monthly) and calculation of growing-season indices are performed to match crop growth stages and modeling time steps. Yield and management records are checked for outliers, coding errors, and inconsistent units across years and locations to avoid bias in training datasets (Archontoulis et al., 2020).

Remote sensing and soil datasets also undergo substantial preprocessing. For optical satellite data, procedures include cloud and shadow masking, compositing, and noise reduction to generate consistent vegetation index and land-surface-temperature time series suitable for yield prediction (Li et al., 2022). Novel image-cleaning techniques, such as quartile-based filtering of local pixel neighborhoods, can reduce sensor noise and atmospheric artifacts, improving the signal-to-noise ratio and enhancing model accuracy when combined with deep learning approaches. Soil property and nutrient data from field sampling or databases are harmonized across sources, interpolated or matched to field or grid units, and normalized or standardized for use in machine learning models that combine soil, climate, and management predictors (Diaz-Gonzalez et al., 2022). Overall, rigorous preprocessing and quality control across all data types are essential to ensure robust, interpretable relationships between soil nutrients, climate variables, and maize yield.

4 Construction and Selection of Feature Variables

4.1 Construction of soil nutrient indicator system

A scientific soil nutrient indicator system should reflect both the supply of key macronutrients and the broader edaphic conditions that control maize response. Long-term omission experiments identify available and total N, P, and K, soil organic carbon, C:N and N:P ratios as primary determinants of yield and nutrient use efficiency, showing that edaphic indicators explain more yield variation than phenological factors in maize systems (Wang et al., 2024). Meta-analysis in northern China further supports including soil organic matter, total N, and available P and K as core indicators, because these properties consistently increase under rational fertilization and are closely aligned with yield gains and water use efficiency (Jiang et al., 2024).

For predictive modeling, soil indicators must also capture spatial heterogeneity and nutrient limitations. Maize nutrient omission trials across 324 farmers' fields in the Eastern Indo-Gangetic Plains showed that soil pH was the most critical variable controlling relative N- and P-limited yields, while soil N and Zn strongly influenced Zn-limited yield (Figure 2) (Ahmed et al., 2024). Post-harvest soil test value prediction equations for N, P, and K demonstrate how pre-sowing soil tests, crop uptake, and fertilizer inputs can be combined to estimate dynamic soil nutrient status, supporting targeted fertilizer recommendations for subsequent crops (Abdel-Salam et al., 2024).

4.2 Extraction of climate variable features

Climate feature construction should represent both mean conditions and stress events during sensitive growth stages. Studies that assessed the relevance of climatic attributes for corn yield found that solar radiation, precipitation, vapor pressure, and maximum and minimum temperature are among the most influential variables, with radiation slightly exceeding precipitation in importance in Neotropical environments (Sierra-Forero et al., 2024). Regional analyses that combine multiple climate time series with yield records confirm that temperature- and water-related indicators together explain a large share of yield variability, especially when evaluated over the growing season (Luthra et al., 2024).

Careful temporal aggregation and transformation of climate variables can greatly improve prediction. Monthly vapor pressure deficit and precipitation expressed with spline functions produced the "best climate-only" model for rainfed corn, with high out-of-sample R^2, and adding satellite vegetation indices further enhanced performance (Li et al., 2019). Similar work on climate-driven yield variability uses downscaled temperature, precipitation, and shortwave radiation, plus extreme-climate indices, to quantify how mean growing-season warming, radiation changes, and counts of hot or dry days affect maize yield projections (Chen et al., 2020).