agronomy

MDPI

Article

Structured Multi-Kernel Heteroscedastic Gaussian Process for Crop Straw-to-Grain Ratio Prediction and Uncertainty Quantification

Ailian Zhou [1,2], Manfu Huang [3], Zirui Wang [3], Jiajia Liu [1,2], Xiaohe Liang [1,2], Qi Wang [1,2], Shuo Xiao [3] and Jiayu Zhuang [1,2,*]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; zhouailian@caas.cn (A.Z.); liujiajia@caas.cn (J.L.); liangxiaohe@caas.cn (X.L.); wangqi03@caas.cn (Q.W.)
[2] Key Laboratory of Agricultural Blockchain Application, Ministry of Agriculture and Rural Affairs, Beijing 100081, China
[3] School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou 221116, China; manfu_huang@163.com (M.H.); zirui_wang@163.com (Z.W.); sxiao@cumt.edu.cn (S.X.)
* Correspondence: zhuangjiayu@caas.cn

Abstract

Crop straw-to-grain ratio (SGR) estimation underpins regional straw resource assessment, yet national inventories rely on fixed coefficients that ignore structured variation across variety, environment, and phenotype. We introduce a Structured Multi-Kernel Heteroscedastic Gaussian Process (GP) framework that models SGR variation through three additive kernels heuristically motivated by the genotype–environment–phenotype (G+E+P) framework—capturing variety-associated variation, spatially structured variation, and environmental and management covariates—and employs an input-dependent noise model for prediction-specific uncertainty quantification. To prevent information leakage, target encoding and feature scaling are recomputed within each cross-validation fold. Evaluated via internal leave-one-out cross-validation on 80 rice samples (42 varieties, six Chinese provinces), the model achieves $R^2 = 0.541$ with a prediction interval coverage probability of 0.95. Ablation identifies variety-associated variation as the largest contributor among the modeled factors ($\Delta R^2 = -0.024$) and the multi-kernel design, by incorporating variety-specific information, substantially improves upon a covariate-only RBF GP ($\Delta R^2 = 0.103$). On point-prediction accuracy, Gradient Boosting achieves $R^2 = 0.58$, slightly ahead of the Heteroscedastic GP ($R^2 = 0.54$), underscoring that the primary advantage of the GP lies in its input-dependent uncertainty quantification. However, leave-one-county-out validation yields $R^2 \approx 0$ (with σ escalating to 24.4), confirming that the model does not yet generalize to unsampled counties; all reported performance is therefore internal to the nine sampled counties. The framework couples an agronomically motivated additive kernel structure with input-dependent uncertainty quantification, offering a path toward uncertainty-aware prediction from small field datasets.

Keywords: straw-to-grain ratio; Gaussian process; multi-kernel learning; heteroscedastic uncertainty; crop residue estimation; precision agriculture

Academic Editor: Jayme Garcia Arnal Barbedo

Received: 3 July 2026
Revised: 5 August 2026
Accepted: 7 August 2026
Published: 9 August 2026

1. Introduction

Crop straw resource quantification supports bioenergy infrastructure planning, soil organic carbon management, and agricultural greenhouse gas inventory compilation [1–3].

The straw-to-grain ratio (SGR)—the mass of harvestable straw produced per unit grain yield—is the primary conversion coefficient used to estimate total straw production from grain yield statistics [4]. In China, the world's largest rice producer, national straw resource assessments have conventionally adopted a fixed SGR coefficient of approximately 0.9 for rice [5], treating it as a stable species-level constant.

Field measurements contradict this assumption: rice SGR ranges from 0.4 to 3.9 across production systems, with a coefficient of variation exceeding 50%. The variation is structured: it reflects the combined influence of varietal characteristics (cultivars differ systematically in how they allocate photosynthate between grain and straw), spatial environment (climate, soil, and management practices vary across production regions), and phenotypic plasticity (within-variety trait expression modulated by local growing conditions) [6]. Applying a single fixed coefficient across all rice-producing regions consequently introduces systematic bias into national straw resource estimates.

Machine learning (ML) methods have increasingly been applied to crop trait prediction, including yield estimation with Random Forest (RF) and Gradient Boosting (GB) [7,8], and SGR estimation in wheat using UAV-derived multispectral features [9]. Recent work published in this journal has demonstrated the viability of RF and XGBoost for assessing paddy-rice biomass and nutrient status from UAV spectral data [10]. Recent reviews note the promise of these methods and the persistent gap in uncertainty quantification for data-scarce agricultural settings [11,12]. For operational SGR estimation, however, existing approaches share a common gap: they predict without explaining, and they estimate without qualifying their confidence. Black-box ensemble methods yield point predictions with no interpretable variance decomposition—an agronomist cannot determine whether a given prediction differs from the fixed coefficient because of variety, environment, or measurement noise. Standard ML models either lack uncertainty quantification entirely or provide only constant-variance prediction intervals, which do not distinguish between well-supported predictions on familiar variety–environment combinations and unsupported extrapolations to novel ones. GP regression addresses several of these constraints: it handles small samples ($n = 80$) where deep learning would overfit, yields analytic predictive uncertainty without post-hoc calibration [13], and supports structured kernel design that encodes domain knowledge directly into the covariance function.

This paper introduces a Structured Multi-Kernel Heteroscedastic Gaussian Process framework for rice SGR prediction, built around two components. First, an additive three-kernel covariance function decomposes SGR variation into varietal similarity (RBF kernel on leave-one-out target encoding), spatial environmental autocorrelation (Matérn-3/2 kernel on Haversine distance), and phenotypic plasticity (ARD kernel on reduced phenotype and external environmental features). This kernel architecture is motivated by the G+E+P variance partitioning framework of quantitative genetics [14] as a modeling heuristic, though we note that the covariate kernel shares information with the spatial kernel (e.g., elevation, weather, soil) and therefore the three kernels do not constitute a strict variance decomposition under the standard quantitative-genetic definition. Second, an input-dependent heteroscedastic noise model, $\sigma^2(\mathbf{x}) = \text{softplus}(\mathbf{W}\mathbf{x}_{\text{cov}} + b)$, replaces the standard constant-noise assumption, enabling the model to express higher uncertainty for samples whose covariate profile deviates from the training distribution.

Evaluated via internal leave-one-out cross-validation (LOOCV) on 80 rice samples spanning 42 varieties across six Chinese provinces, with target encoding and feature scaling recomputed independently within each fold to prevent information leakage, the Structured Heteroscedastic GP achieves $R^2 = 0.541$ with PICP = 0.95. Ablation analysis identifies the variety kernel as the dominant contributor to predictive accuracy ($\Delta R^2 = -0.024$) and the structured multi-kernel design substantially outperforms a single-RBF GP ($\Delta R^2 = 0.103$).

https://doi.org/10.3390/agronomy16161524

On point-prediction accuracy, Gradient Boosting achieves $R^2 = 0.58$, slightly ahead of the Heteroscedastic GP ($R^2 = 0.54$), underscoring that the primary advantage of the GP lies in its input-dependent uncertainty quantification. The heteroscedastic σ is calibrated within-domain and escalates under cross-province extrapolation; within-domain variety novelty alone does not trigger elevated uncertainty (Section 3.5). Under the more stringent leave-one-county-out protocol, however, R^2 drops to approximately zero with σ escalating to 24.4 (log-scale), underscoring that the current evaluation is internal to the nine sampled counties and does not yet support generalization to unsampled counties (Section 3.2). Together, the structured kernel design and input-dependent noise model yield a GP that is both interpretable in agronomic terms and provides input-dependent predictive variance that escalates under geographic extrapolation.

2. Materials and Methods

2.1. Data Constraints and Modeling Strategy

The dataset contains $n = 80$ field-collected samples spanning 42 rice varieties across six provinces. This covers representative rice production systems across major Chinese growing regions, but is too small for data-intensive methods: deep neural networks or high-dimensional learned embeddings would overfit. Our objective is not to chase the highest possible R^2 through model capacity. We aim instead to build an interpretable prior framework—one that encodes what is already known about crop trait variation directly into the model architecture. Under small-sample conditions, the Structured GP compensates for data scarcity through kernel design. Each kernel's functional form and input features reflect a specific component of established agronomic theory: varietal, environmental, and phenotypic. The kernels act as Bayesian priors, constraining the hypothesis space with domain knowledge rather than relying on large datasets to explore it.

2.2. Study Area and Data Collection

The dataset comprises 80 field-collected rice (*Oryza sativa* L.) samples from six provinces in China: Heilongjiang (20 samples), Zhejiang (20), Sichuan (10), Hubei (10), Hunan (10), and Hebei (10). An additional 5 soybean samples from Heilongjiang were collected under the same sampling protocol but excluded from this study, which focuses exclusively on rice. Samples were collected during the 2025 harvest season (September–October) and span 42 distinct rice varieties, with most varieties represented by 1–2 samples (maximum 8). At each sampling plot, five 50 cm × 50 cm sub-quadrats were pooled into one composite sample (full protocol below). Compositional analyses followed national standards: elemental analysis [15], lignocellulosic analysis [16], and ash content [17]. For each sample, the following measurements were recorded: geographic coordinates (longitude, latitude), elevation (5–519 m), cropping system, growth period (105–158 days), plant height (80–110 cm), grain color category, and straw-to-grain ratio (0.4–3.9). These 80 samples represent the subset of a broader national survey (90 rice samples across 8 provinces) for which all compositional measurements were completed at the time of analysis. Oven-dry weight of the whole plant was also recorded but is excluded from all analyses because it is computed from the same raw straw and grain weight measurements as the response variable (SGR), creating procedural leakage. Table 1 summarizes the dataset characteristics.

The sampling protocol followed a two-stage cluster design. Within each county, five administrative villages were selected across at least three townships. Within each village, two uniform fields were chosen, designed to yield ten sampling plots per county, with actual counts ranging from 5 to 10 (mean 8.9) across the nine counties represented in the final dataset. At each plot, five 50 cm × 50 cm sub-quadrats were placed using a five-point layout (four corner-adjacent and one center); above-ground plant material from all five

sub-quadrats was harvested at ground level, pooled, and processed as a single composite sample. Straw and grain were separated by manual threshing (rice hulls were retained with the grain fraction; SGR is reported on a paddy-rice basis). Both fractions were sun-dried, then oven-dried at 65°C to constant weight (consecutive weighings differing by <0.5%). Straw-to-grain ratio was calculated as SGR $= W_{straw}/W_{grain}$, where W_{straw} and W_{grain} denote the oven-dry weights of the straw and grain fractions, respectively. All weighings used an electronic balance with 0.01 g precision.

Table 1. Dataset summary statistics.

Characteristic	Value
Number of samples	80
Number of rice varieties	42
Number of provinces	6
SGR range	0.4–3.9
SGR mean ± SD	1.35 ± 0.60
SGR coefficient of variation	44%
Phenotype features (retained)	7
External weather features	7
External soil features	6
Total feature dimension after encoding	20

External environmental data were obtained from two public sources and matched to each sample by geographic coordinates. Seven annual meteorological variables were retrieved via the NASA POWER API (https://power.larc.nasa.gov/, accessed 15 July 2025) (0.5° × 0.5° grid, version 2.0) for the 2024 calendar year: annual mean temperature (°C), annual maximum temperature (°C), annual minimum temperature (°C), total annual precipitation (mm), daily mean solar radiation (kWh m^{-2}), annual mean relative humidity (%), and ≥0°C accumulated thermal time (°C·d). Because sowing and transplanting dates were not recorded (see Section 4.4), variety-specific growing-season windows could not be determined; annual aggregates were used as a coarser proxy, acknowledging that they include non-growing-season months and may dilute the climate signal relevant to SGR. Six soil properties were extracted from the SoilGrids 2.0 database (250 m resolution, WGS84 CRS, accessed July 2025) as depth-weighted averages of the 0–5, 5–15, and 15–30 cm standard layers (weighted by layer thickness: 5, 10, and 15 cm, respectively) [18]: soil organic carbon (g kg^{-1}), total nitrogen (g kg^{-1}), pH (H_2O), clay content (%), sand content (%), and silt content (%). Environmental data were pre-extracted at the county level for eight administrative units across the six provinces and matched to individual field samples via a three-tier procedure (county-exact match → city-level fallback → province-level fallback; see Table A1 for county-level sample composition). Under this scheme, 40 of 80 samples received county-exact environmental data, 30 received city-level matches, and 10 received province-level matches; no environmental data were missing for any field sample after matching. One external-data extraction point (Fujin, Heilongjiang, 85 km from the nearest field-sampled county) had incomplete SoilGrids coverage at the time of retrieval and was imputed from the nearest available point (Baoqing, Heilongjiang); this extraction point was not assigned to any field sample under the matching procedure and therefore does not affect the analysis.

Of the 90 rice samples originally collected across 8 provinces, 10 samples from Jiangsu and Anhui provinces were excluded because lignocellulosic and elemental compositional measurements had not been completed at the time of analysis. The remaining 6 provinces (Heilongjiang, Hebei, Hubei, Hunan, Zhejiang, Sichuan) span single-cropping and double-cropping rice systems across northeastern, central, and southwestern China. However,

Jiangsu and Anhui represent important rice-producing areas in the lower Yangtze basin, and their absence may introduce a geographic coverage gap; this should be considered when interpreting the model's regional representativeness.

The 80 retained samples are distributed across nine counties (mean 8.9 samples per county; range 5–10). Within each county, five villages were selected across at least three townships; within each village, two uniform fields were chosen. The resulting hierarchical structure—county, village, field—means that samples from the same county share local climate, soil, and management practices, and are not statistically independent spatial replicates. Of the 42 rice varieties, 26 are represented by a single sample and 16 by two to eight samples, with only 8 varieties having three or more samples. This imbalanced variety representation limits the model's ability to learn variety-specific effects for most cultivars and means that standard LOOCV mixes predictions for well-represented varieties (where a training-set peer exists) with predictions for singletons (where the variety TE collapses to the global mean).

The NASA POWER 0.5° × 0.5° grid corresponds to approximately 50 km at mid-latitudes; samples from the same county (typically 10–30 km across) may therefore share identical weather inputs, reducing effective weather-dimension sample size. Because county-level environmental data were matched to field samples via a three-tier fallback (county → city → province; see Table A1), 50% of samples received city- or province-level environmental proxies rather than county-exact matches. This introduces spatial smoothing of environmental covariates for samples from counties without dedicated external-data extraction, though the resulting bias is conservative: province-level fallback assigns the nearest available same-province environmental profile, which preserves broad climatic gradients while attenuating fine-scale variation.

Most varieties appear only once or twice in the dataset—a common feature of multi-site field surveys that prioritize breadth over depth. This shapes the modeling strategy in several ways. One-hot encoding of variety identity is ruled out: it would add 42 sparse binary dimensions to an 80-sample dataset. Learned embeddings are similarly impractical, since they need multiple examples per category to produce useful representations. The uneven distribution of samples across varieties makes leave-one-out cross-validation the natural choice: k-fold partitioning risks dumping all samples of a given variety into the same fold. The feature engineering and validation design described below work within these constraints.

2.3. Feature Engineering

2.3.1. Target Encoding of Variety

Variety identity was encoded using leave-one-out target encoding (LOO TE). For a training sample i belonging to variety v, the encoding value is the empirical-Bayes-shrunken mean log-SGR of all other samples of variety v:

$$\mathrm{TE}_i^{(tr)} = \frac{n_v \cdot \bar{y}_v^{(-i)} + k \cdot \bar{y}^{(tr)}}{n_v + k}, \quad \lambda_v = \frac{n_v}{n_v + k} \tag{1}$$

where n_v is the number of training samples of variety v, $\bar{y}_v^{(-i)} = \frac{1}{n_v - 1} \sum_{j \in S_v \setminus \{i\}} \log(y_j)$ is the LOO mean of variety v, $\bar{y}^{(tr)}$ is the training-set global mean of log-SGR, and k is a smoothing parameter ($k \to 0$: unshrunken within-variety mean; $k \to \infty$: global mean). For a test sample j, the encoding is computed from training-set statistics *without* leave-one-out adjustment:

$$\mathrm{TE}_j^{(te)} = \frac{n_v^{(tr)} \cdot \bar{y}_v^{(tr)} + k \cdot \bar{y}^{(tr)}}{n_v^{(tr)} + k} \tag{2}$$

where $n_v^{(tr)}$ and $\bar{y}_v^{(tr)}$ are the count and mean of variety v in the training set. Two boundary cases collapse to the global mean: (i) a singleton training variety with $k = 1$ where $n_v^{(-i)} = 0$ after own-exclusion, yielding $\mathrm{TE}_i^{(tr)} = \bar{y}^{(tr)}$; and (ii) a test variety absent from training ($n_v^{(tr)} = 0$), yielding $\mathrm{TE}_j^{(te)} = \bar{y}^{(tr)}$.

Consequently, for a variety not present in the training set, the TE value defaults to the training-set global mean of log-SGR—equivalent to an empirical-Bayes shrinkage estimator with the grand mean as prior and zero variety-specific observations. This directly addresses a practical concern: when a new rice cultivar enters production, the model predicts using the global average SGR until variety-specific data accumulate.

To prevent information leakage, TE statistics ($\bar{y}^{(tr)}$, $\{n_v^{(tr)}, \bar{y}_v^{(tr)}\}$) are recomputed within each cross-validation fold using only the training partition. The smoothing parameter $k = 1$ was selected via flat 5-fold cross-validation over $\mathcal{K} = \{1, 2, 3, 5, 8, 10, 15\}$ with nested TE recomputation per fold. The low optimal k reflects the predominance of singleton varieties: 26 of 42 varieties appear only once, and larger k would shrink them strongly toward the global mean, diluting the limited variety-specific signal. Feature scaling (StandardScaler) was similarly fitted on each training fold and applied to the held-out sample. LOO encoding was chosen over one-hot encoding (42 sparse binary dimensions for 80 samples) and learned embeddings (overfitting risk with 1–2 samples per variety). The complete nested encoding procedure is specified in Appendix A.

2.3.2. Spatial Features

Pairwise geographic distances between sampling locations were computed using the Haversine formula on (longitude, latitude) coordinates, accounting for Earth's curvature. The resulting distance matrix directly parameterizes the spatial kernel.

2.3.3. Phenotype Features and Log Transformation

We retained 7 of the original phenotype dimensions, dropping grain color (limited variation), growth period (collinear with variety), dry weight (procedural leakage; see Section 2.1), and interaction terms. The 7 retained features comprise plant height, elevation, and five one-hot encoded cropping system indicators, concatenated with 7 weather and 6 soil external features to yield a 20-dimensional covariate vector. The target variable (SGR) exhibits right-skew (skewness = 1.62). We modeled $\log(y)$ rather than y directly, with final predictions transformed back to the original scale using the LogNormal correction $\mathbb{E}[y] = \exp(\mu + \sigma^2/2)$. All predictive standard deviations (σ) reported in this paper are expressed in log-scale unless explicitly noted otherwise.

2.4. Structured Multi-Kernel Gaussian Process

2.4.1. Gaussian Process Preliminaries

A Gaussian Process (GP) is a stochastic process defined by a mean function $m(\mathbf{x})$ and a covariance (kernel) function $k(\mathbf{x}, \mathbf{x}')$, such that any finite collection of function values follows a joint Gaussian distribution: $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. Given n training observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i = f(\mathbf{x}_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$, the posterior predictive distribution at a test point $\mathbf{x}_*$ is:

$$p(f_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \mathcal{N}(\mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{y},\ k_{**} - \mathbf{k}_*^\top \mathbf{K}^{-1} \mathbf{k}_*) \tag{3}$$

following standard notation [13], where $\mathbf{K}$ denotes the covariance matrix of the training observations (including the noise variance on the diagonal). All GP models were implemented in GPyTorch v1.15.2 [19].

2.4.2. Kernel Design

We construct an additive composite kernel whose three components capture the main sources of SGR variation:

$$K = K_{\text{variety}} + K_{\text{spatial}} + K_{\text{covariate}} \tag{4}$$

Observation noise is modeled separately through the heteroscedastic noise component (Section 2.5).

Variety kernel. Variety-associated similarity is modeled through an RBF kernel on the one-dimensional target encoding:

$$K_{\text{variety}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_v^2 \exp\left(-\frac{(\text{TE}_i - \text{TE}_j)^2}{2\ell_v^2}\right) \tag{5}$$

RBF was chosen over linear for the variety kernel because the similarity between two varieties' SGR may decay nonlinearly with their TE distance: varieties with very different mean SGR should be substantially less similar than those with moderately different mean SGR. One-dimensional TE was chosen over 42-dimensional one-hot encoding, which would be prohibitively sparse at $n = 80$.

Spatial kernel. Environmental autocorrelation is captured through a Matérn-3/2 kernel on Haversine distance:

$$K_{\text{spatial}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_s^2 \left(1 + \frac{\sqrt{3}\, d_{ij}}{\rho_s}\right) \exp\left(-\frac{\sqrt{3}\, d_{ij}}{\rho_s}\right) \tag{6}$$

where d_{ij} is the Haversine distance (km). Matérn-3/2 was preferred to the smoother RBF because environmental covariates often exhibit abrupt rather than infinitely smooth spatial transitions at the regional scale. For the spatial extent considered here (≈2000 km maximum separation), the Matérn-3/2 kernel with Haversine distance yields positive definite covariance matrices in all LOOCV folds, as verified by Cholesky decomposition (with numerical jitter of 10^{-6}). For global-scale applications, chordal-distance formulations would be recommended [13].

Covariate kernel. Environmental and management covariates are modeled through an RBF kernel with automatic relevance determination (ARD):

$$K_{\text{covariate}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_p^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \frac{(x_{i,d} - x_{j,d})^2}{\ell_d^2}\right) \tag{7}$$

where $D = 20$ (7 covariate + 7 weather + 6 soil features), and each dimension d has an independent lengthscale ℓ_d. ARD was preferred to an isotropic RBF because the 20 input dimensions span heterogeneous data sources with fundamentally different characteristic scales. While additive GP structures decompose the predictive function into interpretable components, individual kernel contributions are not statistically identifiable in the variance-decomposition sense without orthogonality constraints [20]; we therefore treat the additive structure as a modeling heuristic rather than a formal decomposition, consistent with the interpretation in Section 2.3.

2.4.3. Agronomic Motivation of the Additive Kernel Structure

The additive kernel structure is motivated by the G+E+P framework widely used in quantitative genetics and plant breeding [14,21] as a modeling heuristic. Conceptually, K_{variety} captures variety-associated (G) variation, K_{spatial} captures spatially structured (E) variation, and $K_{\text{covariate}}$ captures covariate-driven variation (subsuming the phenotype-

mediated P component in the G+E+P framework along with environmental and management covariates). These three kernels are not strictly orthogonal: the covariate kernel includes elevation, cropping system, weather, and soil features that overlap with the spatial kernel, and the TE-based variety kernel may indirectly capture environment effects when varieties are geographically clustered. The additive structure is therefore best understood as a structured prior that organizes domain knowledge, rather than a formal variance decomposition in the quantitative-genetic sense. Figure 1 illustrates the full model architecture. The ablation experiment (Section 3.3) directly tests whether removing each component degrades prediction in the direction predicted by agronomic knowledge.

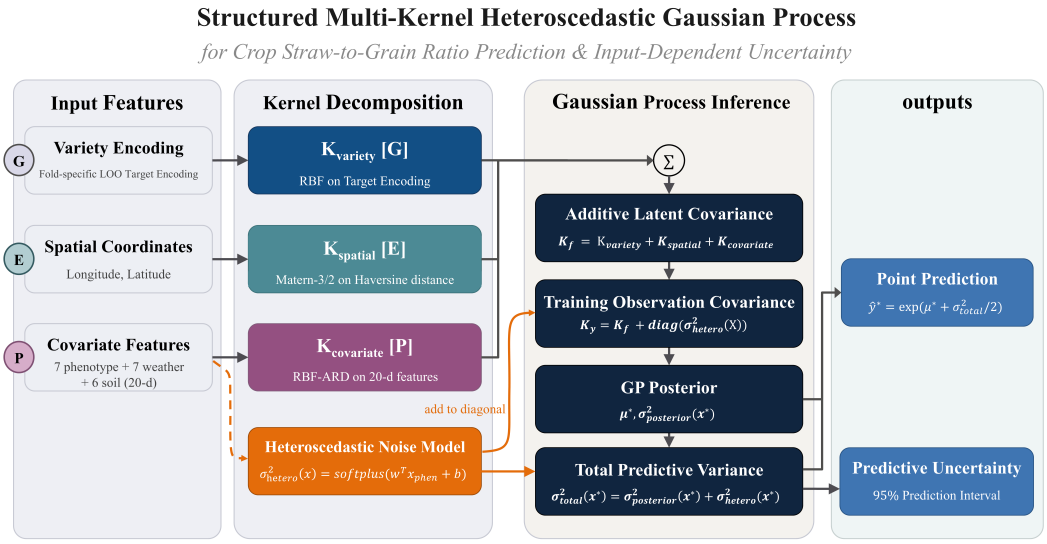


Figure 1. Architecture of the Structured Multi-Kernel Heteroscedastic Gaussian Process. Input features are routed to three additive kernels (K_{variety}, K_{spatial}, $K_{\text{covariate}}$) whose sum forms the GP covariance function. Covariate features separately feed a heteroscedastic noise model $\sigma^2(\mathbf{x})$ that provides input-dependent predictive variance. The additive kernel structure is heuristically motivated by the agronomic G+E+P framework; the kernels are not orthogonal and do not constitute a formal variance decomposition.

2.5. Heteroscedastic Noise Model

The standard homoscedastic GP assumes constant observation noise σ_n^2 across all inputs. Our data violate this assumption: a Standard GP with constant noise achieves PICP = 0.650 (Section 3.4), with a predicted σ 32% below the observed MAE (calibration ratio = 0.68; for reference, well-calibrated Gaussian errors yield a ratio of ≈1.25).

We replace the constant noise with an input-dependent noise variance:

$$\sigma^2(\mathbf{x}_i) = \text{softplus}(\mathbf{w}^\top \mathbf{x}_i^{\text{cov}} + b) \tag{8}$$

where $\mathbf{x}_i^{\text{cov}} \in \mathbb{R}^{20}$ is the covariate feature vector and $\text{softplus}(z) = \log(1 + e^z)$ ensures positivity. This linear-in-parameters noise model was chosen over a neural network because, at $n = 80$, a deeper noise model would risk overfitting the variance function [22,23]. Recent advances in efficient Heteroscedastic GP inference via kernel smoothing lend further support to this modeling strategy [24]. Covariate features were chosen as the noise predictor because variety novelty typically appears as covariate profiles that depart from the training distribution.

Predictive Variance Structure

Throughout this paper, the reported predictive standard deviation σ for the Heteroscedastic GP refers to the total predictive uncertainty. The additive kernel is $K = K_{\text{variety}} + K_{\text{spatial}} + K_{\text{covariate}}$. The observation noise consists of two components: a constant base noise from the Gaussian likelihood (fixed at 10^{-3} for numerical stability) and an input-dependent term $\sigma^2_{\text{noise}}(\mathbf{x}) = \text{softplus}(\mathbf{w}^\top \mathbf{x}^{\text{cov}} + b)$ (Equation (7)) added to the diagonal of the training and test covariance matrices. The total predictive variance is:

$$\sigma^2_{\text{total}}(\mathbf{x}_*) = \sigma^2_{\text{posterior}}(\mathbf{x}_*) + \sigma^2_{\text{noise}}(\mathbf{x}_*) \tag{9}$$

where $\sigma^2_{\text{posterior}}$ is the GP posterior variance derived from the three-kernel covariance structure, capturing **posterior uncertainty** —the model's uncertainty about the latent function value due to sparse sampling in variety–spatial–covariate space. The input-dependent term σ^2_{noise} captures **input-dependent observation variance** that varies with the covariate profile of each sample. The constant likelihood noise (10^{-3}) is negligible relative to both components and serves only to ensure numerical stability of the Cholesky decomposition; it does not contribute to the reported σ. Without replicate observations, the present data cannot empirically separate measurement variability from function misspecification. The variance components reported here should be interpreted as model constructs rather than identified aleatoric and epistemic uncertainties. A high σ may reflect posterior uncertainty (the sample lies in a sparsely populated region of variety–spatial–covariate space), input-dependent observation variance (the sample has an atypical covariate profile), or both. We report the prediction interval for the observation (future sample), which includes both uncertainty components.

Kernel hyperparameters and noise model parameters are jointly optimized by maximizing the log marginal likelihood with the Adam optimizer [25] (learning rate 0.05, 300 iterations). A Cholesky jitter of 10^{-3} is applied for numerical stability.

2.6. Model Training and Evaluation

All models were evaluated using leave-one-out cross-validation (LOOCV). We used LOOCV rather than k-fold cross-validation because (1) at $n = 80$, maximizing the training-set size for each fold is critical for stable GP hyperparameter estimation; and (2) k-fold partitioning risks systematic bias when folds inadvertently concentrate or exclude specific varieties. Within each outer LOOCV fold, all preprocessing steps—target encoding, feature standardization, and external data matching—are performed exclusively on the 79 training samples and then applied to the held-out sample. This nested design ensures that the evaluation provides an honest estimate of generalization performance conditional on the sampled counties.

We employed two complementary sets of metrics to assess point prediction accuracy and uncertainty quality:

- **Point prediction:** coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE).
- **Uncertainty quality:** prediction interval coverage probability (PICP)—the fraction of true values falling within the 95% prediction interval [26]; mean prediction interval width (MPIW); and calibration ratio—the ratio of mean predicted standard deviation to mean absolute error. Recent work has demonstrated that dual-objective optimization of prediction interval width and coverage can produce narrower, more informative intervals for agricultural prediction tasks [27].

2.7. Baseline Models

Ten models were evaluated under the identical LOOCV protocol: (1) Empirical coefficient (fixed 0.9); (2) Ridge regression; (3) SVR-RBF; (4) Random Forest (200 trees); (5) RF with tree standard deviation intervals; (6) Gradient Boosting (HistGradientBoostingRegressor); (7) Standard GP (single RBF+ARD kernel, constant noise); (8) Structured GP (three-kernel additive, constant noise); (9) Heteroscedastic GP (three-kernel additive with input-dependent noise, proposed); and (10) Mixed-Effects model (random intercept by county, REML estimation).

All baseline models were evaluated with fixed scikit-learn hyperparameters (Ridge: $\alpha = 1.0$; SVR: RBF kernel, $C = 1.0$, $\gamma =$ scale; Random Forest: 200 trees; Gradient Boosting: 200 iterations, no early stopping). No hyperparameter optimization was performed for any baseline model; the comparison therefore reflects out-of-the-box performance rather than each model's fully tuned potential.

3. Results

3.1. Comprehensive Model Comparison

Table 2 presents the internal LOOCV results with all preprocessing steps nested within each fold to prevent information leakage. The fixed empirical coefficient (0.9) yields $R^2 \approx 0.00$, confirming that a constant SGR provides no predictive information for individual field samples. Gradient Boosting achieves the highest R^2 (0.577) among all models, marginally ahead of the proposed Heteroscedastic GP ($R^2 = 0.541$, $\Delta R^2 = -0.036$), followed by Random Forest (0.477). The R^2 difference between GP and GB is not statistically significant: bootstrap 95% CI for ΔR^2 (GP − GB) from 10,000 paired resamples of the 80 LOOCV folds is $[-0.111, +0.057]$, and a Wilcoxon signed-rank test on paired absolute errors yields $p = 0.75$. The primary advantage of the GP lies not in point-prediction accuracy but in its simultaneous delivery of input-dependent uncertainty with near-nominal coverage for within-distribution samples. The distinction is further underscored in uncertainty quality: the Standard GP yields miscalibrated uncertainty (PICP = 0.650, calibration ratio = 0.68), with its single-RBF kernel capturing substantially less predictive structure than the structured variant ($R^2 = 0.407$ vs. 0.508 for the Structured GP). The proposed Heteroscedastic GP is the only model that simultaneously delivers competitive point accuracy ($R^2 = 0.541$) and input-dependent uncertainty with near-nominal marginal coverage (PICP = 0.95), though coverage degrades under LOVO (PICP = 0.788; Section 3.5). In original SGR units (log-normal corrected for GP models, naive exponentiation for others), the Heteroscedastic GP achieves $\text{RMSE}_{\text{orig}} = 0.491$ and $\text{MAE}_{\text{orig}} = 0.338$, compared to Gradient Boosting $\text{RMSE}_{\text{orig}} = 0.450$ and $\text{MAE}_{\text{orig}} = 0.306$. The relative ordering of models is preserved across transformations.

As a classical statistical baseline, a linear mixed-effects model with a random intercept by county ($\log \text{SGR} \sim 1 + (1 \mid \text{county})$) achieves $R^2 = 0.470$ under the same nested LOOCV protocol—below Gradient Boosting and the Structured GP variants, but above Standard GP ($R^2 = 0.407$), confirming that county-level grouping explains substantial SGR variation. The mixed model's prediction intervals, derived from REML variance components, are homoskedastic (same width for all samples) and over-cover substantially (PICP = 0.988, MPIW = 1.805). This comparison highlights a key limitation of classical mixed models for SGR estimation: while they capture grouped mean structure parsimoniously, they cannot express the input-dependent predictive uncertainty that the Heteroscedastic GP provides.

Table 2. Comprehensive model comparison under LOOCV ($n = 80$, log-transformed target). Dashes indicate models without native prediction intervals. All GP variants use $k = 1$ target encoding for like-for-like kernel comparison. MAE = mean absolute error; MPIW = mean prediction interval width (95%).

Model	R^2	RMSE	MAE	PICP	MPIW	Calib. Ratio
Empirical coefficient (0.9)	0.000	0.430	0.353	–	–	–
Ridge regression	0.516	0.299	0.228	–	–	–
SVR-RBF	0.311	0.357	0.283	–	–	–
Random Forest	0.477	0.311	0.221	–	–	–
+Tree Std	0.477	0.311	0.221	0.688	0.675	0.78
Gradient Boosting	0.577	0.280	0.211	–	–	–
Mixed-Effects (1|county)	0.470	0.313	0.242	0.988	1.805	1.90
Standard GP	0.407	0.331	0.220	0.650	0.587	0.68
Structured GP	0.508	0.302	0.225	0.475	0.358	0.41
Heteroscedastic GP	**0.541**	**0.291**	**0.215**	**0.950**	1.422	1.63

Bold values denote the proposed method (Heteroscedastic GP).

3.2. Internal vs. External Validation Scope

The LOOCV results reported above assess the model's ability to predict held-out samples from the same set of nine counties. Because samples from the same county share local climate, soil, and management regimes, LOOCV evaluates interpolation within the sampled spatial domain rather than generalization to new locations. To delineate the model's current generalization boundary, we conducted two additional validation protocols that remove entire geographic units: leave-one-province-out and leave-one-county-out (LOCO).

Table 3 reports per-province spatial holdout results for the Heteroscedastic GP. All six provinces yield negative R^2 under cross-province prediction, confirming that the model cannot generalize to entirely unseen provinces at $n = 80$. The degree of degradation varies: Hubei ($R^2 = -0.004$, RMSE $= 0.321$) and Hunan ($R^2 = -0.362$, RMSE $= 0.243$) approach the constant baseline, while Heilongjiang ($R^2 = -6.97$, 16 of 16 test varieties novel) and Hebei ($R^2 = -2.88$) represent worst-case extrapolation scenarios where varietal and climatic overlap with the training provinces is effectively zero. Hebei exhibits GP numerical instability (mean $\sigma = 105.0$ on log-scale), reflecting the complete absence of any training sample from the North China Plain rice system. Zhejiang ($n = 20$, PICP $= 0.35$) shows that cross-province prediction intervals can be severely under-confident as well as over-conservative—the direction of miscalibration depends on the specific extrapolation direction.

Table 3. External validation: leave-one-province-out results for the Heteroscedastic GP ($n = 80$, log-scale). Provinces are sorted by R^2 (descending). Leave-one-county-out (LOCO) pooled metrics: $R^2 = 0.006$, RMSE $= 0.429$, PICP $= 0.788$, mean $\sigma = 24.4$ ($n = 80$, 9 counties; per-county R^2 is not reported individually due to test-set sizes of 5–10).

Province	n	R^2	RMSE	PICP	Mean σ
Hubei	10	−0.004	0.32	0.80	0.22
Hunan	10	−0.36	0.24	1.00	1.01
Sichuan	10	−0.84	0.63	0.90	0.69
Zhejiang	20	−2.70	0.40	0.35	0.14
Hebei [a]	10	−2.88	0.67	1.00	105.0
Heilongjiang	20	−6.97	0.67	1.00	0.95

[a] All test varieties and the North China Plain rice production system are absent from the five training provinces. The extreme σ reflects GP numerical instability under complete varietal and climatic non-overlap.

The leave-one-county-out (LOCO) experiment complements this picture at finer spatial resolution. Holding out each of the nine counties in turn, the Heteroscedastic GP achieves LOCO $R^2 \approx 0.006$ (pooled, $n = 80$), confirming that the model cannot generalize to unsampled counties. Mean σ reaches 24.4, and MPIW reaches 95.5 (log-scale), two orders

of magnitude above the internal LOOCV values, driven largely by one county where all test varieties are novel and spatial extrapolation is extreme. Per-county R^2 values are not reported individually because test-set sizes of 5–10 samples per county make county-level R^2 statistically unreliable; the pooled LOCO metrics provide the only stable summary at this spatial resolution.

These results establish the scope of the present validation: the LOOCV metrics reported in Sections 3.1–3.3 constitute **internal validation** —performance conditional on the nine sampled counties and their specific variety–environment combinations. External validation on independent counties and provinces remains necessary before the model can support operational SGR estimation beyond the sampled domain.

3.3. Ablation Study: Kernel Contribution Analysis

To isolate the contribution of the structured kernel design from the effect of target encoding, we conducted a like-for-like ablation study in which all five configurations share identical TE preprocessing ($k = 1$, recomputed within each LOOCV fold) and homoscedastic Gaussian noise. The configurations differ only in kernel structure (Table 4):

(i) TE + monolithic ARD single kernel—all feature blocks (TE, spatial, covariate) stacked into one RBF+ARD;
(ii) TE + three-kernel additive—$K_{\text{var}} + K_{\text{spat}} + K_{\text{cov}}$ (current full model, baseline);
(iii) TE + variety kernel only (K_{var}, 1D RBF on TE scalar);
(iv) TE + spatial kernel only (K_{spat}, Matérn-3/2 on Haversine distance);
(v) TE + covariate kernel only (K_{cov}, RBF+ARD on 20-dimensional feature block).

Table 4. Ablation study: like-for-like LOOCV comparison with TE controlled across all configurations.

Configuration	R^2	ΔR^2
Three-kernel add ($K_{\text{var}} + K_{\text{spat}} + K_{\text{cov}}$)	0.509	(baseline)
TE + ARD monolithic single kernel	0.407	**−0.103**
TE + variety kernel only	0.320	**−0.189**
TE + spatial kernel only	0.464	**−0.045**
TE + covariate kernel only	0.490	**−0.019**

All configurations use identical nested TE preprocessing ($k = 1$, recomputed per LOOCV fold) and homoscedastic Gaussian noise for clean kernel-structure comparison. Bold ΔR^2 values indicate performance degradation relative to the three-kernel full model.

The structured three-kernel additive model achieves $R^2 = 0.509$ (baseline). The covariate-only kernel provides the strongest single-channel signal ($R^2 = 0.490$, $\Delta R^2 = -0.019$), closely followed by the spatial-only kernel ($R^2 = 0.464$, $\Delta R^2 = -0.045$). The variety-only kernel, constrained to a 1D RBF on the TE scalar, yields modest standalone performance ($R^2 = 0.320$, $\Delta R^2 = -0.189$). Critically, the monolithic ARD single kernel—which receives identical features and TE but lacks the structured additive decomposition—achieves only $R^2 = 0.407$ ($\Delta R^2 = -0.103$). The $+0.103$ gain of the three-kernel additive design over the monolithic baseline isolates the value of the structured prior, independent of TE.

The covariate and spatial kernels are the primary predictive drivers, together nearly matching the full model. The variety kernel alone provides limited predictive power due to its 1D structure, but its inclusion in the additive ensemble yields a positive synergistic contribution (by $\Delta R^2 = +0.019$ over covariate-only and $\Delta R^2 = +0.045$ over spatial-only). The large gap between the additive decomposition and the monolithic ARD baseline ($\Delta R^2 = +0.103$) confirms that architecting domain knowledge into the kernel design provides predictive value beyond feature engineering alone.

3.4. Uncertainty Calibration Analysis

We assess the probabilistic calibration of the Heteroscedastic GP beyond the single-point PICP metric. Figure 2 (left) shows the reliability diagram at five nominal coverage levels (0.50, 0.68, 0.80, 0.90, 0.95), each with its binomial 95% confidence interval. The empirical coverage tracks the nominal level across all five targets: at the 95% level, PICP = 0.94 (95% CI [0.86, 0.98]); at the 90% level, PICP = 0.86 (95% CI [0.77, 0.93]). All nominal values fall within their respective CIs, confirming that the prediction intervals are well-calibrated at multiple coverage levels, not only at 95%.

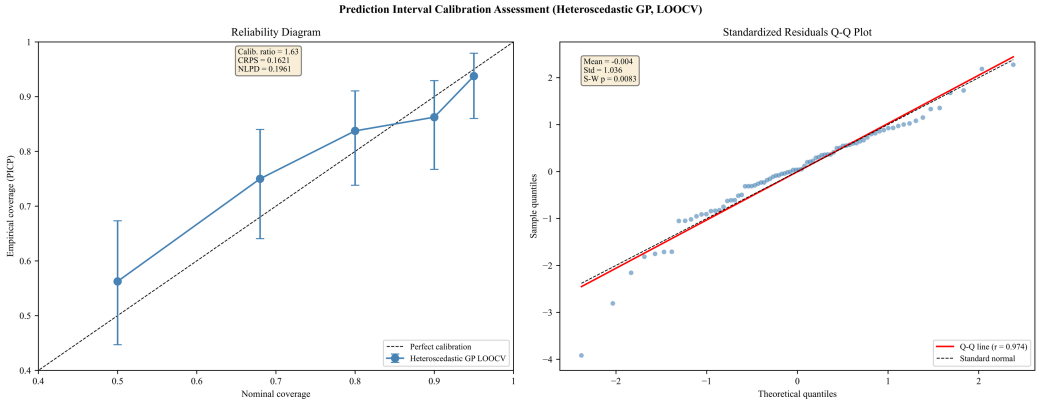


Figure 2. Probabilistic calibration assessment of the Heteroscedastic GP under LOOCV ($n = 80$, log-scale). (**Left**): Reliability diagram at five nominal coverage levels with binomial 95% confidence intervals. (**Right**): Q-Q plot of standardized residuals $(y_i - \hat{\mu}_i)/\hat{\sigma}_i$ against theoretical standard normal quantiles. The close alignment to the diagonal (std = 1.04, mean ≈ 0) confirms well-calibrated predictive distributions.

As a proper scoring rule, the Continuous Ranked Probability Score (CRPS) is 0.162 (log-scale), and the Negative Log Predictive Density (NLPD) is 0.196, lower than a constant-variance baseline (NLPD = 0.208), confirming that the input-dependent variance model improves density forecasts. The calibration ratio is 1.63, indicating mild conservatism (predicted σ exceeds mean absolute error by 63%), which is preferable to the under-confidence of the Standard GP (calibration ratio = 0.68).

Figure 2 (right) shows the Q-Q plot of standardized residuals $(y_i - \hat{\mu}_i)/\hat{\sigma}_i$. The standardized residuals have mean -0.004 and standard deviation 1.036, closely matching the standard normal distribution (Shapiro–Wilk $p = 0.008$, indicating slight heavy-tail deviation). A standard deviation near 1.0 is the defining property of probabilistic calibration: the predictive distribution's spread matches the observed error magnitude. The mild over-coverage at lower nominal levels (PICP 0.56 at nominal 0.50) reflects the model's conservative tendency, which is practically acceptable for agricultural decision support where under-confidence is costlier than over-confidence.

Stratifying by variety frequency, known varieties ($n = 54$) achieve PICP = 0.94 at nominal 0.95, while single-sample varieties ($n = 26$) achieve PICP = 0.92 (95% CI [0.75, 0.99]). The slightly lower coverage for singletons reflects the reduced information available for these varieties, but the nominal rate remains within the confidence interval.

3.5. Leave-One-Variety-Out Stress Test and Known-vs-Novel Generalization

Standard LOOCV mixes two prediction scenarios: **known-variety** prediction (54 samples from 16 multi-sample varieties, where at least one same-variety peer remains in training) and **single-sample-variety** prediction (26 samples from 26 singletons, where TE collapses to the global mean). Table 5 reports these two subsets separately alongside LOVO results.

Table 5. Validation results by novelty level for the Heteroscedastic GP ($n = 80$, log-scale). LOOCV is split into known-variety (multi-sample, at least one same-variety peer in training) and single-sample-variety subsets. LOVO holds out all samples of each variety. Cross-province results (Section 3.6) are shown for comparison of within- and cross-domain uncertainty.

Validation Level	R^2	RMSE	PICP	MPIW	Mean σ
LOOCV, known-variety ($n = 54$)	0.480	0.280	0.944	1.45	0.369
LOOCV, single-sample variety ($n = 26$)	0.480	0.305	0.923	1.22	0.312
LOVO, all varieties ($n = 80$, 42 folds)	0.318	0.355	0.788	1.18	0.302
Cross-province holdout (mean of 6 provinces) [a]	−2.29	0.489	0.84	70.6	18.0

[a] Per-province breakdown provided in Table 3. Mean MPIW and mean σ are heavily influenced by Hebei province (MPIW = 411.8, mean $\sigma = 105.0$), where GP numerical instability occurs under extreme extrapolation; excluding Hebei, the remaining five provinces yield mean MPIW = 2.35 and mean $\sigma = 0.60$.

To evaluate prediction for completely unseen varieties, we conducted leave-one-variety-out (LOVO) validation: all samples of each variety were held out in turn (42 folds), with target encoding and feature scaling recomputed from the training varieties only. For held-out varieties, the TE encoding collapses to the global mean, and the model predicts SGR from spatial and covariate features alone. The Heteroscedastic GP achieves LOVO $R^2 = 0.318$, RMSE $= 0.355$ (log-scale) (Table 5), above the constant baseline and comparable to SVR ($R^2 = 0.31$, Table 2), confirming that spatial and covariate features carry predictive signal without variety information.

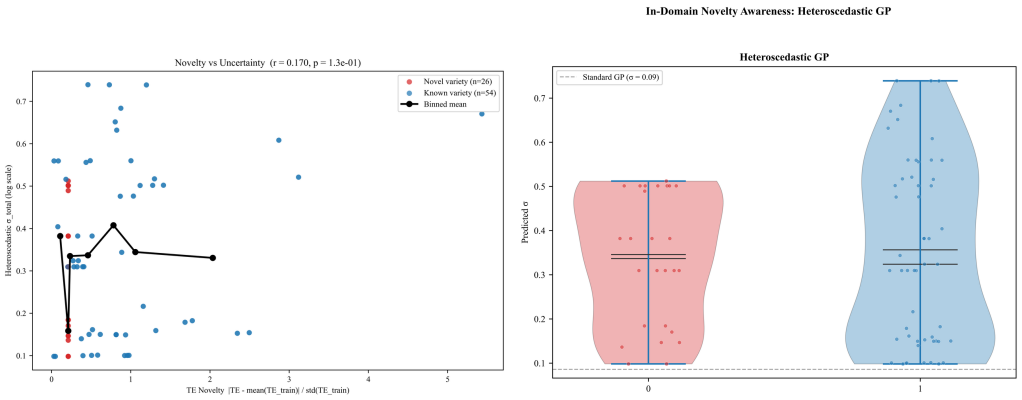
The heteroscedastic σ does *not* increase with variety novelty within the training domain (Table 5). Within LOOCV, single-sample varieties show lower mean σ (0.312) than known varieties (0.369), with narrower intervals (MPIW $= 1.22$ vs. 1.45). Under LOVO, mean σ decreases further (0.302) and coverage drops below nominal (PICP $= 0.788$). This is the opposite of the monotonic novelty–uncertainty relationship that the model design anticipates: the input-dependent noise model $\sigma^2(\mathbf{x}) = \text{softplus}(\mathbf{w}^\top \mathbf{x}^{\text{cov}} + b)$, trained on covariate features alone, does not encode variety novelty within the training domain. The σ escalation observed under cross-province holdout (Section 3.6) is therefore likely driven by environmental and spatial extrapolation rather than varietal unfamiliarity.

These results delineate the model's current capability boundary. For varieties with training representatives, point predictions and coverage are calibrated. For unseen varieties within the same provincial pool, point predictions remain above baseline, but σ does not provide a reliable novelty signal. For cross-province extrapolation, σ escalation is substantial, but intervals become impractically wide. These limitations are discussed further in Section 4.4.

3.6. Variety Novelty and Predictive Uncertainty

Figures 3 and 4 show the heteroscedastic model's two-phase behavior. For **in-domain** predictions (LOOCV, Figure 3), σ averages 0.350 (log-scale) with PICP = 0.95—the model is adequately calibrated for within-distribution samples. Decomposing this total: the posterior component ($\sigma_{\text{posterior}}$, mean = 0.261) and the input-dependent noise component (σ_{noise}, mean = 0.252) contribute roughly equally, indicating that both sparsity in the training input space and covariate-driven observation variance play material roles in predictive uncertainty. For **out-of-domain** predictions (leave-one-province-out, where 100% of test varieties are absent from training, Figure 4), σ increases substantially—often by more than one order of magnitude—while PICP remains close to 1.00. However, cross-province prediction accuracy is poor (all six provinces $R^2 < 0$, with mean $R^2 = -2.29$; per-province breakdown in Table 3), consistent with the dual challenge of geographic and varietal extrapolation at $n = 80$. The σ escalation observed under cross-province holdout is consistent with the expectation that geographic extrapolation inflates predictive variance, though the wide intervals suggest that the magnitude may overstate true uncertainty.

This behavior is triggered by environmental and spatial extrapolation, not by varietal unfamiliarity alone (see Section 3.5): when a test sample originates from a geographically distant province, its spatial coordinates, climate, and soil features deviate from the training distribution, and the noise model responds with substantially elevated σ. The Standard GP, by contrast, assigns a near-constant $\sigma \approx 0.09$ regardless of novelty.



(**a**) TE encoding deviation vs. σ (**b**) Novel vs. known variety σ distribution

Figure 3. Within-domain variety novelty and predictive uncertainty. (**a**) Leave-one-out TE encoding deviation from the training mean versus heteroscedastic σ (log-scale), with a binned trend line (Pearson $r = 0.17$, $p = 0.13$). (**b**) Violin plot with overlaid individual-sample scatter of σ for novel varieties ($n = 26$, single-sample, no training peer) versus known varieties ($n = 54$, multi-sample), Heteroscedastic GP (t-test $p = 0.63$). The dashed line marks the Standard GP's constant σ level ($\sigma \approx 0.09$), which is near-identical regardless of novelty status.

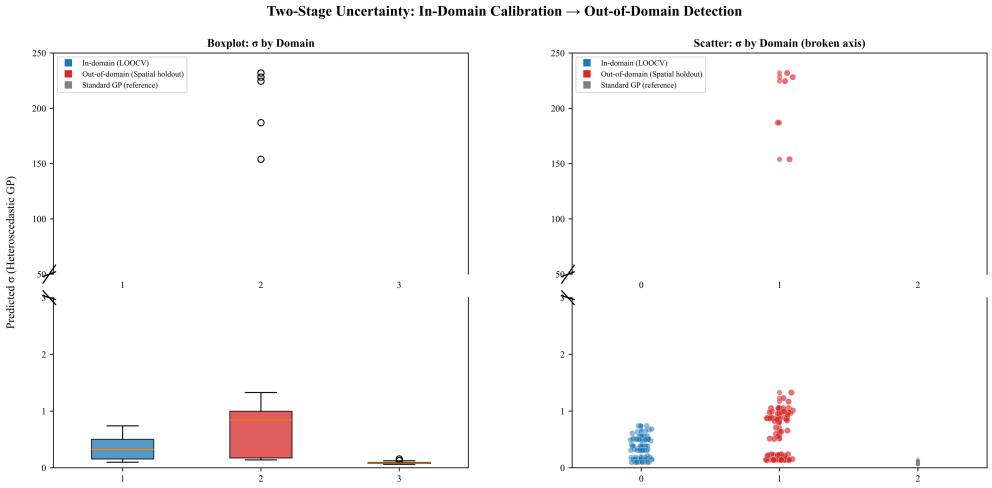


Figure 4. In-domain versus out-of-domain σ (log-scale) contrast with broken y-axis. (**Left**): boxplot comparison of predictive σ for in-domain (LOOCV), out-of-domain (spatial holdout), and Standard GP reference. The y-axis is split at $\sigma = 3$ to compress the empty span between the tightly clustered in-domain values and the dispersed out-of-domain extremes. (**Right**): scatter view of individual-sample σ values across the three groups, with the same broken-axis treatment. The Heteroscedastic GP exhibits a two-stage behavior—calibrated within-domain, substantially elevated out-of-domain—while the Standard GP remains flat.

As a preliminary indication based on the current dataset, lower σ values tend to identify predictions well-supported by similar training samples; higher σ values tend to identify predictions that warrant field verification. No quantitative σ threshold has been validated for operational screening; the illustrative examples below are intended

to demonstrate the qualitative relationship between σ and predictive reliability, not to prescribe decision rules.

3.7. Application Cases

The three cases were selected by deterministic rules from the full set of 80 LOOCV predictions: Case A = the sample where the fixed coefficient 0.9 most overestimates the model prediction; Case B = the sample where 0.9 most underestimates the model prediction (with Case A excluded); Case C = the sample with the highest model σ (with Cases A and B excluded). All 80 LOOCV predictions, σ values, and errors are listed in Table A1.

Table 6 presents three scenarios comparing the fixed 0.9 coefficient with the Heteroscedastic GP. Case A (Chunyu 1819 from Heilongjiang, actual SGR = 0.96) illustrates a near-unit SGR scenario: the fixed 0.9 coefficient is only 6% off, yet the model's prediction (0.70, $\sigma = 0.42$) shows moderately elevated uncertainty. Case B (Quanyou E153 from Sichuan, actual SGR = 2.95) shows the model capturing an extreme value: the fixed coefficient underestimates SGR by 69%, while the model predicts 3.78 ($\sigma = 1.06$)—overcorrecting upward, but with the highest σ among all 80 samples, consistent with the extreme SGR value and suggesting that the prediction warrants caution. Case C (Yanfeng from Hebei, actual SGR = 2.10) demonstrates discriminatory σ: the model predicts 2.14 (2% error versus 57% for the fixed coefficient) yet $\sigma = 0.83$ remains elevated, indicating that the prediction warrants caution despite the accurate point estimate. Approximately 50% of samples in this dataset exhibit relatively low σ, suggesting that model-based SGR estimates could serve as a screening tool for field-level estimation pending external validation.

Table 6. Application cases: fixed coefficient (0.9) versus Heteroscedastic GP predictions. Model σ values are reported on a log scale.

Case	Variety	Actual SGR	Fixed 0.9	Error	Model Pred.	Model σ
A	Chunyu 1819	0.96	0.90	−6%	0.70	0.42
B	Quanyou E153	2.95	0.90	**−69%**	3.78	1.06
C	Yanfeng	2.10	0.90	**−57%**	2.14	0.83

Bold errors indicate severe underestimation by the fixed 0.9 coefficient.

3.8. Sample Efficiency: How Many Data Are Enough?

A practical concern for deployment is whether the model requires the full 80 samples to deliver useful predictions. Field data collection is expensive, and agricultural datasets are often smaller than this. To assess sample efficiency, we conducted a learning curve analysis: for each sample size $n \in \{10, 20, 30, 40, 50, 60, 70, 80\}$, we drew 20 stratified random subsets (province-balanced) and evaluated all models via LOOCV on each subset.

Figure 5 shows the results. At $n = 10$, all models yield $R^2 \approx 0$. The Heteroscedastic GP achieves PICP $= 0.72 \pm 0.12$ at this sample size, indicating that the noise model begins to express caution before the mean function is well-constrained, though calibration remains below the nominal 0.95 target. The Heteroscedastic GP overtakes Random Forest in point-prediction accuracy at approximately $n = 50$ (HGP $R^2 = 0.381 \pm 0.147$ versus RF $R^2 = 0.268 \pm 0.102$), substantially later than ensemble baselines that benefit from empirical averaging at small n. At $n = 20$, the cross-draw variance remains high (HGP R^2 ranging from -0.50 to 0.41), confirming that variety coverage—not merely sample count—governs performance at small n and that 20 random draws are needed for stable variance estimates at these sample sizes. The GP-based methods follow different learning trajectories from the ensemble baselines: Random Forest improves through empirical averaging, whereas the Structured GP kernel compensates for data scarcity with agronomic prior knowledge. The

heteroscedastic noise model achieves stable uncertainty calibration (PICP ≥ 0.85) at $n \geq 40$, and approaches the nominal level (PICP ≥ 0.93) at $n \geq 60$.

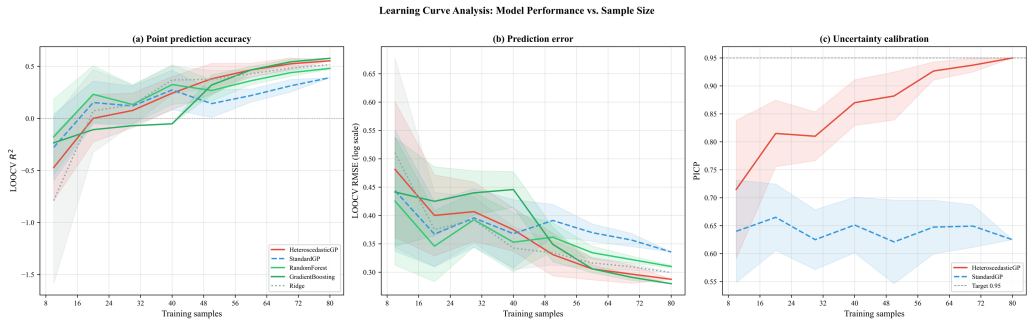


Figure 5. Learning curve analysis under stratified subsampling (20 random draws per sample size). (**Left**) R^2 versus training-set size. (**Centre**) RMSE versus training-set size. (**Right**) PICP for GP-based models. Shaded bands indicate ± 1 standard deviation across draws. The Heteroscedastic GP converges faster than ensemble baselines, and its uncertainty calibration (PICP) approaches the nominal level at $n \geq 60$.

With 20 random draws per sample size, the variance estimates are more reliable than in the earlier 5-draw analysis. The Structured GP framework may be informative with fewer than 80 samples. A pilot study could begin with 40–50 strategically collected samples spanning the main variety–environment combinations and obtain reasonably calibrated uncertainty estimates (PICP ≥ 0.85), with point predictions improving as the sample grows. These sample-size thresholds remain provisional, as they are conditional on the diversity of variety–environment combinations in the training sample rather than on sample count alone. This sample efficiency follows from the GP formalism: the Bayesian prior encoded in the kernel design compensates for data scarcity with domain structure, unlike deep learning approaches that need orders of magnitude more data to avoid overfitting [11].

4. Discussion

4.1. Agronomic Motivation and Predictive Value of the Additive Kernel Structure

The structured multi-kernel GP is motivated by the agronomic G+E+P framework as a modeling heuristic. The structured kernel design does more than boost prediction accuracy—it makes the model's reasoning traceable. The variety kernel captures systematic SGR differences across cultivars that reflect known patterns in rice biology: indica varieties tend toward higher SGR than japonica types due to differences in tillering architecture and harvest index, and modern semi-dwarf cultivars bred during the Green Revolution partition more dry matter to grain than traditional tall varieties, lowering their SGR relative to landraces. These varietal-level differences are substantial and have been documented in multi-site field surveys across Asian rice systems. SGR is not a species-level constant—it is a plastic trait under strong varietal control—and as new plant-type cultivars replace older ones across China's rice-growing regions, the bias from fixed conversion coefficients will keep widening.

The spatial kernel captures more than latitudinal or longitudinal trends. It captures the joint spatial structure of temperature, solar radiation, water availability, soil type, and fertilization regimes—environmental covariates that co-vary across geography. This spatial pattern is agronomically interpretable: latitudinal gradients in temperature and solar radiation drive differences in grain-filling duration, which in turn affect the proportion of total biomass allocated to grain versus straw; single-cropping rice regions at higher latitudes (e.g., Heilongjiang) exhibit systematically different SGR profiles from double-cropping systems at lower latitudes (e.g., Hunan). This aligns with recent work applying spatial

kernels to grain yield prediction [14]. There is a practical complementarity here: spatial interpolation methods characterize the structure of known environmental variables well, while machine learning can encode environmental effects implicitly through coordinates when covariates are missing.

The covariate kernel's contribution is modest. Most SGR-relevant phenotypic information is already carried indirectly by the variety kernel, and within-variety phenotype expression is relatively stable, leaving limited marginal signal for an explicit high-dimensional covariate kernel to capture. The finding itself is useful: under small-sample conditions, over-parameterized kernels risk contributing more noise than signal. For crop modeling with limited field data, the practical rule is to prioritize a few traits with clear agronomic justification and constrain ARD kernels with regularization or Bayesian priors. Future work could fuse variety and covariate kernels to resolve genotype × environment (G × E) interaction effects [21].

The additive kernel framework should transfer to SGR prediction for other crops, though the contribution profile will shift. Maize exhibits lower within-variety heterogeneity, so the variety kernel would likely contribute less. Wheat shows stronger environmental responsiveness under drought, so the covariate kernel may matter more. Cross-crop comparisons using this framework could uncover agronomic regularities while building toward a unified method for straw resource accounting.

4.2. Practical Value of Uncertainty-Aware Estimation

The Heteroscedastic GP's most practically useful output is the per-sample σ (log-scale)—an indicator of predictive reliability that reflects both the density of training data in the local input space and the typicality of the sample's phenotype profile.

In practice, the per-sample σ supports a tiered workflow: predictions with low σ (within the range observed for well-represented training samples) can be used directly in resource accounting or breeding decisions with reasonable confidence, while predictions with elevated σ—for instance, those exceeding the 90th percentile of the LOOCV σ distribution ($\approx$0.61 in log-scale)—indicate that the prediction is likely less reliable and the sample warrants priority field verification before use. For county-level straw resource surveys and breeding program variety screening, this means that the model can distinguish regions or varieties where predictions are likely to be adequate from those where targeted field measurement is still needed. For agronomists and extension workers, the model output is interpretable: a high σ flags a specific field or variety for which the model lacks sufficient local or variety-specific information, guiding data-collection priorities in iterative survey design.

Straw resource accounting. National straw assessments currently apply a single SGR coefficient to all rice-producing regions. Our application cases (Table 6) illustrate that this produces substantial errors where low- or high-SGR varieties dominate, and even for a near-unit SGR variety the fixed coefficient still deviates noticeably. These are retrospective illustrations from LOOCV predictions within the nine sampled counties; they demonstrate qualitative patterns, not prospectively validated screening rules. With external validation on independent datasets, the model could augment existing data pipelines—variety composition, location, and environmental covariates could return an SGR estimate with per-sample σ—but at present the framework remains a proof-of-concept whose operational value is contingent on validation across multiple seasons and broader geographic coverage.

Biomass energy planning. Siting biomass power plants and biofuel facilities depends on reliable feedstock forecasts within a collection radius. Overestimating straw supply strands assets; underestimating deters investment. If validated at the field-to-county scale, the model would deliver spatially resolved feedstock estimates with confidence bands,

letting investors build supply scenarios under different risk tolerances. Combined with robust optimization, these predictions and intervals may inform plant siting and collection radius design.

Agricultural carbon accounting. National greenhouse gas inventories submitted under the UNFCCC must report crop residue carbon flows. IPCC good practice guidance encourages reporting uncertainty ranges alongside emission factor point estimates. The variety kernel quantifies differences in dry matter partitioning across cultivars—something a fixed SGR coefficient cannot do. If externally validated, the per-sample σ supplies the confidence intervals required for Tier 2–3 reporting, enabling formal error propagation consistent with IPCC inventory guidance [1].

Smart breeding and variety deployment. Because variety identity is the largest single contributor among the modeled factors, differentiated variety deployment strategies could target grain yield, straw availability, and carbon sequestration simultaneously, feeding into multi-objective breeding programs. The variety kernel's learned lengthscale offers an interpretable summary of SGR diversity: a short lengthscale indicates that even small changes in TE correspond to substantial SGR differences between varieties, while a long lengthscale suggests that SGR is relatively uniform across a variety group. For crop residue policy, this means that regions dominated by short-lengthscale variety groups—e.g., regions cultivating both traditional tall and modern semi-dwarf rice—need variety-specific coefficients more urgently than regions dominated by genetically homogeneous cultivars.

Comparison with alternative uncertainty methods. RF with tree-level standard deviation provides prediction intervals of moderate coverage; however, this uncertainty estimate is constant-width across all predictions and cannot distinguish well-supported predictions on common varieties from extrapolations to novel ones. The Heteroscedastic GP's input-dependent σ supplies this discrimination without sacrificing point accuracy. Among prediction interval methods, GP-based intervals provide analytic uncertainty estimates without post-hoc calibration [28,29], and the σ escalation observed under cross-province holdout is consistent with the interpretation that predictive uncertainty magnitude can serve as a model-extrapolation warning.

Relationship to linear mixed models. The Structured GP kernel has a formal connection to linear mixed-effects models. The variety kernel with target encoding approximates a random intercept by variety, while the spatial kernel (Matérn-3/2 over Haversine distance) generalizes a random intercept by county to a continuous spatial field that allows similarity between geographically proximate counties rather than treating them as exchangeable categories. The key difference is that the GP formulation (i) captures nonlinear interactions through kernel composition, (ii) supports input-dependent noise via the heteroscedastic likelihood, and (iii) provides analytic predictive distributions without assuming homoskedasticity. The mixed-effects baseline confirms that grouped random effects capture substantial SGR variation, but its homoskedastic intervals are too wide to discriminate reliable from unreliable predictions—precisely the gap the Heteroscedastic GP fills.

4.3. Comparison with Existing Machine Learning Crop Modeling Methods

Existing ML approaches to crop modeling fall into three broad groups. Tree-based ensembles (RF, Gradient Boosting) deliver strong point predictions but lack native uncertainty quantification and interpretable variance structure. Deep learning handles multimodal data fusion well but struggles with the small sample sizes typical of field-collected agricultural data [11]. Traditional statistical models are interpretable but cannot capture nonlinear, high-dimensional feature relationships.

The GP framework here addresses these constraints in small-data settings: sample efficiency, uncertainty quantification, and kernel interpretability. Most ML studies on

agricultural straw estimation predict total biomass directly; this framework targets the SGR conversion coefficient instead, which adapts more flexibly to different spatial and temporal scales of resource accounting. While existing work tends to report only point-prediction accuracy, this framework provides adequately calibrated prediction intervals and decomposable uncertainty sources. The structured kernels motivated by variety-associated, spatial, and covariate-driven variation yield conclusions with agronomic interpretability—more practically useful than a generic feature importance ranking. For deterministic optimization models, the uncertainty quantification supports robust optimization of straw resource allocation under data scarcity and varietal turnover.

The approach fits a broader shift in crop ML modeling: embedding agronomic prior knowledge into model architecture to balance predictive performance with interpretability. It methodologically complements existing agronomic kernel design work [14,21]. The framework may be adaptable to other crops with appropriate kernel and hyperparameter adjustments, accommodating the different varietal and environmental characteristics of maize, wheat, and other staple crops, though the kernel contribution profile and predictive performance would need to be empirically re-evaluated for each crop.

4.4. Limitations and Future Research Directions

Several limitations qualify these findings.

Sample size and covariate kernel utility. At $n = 80$, the covariate kernel contributes modestly ($\Delta R^2 = -0.013$) and the spatial kernel's marginal contribution ($\Delta R^2 = +0.005$) suggests that, at this sample size, not all kernel components independently improve prediction. Over-parameterized kernels in small-sample settings risk adding noise, not signal—a caution for agricultural ML more broadly. Models working with limited field data should favor parsimonious architectures over exhaustive feature parameterization. As active learning cycles add samples, the covariate kernel may become more informative once enough observations per variety–environment combination accumulate. However, GP training differs fundamentally from models that select parameters by minimizing holdout error: maximizing the marginal likelihood automatically balances data fit against model complexity, a property formalized as an "automatic Occam's razor" [13] (§5.4) . Empirically, the 80-fold LOOCV provides the strictest possible out-of-sample test—if the model were substantially overfitting, R^2 would approach or fall below zero, not reach 0.541.

Variety representation. The variety kernel collapses varietal similarity onto a single dimension via target encoding. This is statistically pragmatic—the learning curve analysis shows it already carries useful signal at modest sample sizes—but richer representations are possible. An inherent limitation of TE is that varieties represented by a single sample in the training fold receive the global mean encoding, providing no variety-specific information. With 26 of 42 varieties (62%) represented by a single sample in the full dataset, the TE signal for most varieties relies on the smoothing prior rather than empirical variety-level statistics. As data accumulate, replacing TE with a Linear Model of Coregionalization (LMC) that learns a 2–3-dimensional latent variety representation [30] would capture more nuanced varietal relationships and genotype–environment correlation structures [31]. For deployment scenarios where novel varieties constitute a large fraction of predictions—for instance, in hybrid rice systems with rapid cultivar turnover—the TE approach would provide limited variety-specific information. In such settings, replacing TE with a pedigree-based or genomic relationship matrix would offer a more structurally informed variety representation that does not depend on having observed the specific variety in training.

Known-variety vs. novel-variety prediction. The validation results (Section 3.5) reveal a counterintuitive pattern: within LOOCV, known-variety samples ($n = 54$) and single-sample-variety samples ($n = 26$) achieve comparable within-group R^2 (0.480 for

both subsets), but the model assigns higher mean σ to known-variety samples (0.369 vs. 0.312) and wider prediction intervals (MPIW = 1.45 vs. 1.22). Under LOVO, mean σ decreases further (0.302) and coverage drops below nominal (PICP = 0.788)—the opposite of what a novelty-detection mechanism would produce. This indicates that the current heteroscedastic noise model, trained on phenotype features alone, does not reliably encode variety novelty; the elevated σ observed in cross-province experiments (Section 3.6) may be driven by environmental and spatial extrapolation rather than variety novelty. All application claims in this paper therefore refer to within-distribution prediction; operational use for novel varieties requires external validation on independent variety collections.

Geographic coverage and sample dependence. The six surveyed provinces cover only part of China's rice production region. Spatial holdout experiments, in which each province is held out in turn, confirm that cross-province generalization is currently poor (all six provinces $R^2 < 0$, with mean $R^2 = -2.29$; per-province breakdown in Table 3, with all test varieties being novel to the training set). This reflects the dual challenge of geographic extrapolation and variety novelty combined, and positions the current LOOCV evaluation as internal validation. External validation on independent datasets is necessary before national-scale use. Additionally, the dataset represents a single harvest season (2025); year-to-year variation in genotype–environment interaction is not captured.

Clustered sampling design. The 80 samples are drawn from only nine counties (mean 8.9 samples per county), with five villages per county and two fields per village. Samples from the same county share local climate, soil, and management regimes, and are not independent spatial replicates. Standard sample-level LOOCV may therefore yield optimistic performance estimates because held-out test samples are not fully independent of spatially or environmentally correlated training samples from the same county.

To assess the impact of this clustering, we conducted a leave-one-county-out (LOCO) experiment: each of the nine counties was held out in turn, with target encoding and feature scaling recomputed from the training counties only. The Heteroscedastic GP achieves LOCO $R^2 \approx 0.006$ (see Table 3), confirming that the model cannot generalize to entirely unseen counties at the current sample size and county coverage. Mean σ reaches 24.4 (log-scale)—two orders of magnitude above the LOOCV mean—driven largely by one county where all test varieties are novel and spatial extrapolation is extreme. These results underscore that the LOOCV performance reported in this paper is conditional on the specific set of counties sampled, and that county-level generalization would require either broader geographic coverage or a fundamentally larger dataset.

Crop specificity. The model is specific to rice. Extending the framework to wheat and maize would test whether the G+E+P-motivated additive kernel structure holds across species with different varietal architectures and environmental responsiveness.

Unmeasured management covariates. Irrigation regime (continuous flooding vs. alternate wetting–drying), fertilizer application rates (N, P, K), and sowing or transplanting dates were not recorded in the current survey. These management practices are known to influence SGR through effects on tillering, biomass partitioning, and grain filling. Their omission may introduce residual confounding into the phenotype and spatial kernel components—for example, coordinated management regimes within a county could inflate apparent spatial similarity. Incorporating these covariates in future surveys would allow the covariate kernel to be further decomposed into management and environmental sub-components, improving both predictive accuracy and mechanistic interpretability.

A promising future direction is closing the loop: use σ to prioritize samples for field verification in an iterative survey design. Prioritize high-σ counties for the next field campaign; feed new samples back into the model; re-estimate σ; repeat. This "predict–sample–update" cycle uses σ as a prioritization heuristic—it indicates where new data

would be most informative. Two further directions are extending to wheat and maize, and incorporating management covariates (irrigation, fertilizer, sowing date) to split the covariate kernel into management and environmental sub-components.

5. Conclusions

We have shown that a structured multi-kernel GP with heteroscedastic noise addresses two persistent failures in SGR estimation: the use of fixed coefficients that ignore structured variation, and the absence of calibrated uncertainty in ML-based crop trait models. All evaluations reported here are based on a single dataset of 80 rice samples collected across six Chinese provinces during the 2025 growing season. Independent external validation on data from different years, geographic regions, or management conditions has not been performed and is necessary before these findings can be generalized beyond the nine sampled counties. The kernel architecture, motivated by the agronomic G+E+P framework as a modeling heuristic, renders the GP's learned covariance structure interpretable in breeding and agronomic terms. The heteroscedastic noise model improves uncertainty coverage relative to the Standard GP (calibration ratio 0.68 $\rightarrow$ 1.63, PICP 0.650 $\rightarrow$ 0.95; the calibration ratio of 1.63 indicates a conservative bias—the predicted σ overestimates the mean absolute error by approximately 63%, compared with a theoretical ratio of $\approx$1.25 for well-calibrated Gaussian errors) and the structured multi-kernel design, by incorporating variety information, substantially outperforms a covariate-only RBF GP ($\Delta R^2 = 0.103$). Under a nested internal cross-validation protocol that prevents information leakage, the model shows that Structured GP prediction is viable at $n = 80$ within the nine sampled counties. For varieties with training representatives, coverage is near the nominal level; under cross-province extrapolation, σ escalation is substantial; within-domain variety novelty alone does not trigger elevated uncertainty, and prediction intervals are under-covered for unseen varieties (PICP = 0.788 in LOVO). Leave-one-county-out validation ($R^2 \approx 0$, Table 3) confirms that the model does not yet generalize to unsampled counties; external validation on independent datasets is required before operational deployment. On 80 rice samples across six Chinese provinces, the model provides a proof-of-concept for variety-informed SGR prediction under internal validation, with input-dependent predictive variance that can help prioritize samples for targeted field verification. The σ escalation under geographic extrapolation should be interpreted as a model-extrapolation warning rather than a validated out-of-distribution detector; external validation across multiple seasons and broader geographic coverage is needed to establish its operational reliability.

Author Contributions: Conceptualization, M.H., A.Z., S.X. and J.Z.; methodology, M.H., S.X. and J.Z.; software, M.H. and Z.W.; validation, M.H. and J.L.; formal analysis, M.H. and Z.W.; investigation, A.Z., J.L., X.L. and Q.W.; resources, J.Z.; data curation, A.Z., J.L., X.L. and Q.W.; writing—original draft preparation, M.H.; writing—review and editing, M.H., A.Z., Z.W., J.L., X.L., Q.W., S.X. and J.Z.; visualization, M.H.; supervision, S.X. and J.Z.; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program, grant number 2025YFD1700304.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Acknowledgments: The authors thank the field sampling teams at the Chinese Academy of Agricultural Sciences for their assistance with data collection.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARD	Automatic Relevance Determination
GB	Gradient Boosting
GP	Gaussian Process
LOOCV	Leave-One-Out Cross-Validation
PICP	Prediction Interval Coverage Probability
MPIW	Mean Prediction Interval Width
RBF	Radial Basis Function
RF	Random Forest
SGR	Straw-to-Grain Ratio
SVR	Support Vector Regression
TE	Target Encoding

Appendix A. Nested Target Encoding Algorithm

Table A1 (available in the Chinese manuscript and upon request from the corresponding author) lists the complete LOOCV predictions ($n = 80$) from the Heteroscedastic GP model, sorted by descending model σ. All predictions were generated under the nested cross-validation protocol described in Section 2.3. Cases A–C from Table 6 are highlighted.

The following algorithm specifies the complete nested target encoding (TE) procedure within leave-one-out cross-validation, as referenced in Section 2.3. The smoothing parameter k is selected once via flat 5-fold cross-validation (Section 2.3) and held fixed during LOOCV.

Algorithm A1: Nested TE within LOOCV (fixed k, pre-selected via flat 5-fold CV

Input: dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^{n}$ with variety labels v_i, pre-selected k
Output: LOOCV predictions $\{\hat{y}_i\}_{i=1}^{n}$

for $i = 1$ **to** n **do**
  $\mathcal{D}_{tr} \leftarrow \mathcal{D} \setminus \{i\}$ // hold out sample i
  $\bar{y}^{(tr)} \leftarrow \frac{1}{n-1} \sum_{j \in \mathcal{D}_{tr}} y_j$ // training-set global mean
  for each variety v in $\mathcal{D}_{tr}$: compute $n_v^{(tr)}, \bar{y}_v^{(tr)}$
  for each training row $j \in \mathcal{D}_{tr}$ **do**
    $n_{v_j}^{(-j)} \leftarrow n_{v_j}^{(tr)} - 1$ // LOO: exclude row j
    if $n_{v_j}^{(-j)} > 0$ **then**
      $\bar{y}_{v_j}^{(-j)} \leftarrow (n_{v_j}^{(tr)} \bar{y}_{v_j}^{(tr)} - y_j) \,/\, n_{v_j}^{(-j)}$
    else $\bar{y}_{v_j}^{(-j)} \leftarrow \bar{y}^{(tr)}$
    $\mathrm{TE}_j^{(tr)} \leftarrow (n_{v_j}^{(-j)} \bar{y}_{v_j}^{(-j)} + k\bar{y}^{(tr)}) \,/\, (n_{v_j}^{(-j)} + k)$
  end for
  for the test sample i **do**
    $\mathrm{TE}_i^{(te)} \leftarrow (n_{v_i}^{(tr)} \bar{y}_{v_i}^{(tr)} + k\bar{y}^{(tr)}) \,/\, (n_{v_i}^{(tr)} + k)$
    if $n_{v_i}^{(tr)} = 0$ **then** $\mathrm{TE}_i^{(te)} \leftarrow \bar{y}^{(tr)}$ // unseen variety
  end for
  Fit GP($K_{\mathrm{var}} + K_{\mathrm{spat}} + K_{\mathrm{cov}}$) on $\{(\mathrm{TE}_j^{(tr)}, \mathbf{x}_j, y_j)\}_{j \in \mathcal{D}_{tr}}$
  Compute predictive mean $\hat{y}_i$ at encoded test point $\mathrm{TE}_i^{(te)}$
end for
return $\{\hat{y}_i\}_{i=1}^{n}$

Table A1 provides the county-level sample composition referenced in Section 2.1.

Table A1. County-level sample composition.

Province	County	Samples	Varieties	SGR Range	Cropping System
Sichuan	Xindu	10	5	0.95–3.94	single, rice–wheat rotation
Hubei	Qichun	10	2	0.75–1.86	single
Heilongjiang	Baoqing	10	10	0.40–0.96	single
Heilongjiang	Dongshan	5	5	0.75–0.92	single
Heilongjiang	Luobei	5	5	0.84–1.10	single
Hebei	Changli	10	2	0.85–2.83	single
Zhejiang	Wuxing	10	8	1.27–2.27	single, rice–wheat rotation
Zhejiang	Nanxun	10	7	0.91–1.94	single, rice–wheat rotation
Hunan	Xiangyin	10	5	1.06–1.84	double

References

1. Fang, Y.R.; Zhang, S.; Zhou, Z.; Shi, W.; Xie, G.H. Sustainable development in China: Valuation of bioenergy potential and CO_2 reduction from crop straw. *Appl. Energy* **2022**, *322*, 119439.
2. Zhang, J.; Wei, J.; Guo, C.-L.; Tang, Q.; Guo, H. The spatial distribution characteristics of the biomass residual potential in China. *J. Environ. Manag.* **2023**, *338*, 117777.
3. Tang, Z.; Zhang, X.; Chen, R.; Ge, C.; Tang, J.; Du, Y.; Jiang, P.; Fang, X.; Zheng, H.; Zhang, C. A comprehensive assessment of rice straw returning in China based on life cycle assessment method: Implications on soil, crops, and environment. *Agriculture* **2024**, *14*, 972.
4. Yang, C.-W.; Xing, F.; Zhu, J.-C.; Li, R.-H.; Zhang, Z.-Q. Temporal and spatial distribution, utilization status, and carbon emission reduction potential of straw resources in China. *Huanjing Kexue* **2023**, *44*, 1149–1162.
5. Xie, G.H.; Han, D.; Wang, X.; Lü, R. Harvest index and straw coefficient of cereal crops in China. *J. China Agric. Univ.* **2011**, *16*, 1–8.
6. Ludemann, C.I.; Hijbeek, R.; van Loon, M.P.; Murrell, T.S.; Dobermann, A.; van Ittersum, M.K. *Field Experimental Data of Crop Yields, Nutrient Concentrations, and Harvest Indices From Around the World*; Dryad Digital Repository: Davis, CA, USA, 2024.
7. Wang, Y.; Zhang, Q.; Yu, F.; Zhang, N.; Zhang, X.; Li, Y.; Wang, M.; Zhang, J. Progress in research on deep learning-based crop yield prediction. *Agronomy* **2024**, *14*, 2264.
8. Liu, F.; Su, L.; Luo, P.; Tao, W.; Wang, Q.; Deng, M. Prediction models of growth characteristics and yield for Chinese winter wheat based on machine learning. *Agronomy* **2024**, *14*, 839.
9. Wei, L.; Yang, H.; Niu, Y.; Zhang, Y.; Xu, L.; Chai, X. Wheat biomass, yield, and straw-grain ratio estimation from multi-temporal UAV-based RGB and multispectral images. *Biosyst. Eng.* **2023**, *234*, 187–205.
10. Allu, A.R.; Mesapam, S. Crop health assessment from predicted AGB and NPK derived from UAV spectral indices and machine learning techniques. *Agronomy* **2025**, *15*, 2059.
11. Oikonomidis, A.; Catal, C.; Kassahun, A. Deep learning for crop yield prediction: A systematic literature review. *N. Z. J. Crop Hortic. Sci.* **2023**, *51*, 1–26.
12. Aslan, M.F.; Sabanci, K.; Aslan, B. Artificial intelligence techniques in crop yield estimation based on Sentinel-2 data: A comprehensive survey. *Sustainability* **2024**, *16*, 8277.
13. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
14. Hu, X.; Carver, B.; El-Kassaby, Y.A.; Zhu, L.; Chen, C. Weighted kernels improve multi-environment genomic prediction. *Heredity* **2023**, *130*, 82–91.
15. *NY/T 3498–2019*; Determination of Elemental Composition of Crop Straw. China Agriculture Press: Beijing, China, 2019.
16. *NY/T 3494–2019*; Determination of Lignocellulosic Composition of Crop Straw. China Agriculture Press: Beijing, China, 2019.
17. *GB/T 28731–2012*; *Determination of Ash Content of Solid Biofuels.* Standards Press of China: Beijing, China, 2012.
18. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL* **2021**, *7*, 217–240.
19. Gardner, J.R.; Pleiss, G.; Bindel, D.; Weinberger, K.Q.; Wilson, A.G. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In Proceedings of the Neural Information Processing Systems 31 (NeurIPS), Montréal, QC, Canada, 3–8 December 2018 ; pp. 7587–7597.
20. Lu, X.; Boukouvalas, A.; Hensman, J. Additive Gaussian processes revisited. In Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022; pp. 14358–14383.
21. Sousa, M.B.; Cuevas, J.; de Oliveira Couto, E.G.; Pérez-Rodríguez, P.; Jarquín, D.; Fritsche-Neto, R.; Burgueño, J.; Crossa, J. Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3 Genes Genomes Genet.* **2017**, *7*, 1995–2014.
22. Kersting, K.; Plagemann, C.; Pfaff, P.; Burgard, W. Most likely heteroscedastic Gaussian process regression. In Proceedings of the 24th International Conference on Machine Learning (ICML), Corvallis, OR, USA, 20–24 June 2007; pp. 393–400.

23. Goldberg, P.W.; Williams, C.K.I.; Bishop, C.M. Regression with input-dependent noise: A Gaussian process treatment. In Proceedings of the Neural Information Processing Systems 10 (NIPS), Denver, CO, USA, 1–6 December 1997; pp. 493–499.
24. Faza, G.A.; Loka, N.R.B.S.; Shariatmadar, K.; Hallez, H.; Moens, D. Most likely heteroscedastic Gaussian process via kernel smoothing. *Knowl.-Based Syst.* **2025**, *328*, 114202.
25. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
26. Pernot, P. On the good reliability of an interval-based metric to validate prediction uncertainty for machine learning regression tasks. *arXiv* **2024**, arXiv:2408.13089.
27. Morales, G.; Sheppard, J.W. Dual accuracy-quality-driven neural network for prediction interval generation. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 2843–2853.
28. De Wolf, N.; De Baets, B.; Waegeman, W. Valid prediction intervals for regression problems. *Artif. Intell. Rev.* **2023**, *56*, 577–613.
29. Psaros, A.F.; Meng, X.; Zou, Z.; Guo, L.; Karniadakis, G.E. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *J. Comput. Phys.* **2023**, *477*, 111902.
30. Álvarez, M.A.; Rosasco, L.; Lawrence, N.D. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.* **2012**, *4*, 195–266.
31. Friedli, L.; Steinert, T.; Wuyts, N.; Guignard, F.; Levy Häner, L.; Pellet, D.; Herrera, J.M.; Ginsbourger, D. Gaussian process modeling with genotype × environment kernels for wheat performance prediction. *arXiv* **2025**, arXiv:2508.17730.