***ML based forecasting of rice yield in Indo-Gangetic plains using soil, CO2 emissions and micro-climate patterns from satellite imagery***

India is enlisted in the most vulnerable countries to climate change by Global Climate Risk Index. At least one of the states experienced extreme weather events on 314 days in 2022 which affected about 2 million hectares (ha) crop area. Amongst food crop, rice is the staple for about 60% of the population and Uttar Pradesh is the second highest producer with an yield of 16.5Mt (2022). The yield is highly dependent on climate changes because solar radiation, temperature, and precipitation are the main drivers of rice growth. Most papers on yield prediction are done at a macro level and hence localized climate changes are not taken into account. In this paper, the 75 districts of UP are further divided into 25x25 km grids and the Kharif rice yield is forecasted at a grid level.

Kharif rice yield data is collected from Directorate of Economics and Statistics (Govt. of India) at a district level from 2010-2022. The yield is adjusted historically for external qualitative factors like seed variety changes, productivity increase measures over time, farming techniques, irrigation practices etc. Further, the yield is disaggregated to a 25x25 km grid based on the LULC classification done from the satellite imagery. Additionally, time series weather data, CO2 emissions and soil data from Landsat, Sentinel and other satellites is also processed at the same granularity and mapped to the yield. The overall crop production cycle is split into sowing, booting(including pre-booting) and harvesting. Based on the expected impact at each stage of growth, time-series meteorology data is utilized to create change variables – average/total 1 month precipitation, daily sunshine time, daily maximum temperature, relative humidity changes etc. Similarly soil parameters like moisture, temperature and chemical composition etc. is processed to create derived attributes. Also, soil pH, normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), near-infrared reflectance of vegetation (NIRV), and solar-induced chlorophyll fluorescence (SIF) and salinity which are key to crop growth and a healthy production is extracted using hyperspectral imagery.

Based on the year wise(2010-2020) grid level dataset with yield and other predictive variables created above, a baseline XGBoost model is trained. The model results are validated on 2021 and 2022 yield data and shows a high accuracy (R2 = 0.75, RMSE = 0.51 t/ha). Precipitation, temperature, and soil pH parameters closer to the sowing are the most significant variables in the model. This prediction model is also validated against the DSSAT software designed to estimate rice yield and remains fairly close.

This prediction model is further used to predict crop yield for the subsequent years. Also, simulations are run for the IPCC climate change scenarios (RCP 2.6, 4.5 and 8.5) and the yield estimates are forecasted. The rice yield is expected to drop by 3-5% in the medium emission scenario and about 4-10% in the high emission scenario.