Developing a machine-learning Surrogate Model to rapidly predict wheat yields under Soil-Moisture-Triggered irrigation strategies in Northwest India

Divyam Garg (divyam_g@ce.iitr.ac.in)[1], Hemant Kumar (hemant@ce.iitr.ac.in)[1]

[1]Civil Engineering Department, Indian Institute of Technology Roorkee, Uttarakhand, India

1. Introduction

- Process-based crop models are widely used for accurate simulation of crop response to irrigation management and climate variability.
- These models require elaborate calibration procedures whenever used for a new area or region making it a tedious and challenging task, especially for users operating in data-scarce environments.
- Further, repeated simulations are often required for large-scale multiple scenario evaluation which is time consuming and computationally expensive.
- Machine-learning based surrogate models act as emulator, learning the underlying relationships to predict outcome of interest (yield) with relatively less inputs and much faster run times.
- These surrogates, if developed and deployed at regional scales, can enable rapid evaluation of irrigation strategies, uncertainty assessment, and future climate scenarios.

2. Study Area and Data

- Two districts (in red in Fig. 1) lying in irrigation-dominated north-western (NW) Indian region were selected to develop and test the yield surrogate for wheat crop grown during rabi season (late October to early April).

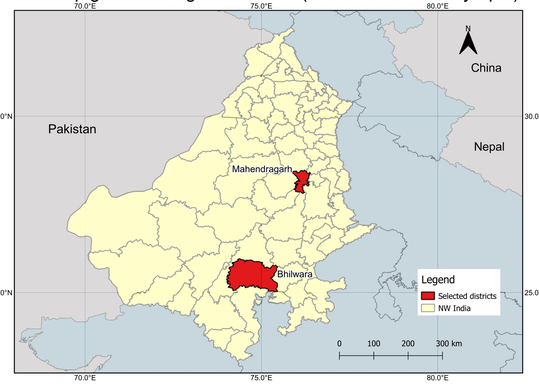

Fig. 1. Map showing location of selected districts (red) in NW India.

- Semi-arid climate with hot dry summer and cold winter, about 80% of precipitation occurs during monsoon period (June to September).

District	Average rabi precipitation	Dominant soil texture
Mahendragarh (MAH)	35 mm	Loamy sand
Bhilwara (BHIL)	17 mm	Sandy loam

Study period: 26 seasons, 1997-98 to 2022-23

Data source:

Min. and max. temperature, Rainfall	India Meteorological Department (IMD)
Observed crop yield data	Ministry of Agriculture & Farmers' Welfare

3. Methodology

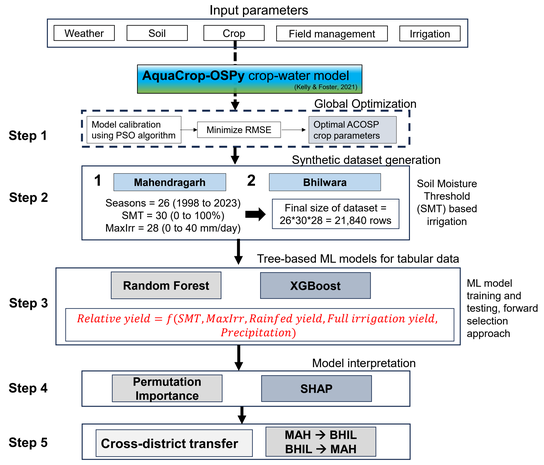

Fig. 2. Schematic overview of methodological framework.

4. Results and Discussion

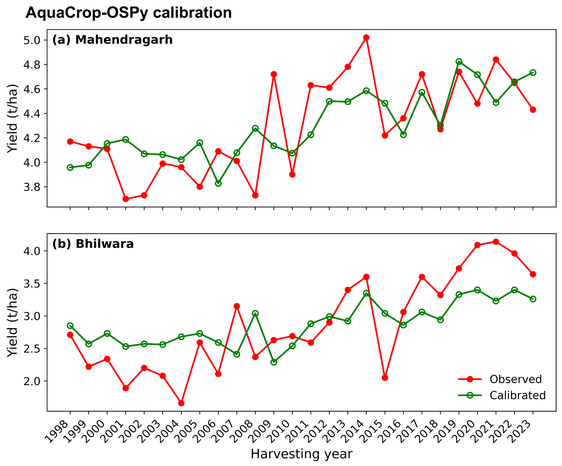

Fig. 3. Time series plots showing observed and calibrated ACOSP model wheat yield values from 1998 to 2023 for (a) Mahendragarh and (b) Bhilwara

Table 1. PSO-based calibration performance metrics for AquaCrop-OSPy model

District	RMSE (t/ha)	NRMSE (%)	R^2	Pearson's correlation (r)
MAH	0.28	6.6	0.44	0.66
BHIL	0.53	18.4	0.45	0.73

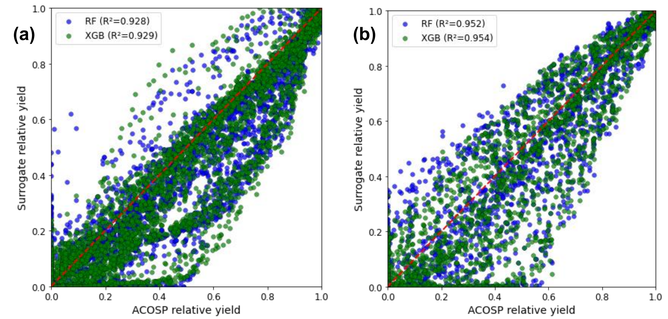

Fig. 4. Same-district performance of RF and XGBoost models to predict relative yield for (a) Mahendragarh and (b) Bhilwara.

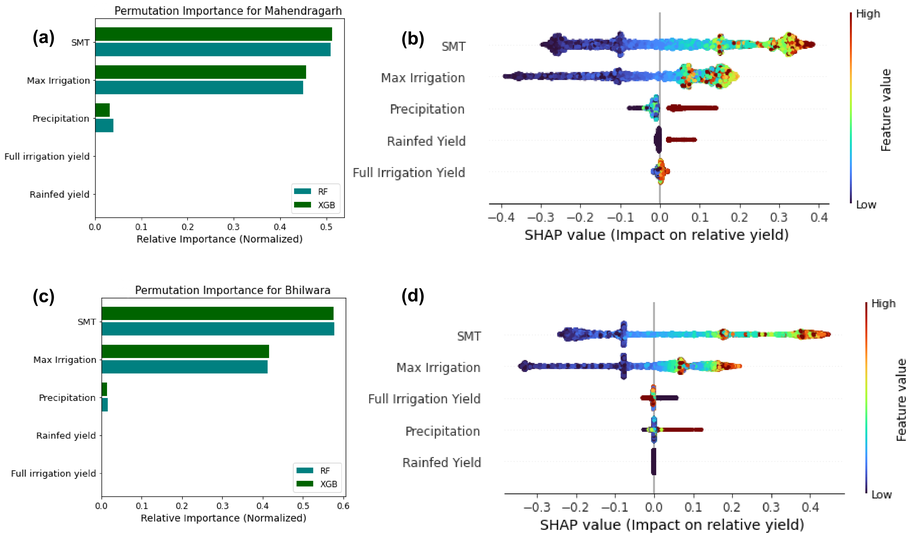

Fig. 5. Interpreting RF and XGBoost models using permutation importance and SHAP methods.

Table 2. Cross-district transfer performance of RF and XGBoost surrogate models. Test set included entire dataset of the other district as all of it was unseen domain for transferred model.

Train set	Test set	R^2	RMSE (fraction)
RF			
MAH	BHIL	0.927	0.106
BHIL	MAH	0.81	0.17
XGBoost			
MAH	BHIL	0.928	0.105
BHIL	MAH	0.80	0.17

- While training, RF achieved a higher performance score much faster as compared to XGBoost whose improvements were more gradual.
- Both RF and XGBoost models generalized well as they achieved a test set R^2 values of about 0.93 (MAH) and 0.95 (BHIL) (Fig. 4).
- Both permutation importance and SHAP values indicate irrigation management features, SMT and MaxIrr were found to be the dominant drivers, most important parameters influencing yield variability, for both districts (Fig. 5).
- But for drier BHIL, SMT and MaxIrr contribute more heavily as compared to other features than MAH (Fig. 5c, 5d). Little rabi precipitation in these districts can not fulfill wheat crop water demand of around 250 to 350 mm.
- Yield features (rainfed and full) were added to set production constraints but contributed least to predictions because these features had little variability themselves. They were same for a particular season for all SMT-MaxIrr combinations.
- The surrogate model that can be deployed for regional use cannot be selected solely on the basis of same-district performance. Its transferability has to be checked.
- Cross-district transfer of models help us assess their capability to predict for unseen domains.
- Models trained on MAH performed better, though XGBoost was marginally better (Table 2).
- Therefore, we suggest using XGBoost based surrogate as it better understands yield variability and it could be used for prediction of relative yield across the NW Indian region with reasonable error rates.
- Multiplying relative yield with potential yield for a district (if known) will give the absolute yield values.

5. Conclusions

- A hybrid approach to develop a parsimonious tree-based ML model that can predict wheat yields with good accuracy with feature contributions was shown.
- Yield surrogate can be combined with irrigation surrogate model to create a local level decision support system to provide yield as well as irrigation water estimates for different irrigation strategies.
- Faster simulations are critical enabler of scenario-based optimizations which allows stakeholders to take data-driven decisions.

Acknowledgment

We acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM Ganga' at the Indian Institute of Technology Roorkee, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.