over five years, followed by soil pH, clay content, electrical conductivity and NDVI, again emphasizing the high influence weight of water-related variables alongside key soil properties.

### 8.3 Discussion on model applicability and uncertainty

The applicability of soil-nutrient- and climate-based yield models depends critically on how uncertainty is handled across space, time and scenario conditions. A recent meta-analysis of crop yield responses to projected climate change combined mixed-effects modeling with block bootstrapping to partition uncertainty arising from model structure, climate projections (CMIP6) and emissions pathways, showing that simple pooled OLS tends to underestimate yield losses and under-represent uncertainty ranges (Li et al., 2025). Similarly, a crop-model and ML ensemble for maize and soybean across China demonstrated that coupling GGCMs with Random Forest greatly improved correlation (r up to 0.77 for maize) and reduced normalized RMSE, while variance decomposition revealed that the dominant uncertainty source shifted from crop models in the baseline GGCM runs to global climate models and then scenarios as projections extended further into the century (Li et al., 2023). These results imply that model applicability under future climates requires explicit accounting for structural, climate and scenario uncertainties rather than relying on single-model projections.

Transferability across domains and scales introduces additional uncertainty dimensions for data-driven yield models. Domain-adaptation work on maize in the US Corn Belt, using DANN, KLIEP and RTNN, found that models trained in temperate regions with medium-high growing degree days and moderate vapor pressure deficit generalized well, whereas strong dependence on vegetation indices (GCI) reduced transferability when source and target domains had limited overlap (Priyatikanto et al., 2023). Independent evaluations of cross-validation strategies in UAV-based yield prediction further showed that random CV can substantially overestimate performance when models are applied outside their training spatial domain, whereas spatial or leave-one-field-out CV and simpler, regularized models gave more realistic extrapolation accuracy (Habibi et al., 2024). Together with county-scale ensemble studies that link large prediction errors to low cropland ratios and extreme weather events (Sajid et al., 2022), these findings stress that robust maize yield prediction demands careful validation design, domain-aware training, and transparent uncertainty quantification before models are applied for management or policy decisions in new regions or under novel climate conditions.

## 9 Conclusions and Future Research Directions

Existing studies confirm that integrating soil nutrients, soil physical properties, and climate variables can explain a substantial share of maize yield variability across diverse agroecological zones. Soil indicators such as nitrogen fertilizer rate, soil organic carbon, pH, bulk density, and exchangeable bases consistently emerge among the most influential predictors, often exceeding the importance of individual climate variables for yield prediction in tropical and semi-arid environments. At the same time, temperature, rainfall, and related weather indices remain key drivers of interannual variation, especially when combined with management and genotype information in large datasets. From a modeling perspective, tree-based and boosting algorithms (Random Forest, XGBoost, Gradient Boosting) generally outperform linear methods and many deep architectures for maize yield prediction using soil-climate feature sets. Meta-modeling of process-based simulations and large empirical trial datasets shows that these methods can achieve relative errors around 10-15% when sufficient training samples and well-designed features are available. Systematic reviews across maize and other crops further indicate that these algorithms are among the most frequently adopted and robust options, particularly when coupled with feature engineering and multimodal data integration.

High-accuracy soil-climate yield models provide actionable information for fertilizer management and nutrient efficiency. In Ghana, Random Forest and XGBoost models trained on long-term maize trials successfully predicted both yield and agronomic efficiency, highlighting nitrogen rate, rainfall, and key soil properties as dominant management levers. Such models support the design of site-specific recommendations that can raise productivity while reducing the environmental costs of blanket fertilizer application. Similar ML-process-model hybrids using APSIM outputs demonstrate that meta-models can rapidly explore genotype-environment-management scenarios for preseason planning. At larger scales, integrating soil maps,