becomes difficult, and specialized cross-validation (CV) schemes such as leave-one-out (LOO) or nested CV are recommended to obtain unbiased generalization estimates (Dinh and Aires, 2022).

For crop yield prediction with strong spatial and temporal dependence, the choice of CV strategy affects both apparent skill and interpretability. Studies using simulated or observed yields show that random CV can give overly optimistic accuracy when neighboring samples are highly correlated, while spatial or cluster-based CV provides more realistic estimates on held-out regions (Radočaj et al., 2025). Nested CV or nested leave-two-out schemes further separate inner folds for model selection from outer folds for performance estimation, preventing overly complex models from being chosen and improving transferability across years and locations (Sweet et al., 2023).

6.2 Model parameter optimization methods

Hyperparameters of machine learning models, such as the number of trees in random forests or learning rates in gradient boosting, strongly influence predictive performance and must be tuned systematically rather than by ad-hoc trial-and-error (Bischl et al., 2021). Classical search strategies include grid search and random search, which evaluate candidate configurations on resampling-based performance estimates, but they become inefficient as the hyperparameter space grows.

More advanced approaches treat hyperparameter tuning as a black-box optimization problem and use probabilistic surrogate models. Bayesian optimization with Gaussian processes or related surrogates iteratively proposes promising configurations based on past evaluations and has been shown to find better settings than random search under comparable budgets (Wu et al., 2019). In crop yield estimation, Bayesian optimization frameworks applied to tree-based models such as LightGBM achieve high coefficients of determination and low mean squared error across several agricultural datasets, demonstrating the gains from automated hyperparameter optimization. Random forest-specific tuning via model-based optimization (e.g., tuning mtry, node size, sample size) can further increase accuracy over default settings while controlling runtime (Probst et al., 2018).

6.3 Model evaluation indicator system

Because maize yield prediction is a regression problem, a comprehensive indicator system is needed to evaluate both accuracy and explanatory power. Error-based metrics such as root mean square error (RMSE), mean absolute error (MAE), and related deviations are widely used in crop model evaluation because they directly characterize the magnitude of prediction errors in yield units (Yang et al., 2014). RMSE is particularly sensitive to large errors and is appropriate when error distributions are approximately Gaussian, whereas MAE provides a more robust and interpretable measure of average error and is less influenced by outliers (Chai and Draxler, 2014).

To complement absolute error measures, goodness-of-fit and efficiency statistics assess how much of the observed variance is explained by the model. The coefficient of determination (R^2) is often preferred as a standard metric in regression because it relates performance to the variance of ground-truth yields and is more informative than stand-alone error magnitudes in many applications (Chicco et al., 2021). In process-based crop modeling, additional indices such as modeling efficiency (EF) and the index of agreement (d) are used alongside RMSE and MAE to provide a balanced view of model bias, dispersion, and agreement with observations (Yang et al., 2014). For maize yield prediction models based on soil nutrients and climate variables, combining R^2 (or EF) with RMSE and MAE yields a robust evaluation framework that captures both accuracy and reliability across different environments.

7 Case Study: Empirical Analysis of Regional Maize Yield Prediction

7.1 Study area and sample construction

In many recent maize yield prediction studies, the study area is defined to capture both environmental gradients and management diversity so that models generalize beyond a single field or season. For example, plot-scale work integrates multi-year trials under contrasting fertilizer systems, combining climate, soil, and satellite data to represent heterogeneous growing conditions across years and treatments (Meng et al., 2021). Similar multi-farm designs in Western Australia aggregate yield monitor data from thousands of hectares over several seasons, then