probability distribution (Breiman, 2001). We used the 'ranger' function with default parameters from the R package 'ranger' for random forests (Wright & Ziegler, 2017).

We used an 80-20 train-test split on the barley yield data where 248 yield records were randomly selected and used to train the RF model and the remaining 61 yield records were used to assess the model accuracy. Accuracy is measured using the coefficient of determination (R2) between observed and predicted yields, mean absolute error (MAE), root mean square error (RMSE) and normalised root mean square error (NRMSE).

The coefficient of determination provides the proportion of variance in the observed data explained by a model, relative to observed mean with larger values being

better. The MAE and RMSE provide measures of the model error. NRMSE is the ratio of the model error and mean observed value. Lower error values are better.

2 Results

Observed yields ranged from 2.2 to 8.4 t ha-1 with substantial spatial and year to year variation. Figure 1 shows the correlation between barley yield, topography and vegetation indices across all fields. Aspect has a significant correlation with yield (R = 0.15). Even though the correlation coefficient is low, it was found that fields facing south presented a higher yield as compared to fields facing East. This indicates that Aspect has potential usefulness for explaining yield differences. However, elevation and slope did not correlate with yield and were therefore excluded. Maximum NDRE over the entire barley growing season has a significant positive correlation with yield and has the highest R coefficient (R = 0.41) followed by WDVI, NDVI and NDMI with a significant correlation of 0.36, 0.33 and 0.15, respectively. The significant correlation with NDMI indicates that NDMI has potential usefulness for explaining yield variability due to water stress. NDVI and WDVI has a significant strong correlation with NDRE (R > 0.6) and since NDRE doesn’t suffer from the saturation effect, it was decided to only use NDRE and NDMI as explanatory variables for the yield forecasting model along with field Aspect.

Figure 1. Multi-panel scatterplots of tested field topography (top), vegetation indices (bottom) and barley yield. The lower left panels show Pearson correlation coefficients, and the upper right panels show pairwise scatterplots among variables. Symbols***, **and * indicate significance level of p < .001, 0.01 and 0.05 respectively. The red lines are fitted linear regression among yield and tested variables.

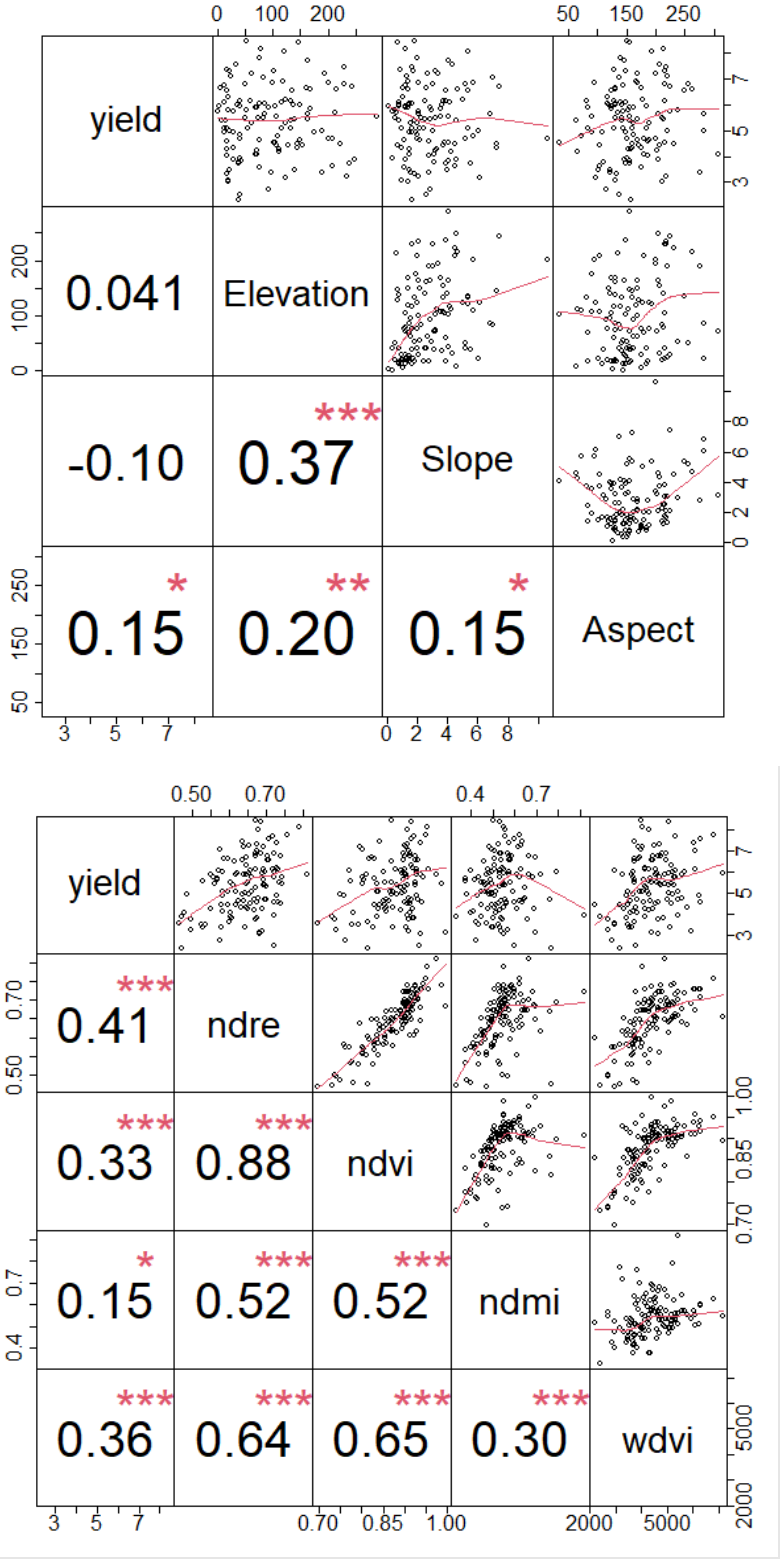


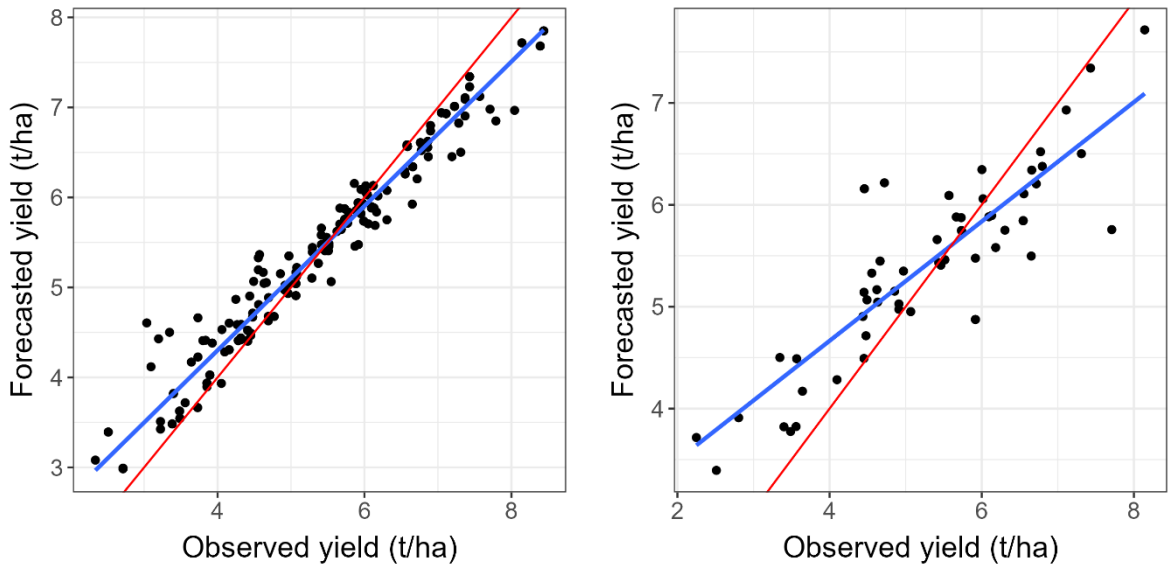
Table 1 shows the accuracy statistics for estimating yield, calculated for the training and the independent test data for the random forest model. Overall model performance was high for training and testing with high R^2 values (>0.7) and low error metrics.

Table 1. Accuracy statistics for the Random Forest model.

Dataset	R^2 (-)	RMSE (t ha^{-1})	MAE (t ha^{-1})	NRMSE (%)
Train	0.96	0.34	0.22	6.17
Test	0.73	0.61	0.55	13.37

A comparison of the predicted and measured yield values are shown in Figure 2. In both training and testing steps of the RF model, the scattered points were evenly distributed around the 1:1 line, the RMSE were 0.34 and 0.61 t ha-1 for training and testing of the model, respectively, and the NRMSE was lower than 13 %. This showed that the predicted yield values were close to the measured values, which indicated the reliability of the developed RF yield forecasting model. The developed RF model tends to slightly overestimate low yield and slightly underestimate high yield.

Figure 2. Comparison between predicted and observed grain yield for barley using training data (left) and testing data (right). The blue lines show fitted linear regressions and the red lines show 1:1 line.



3 Next Steps and Recommendations

In this study the seasonal maximum NDVI was used as the main remotely sensed input parameter. Some studies (e.g. Rojas, 2007) have shown that seasonally integrated NDVI, can predict yields more accurately than measures such as the seasonal maximum NDVI, since it can capture the effect of adverse events which occur after flowering. Nevertheless, in this study the seasonal maximum VI was chosen since it enabled a timely prediction of production approximately a month and a half prior to harvest.

The next step would be to investigate the performance of using seasonally integrated VI. However, calculating seasonally integrated VI implies that starting of the growing season (i.e. sowing date) and key phenological stages (e.g. flowering, senescence) are known which is usually not the case. So, further work is needed to try to infer phenology from remote sensing time series. Moreover, crop sowing dates is one of the main factors affecting yield and late sowing can have a negative effect on barley yield. Incorporating sowing date as an explanatory variable (as well as other phenological stages) in a yield forecasting model will likely improve the model accuracy.

It is worth mentioning that the maximum VIs did not pertain to a singular date during the barley growing season but rather varied in time based on the crop and unique growing conditions, as expressed with the VI temporal profile of that year. For barley the peak VIs tended to occur in June to early July. Because of persistent cloud cover, this period can be missed and in that case the yield forecasting model cannot be applied. The synergistic use of Synthetic-Aperture Radar and optical data is expected to alleviate this problem. However, due to the different characteristics of optical and SAR sensors, it is difficult to build a relationship between the two but successful applications are found in the literature (Bai et al., 2020, Li et al., 2022) which are worth exploring. Another alternative is the utilization of multi-sensor data through data fusion to produce more frequent cloud free observations, e.g. fusion of Landsat 8 and Sentinel-2 data (Wang et al., 2017).

The developed yield forecasting (Random Forest) model showed a relatively high accuracy but additional yield observations across the whole of Scotland spanning a wide range of weather conditions and soils would be needed to further test and improve the model.

Besides empirical yield forecasting models, process-based crop simulation models offer powerful tools to simulate crop yield at the field scale based on the interactions among environmental characteristics (i.e. the climate, crop management, and soil conditions). However, their practical application at a regional scale is restricted by uncertainties in the model's input parameters and initial conditions. To ensure better estimates of model input parameters, remotely sensed data (e.g. biomass, leaf area index, soil moisture) which provide up-to-date overview of actual crop growing conditions over large areas have been widely utilized in conjunction with crop models through data assimilation to improve crop yields prediction at large scales. This approach for crop yield forecasting has been successfully developed and used for

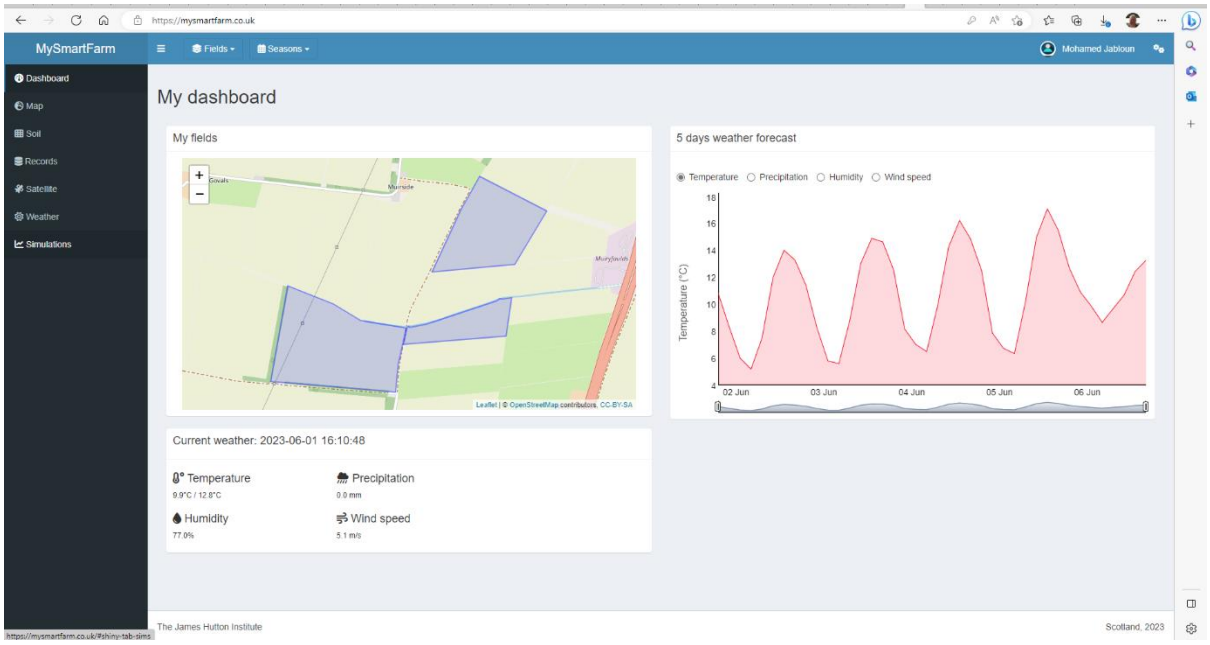
several crops in Europe (Van der Velde & Nisini, 2019) and worldwide and can constitute another potential research area for developing a yield monitoring system for Scotland.

3.1 OurSmartFarm

A key solution to improving forecasting model skill is to be able to access field specific data from farmers. To address this there is need to develop facilities to enable farmers to provide data for the growing season.

MySmartFarm (https://mysmartfarm.co.uk), a new two-way data exchange and simulation modelling research platform currently being developed by JHI scientists, can potentially serve as a tool to collect timely field observations (e.g. crop management, yield) that can be used to improve the RF yield forecasting model and as a national scale yield monitoring platform through the integration of remotely sensed data and crop model outputs.

Figure 3. An example of OurSmartFarm



OurSmartFarm provides a state-of-the-art research platform that is also a decision support system, crop growth monitoring and a data management system allowing farmers to upload their own field operations (e.g. tillage, sowing, fertilization, harvest,…) and observations (e.g. flowering date, yield, pest/disease,…), utilize multiple spatial data (field topography, satellite multispectral vegetation indices) and spatial data analysis methods, and crop model outputs to enable improved data driven decision making. Therefore, combining the use of earth observation data, crop models and farmer supplied field observations in OurSmartFarm creates a bridge between farmers and scientists to help resolve the challenge of increasing crop production while reducing agriculture's environmental impact.

4 References

Bai, Z., Fang, S., Gao, J., Zhang, Y., Jin, G., Wang, S., Zhu, Y. and Xu, J., 2020. Could vegetation index be derive from synthetic aperture radar?–the linear relationship between interferometric coherence and NDVI. Scientific Reports, 10(1), pp.1-9. https://doi.org/10.1038/s41598-020-63560-0

Becker-Reshef, I., Vermote, E., Lindeman, M. and Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote sensing of environment, 114(6), pp.1312-1323. https://doi.org/10.1016/j.rse.2010.01.010

Breiman, L., 2001. Random Forests. Machine Learning, 45, 5–32.

Clevers, J.G.P.W., 1991. Application of the WDVI in estimating LAI at the generative stage of barley. ISPRS journal of photogrammetry and remote sensing, 46, 1, 37-47. https://doi.org/10.1016/0924-2716(91)90005-G

Franch, B., Vermote, E., Becker-Reshef, I., Claverie, M., Huang, J., Zhang, J., Justice, C. & Sobrino, J. 2015. Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. Remote Sensing of Environment 161, 131–148. https://doi.org/10.1016/j.rse.2015.02.014

Gao, B.C., 1996. NDWI - a normalized difference water index for remote sensing of vegetation liquid water from space. Rem. Sens. Environ. 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 2017, 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hollister J, Shah T, Robitaille A, Beck M, Johnson M (2021). elevatr: Access Elevation Data from Various APIs. doi:10.5281/zenodo.5809645, R package version 0.4.2, https://github.com/jhollist/elevatr/.

Johnson, D.M., Rosales, A., Mueller, R., Reynolds, C., Frantz, R., Anyamba, A., Pak, E. & Tucker, C. 2021. USA Crop Yield Estimation with MODIS NDVI: Are Remotely Sensed Models Better than Simple Trend Analyses? Remote Sensing 13, 4227. https://doi.org/10.3390/rs13214227

Li, J., Li, C., Xu, W., Feng, H., Zhao, F., Long, H., Meng, Y., Chen, W., Yang, H. and Yang, G., 2022. Fusion of optical and SAR images based on deep learning to reconstruct vegetation NDVI time series in cloud-prone regions. International Journal of Applied Earth Observation and Geoinformation, 112, p.102818. https://doi.org/10.1016/j.jag.2022.102818

Rojas, O. (2007). Operational maize yield model development and validation based on remote sensing and agro-meteorological data in Kenya. International Journal of Remote Sensing, 28, 3775−3793. https://doi.org/10.1080/01431160601075608

Van der Velde, M. & Nisini, L. 2019. Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015. Agricultural Systems 168, 203–212. https://doi.org/10.1016/j.agsy.2018.06.009

Wang, Q., Blackburn, G.A., Onojeghuo, A.O., Dash, J., Zhou, L., Zhang, Y. and Atkinson, P.M., 2017. Fusion of Landsat 8 OLI and Sentinel-2 MSI data. IEEE Transactions on Geoscience and Remote Sensing, 55(7), pp.3885-3899. 10.1109/TGRS.2017.2683444

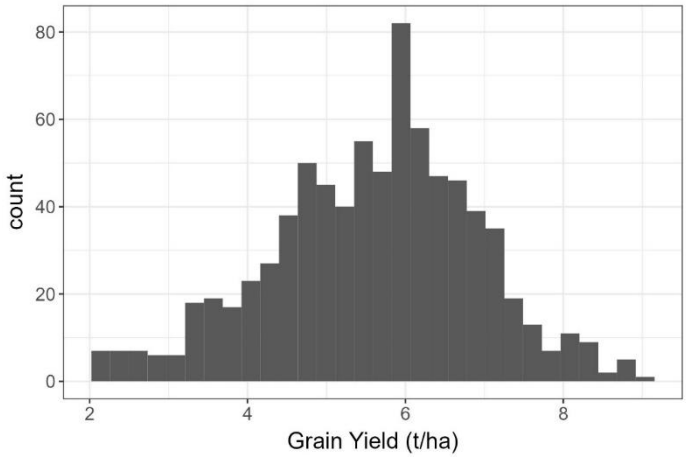
Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17. https://doi.org/10.18637/jss.v077.i01

5 Appendix

5.1 Methods and Data

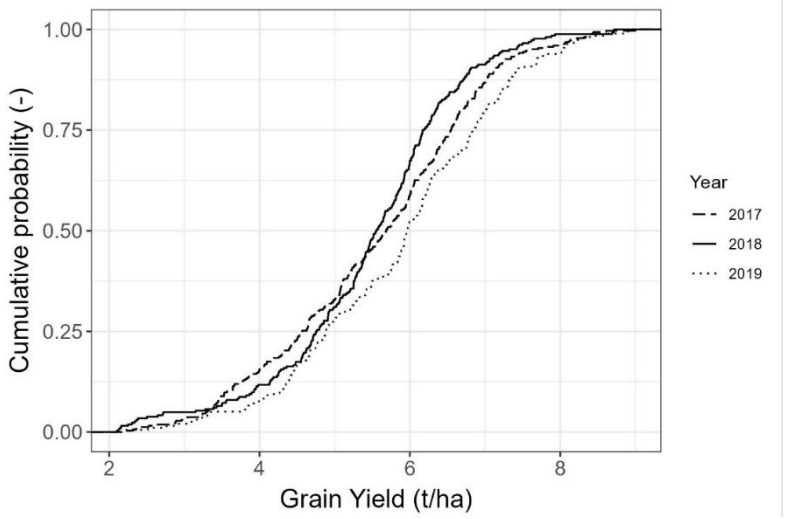
5.1.1 Yield data overview

Figure 4. Distribution of holding level barley yield data for the years 2017-2019.



The observed yield distribution is shown in Figure 4 for the years 2017-2019, indicating a near-normal distribution but with a marginal skewness to lower yields. A constraint on the utility of the forecasting model is the accuracy of the observations.

Figure 5. Cumulative distribution function of the holding level barley yield data for the years 2017-2019.

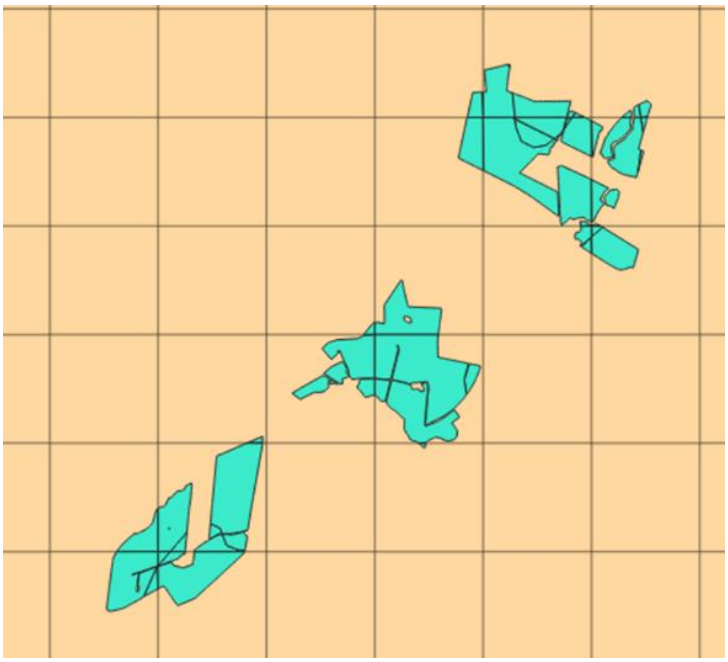


Differences in yield cumulative probability between years are shown in Figure 5. There is a large difference in cumulative probability for the yields above the average (c. 6 t ha^{-1}).

Alignment of holding level yield values with fields, soil type and weather cell:

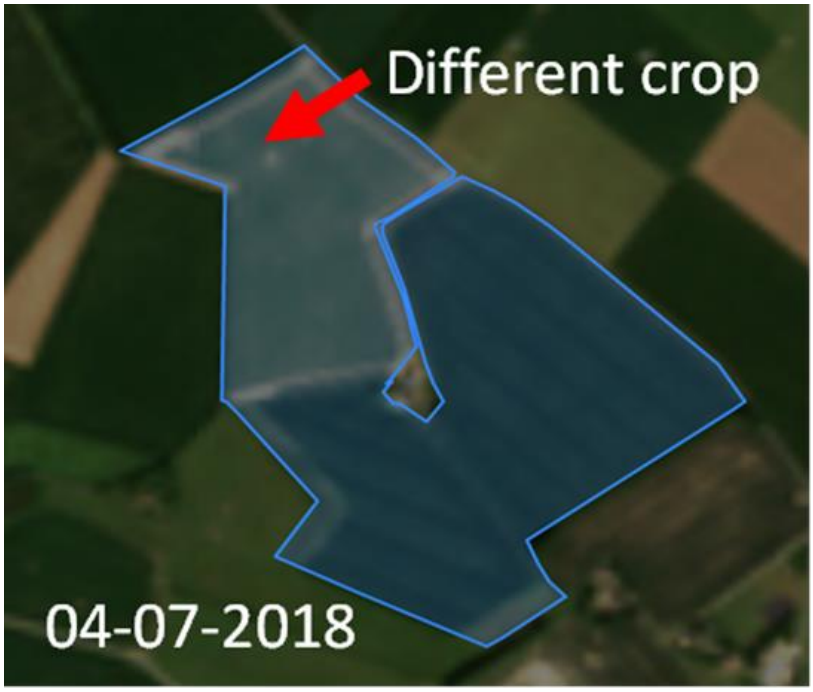
Figure 6 illustrates the challenges of aligning yield values with specific data types that can enable improved forecasting skill. Values for observed yield are provided at a business holding level, yet the spatial distribution of fields can be widely separated meaning that yields cannot be directly aligned to specific soil types or weather grid cell. An objective or OurSmartFarm is to overcome this by enabling farmers to provide field-specific yield (and management) data.

Figure 6. Example of fields within a holding which are several kilometres apart spread across 18 (1km x 1 km) climate grids.



Aligning observed yields with fields and Remote Sensed data:

Figure 7. Two different crops in one holding as depicted by the RGB sentinel image which can produce misleading maximum vegetation indices when averaged across the holding.



As above, the use of OurSmartFarm can help overcome this problem by enabling field specific data entry.