Can Machine Learning and Crop Models Help Farmers? A Case Study of Oyster Cove Potato Farms

Mariaelisa Polsinelli[a,b], Morteza Mesbah[b], Zhiming Qi[a], Matt Ramsay[c]

[a.] Department of Bioresource Engineering, McGill University Macdonald Campus, St-Anne-De-Bellvue; [b.] Agriculture and Agri-Food Canada, Charlottetown Research and Development Centre; [c.] Kensington North Watersheds Association and Oyster Cove Farms

Introduction

Climate change can put the potato crop (the major crop in PEI) at risk. Yield prediction can help farmers plan ahead with crop management decisions and adaptation strategies. Yield can be predicted by process-based crop models and machine learning. In this study we used the process-based model STICS (Simulateur mulTIdisciplinaire pour les Cultures Standard / Multidisciplinary Simulator for Standard Crops) (Brisson et al., 2003) and **Random Forest (RF) Machine Learning (ML)** model (Breiman, 2001) in a hybrid approach to predict the yields of industrial farm fields in Prince County, PE.

How Does the STICS Process Based Crop Model Work?

STICS simulates crop growth by taking climate and soil data, management information and specific cultivar and plant traits as inputs. It simulates the physical, chemical and biological processes in a soil-crop-atmosphere system for a given location over a daily time step. It predicts farming (e.g., fresh/dry yield) and environmental variables (e.g., N loss).

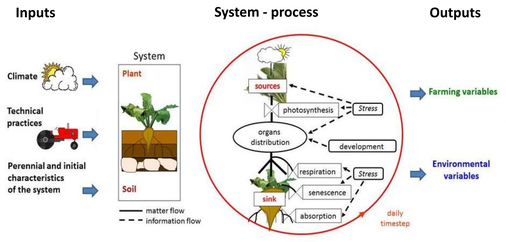

Figure 1. Overview of STICS crop model (Brisson et al., 2003)

How Does a Random Forest Model Work?

RF is a type of ML that combines the output of multiple decision trees (Breiman, 2001). Figure 2 represents a classification problem (e.g., A-high yield, B-low yield). Each tree uses a portion of the dataset and makes a decision on output by going downward to left or right at each node and answering a yes/no question based on randomly selected features (e.g., temperature, rainfall). The final result is chosen by "majority voting". For a regression problem with continuous data (e.g., yield) the final result is the average over all trees.

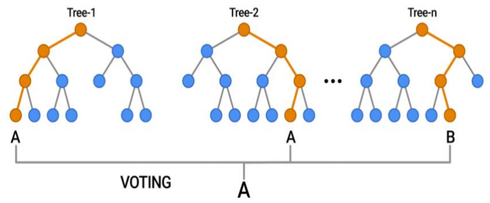

Figure 2. Overview of a Random Forest ML model performing a classification problem (Source: TensorFlow).

Objectives

- ❑ To determine if a crop model calibrated on research farm data can be used for industrial fields.
- ❑ To determine the performance of RF models on predicting the yield of unseen field-years.
- ❑ Investigate if a hybrid approach will improve model performance.

Methodology

Industrial farm data (Figure 3) from Prince County, PE was obtained for years 2015-2021. A subset of 20 field-years was taken from this data which were growing Russet Burbank and had soil sample data from the top 20cm (this subset excluded 2018). This dataset contained ~2000 gridded yield data points.

STICS

STICS was calibrated for Russet Burbank by Morissette et al. (2016) in research fields in Ste. Foy, QC and Fredericton, NB. Evapotranspiration was calculated using the simplified **Crop Coefficient Approach (CSA)** . The evapotranspiration can also be calculated using the more sophisticated **Resistance Approach (RA)** using a daily time step. The difference of the two approaches is summarized below:

1. CSA: Reference evapotranspiration x Crop Coefficient
2. RA: Estimates soil evaporation and crop water requirement separately

Random Forest Alone

Feature Selection:

- 43 initial features.
- Obtained from climate, soil and management data.
- All data points with an 80/20 train/test split.
- Obtained 10 most important features.

Yield Prediction:

- Individual field-year set aside as test set, trained on the rest.
- 6 features sampled at each node split with 300 trees generated.
- The gridded predictions were then averaged to obtain one value for each field-year.

Hybrid

The method for the RF Alone was repeated including outputs from STICS as features for training in a "hybrid approach" (Figure 4). The feature importance resulted in 18 top features with 11 from STICS and 7 from the original RF Alone top features (Figure 5). 10 features were sampled at each node split and 300 trees were generated.

Figure 3. Example of potato yield map at Oyster Cove Farms, Prince County, PE.

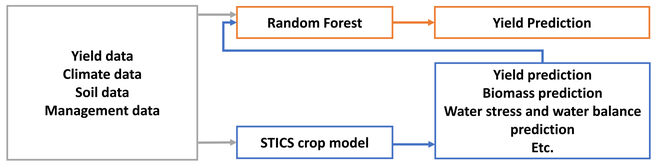

Figure 4. Hybrid approach framework based on concept by Shahhosseini et al. (2021).

Top 18 from 68 Total Features (25 from STICS):

1. Cumulative growing season GDD (base 5)
2. Total September precipitation (mm)
3. Stomatic water stress index (0-1)
4. Plant N uptake (kg/h)
5. Start date of plant physiological maturity (DOY)
6. DEM Slope
7. Cumulative crop cycle transpiration (mm)
8. Cumulative crop cycle evapotranspiration (mm)
9. Soil pH (Top 20 cm)
10. Average denitrification rate (kg/h/d)
11. Ratio of water content in top 30cm soil layer to field capacity (mm)
12. Above ground fresh matter (t/ha)
13. Cumulative growing season radiation (MJ/m^2)
14. 15. % clay in soil (top 20cm)
15. Max depth of root system (cm)
16. % Organic matter in soil (top 20cm)
17. Cumulative NO_3-N leached over crop cycle (kg/h)
18. Amount of N in harvest organs (kg/h)

Figure 5. Feature importance obtained with RF using the hybrid approach.

Results and Discussion

STICS

STICS with CSA underpredicted yield (RMSE: 285.4 ctw/ac. nRMSE 74.3%, nMBE: -73%). A nRMSE (normalized root mean squared error) >30% has poor performance (<10% is excellent, 10-20% is good) (Jamieson et al., 1991). The water stress was not well predicted under CSA. We tested STICS with the RA by recalibrating two parameters: *aclim* – the climatic component for calculation of actual soil evaporation, which is dependent on windspeed and is site specific, and *rsmin* – the minimal stomatal resistance of leaves and is cultivar specific.

STICS with RA:

- Performed well overall (Figure 6. A).
- A normalized mean bias error (nMBE) of -6.95% indicates a slight underprediction.
- The underprediction was larger in higher yielding years.
- Simulated yields for 2019 (cool year with average growing season precipitation levels) and 2020 (drought year) were quite accurate.

Random Forest Alone

- Led to highest nRMSE (i.e., 16.6 %)
- Improved nMBE compared to STICS (e.g., from -6.95% to -1.88%)
- Predicted yield in 2021 relatively well, for which RF was trained entirely on previous years' data.
- Overpredicted yield in few years greatly (e.g., 2020).
- Predicted drought years poorly. This was also reported by other ML models (e.g., Shahhosseini et al. 2021). The drought year yield prediction is particularly important under changing climate.

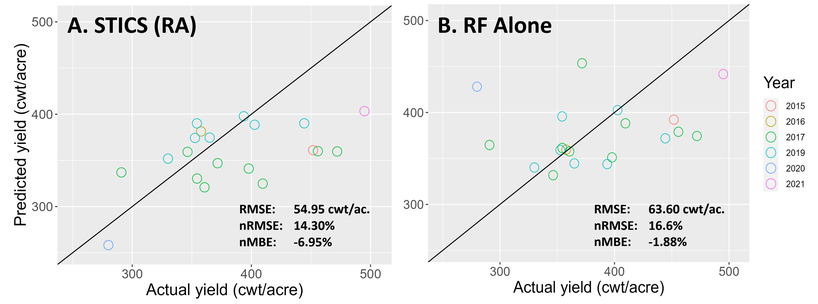

Figure 6. Performance of yield prediction by STICS, RF Alone and Hybrid approach. Points closer to the one-to-one line indicate better prediction.

Hybrid

- Outperformed both STICS and RF alone (i.e., lowest nRMSE and nMBE).
- Improved drought year (i.e., 2020) yield prediction compared to RF Alone.
- Predicted underpredicted (e.g., 2015) and overpredicted (e.g., 2020) values better compared to RF alone. The improvements were due to inclusion of some important STICS features related to water balance, which highlights the importance of precise estimation of evapotranspiration by STICS with RA.

Conclusion

- RA enhanced STICS's yield prediction performance significantly compared to CSA.
- RF Alone performed well in predicting unseen field-years.
- STICS predicted drought year best and underpredicted higher yielding years.
- Both RF Alone and Hybrid overpredicted drought year and predicted most other years well.
- Hybrid approach had best performance overall.
- Water (e.g., water stress, evapotranspiration) and N balance (e.g.,. plant N uptake) variables were among most important features for the hybrid approach.
- Combining the strengths of STICS and Machine Learning models provides a good yield prediction tool for farmers to make crop management decisions under changing climate.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., Bussière, F., Cabidoche, Y. M., Cellier, P., Debaeke, P., Gaudillère, J. P., Hénault, C., Maraux, F., Seguin, B., & Sinoquet, H. (2003). An overview of the crop model STICS. *European Journal of Agronomy*, *18*(3–4), 309–332. https://doi.org/10.1016/S1161-0301(02)00110-7

Jamieson, P. D., Porter, J. R., & Wilson, D. R. (1991). A test of the computer simulation model ARCWHEAT 1 on wheat crops grown in New Zealand. *Field Crops Research*, *27*, 337–350. https://doi.org/https://doi.org/10.1016/0378-4290(91)90040-3

Morissette, R., Jégo, G., Bélanger, G., Cambouris, A. N., Nyiraneza, J., & Zebarth, B. J. (2016). Simulating potato growth and nitrogen uptake in eastern Canada with the STICS model. *Agronomy Journal*, *108*(5), 1853–1868. https://doi.org/10.2134/agronj2016.02.0112

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. v. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, *11*(1606). https://doi.org/10.1038/s41598-020-80820-1

Acknowledgements

The funding for this project was provided by Agriculture and Agri-Food Canada. We thank René Morissette and Guillaume Jégo for their help and assistance with STICS, and Kristen Murchison for climate data processing, gap filling and technical support.

For More Information Contact:

Mariaelisa Polsinelli: *mariaelisa.polsinelli@mail.mcgill.ca, mariaelisa.polsinelli@agr.gc.ca*
Morteza Mesbah: *morteza.mesbah@agr.gc.ca*