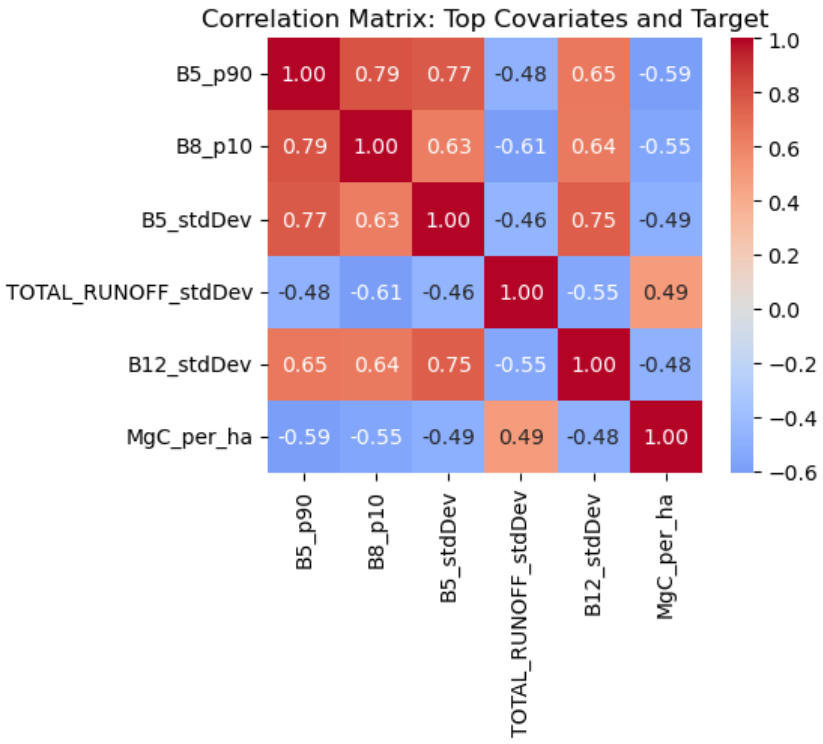
Soil Organic Carbon Stock Modeling in the Rangelands North of Mount Kenya

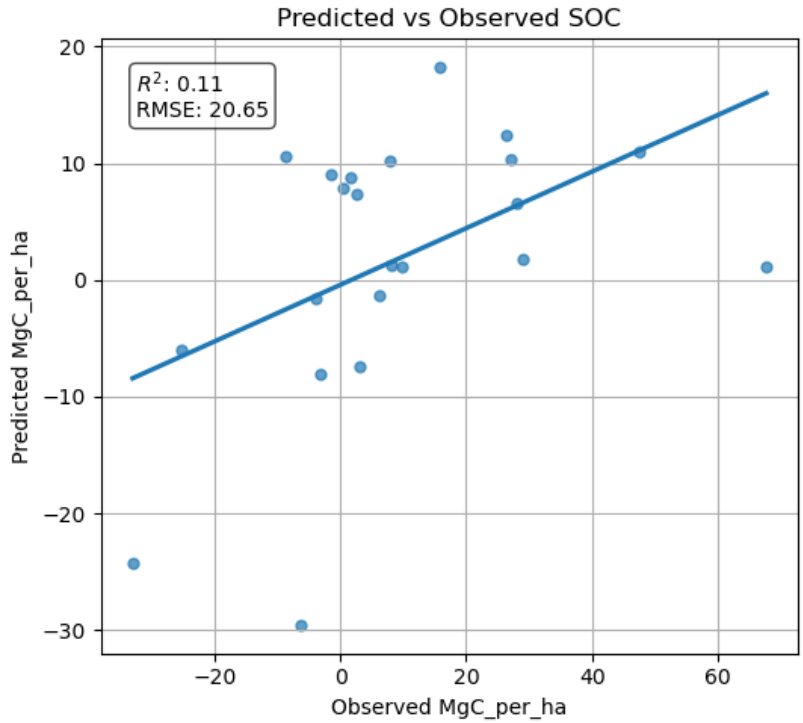
To model spatial variation in soil organic carbon (SOC) stocks, I implemented a Random Forest regression using the Google Earth Engine ee.Classifier.smileRandomForest. This model was trained on 80 field samples collected between March 2023 and February 2024 and covariate features generated from Sentinel-2 SR, TerraClimate, Copernicus DEM, and ESA WorldCover datasets. For this analysis, I assumed that SOC values from the sample period largely depend on the environmental covariate space from the previous five years (2018-2013). The chosen covariates included surface reflectance bands, vegetation/soil indices, topographic metrics, and monthly climate summaries. Each covariate was summarized using the mean, 10th and 90th percentiles, and standard deviation over the sampling period. Predictor variables were selected using a two-stage filtering process: (1) ranking by Pearson correlation with the target (MgC_per_ha), and (2) removing predictors with high multicollinearity. The final predictors were:

- B5_p90 (Near-infrared reflectance 90th percentile)
- B8_p10 (Red-edge reflectance 10th percentile)
- B5_stdDev (NIR reflectance variability)
- TOTAL_RUNOFF_stdDev (Runoff variability)
- B12_stdDev (Shortwave infrared reflectance variability)



These variables were not only statistically correlated with the target variable but are also ecologically meaningful indicators of processes that influence SOC accumulation and variability in semi-arid rangelands. For example, vegetation indices represent carbon inputs from photosynthesis and biomass, SWIR bands relate to soil conditions and water stress, and runoff metrics highlight hydrological redistribution, which governs erosion, transport, and deposition of organic carbon.

The Random Forest model was selected for its ability to capture non-linear interactions and handle high-dimensional, well-suited for environmental modeling with limited ground truth data. Modeling was performed in GEE, allowing integration of covariates, training data, and large-scale image processing within a cloud-native environment. The model was validated via an 80/20 train-test split yielding an R^2 of 0.11 and an RMSE of 20.65 MgC/ha, suggesting moderate predictive skill but substantial unexplained variance, likely due to limited training data and spatial heterogeneity.



Given the modest R^2 and high RMSE, I do not believe this model would be suitable for use in carbon crediting frameworks without further refinement. However, the workflow is scientifically sound and reproducible, providing a strong foundation for future improvements. Incorporating spatial cross-validation,

simulating SOC ranges using field standard errors, and expanding sampling density would enhance confidence in the predicted SOC estimates. This work demonstrates best practices in covariate selection, model transparency, and remote sensing integration.