Model averaging for estimation of optimum rate and critical values under model uncertainty

by

Gustavo Adolfo Roa Acosta

B.S., EARTH University, 2017
Master's, UnADM, 2020
M.S., Kansas State University, 2024

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Dr. Christopher I. Vahl

Copyright

Abstract

Fertilizer recommendations rely on fitted yield response models to estimate optimum rates and critical values, but these estimates are sensitive to the assumed functional form. This study evaluated how model uncertainty affects estimates of agronomic optimum nitrogen (N) rate (AONR) and critical soil test values (CSTV) for phosphorus (P), and whether model averaging improves inference. Simulation experiments were conducted using linear plateau, quadratic plateau, quadratic, and Mitscherlich functions as data generating processes. Estimation approaches included single model fits, Akaike information criterion (AIC) weighted model averaging, bootstrap aggregation (bagging), and Bayesian model averaging (BMA). Performance was assessed using bias, variance, and mean squared error (MSE). Two field case studies were used to evaluate practical implications. Single-model estimators performed well only under correct specification but showed large bias under misspecification. When the true response was linear plateau, the Mitscherlich model overestimated AONR by 289 kg N ha^{-1} and produced a MSE of 173,423, compared to 352 for the true model. Model averaging reduced error across approaches, although no single method was uniformly superior. A reduction in (MSE to 4,765 and bias to 55 kg N ha^{-1} was achieved using BMA, while AIC-weighted averaging and bagging also reduced bias and variance relative to single model estimates. In field data, AONR estimates ranged from 121 to 241 kg N ha^{-1} across models, whereas model averaged estimates were more stable (163–188 kg N ha^{-1}). Similarly, CSTV estimates ranged from 12 to 27 mg P kg^{-1} for individual models and were more consistent under model averaging (18–20 mg P kg^{-1}). Model uncertainty is a major source of variability in fertilizer recommendations. Model averaging reduces sensitivity to functional form and provides more stable and reliable estimates of optimal rates and critical values.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I am grateful for the unwavering support of my partner, family, and friends throughout this journey. Their encouragement and understanding have been crucial to my success. I am especially grateful to Dr. Dorivar Ruiz Diaz for encouraging and supporting my pursuit of this degree. I would also like to extend my sincere thanks to Dr. Christopher Vahl for serving as my major advisor and for his guidance and support throughout this project.

I sincerely appreciate the mentorship provided by my committee members, Dr. Trevor Hefley and Dr. Josefina Lacasa, whose expertise and insights have been instrumental in shaping my academic growth. I am also indebted to my fellow graduate students for their support and collaboration throughout this project.

I am deeply thankful to all individuals who have played a role in my academic journey and honored to have had the opportunity to work alongside such dedicated colleagues.

Introduction

Mineral fertilization underpins modern crop production, yet misplaced nutrient rates remain a major source of both yield gaps and environmental losses across a wide range of cropping systems (Cassman et al., 2002; Galloway et al., 2003; Sutton et al., 2011). For mobile nutrients in the soil such as nitrogen (N) and sulfur, excessive applications can lead to leaching and gaseous emissions, whereas under application limits yield and profitability; for relatively immobile nutrients in the soil such as phosphorus (P) and potassium, over accumulation in soil elevates runoff risk while inadequate supply constrains crop growth and nutrient use efficiency (Mallarino & Blackmer, 1992; Lyons et al., 2019). These agronomic and environmental trade-offs motivate decision oriented quantities derived from crop response curves, including optimum fertilizer rates and critical soil test values, that can in principle be generalized across nutrients with appropriate interpretation (Zou et al., 2022).

Within this decision framework, agronomists distinguish between optimum rate and soil test thresholds. For nitrogen, the agronomic optimum nitrogen rate (AONR) is often defined as the smallest fertilizer rate beyond which further additions do not meaningfully increase yield, while the economic optimum nitrogen rate (EONR) occurs where marginal yield gains no longer offset fertilizer costs (Cerrato & Blackmer, 1990; Miguez & Poffenbarger, 2022; Scharf et al., 2005). For less mobile nutrients, recommendation systems frequently target a critical soil test value (CSTV), the soil test level above which additional nutrient application is unlikely to enhance yield, so that management focuses on building and maintaining soil tests near that threshold (Lyons et al., 2021; Mallarino & Blackmer, 1992). Although AONR, EONR, and CSTV differ conceptually (rate based versus soil test based criteria), their estimation is typically grounded in the same underlying response models, with the difference residing in the functional

mapping from the fitted curve to the chosen agronomic or economic metric (Slaton, Pearce, Gatiboni, et al., 2024).

Most fertilizer recommendation studies therefore begin with experiments spanning a range of nutrient supply either a series of fertilizer rates or a gradient of soil test values and then fit a parametric model to describe the relationship between grain yield and nutrient availability (Correndo et al., 2021; Lyons et al., 2019). Across nutrients, a small set of functional forms is used repeatedly, including linear plateau, quadratic plateau, Mitscherlich type exponential, and purely quadratic models, each encoding different assumptions about diminishing marginal returns, yield plateaus, or possible yield decline at high inputs (Cerrato & Blackmer, 1990; Cook & Weisberg, 1982). Comparative analyses across crops show that these models can provide similarly adequate fits to observed data yet imply quite different estimates of optimum rates or critical soil test values, underscoring that recommendation relevant quantities are often more sensitive to functional form choice than simple goodness of fit metrics suggest (Bullock & Bullock, 1994; Dhakal & Lange, 2021; Jaynes, 2011).

The implications of model choice for decision making have been explored from several methodological perspectives. Early work comparing quadratic and plateau type models found that quadratic forms could substantially overestimate both yield and optimum nutrients rates relative to quadratic plateau models in field trials (Bullock & Bullock, 1994; Cerrato & Blackmer, 1990). Subsequent studies used profile likelihood, Monte Carlo, and bootstrap techniques to quantify uncertainty in EONR under individual models, revealing very wide confidence intervals and highlighting that combining model choice and parameter uncertainty yields a broad range of plausible optimum rates (Bachmaier & Gandorfer, 2012; Hernandez & Mulla, 2008; Jaynes, 2011; Nigon et al., 2019). More recent work in phosphorus calibration and

other relatively immobile nutrients has demonstrated analogous sensitivity of CSTV estimates to model choice, with direct consequences for both economic returns and environmental risk (Filippi et al., 2025).

These findings have prompted increasing recognition that treating a single "best" response model as known is rarely justified and that model uncertainty should be explicitly incorporated into fertilizer decision support across nutrients (Burnham & Anderson, 2004; Henke et al., 2007; Miguez & Poffenbarger, 2022). Within the multimodel inference paradigm, all candidate functions are regarded plausible to describe the yield nutrient relationship, and derived quantities such as AONR, EONR, or CSTV can be obtained by aggregating over models rather than conditioning on a single selected form (Burnham & Anderson, 2004; Mac Nally et al., 2018). Information criterion based model averaging, bootstrap aggregation (bagging), and Bayesian model averaging (BMA) offer complementary strategies to implement this idea: Akaike Information Criterion (AIC) based averaging assigns weights to candidate models based on their relative AIC values, giving greater influence to models with stronger empirical support (Matavel et al., 2025; Miguez & Poffenbarger, 2022), bagging combines bootstrap resampling with random model selection to stabilize estimators (Breiman, 1996; Palmero, 2025), and BMA integrates over model space using posterior model probabilities derived from marginal likelihoods (Hoeting et al., 1999; Palmero et al., 2026).

Recent nitrogen response studies have shown that AIC weighted averaging can reduce bias in AORN estimates relative to selecting a single plateau model (Miguez & Poffenbarger, 2022), while bagging provides a flexible ensemble learner that performs well across experimental designs and data qualities (Palmero, 2025). BMA further extends the multimodel framework by combining posterior distributions of the optimum rate across a set of plausible

yield nutrient response models, allowing prior knowledge and model probabilities to inform inference while explicitly accounting for model uncertainty (Palmero et al., 2026). Bayesian optimized experimental designs also demonstrate how BMA can be used not only to analyze fertilizer trials but also to plan experiments that are more informative about optimum rates under competing response functions (Matavel et al., 2025). However, there remains a need for a unified assessment of these model averaging strategies across both mobile and relatively immobile nutrients, treating optimum rates and critical values as model derived estimands and evaluating performance under realistic designs and variance structures. To address this gap, the objective of this study is to evaluate these approaches through simulation experiments and applied case studies. Specifically, we (i) evaluate the impact of model misspecification on the estimation of AONR and CSTV; (ii) compare individual response models with model averaging approaches, including AIC-weighted bootstrap, bagging, and BMA, in terms of bias, variance, and mean squared error; and (iii) assess the practical implications of these methods using nitrogen response data in maize (*Zea mays* L.) and phosphorus response data in soybean (*Glycine max* (L.) Merr.). The proposed framework is directly extendable to other nutrients where yield nutrient response relationships can be represented by the same family of models.

Methodology

Set of candidate models

Several mathematical models can be used to describe grain yield response to nutrient supply. In this study we considered a finite set of parametric models for the response to fertilizer nutrient rate and soil test values, denoted $\mathcal{M} = \{M_1, \ldots, M_K\}$. The candidate set included the linear plateau (LP), quadratic plateau (QP), quadratic (QUA), and Mitscherlich (MIT) models, which are commonly used in fertilizer response analyses (Cerrato & Blackmer, 1990; Matavel et al., 2024; Palmero et al., 2026).

For all models, we assumed a Gaussian sampling distribution,

$$\mathbf{y} \sim \text{Normal}\ (\boldsymbol{\mu}, \sigma^2\mathbf{I}),$$

where bold lowercase letters denote vectors and bold uppercase letters denote matrices. Here, $\mathbf{y}$ is the $n \times 1$ vector of observed responses, $\boldsymbol{\mu}$ is the corresponding vector of expected values, σ^2 is the residual variance, and $\mathbf{I}$ is the $n \times n$ identity matrix. The candidate models differ only in the functional form used to define each component μ_i as a function of the predictor x_i. A model specific estimand θ is defined, representing the AONR or CSTV depending on the application. We distinguish two application specific candidate model sets: $M^{(N)}$, used for nitrogen analyses, and $M^{(P)}$, used for phosphorus analyses. For nitrogen analyses, the estimand corresponds to AONR, whereas for phosphorus analyses, the estimand corresponds to CSTV. For nitrogen, the candidate model set is

$$M^{(N)} = \{\text{LP, QP, MIT, QUA}\},$$

whereas for phosphorus, the candidate model set is

$$M^{(P)} = \{\text{LP, QP, MIT}\}.$$

Linear plateau model

$$\mu_i = \begin{cases} \beta_0 + \beta_1 x_i, & x_i < \theta_{\text{LP}}, \\ \beta_0 + \beta_1 \theta_{\text{LP}}, & x_i \geq \theta_{\text{LP}}, \end{cases}$$

where θ_{LP} is the breakpoint at which the response reaches the plateau.

Quadratic plateau model

$$\mu_i = \begin{cases} \beta_0 + \beta_1 x_i + \beta_2 x_i^2, & x_i < \theta_{\text{QP}}, \\ \beta_0 + \beta_1 \theta_{\text{QP}} + \beta_2 \theta_{QP}^2, & x_i \geq \theta_{\text{QP}}, \end{cases}$$

where

$$\theta_{\text{QP}} = -\frac{\beta_1}{2\beta_2}.$$

which is the location of the quadratic maximum.

Mitscherlich model

$$\mu_i = \beta_0 + (\beta_1 - \beta_0)\exp(-\beta_2 x_i),$$

where β_0 is the asymptotic yield, β_1 is the yield at zero input, and $\beta_2 > 0$ controls curvature. Because this model does not attain a finite plateau, θ_{MIT} was defined as the predictor level at which expected response reaches a specified proportion of the asymptotic yield. In this study, this proportion was set to 0.95, such that

$$\mu(\theta_{\text{MIT}}) = 0.95\,\beta_0,$$

which implies

$$\theta_{\text{MIT}} = -\frac{\log\left((1 - 0.95)\,\frac{\beta_0}{\beta_0 - \beta_1}\right)}{\beta_2}.$$

Quadratic model

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2.$$

This function is concave when $\beta_2 < 0$ and has a unique maximum at the vertex of the parabola. The estimand is defined as

$$\theta_{\text{QUA}} = -\frac{\beta_1}{2\beta_2}.$$

The quadratic model was excluded for CSTV estimation.

Estimation procedures

Let $D = \{(y_i, x_i)\}_{i=1}^{n}$ denote a dataset, where x_i is either fertilizer nutrient rate or soil test value, depending on context, and y_i is grain yield or relative yield. Here, y_i and x_i denote scalar observations corresponding to the vector quantities **y** and **x**. For a given candidate model $M_k, k = 1, \ldots, K$, let θ_k denote the derived quantity of interest (AONR in nitrogen analyses or CSTV in phosphorus analyses) and let $\hat{\theta}_k$ be its estimator. For each simulated or real dataset, we applied four estimation strategies: (i) individual models, (ii) Akaike information criterion - weighted bootstrap (AIC-WB), (iii) bootstrap aggregating (Bagging), and (iv) Bayesian model averaging (BMA).

AIC-weighted bootstrap

The AIC weighted model averaging procedure followed the general approach of Burnham & Anderson (2004), with the application to optimum rate estimation following Miguez & Poffenbarger (2022). Akaike weights were computed from the original dataset using the fitted candidate models. For each bootstrap sample, model specific estimators were recomputed and combined using the weights. For a given dataset D, all models in the relevant candidate set $\mathcal{M}^{(N)}$ or $\mathcal{M}^{(P)}$ were fitted by maximum likelihood, and the AIC was computed for each model,

$$\text{AIC}_k = -2\log\, L_k + 2p_k,$$

where L_k is the maximized likelihood under model M_k and p_k is the number of free parameters.

We defined the usual AIC differences

$$\Delta_k = \mathrm{AIC}_k - \min_j \mathrm{AIC}_j,$$

and AIC weights

$$w_k = \frac{\exp\left(-\frac{1}{2}\Delta_k\right)}{\sum_j \exp\left(-\frac{1}{2}\Delta_j\right)}.$$

Non convergent models were assigned $\mathrm{AIC}_k = \infty$ and weight zero.

Let $\hat{\theta}_k$ denote the AONR (nitrogen) or CSTV (phosphorus) obtained from model M_k. The AIC weighted point estimator was

$$\hat{\theta}_{\mathrm{AIC}} = \sum_k w_k \hat{\theta}_k.$$

To quantify uncertainty, we used a nonparametric bootstrap. For each model M_k, we generated B bootstrap samples by resampling the original observations with replacement and refitting the model to each resampled dataset. This procedure produced replicate estimators $\hat{\theta}_k^{(b)}$ for $b = 1, \ldots, B$. For each bootstrap replicate we then formed the AIC weighted estimator. AIC weights w_k were computed once from the original dataset and held fixed across bootstrap replicates.

$$\hat{\theta}_{\mathrm{AIC}}^{(b)} = \sum_k w_k \hat{\theta}_k^{(b)}.$$

The empirical 0.025 and 0.975 quantiles of the bootstrap distribution $\{\hat{\theta}_{\mathrm{AIC}}^{(b)}\}_{b=1}^{B}$ were used as a 95% confidence interval for $\hat{\theta}_{\mathrm{AIC}}$.

Bagging

Bagging was implemented following the bootstrap aggregation framework introduced by Breiman (1996), with the adaptation for optimum nitrogen rate estimation following (Palmero, 2025). Bagging combines two components: (i) bootstrap resampling of the data and (ii) aggregation of estimators obtained from multiple learners (here, the candidate response models).

Let

$$\mathbf{D} = \begin{bmatrix} y_1 & x_1 \\ \vdots & \vdots \\ y_n & x_n \end{bmatrix}$$

denote the original $n \times 2$ data matrix. The bagging algorithm for a given dataset proceeded as follows (Palmero, 2025):

1. Bootstrap phase (B training sets): Generate $B = 5000$ bootstrap training sets $\mathbf{D}^{(b)}$ by sampling n rows from $\mathbf{D}$ with replacement, for $b = 1, \ldots, B$.
2. Model selection for each bootstrap sample: For each b, randomly select one model from the candidate set with equal probability, that is $1/|\ \mathcal{M}^{(N)}\ |$ for nitrogen or $1/|\ \mathcal{M}^{(P)}\ |$ for phosphorus.
3. Model fitting: Fit the selected model to $\mathrm{D}^{(b)}$ and obtain parameter estimates.
4. Derived quantity: Compute the corresponding AONR (for N data) or CSTV (for P data) according to the closed form expression for the fitted model, yielding an estimate $\hat{\theta}^{(b)}$.
5. Aggregation: Repeat steps 1–4 until B bootstrap estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$ have been obtained.

The resulting B values $\{\hat{\theta}^{(b)}\}$ form an empirical distribution for the estimand under bagging. The bagging point estimator was defined as

$$\hat{\theta}_{\mathrm{BAG}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{(b)},$$

and uncertainty was summarized using the empirical 0.025 and 0.975 quantiles of $\{\hat{\theta}^{(b)}\}$ as a 95% confidence interval.

Bayesian inference and Bayesian model averaging

Bayesian inference was used to obtain model specific posterior distributions for AONR or CSTV and to combine these distributions across models using BMA. The overall framework follows the general formulation of BMA described by Hoeting et al. (1999), with the application to optimum nitrogen rate estimation following (Palmero et al., 2026). Let M_k denote a candidate model in the set M, and let $\theta_k = g(\boldsymbol{\beta}_k)$ denote the derived quantity of interest (AONR or CSTV) under that model, where $\boldsymbol{\beta}_k$ is the vector of model parameters.

Bayesian inference

Following standard Bayesian notation (Hooten & Hobbs, 2015), the joint distribution of observable data $\mathbf{y}$ and model parameters under model M_k can be expressed as

$$p(\mathbf{y}, \boldsymbol{\beta}_k, \sigma^2 \mid M_k) = p(\mathbf{y} \mid \boldsymbol{\beta}_k, \sigma^2, M_k)\, p(\boldsymbol{\beta}_k, \sigma^2 \mid M_k),$$

Where $p(\mathbf{y} \mid \boldsymbol{\beta}_k, \sigma^2, M_k)$ is the likelihood, and $p(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)$ is the prior distribution, built from agronomic prior information following (Correndo et al., 2021; Palmero et al., 2026).

Bayes' rule gives the posterior distribution

$$p(\boldsymbol{\beta}_k, \sigma^2 \mid \mathbf{y}, M_k) \propto p(\mathbf{y} \mid \boldsymbol{\beta}_k, \sigma^2, M_k)\, p(\boldsymbol{\beta}_k, \sigma^2 \mid M_k),$$

which was approximated via Markov chain Monte Carlo (MCMC) using Hamiltonian Monte Carlo with the No-U-Turn Sampler.

Because AONR or CSTV is a deterministic function of the parameters $\boldsymbol{\beta}_k$, the posterior distribution of θ_k under model M_k is directly obtained by computing $g(\boldsymbol{\beta}_k)$ at each MCMC iteration using the generated quantities block. This yields a sample from

$$p(\theta_k \mid \mathbf{y}, M_k),$$

the model specific posterior distribution of the estimand.

Prior specification

For each candidate model M_k, prior distributions were assigned to the model parameters $\boldsymbol{\beta}_k$ and the residual variance σ^2. Priors were specified independently across parameters and followed the same distributional families across models.

All parameters constrained to be positive were assigned Gamma prior distributions,

$$\beta_{kj} \sim \text{Gamma}(\alpha, \lambda),$$

with shape parameter fixed at $\alpha = 6$ and rate parameter $\lambda = 6/\mu$, where μ denotes the prior mean. Fixing the shape parameter yields weakly informative priors with coefficient of variation $1/\sqrt{6} \approx 0.41$, allowing substantial information to be learned from the data while maintaining realistic parameter scales.

Prior means were informed by parameter values reported in (Correndo et al., 2021) and (Palmero et al., 2026) for nitrogen response models, and by values reported by Dodd & Mallarino (2005) for phosphorus response models. The same prior construction was used for all models to ensure consistency across candidate functional forms and between nitrogen and phosphorus analyses.

Residual variance parameters were assigned inverse-gamma priors to ensure positivity. Specifically, the variance parameter was modeled as

$$\sigma^2 \sim \text{Inv-Gamma}(a_\sigma, b_\sigma),$$

and the standard deviation σ was used in the likelihood through the relation $\sigma = \sqrt{\sigma^2}$. Hyperparameters were chosen so that the prior mean matched the expected magnitude of residual variability for each response type. Complete prior specification values for each candidate model, are provided in Appendix B.

Bayesian model averaging

Bayesian model averaging was used to correct for model uncertainty by weighting the model specific posteriors by the posterior probability of each model (Hoeting et al., 1999). By the law of total probability, the posterior distribution of θ accounting for model uncertainty is

$$p(\theta \mid \mathbf{y}) = \sum_{k=1}^{K} p(\theta \mid \mathbf{y}, M_k)\, P(M_k \mid \mathbf{y}),$$

where $P(M_k \mid \mathbf{y})$ is the posterior model probability.

Posterior model probabilities were obtained using Bayes' rule

$$P(M_k \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_k)P(M_k)}{\sum_{j=1}^{K} p(\mathbf{y} \mid M_j)P(M_j)},$$

where $P(M_k)$ is the prior model probability (set to $1/K$ in simulations) and $p(\mathbf{y} \mid M_k)$ is the marginal likelihood,

$$p(\mathbf{y} \mid M_k) = \int p(\mathbf{y} \mid \boldsymbol{\beta}_k, \sigma^2, M_k)\, p(\boldsymbol{\beta}_k, \sigma^2 \mid M_k)\, d\boldsymbol{\beta}_k\, d\sigma^2.$$

Because this integral is typically high dimensional and intractable, we used bridge sampling (Gronau et al., 2020; Meng & Wong, 1996), to approximate $p(\mathbf{y} \mid M_k)$.

Monte Carlo approximation of the BMA posterior

Let $\{\theta_k^{(s)}\}$ denote posterior samples of the estimand under model M_k.

A Monte Carlo approximation of $p(\theta \mid \mathbf{y})$ is obtained by drawing model labels $M^{(s)}$ from a categorical distribution with probabilities $\{P(M_k \mid \mathbf{y})\}$, and for each s, sampling $\theta^{(s)} \sim p(\theta \mid \mathbf{y}, M^{(s)})$. The resulting Monte Carlo sample $\{\theta^{(s)}\}_{s=1}^{S}$ approximates the full BMA posterior distribution.

The BMA point estimator was defined as

$$\hat{\theta}_{\mathrm{BMA}} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)},$$

and the empirical 0.025 and 0.975 quantiles of $\{\theta^{(s)}\}$ were used to form a 95% credible interval.

Data simulation

Simulation experiments were conducted to evaluate the performance of individual models and model averaging procedures under controlled conditions in which the true AONR and CSTV were known by construction. All simulations assumed a completely randomized design. For each observation i, responses were generated as

$$y_i \sim \mathrm{Normal}(\mu_i, \sigma^2),$$

where μ_i was determined by a specified true response function and σ^2 was a constant residual variance. Nitrogen simulations used fertilizer N rate x_i as the predictor, whereas phosphorus simulations used soil test P level x_i; in each case, x_i denotes the relevant predictor.

AONR simulations

For AONR simulations, seven N rates with four replications per rate were specified, yielding 28 observations:

$$x_i \in \{0, 50, 100, 150, 200, 250, 300\} \text{ kg N ha}^{-1}.$$

The candidate nitrogen model set was

$$\mathcal{M}^{(N)} = \{\text{LP, QP, MIT, QUA}\}$$

Each model served as the true data generating process in a separate simulation scenario. Model parameters were selected such that all scenarios shared a common true agronomic optimum,

$$\theta_{\text{AONR}}^{\text{true}} = 200 \text{ kg N ha}^{-1},$$

ensuring that differences in estimator performance reflected model misspecification rather than differences in the true optimum. Mean response functions were parameterized to produce realistic maize yield levels and curvature while satisfying the above constraint. Residual variability was fixed at

$$\sigma = 1000 \text{ kg ha}^{-1}.$$

CSTV simulations

Simulations used soil test P concentration x_i as the predictor. Input levels were generated from a right-skewed distribution to reflect the higher sampling density at low soil test values typical of calibration datasets. Each simulated dataset contained 30 unique soil test P values with one observation per level.

The phosphorus candidate model set was

$$\mathcal{M}^{(P)} = \{\text{QP, LP, MIT}\}.$$

Each model was used as the true generating mechanism in a separate scenario, with parameters chosen so that all scenarios shared the same true CSTV:

$$\theta_{\text{CSTV}}^{\text{true}} = 20 \text{ mg kg}^{-1}.$$

Residual variability for phosphorus simulations was set to

$$\sigma = 5.$$

In each scenario, we generated $W = 500$ independent datasets.

A fixed random seed ensured that all estimation procedures within a scenario (individual models, AIC-WB, Bagging, and BMA) were applied to the exact same 500 datasets.

Performance metrics

Because the true AONR and the true CSTV were known in all simulation scenarios, the properties of each estimator could be evaluated directly. We summarized estimator performance using bias, variance, and mean squared error (MSE).

Let

- W be the number of simulated datasets in a scenario ($W = 500$),
- θ be the true value of the estimand (AONR or CSTV), and
- $\hat{\theta}_{w,m}$ be the estimate obtained from method m for dataset w, where m indexes the individual models and the three averaging methods.

The Monte Carlo mean of the estimator under method m was

$$E[\hat{\theta}_m] = \frac{1}{W}\sum_{w=1}^{W} \hat{\theta}_{(w,m)}.$$

The bias of method m was

$$\text{Bias}(\hat{\theta}_m) = E[\hat{\theta}_m] - \theta.$$

The variance of the estimator was

$$\text{Var}(\hat{\theta}_m) = \frac{1}{W-1}\sum_{w=1}^{W} (\hat{\theta}_{(w,m)} - E[\hat{\theta}_m])^2.$$

The mean squared error was

$$\mathrm{MSE}(\hat{\theta}_m) = \mathrm{Var}(\hat{\theta}_m) + (\mathrm{Bias}(\hat{\theta}_m))^2.$$

The MSE therefore decomposes into a component measuring precision (variance) and a component measuring accuracy (bias). Estimators with favorable performance have both small variance and small bias, leading to a low MSE. These metrics were computed separately for each method and each simulation scenario for both AONR and CSTV.

Applied case studies

Applied case studies were conducted for nitrogen and phosphorus to illustrate the application of individual models and model averaging procedures using real datasets. For nitrogen, we analyzed a maize grain yield response dataset described in Ransom et al. (2021). A single site–year experiment was randomly selected (Trial 27, Amenia, ND, 2015), including eight total N rates from 0 to 315 kg ha^{-1} in a four-block design, with unequal replication across rates. The candidate nitrogen model set was

$$\mathcal{M}^{(N)} = \{\mathrm{LP}, \mathrm{QP}, \mathrm{MIT}, \mathrm{QUA}\}.$$

The estimand was the AONR, computed using the model specific definitions. For the quadratic model, θ_{QUA} was obtained using a numerically stable vertex-based estimator.

For phosphorus, we analyzed a previously described multisite soybean correlation and calibration dataset relating soil test phosphorus to grain yield (Slaton, Pearce, Lyons, et al., 2024). Analyses were conducted separately by site year. The candidate phosphorus model set was

$$\mathcal{M}^{(P)} = \{\mathrm{LP}, \mathrm{QP}, \mathrm{MIT}\}.$$

The estimand was CSTV, defined as the breakpoint for LP and QP and as the soil test level achieving 95% of the asymptotic yield for the MIT model.

Software and reproducibility

All statistical analyses were conducted using R version 4.5.1 (R Core Team, 2025) within RStudio version 2025.05.1+513 (RStudio Team, 2025). Nonlinear response models (LP, QP and MIT) were fit under a frequentist framework by nonlinear least squares using the Levenberg–Marquardt implementation nlsLM() from the minpack.lm package (Elzhov et al., 2022). The QUA model was fit separately by ordinary least squares using lm() in base R.

Uncertainty for individual model estimands was quantified using nonparametric bootstrap resampling with $B = 5000$ replicates of the original observations. For each bootstrap sample, the model was refitted and the derived quantity of interest (θ) recomputed to obtain an empirical distribution of the estimator for each model. The AIC-WB estimators were computed by forming weighted averages of model specific estimates using Akaike weights and applying these weights to the bootstrap distributions of the estimands.

Bagging was implemented as an ensemble procedure in which, for each of $B = 5000$ bootstrap resamples, one candidate model was selected at random with equal probability, refitted to the resampled data, and used to compute the corresponding estimand. The bagging estimator was obtained by aggregating the resulting distribution of bootstrap estimates.

Bayesian analyses were conducted using Stan (version 2.32) via the rstan package (Guo et al., 2015). Posterior sampling was performed using Hamiltonian Monte Carlo with the No-U-Turn Sampler. For each candidate model, four chains were run with 4,000 iterations per chain, including 2,000 warmup iterations, and thinning set to 5. The target acceptance probability was set to 0.95, and a fixed random seed was used for reproducibility. Model specific posterior distributions of AONR or CSTV were obtained as deterministic functions of sampled model

parameters within the generated quantities block. Bayesian model averaging was implemented using marginal likelihoods approximated via bridge sampling, as implemented in the bridgesampling package (Gronau & Singmann, 2017). Posterior model probabilities were used to construct a mixture of posterior distributions of the estimand across candidate models.

Portions of the implementation for bagging and BMA were informed by the workflows described in Palmero (2025) and Palmero et al. (2026). The AIC weighted model averaging framework follows Miguez & Poffenbarger (2022), and the bootstrap procedure follows Correndo et al. (2023).

Reproducible R code illustrating key steps of the analysis is provided in Appendix C, including the implementation of all estimation procedures and applied case study workflows.

Results

Simulations

For AONR, when the LP is the true response, the three nonlinear alternatives show pronounced upward bias. QP and QUA overestimate AONR by 121 kg N ha^{-1}, while the MIT model overestimates by 289 kg N ha^{-1}, resulting in large MSE (24278–173423). In contrast, the LP model is essentially unbiased (Bias = 1) with low variance (Var = 350), as expected under correct specification. Among the averaging approaches, BMA yields the smallest MSE (4765), driven by a substantial variance reduction relative to AIC-WB and bagging (Var = 1751 vs. 8568 and 3525).

When the QP is the true response, misspecification affects models differently, primarily through changes in bias. LP is negatively biased (-57), while QP shows minimal bias (Bias = 5). Here, all three ensemble estimators perform similarly in MSE (1018–1557), with bagging providing the smallest MSE (1018) and BMA close behind (1039). When the MIT response is the true model, the MIT model shows minimal bias (Bias = 3), whereas LP remains strongly negatively biased (Bias = -58). In this scenario, bagging again yields the smallest MSE (973), slightly outperforming BMA (1315), primarily due to lower variance (973 vs. 1207), with comparable bias (0 vs. 10).

When the QUA model is the true response, QUA dominates (MSE = 59). However, the key comparison is among misspecified and averaging estimators: LP is highly negatively biased (-82; MSE = 6885), and the nonlinear plateaus and MIT also remain biased. In this case, BMA produces the smallest MSE among the averaging methods (616) and improves substantially on bagging (1540), while AIC-WB yields a smaller MSE (341) than BMA in this scenario, although both remain biased relative to the correctly specified model.

Over the four data-generating processes, the largest errors arise under severe misspecification (e.g., fitting MIT to LP truth, or fitting LP to QUA truth). Averaging methods consistently reduce these extremes. Importantly, these results also show that the best performing ensemble method is context dependent: BMA is particularly competitive when variance control dominates MSE, while bagging can be best when resampling stabilizes the AONR under saturating nonlinearities.

Similar patterns emerge for CSTV, although magnitudes and rankings differ because CSTV is a different derived estimand and the candidate set is smaller.

When LP is the true model, LP and MIT are unbiased (Bias = 0), but QP shows notable positive bias (Bias = 8) and much larger MSE (77). Among averaging methods, BMA achieves the smallest MSE (3), combining negligible bias with the lowest variance (3). When QP is the true model, QP is unbiased (Bias = 0) but has higher variance (22). LP and MIT are negatively biased (Bias = -5 and -8), yielding larger MSE (35 and 64). Here, AIC-WB is the best performer (MSE = 24), while BMA remains competitive (MSE = 30) with low variance (3) but retains negative bias (-5).

When MIT is the true model, MIT is unbiased (Bias = 0) with low MSE (9), and LP is unbiased but higher variance (15). QP again shows positive bias (Bias = 7) and inflated MSE (63). In this scenario, bagging performs well (MSE = 10), close to MIT (9), and better than AIC-WB (17), while BMA remains slightly higher (MSE = 7). Compared with AONR, CSTV appears more robust under correct shape families (LP or MIT), but more sensitive to imposing a QP form when the true response is not plateau with curvature. Overall, averaging again reduces extreme failures, with BMA excelling when variance dominates, and AIC-WB or bagging occasionally leading when bias control is critical under the true QP case.

Table 1. Simulation-based performance of agronomic optimum nitrogen rate (AONR) estimators across alternative true yield response functions. Results summarize 500 simulated datasets generated under distinct data generating processes and estimation approaches. E [$\hat{\theta}$] is the mean estimated AONR (kg N ha^{-1}). Bias [$\hat{\theta}$] and Var [$\hat{\theta}$] denote the empirical bias and variance of the estimator across simulations. MSE is the mean squared error. Models include linear plateau (LP), quadratic plateau (QP), Mitscherlich (MIT) and Quadratic (QUA); averaging methods include Akaike information criterion–weighted bootstrap (AIC-WB), bootstrap aggregation (bagging), and Bayesian model averaging (BMA).

Models & Methods	E [$\hat{\theta}$]	Bias [$\hat{\theta}$]	Var [$\hat{\theta}$]	MSE
LP as the true model				
LP	201	1	350	352
QP	321	121	9678	24341
MIT	489	289	90146	173423
QUA	321	121	9680	24278
AIC-WB	275	75	8568	14256
Bagging	314	114	3525	16633
BMA	255	55	1751	4765
QP as the true model				
LP	143	-57	818	4032
QP	205	5	1155	1176
MIT	200	0	2365	2365
QUA	237	37	453	1801
AIC-WB	190	-10	1463	1557
Bagging	198	-2	1014	1018
BMA	194	-6	998	1039
MIT as the true model				
LP	142	-58	1177	4524
QP	195	-5	1245	1266
MIT	203	3	1964	1974
QUA	240	40	341	1937
AIC-WB	196	-4	1644	1664
Bagging	200	0	973	973
BMA	210	10	1207	1315
QUA as the true model				
LP	118	-82	167	6885
QP	170	-30	294	1212
MIT	157	-43	464	2348
QUA	201	1	59	59
AIC-WB	191	-9	264	341
Bagging	165	-35	326	1540
BMA	184	-16	366	616

Table 2. Simulation-based performance of critical soil test value (CSTV) estimators across alternative true yield–response functions. Results summarize 500 simulated datasets generated under distinct data-generating processes and estimation approaches. E [$\hat{\theta}$] is the mean estimated CSTV (mg P kg^{-1}). Bias [$\hat{\theta}$] and Var [$\hat{\theta}$] denote the empirical bias and variance of the estimator across simulations. MSE is the mean squared error. Models include linear plateau (LP), quadratic plateau (QP), and Mitscherlich (MIT); averaging methods include Akaike information criterion–weighted bootstrap (AIC-WB), bootstrap aggregation (bagging), and Bayesian model averaging (BMA).

Models & Methods	E [$\hat{\theta}$]	Bias [$\hat{\theta}$]	Var [$\hat{\theta}$]	MSE
LP as the true model				
LP	20	0	6	6
QP	28	8	12	77
MIT	20	0	8	8
AIC-WB	22	2	9	14
Bagging	23	3	9	15
BMA	20	0	3	3
QP as the true model				
LP	15	-5	14	35
QP	20	0	22	22
MIT	12	-8	7	64
AIC-WB	16	-4	10	24
Bagging	14	-6	16	52
BMA	15	-5	3	30
MIT as the true model				
LP	20	0	15	15
QP	27	7	17	63
MIT	20	0	9	9
AIC-WB	22	2	13	17
Bagging	21	1	9	10
BMA	20	0	7	7

Case studies

For the case study on AONR, Figure 1 illustrates the observed grain yield response to nitrogen rate together with fitted response functions and AONR estimates obtained from individual models and model averaging approaches.

Estimated AONR values differed markedly among individual response models. The LP model produced the lowest AONR estimate at 121 kg N ha^{-1}, with a relatively narrow confidence interval of 102–163 kg N ha^{-1}. The QP and MIT models yielded higher optima of 180 kg N ha^{-1} and 171 kg N ha^{-1}, respectively, but with substantially wider uncertainty ranges (132–291 kg N ha^{-1} for QP and 98–310 kg N ha^{-1} for MIT). The QUA model resulted in the highest AONR estimate at 241 kg N ha^{-1}, with a comparatively narrow confidence interval (213–320 kg N ha^{-1}).

Model averaging approaches produced intermediate AONR estimates relative to the individual models. The AIC-WB estimate was 163 kg N ha^{-1}, with a confidence interval of 135–222 kg N ha^{-1}. Bagging resulted in a slightly higher estimate of 188 kg N ha^{-1}, but with wide uncertainty with confidence interval (107–299 kg N ha^{-1}). BMA yielded an AONR estimate of 179 kg N ha^{-1}, with a comparatively narrower credible interval of 109–262 kg N ha^{-1}. Overall, model averaging reduced the influence of extreme optima implied by individual models and concentrated AONR estimates within a more agronomically plausible range supported by the observed data.

For the CSTV case study, Figure 2 presents the relative yield response to soil test phosphorus and corresponding estimates of the CSTV. Relative yield increased sharply at low soil test P levels and approached an asymptote at moderate concentrations, with minimal additional response at higher soil test values. Single-model CSTV estimates varied across response functions. The LP model estimated a CSTV of 21 mg P kg^{-1}, with a confidence interval

of 12–34 mg P kg^{-1}. The QP model produced a higher estimate of 27 mg P kg^{-1}, accompanied by substantial uncertainty (13–56 mg P kg^{-1}). In contrast, the MIT model yielded a lower CSTV estimate of 12 mg P kg^{-1}, with a wide confidence interval of 6–36 mg P kg^{-1}.

Model-averaged CSTV estimates were more tightly clustered than those from individual models. The AIC-WB approach resulted in a CSTV estimate of 20 mg P kg^{-1}, with a confidence interval of 14–34 mg P kg^{-1}. Bagging produced a slightly lower estimate of 18 mg P kg^{-1}, but with wider uncertainty with confidence interval (8–37 mg P kg^{-1}). BMA yielded the lowest and most precise CSTV estimate at 12 mg P kg^{-1}, with a comparatively narrow credible interval of 6–21 mg P kg^{-1}. Model averaging approaches moderate differences among individual functional forms and reduced uncertainty, yielding CSTV concentrated within the range most consistently supported by the data.

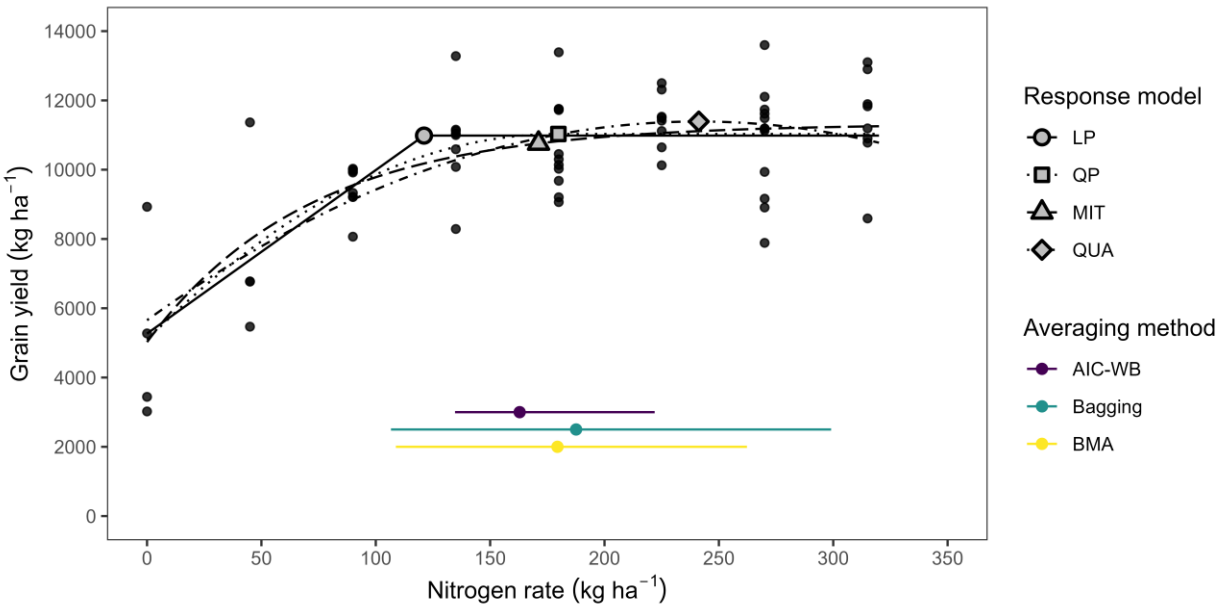


Figure 1. Grain yield response of maize to nitrogen rate and comparison of individual and model averaged optimum nitrogen rate estimates. Response models include linear plateau (LP), quadratic plateau (QP), Mitscherlich (MIT), and quadratic (QUA). Model averaging approaches include Akaike information criterion weighted bootstrap (AIC-WB), bootstrap aggregation (bagging), and Bayesian model averaging (BMA).

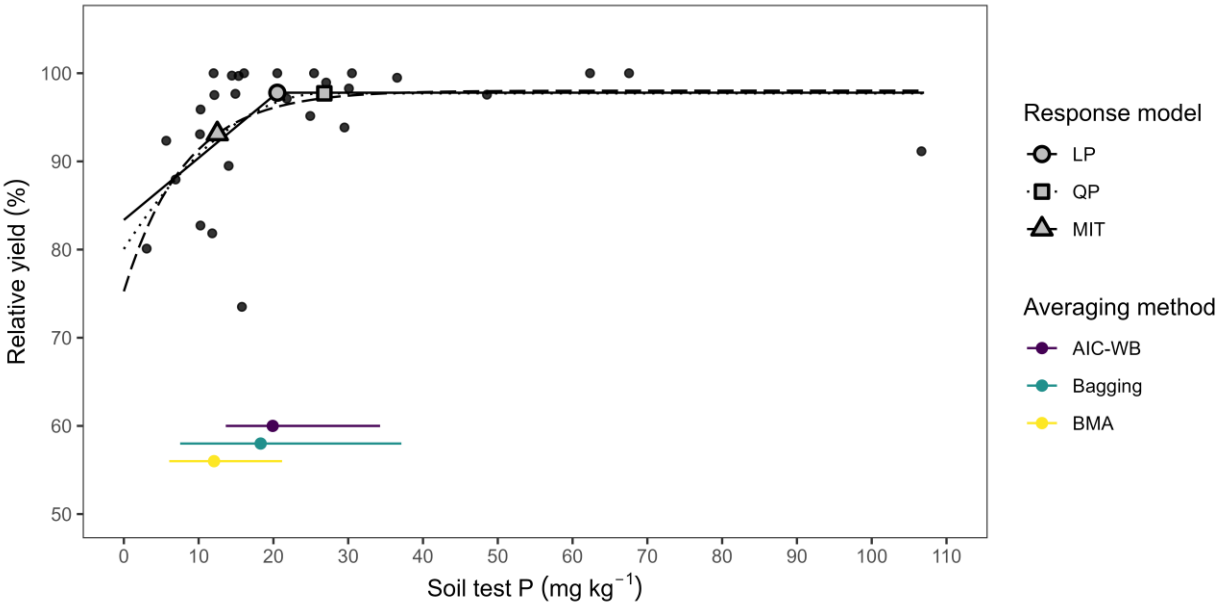


Figure 2. Relative crop yield response of soybean to soil test phosphorus (P) and comparison of individual and model averaged critical soil test value for P estimates. Response models include linear plateau (LP), quadratic plateau (QP) and Mitscherlich (MIT). Model averaging approaches include Akaike information criterion weighted bootstrap (AIC-WB), bootstrap aggregation (bagging), and Bayesian model averaging (BMA).

Discussion

The results from both the simulation study and the applied case studies reinforce that estimation of agronomic optima and critical thresholds is fundamentally constrained by uncertainty in the response functional form. Large differences in AONR and CSTV estimates arose not only from sampling variability, but from structural differences among plausible response models. This distinction between sampling uncertainty and model uncertainty has been emphasized in recent statistical syntheses, which note that conditioning inference on a single selected model implicitly treats structural uncertainty as negligible (Dixon, 2022). Similar sensitivity of agronomic optima to response curve specification has been documented previously, particularly when optima are defined as nonlinear functions of fitted parameters rather than directly observed quantities (Miguez & Poffenbarger, 2022).

In the simulation analysis, single model estimators performed well only when the fitted model matched the true data generating process. Under misspecification, bias and MSE increased sharply, especially for AONR. Flexible nonlinear models, such as quadratic and exponential responses, tended to produce inflated optima when applied to plateau type responses, whereas simpler plateau models underestimated optima when the true response exhibited gradual saturation. These patterns are consistent with previous simulation based evaluations of optimum nitrogen rate estimation, which showed that selecting a single “best” model based on goodness of-fit criteria can lead to systematically biased recommendations when model uncertainty is ignored (Matavel et al., 2024). These patterns reflect systematic bias under misspecification and increased estimator variance when model form is poorly aligned with the underlying response.

Model averaging approaches consistently reduced the magnitude of these errors by integrating information across competing response functions. Importantly, model averaging

depends on a well-defined candidate set; it is not intended to include all possible models, but rather a set of agronomically and statistically plausible response functions, as poorly specified models can degrade performance. Both bagging and BMA attenuated extreme estimates arising from individual misspecified models, leading to lower or comparable MSE across scenarios. Palmero (2025) demonstrated that bagging operates primarily as a variance stabilization technique, with bias reduction depending strongly on the diversity and accuracy of the candidate model set. This behavior is consistent with our results, where bagging substantially reduced variability in AONR estimates but did not uniformly outperform alternative ensemble approaches. In several scenarios, BMA produced the smallest or near smallest MSE by reducing estimator variance, particularly when misspecified nonlinear models generated extreme estimates. However, when the true response exhibited strong nonlinear saturation, bagging occasionally achieved slightly lower MSE through bootstrap stabilization of the estimand. Similar bias-variance trade-offs have been reported in recent evaluations of BMA for estimating optimum nitrogen rates, where BMA often improved estimator precision but its advantage depended on the candidate model set and the underlying response function (Palmero et al., 2026).

The nitrogen case study illustrates the practical implications under real field data. Individual model AONR estimates spanned more than 120 kg N ha^{-1}, a range large enough to alter fertilizer recommendations, economic returns, and environmental losses. This divergence mirrors previous studies, where optimum nitrogen rates varied widely depending on the selected response model (Miguez & Poffenbarger, 2022; Palmero, 2025). Model averaged estimates fell between these extremes, with intervals reflecting both sampling variability and disagreement among plausible response forms.

The phosphorus case study showed smaller dispersion among CSTV estimates relative to AONR, but meaningful differences among models remained, particularly for the QP specification. Similar sensitivity of CSTV to functional form has been reported in soil test calibration studies, where alternative response models imply different critical thresholds despite comparable overall fit. Model averaging moderated these discrepancies, yielding CSTV estimates concentrated within the range supported by the data. The relatively narrow BMA intervals suggest that when candidate models provide similar fits, Bayesian averaging can regularize inference by down-weighting structurally extreme solutions and stabilizing derived quantities such as CSTV. This behavior is consistent with theoretical expectations for model averaging (Dixon, 2022), and recent evaluations (Palmero et al., 2026).

Although computational efficiency was not a primary objective of this study, we observed substantial differences in runtime among model averaging approaches (Appendix Figure A.1). The BMA required substantially longer computation times than bagging and AIC-WB approaches, reflecting the cost of full posterior sampling. These differences may be relevant for large-scale or operational applications, but they do not alter the comparative inference emphasized here.

Taken together, these results demonstrate that model uncertainty should be treated as a first-order source of uncertainty in nutrient recommendation systems. Approaches that rely on selecting a single response function implicitly assume that structural uncertainty is negligible, an assumption not supported by either the simulation results or the applied case studies presented here. Model averaging provides a practical framework for integrating uncertainty across plausible response forms and yields fertilizer recommendations that are less sensitive to model specification.

Conclusion

This study demonstrates that uncertainty in the functional form of yield response relationships is a major driver of bias and variance in estimated agronomic optimum rates and critical soil test values. Simulation results showed that single model estimators perform well only under correct specification, whereas model misspecification leads to substantial bias and increased MSE, resulting in unstable nutrient recommendations.

Across both simulations and applied case studies, model averaging approaches reduced sensitivity to functional form assumptions relative to single model estimation. By integrating information across multiple plausible response models, these methods moderated extreme estimates and produced more stable estimands supported by the data. Although no single averaging method uniformly dominated, BMA was most effective when variance reduction drove performance, whereas bagging and AIC-weighted approaches performed well when bias control or resampling stabilization was more influential.

These results directly address the study objectives by demonstrating the impact of model misspecification on AONR and CSTV estimation, quantifying performance differences among individual and averaged models in terms of bias, variance, and MSE, and illustrating their practical implications in nitrogen and phosphorus case studies. Overall, model uncertainty should be treated as a first order component of uncertainty in nutrient recommendation systems. Model averaging provides a robust and extensible framework for estimating optimal rates and critical thresholds when the true response function is unknown.

References

Bachmaier, M., & Gandorfer, M. (2012). Estimating uncertainty of economically optimum N fertilizer rates. *International Journal of Agronomy*, *2012*(1), 580294. https://doi.org/10.1155/2012/580294

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Bullock, D. G., & Bullock, D. S. (1994). Quadratic and quadratic-plus-plateau models for predicting optimal nitrogen rate of corn: A comparison. *Agronomy Journal*, *86*(1), 191–195. https://doi.org/10.2134/agronj1994.00021962008600010033x

Burnham, K. P., & Anderson, D. R. (Eds.). (2004). *Model selection and multimodel inference*. Springer. https://doi.org/10.1007/b97636

Cassman, K. G., Dobermann, A., & Walters, D. T. (2002). Agroecosystems, nitrogen-use efficiency, and nitrogen management. *AMBIO: A Journal of the Human Environment*, *31*(2), 132–140. https://doi.org/10.1579/0044-7447-31.2.132

Cerrato, M. E., & Blackmer, A. M. (1990). Comparison of models for describing corn yield response to nitrogen fertilizer. *Agronomy Journal*, *82*(1), 138–143. https://doi.org/10.2134/agronj1990.00021962008200010030x

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.

Correndo, A. A., Pearce, A., Bolster, C. H., Spargo, J. T., Osmond, D., & Ciampitti, I. A. (2023). The soiltestcorr R package: An accessible framework for reproducible correlation analysis of crop yield and soil test data. *SoftwareX*, *21*, 101275. https://doi.org/10.1016/j.softx.2022.101275

Correndo, A. A., Tremblay, N., Coulter, J. A., Ruiz-Diaz, D., Franzen, D., Nafziger, E., Prasad, V., Rosso, L. H. M., Steinke, K., Du, J., Messina, C. D., & Ciampitti, I. A. (2021). Unraveling uncertainty drivers of the maize yield response to nitrogen: A Bayesian and machine learning approach. *Agricultural and Forest Meteorology*, *311*, 108668. https://doi.org/10.1016/j.agrformet.2021.108668

Dhakal, C., & Lange, K. (2021). Crop yield response functions in nutrient application: A review. *Agronomy Journal*, *113*(6), 5222–5234. https://doi.org/10.1002/agj2.20863

Dixon, P. M. (2022). *Model averaging in agriculture and natural resources: What is it? When is it useful? When is it a distraction?* https://doi.org/10.26077/234E-1207

Dodd, J. R., & Mallarino, A. P. (2005). Soil-test phosphorus and crop grain yield responses to long-term phosphorus fertilization for corn–soybean rotations. *Soil Science Society of America Journal*, *69*(4), 1118–1128. https://doi.org/10.2136/sssaj2004.0279

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., & Bolker, B. (2022). *minpack.lm: R interface to the Levenberg–Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds* (p. 1.2-4) [Computer software]. https://doi.org/10.32614/CRAN.package.minpack.lm

Filippi, D., Gatiboni, L., Crozier, C., Osmond, D., & Hardy, D. (2025). Effect of model choice on critical soil test value of phosphorus for corn in long-term trials in North Carolina. *Soil Science Society of America Journal*, *89*(4), e70104. https://doi.org/10.1002/saj2.70104

Galloway, J. N., Aber, J. D., Erisman, J. W., Seitzinger, S. P., Howarth, R. W., Cowling, E. B., & Cosby, B. J. (2003). The nitrogen cascade. *BioScience*, *53*(4), 341–356. https://doi.org/10.1641/0006-3568(2003)053%255B0341:TNC%255D2.0.CO;2

Gronau, Q. F., & Singmann, H. (2017). *bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors* (p. 1.2-1) [Computer software]. https://doi.org/10.32614/CRAN.package.bridgesampling

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*, 1–29. https://doi.org/10.18637/jss.v092.i10

Guo, J., Gabry, J., Goodrich, B., Johnson, A., Weber, S., & Badr, H. S. (2015). *rstan: R interface to Stan* (p. 2.32.7) [Computer software]. https://doi.org/10.32614/CRAN.package.rstan

Henke, J., Breustedt, G., Sieling, K., & Kage, H. (2007). Impact of uncertainty on the optimum nitrogen fertilization rate and agronomic, ecological and economic factors in an oilseed rape based crop rotation. *The Journal of Agricultural Science*, *145*(5), 455–468. https://doi.org/10.1017/S0021859607007204

Hernandez, J. A., & Mulla, D. J. (2008). Estimating uncertainty of economically optimum fertilizer rates. *Agronomy Journal*, *100*(5), 1221–1229. https://doi.org/10.2134/agronj2007.0273

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401.

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28. https://doi.org/10.1890/14-0661.1

Jaynes, D. B. (2011). Confidence bands for measured economically optimal nitrogen rates. *Precision Agriculture*, *12*(2), 196–213. https://doi.org/10.1007/s11119-010-9168-3

Lyons, S. E., Arthur, D. K., Slaton, N. A., Pearce, A. W., Spargo, J. T., Osmond, D. L., & Kleinman, P. J. A. (2021). Development of a soil test correlation and calibration database

for the USA. *Agricultural & Environmental Letters*, *6*(4), e20058. https://doi.org/10.1002/ael2.20058

Lyons, S. E., Tang, Z., Booth, J., & Ketterings, Q. M. (2019). Nitrogen response models for winter cereals grown for forage. *Journal of Agronomy and Crop Science*, *205*(2), 248–261. https://doi.org/10.1111/jac.12310

Mac Nally, R., Duncan, R. P., Thomson, J. R., & Yen, J. D. L. (2018). Model selection using information criteria, but is the "best" model any good? *Journal of Applied Ecology*, *55*(3), 1441–1444. https://doi.org/10.1111/1365-2664.13060

Mallarino, A. P., & Blackmer, A. M. (1992). Comparison of methods for determining critical concentrations of soil test phosphorus for corn. *Agronomy Journal*, *84*(5), 850–856. https://doi.org/10.2134/agronj1992.00021962008400050017x

Matavel, C. E., Meyer-Aurich, A., & Piepho, H.-P. (2024). Model-averaging as an accurate approach for ex-post economic optimum nitrogen rate estimation. *Precision Agriculture*, *25*(3), 1324–1339. https://doi.org/10.1007/s11119-024-10113-4

Matavel, C. E., Meyer-Aurich, A., & Piepho, H.-P. (2025). Bayesian-optimized experimental designs for estimating the economic optimum nitrogen rate: A model-averaging approach. *Agronomy Journal*, *117*(3), e70087. https://doi.org/10.1002/agj2.70087

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*, 831–860.

Miguez, F. E., & Poffenbarger, H. (2022). How can we estimate optimum fertilizer rates with accuracy and precision? *Agricultural & Environmental Letters*, *7*(1), e20075. https://doi.org/10.1002/ael2.20075

Nigon, T. J., Yang, C., Mulla, D. J., & Kaiser, D. E. (2019). Computing uncertainty in the optimum nitrogen rate using a generalized cost function. *Computers and Electronics in Agriculture*, *167*, 105030. https://doi.org/10.1016/j.compag.2019.105030

Palmero, F. (2025). *Bagging as a model averaging technique for estimating optimum nitrogen rates in agricultural crops* [Master's Thesis, Kansas State University]. https://hdl.handle.net/2097/44861

Palmero, F., Ciampitti, I. A., & Hefley, T. J. (2026). Bayesian model averaging to estimate economic optimum nitrogen rates in agricultural crops. *Field Crops Research*, *341*, 110397. https://doi.org/10.1016/j.fcr.2026.110397

R Core Team. (2025). *A language and environment for statistical computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Ransom, C. J., Clark, J., Bean, G. M., Bandura, C., Shafer, M. E., Kitchen, N. R., Camberato, J. J., Carter, P. R., Ferguson, R. B., Fernández, F., Franzen, D. W., Laboski, C. A. M., Myers, D. B., Nafziger, E. D., Sawyer, J. E., & Shanahan, J. (2021). Data from a public–industry partnership for enhancing corn nitrogen research. *Agronomy Journal*, *113*(5), 4429–4436. https://doi.org/10.1002/agj2.20812

RStudio Team. (2025). *RStudio: Integrated Development Environment for R* (Version 2025.05.1+513) [Computer software]. RStudio, Inc. http://www.rstudio.com/

Scharf, P. C., Kitchen, N. R., Sudduth, K. A., Davis, J. G., Hubbard, V. C., & Lory, J. A. (2005). Field-scale variability in optimal nitrogen fertilizer rate for corn. *Agronomy Journal*, *97*(2), 452–461. https://doi.org/10.2134/agronj2005.0452

Slaton, N. A., Pearce, A., Gatiboni, L., Osmond, D., Bolster, C., Miquez, F., Clark, J., Dhillon, J., Farmaha, B., Kaiser, D., Lyons, S., Margenot, A., Moore, A., Ruiz Diaz, D., Sotomayor, D., Spackman, J., Spargo, J., & Yost, M. (2024). Models and sufficiency interpretation for estimating critical soil test values for the fertilizer recommendation support tool. *Soil Science Society of America Journal*, *88*(4), 1419–1437. https://doi.org/10.1002/saj2.20704

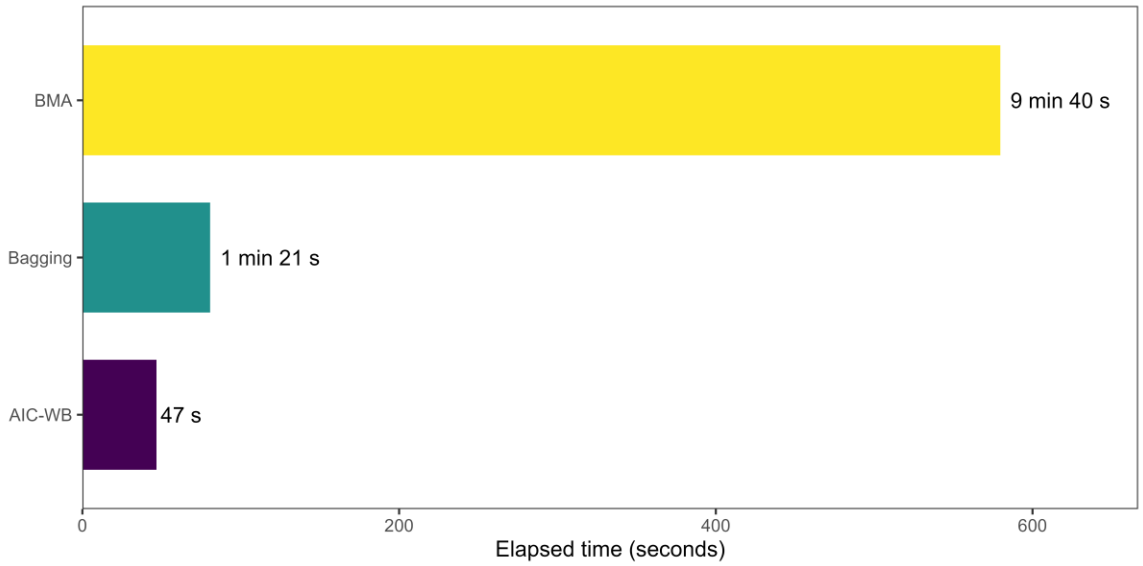
Slaton, N. A., Pearce, A. W., Lyons, S. E., Drescher, G. L., & Smartt, A. D. (2024). *Soybean yield response to fertilizer-phosphorus rate on soils having different Mehlich-3 phosphorus values in Arkansas* [Dataset]. Ag Data Commons. https://doi.org/10.15482/USDA.ADC/1524648

Sutton, M. A., Oenema, O., Erisman, J. W., Leip, A., van Grinsven, H., & Winiwarter, W. (2011). Too much of a good thing. *Nature*, *472*(7342), 159–161. https://doi.org/10.1038/472159a

Zou, T., Zhang, X., & Davidson, E. A. (2022). Global trends of cropland phosphorus use and sustainability challenges. *Nature*, *611*(7934), Article 7934. https://doi.org/10.1038/s41586-022-05220-z

Appendix A - Computational time

Appendix Figure A.1. Total computational time required to estimate AONR and uncertainty on one site using AIC-Weighted Bootstrap (AIC-WB), Bootstrap aggregating (Bagging), and Bayesian model averaging (BMA)



Appendix B - Prior specification

Bayesian models and prior distributions for AONR

Linear Plateau

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \begin{cases} \beta_0 + \beta_1 x_i, & x_i < \theta \\ \beta_0 + \beta_1 \theta, & x_i \geq \theta \end{cases} \\
\beta_0 &\sim \text{Gamma}(6,\ 6/7317), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/33.33), \\
\theta &\sim \text{Gamma}(6,\ 6/150), \\
\sigma &\sim \text{Inv-Gamma}(6,\ 5000).
\end{aligned}
$$

Quadratic Plateau

Let $x_p = \beta_1/(2\beta_2)$.

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \begin{cases} \beta_0 + \beta_1 x_i - \beta_2 x_i^2, & x_i < x_p \\ \beta_0 + \beta_1 x_p - \beta_2 x_p^2, & x_i \geq x_p \end{cases} \\
\beta_0 &\sim \text{Gamma}(6,\ 6/6967), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/50.45), \\
\beta_2 &\sim \text{Gamma}(6,\ 6/0.121), \\
\sigma &\sim \text{Inv-Gamma}(6,\ 5000).
\end{aligned}
$$

Mitscherlich

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \beta_0 + (\beta_1 - \beta_0)\exp(-\beta_2 x_i), \\
\beta_0 &\sim \text{Gamma}(6,\ 6/12544), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/6817), \\
\beta_2 &\sim \text{Gamma}(6,\ 6/0.012), \\
\sigma &\sim \text{Inv-Gamma}(6,\ 5000),
\end{aligned}
$$

Quadratic

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \beta_0 + \beta_1 x_i - \beta_2 x_i^2, \\
\beta_0 &\sim \text{Gamma}(6,\ 6/7214), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/43.0), \\
\beta_2 &\sim \text{Gamma}(6,\ 6/0.09), \\
\sigma &\sim \text{Inv-Gamma}(6,\ 5000).
\end{aligned}
$$

Bayesian models and prior distributions for CSTV

Linear Plateau

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \begin{cases} \beta_0 + \beta_1 x_i, & x_i < \theta \\ \beta_0 + \beta_1 \theta, & x_i \geq \theta \end{cases} \\
\beta_0 &\sim \text{Gamma}(6,\ 6/63.7), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/2.55), \\
\theta &\sim \text{Gamma}(6,\ 6/13.7), \\
\sigma &\sim \text{Inv-Gamma}(3,\ 15).
\end{aligned}
$$

Quadratic Plateau

Let $x_p = \beta_1/(2\beta_2)$.

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \begin{cases} \beta_0 + \beta_1 x_i - \beta_2 x_i^2, & x_i < x_p \\ \beta_0 + \beta_1 x_p - \beta_2 x_p^2, & x_i \geq x_p \end{cases} \\
\beta_0 &\sim \text{Gamma}(6,\ 6/57.25), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/4.295), \\
\beta_2 &\sim \text{Gamma}(6,\ 6/0.1096), \\
\sigma &\sim \text{Inv-Gamma}(3,\ 15).
\end{aligned}
$$

Mitscherlich

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2), \\
\mu_i &= \beta_0 + (\beta_1 - \beta_0)\exp(-\beta_2 x_i), \\
\beta_0 &\sim \text{Gamma}(6,\ 6/100.15), \\
\beta_1 &\sim \text{Gamma}(6,\ 6/40.315), \\
\beta_2 &\sim \text{Gamma}(6,\ 6/0.1745), \\
\sigma &\sim \text{Inv-Gamma}(3,\ 15),
\end{aligned}
$$

Appendix C - Reproducible codes

R code for AIC-WB

```
# Fit candidate models

# Linear Plateau (LP)
fm.LP <- nlsLM(response ~ ifelse(input_level < xs,
                 a + b * input_level,
                 a + b * xs),
  data = dat,
  start = list(
    a  = unname(coef(lm(response ~ input_level, data = dat))[1]),
    b  = unname(coef(lm(response ~ input_level, data = dat))[2]),
    xs = unname(quantile(dat$input_level, 0.6, na.rm = TRUE))
  ),
  control = nls.lm.control(maxiter = 1024))

# Quadratic Plateau (QP)
fm.QP <- nlsLM(response ~ ifelse(input_level < (b1/(2*b2)),
                 b0 + b1*input_level - b2*input_level^2,
                 b0 + b1*(b1/(2*b2)) - b2*(b1/(2*b2))^2),
  data = dat,
  start = list(
    b0 = unname(coef(lm(response ~ input_level, data = dat))[1]),
    b1 = unname(coef(lm(response ~ input_level, data = dat))[2]),
    b2 = unname(abs(coef(lm(response ~ input_level, data = dat))[2]) /
           (2*unname(stats::quantile(dat$input_level, 0.6, na.rm=TRUE)) + 1e-8))
  ),
  lower = c(b0 = -Inf, b1 = -Inf, b2 = 1e-10),
  control = nls.lm.control(maxiter = 1024))

# Mitscherlich (MIT)
fm.MIT <- nlsLM(response ~ b0 + (b1 - b0) * exp(-b2 * input_level),
  data = dat,
  start = list(
    b0 = unname(max(dat$response, na.rm = TRUE)),
    b1 = unname(min(dat$response, na.rm = TRUE)),
    b2 = 1 / unname(max(dat$input_level, na.rm = TRUE))
  ),
  lower = c(b0 = -Inf, b1 = -Inf, b2 = 1e-10),
  control = nls.lm.control(maxiter = 1024))

# Quadratic (QUA)
fm.QUA <- lm(response ~ input_level + I(input_level^2), data=dat)
```

```
# Compute model specific optimum estimates
LP.est   <- coef(fm.LP)[["xs"]]
QP.est   <- unname(coef(fm.QP)[["b1"]] / (2 * coef(fm.QP)[["b2"]]))
MIT.est  <- -log((1 - 0.95) * coef(fm.MIT)[["b0"]] / (coef(fm.MIT)[["b0"]] -
coef(fm.MIT)[["b1"]])) / coef(fm.MIT)[["b2"]]
QUA.est  <- unname(-coef(fm.QUA)[["input_level"]] / (2 *
coef(fm.QUA)[["I(input_level^2)"]]))

# Results
LP.est; QP.est; MIT.est; QUA.est


# Bootstrap uncertainty for each model

R <- 5000
n <- nrow(dat)

# Starts from full-data fits (more stable than re-deriving every time)
LP_start <- as.list(coef(fm.LP))
QP_start <- as.list(coef(fm.QP))
MIT_start <- as.list(coef(fm.MIT))
QUA_start <- as.list(coef(fm.QUA))

# LP  bootstrap: xs
LP.bt <- replicate(R, {
  db <- dat[sample.int(n, replace = TRUE), , drop = FALSE]

  fitb <- try(
    nlsLM(
      response ~ ifelse(input_level < xs,
                        a + b * input_level,
                        a + b * xs),
      data  = db,
      start = LP_start,
      control = nls.lm.control(maxiter = 1024)
    ),
    silent = TRUE
  )

  if (inherits(fitb, "try-error")) return(NA_real_)
  as.numeric(coef(fitb)[["xs"]])
})

LP.ci <- quantile(LP.bt, c(0.025, 0.975), na.rm = TRUE)
```

```
# QP bootstrap: xp = b1/(2*b2)
QP.bt <- replicate(R, {
  db <- dat[sample.int(n, replace = TRUE), , drop = FALSE]

  fitb <- try(
    nlsLM(
      response ~ ifelse(input_level < (b1/(2*b2)),
                        b0 + b1*input_level - b2*input_level^2,
                        b0 + b1*(b1/(2*b2)) - b2*(b1/(2*b2))^2),
      data  = db,
      start = QP_start,
      lower = c(b0 = -Inf, b1 = -Inf, b2 = 1e-10),
      control = nls.lm.control(maxiter = 1024)
    ),
    silent = TRUE
  )

  if (inherits(fitb, "try-error")) return(NA_real_)

  cf <- coef(fitb)
  b1 <- as.numeric(cf[["b1"]])
  b2 <- as.numeric(cf[["b2"]])

  if (!is.finite(b1) || !is.finite(b2) || b2 <= 0) return(NA_real_)

  xp <- b1 / (2 * b2)
  if (!is.finite(xp)) return(NA_real_)

  xp
})

QP.ci <- quantile(QP.bt, c(0.025, 0.975), na.rm = TRUE)

# MIT bootstrap: x at targ% of plateau
MIT.bt <- replicate(R, {
  db <- dat[sample.int(n, replace = TRUE), , drop = FALSE]

  fitb <- try(
    nlsLM(
      response ~ b0 + (b1 - b0) * exp(-b2 * input_level),
      data  = db,
      start = MIT_start,
      lower = c(b0 = -Inf, b1 = -Inf, b2 = 1e-10),
      control = nls.lm.control(maxiter = 1024)
    ),
    silent = TRUE
```

```
  )

  if (inherits(fitb, "try-error")) return(NA_real_)

  cf <- coef(fitb)
  b0 <- as.numeric(cf[["b0"]])
  b1 <- as.numeric(cf[["b1"]])
  b2 <- as.numeric(cf[["b2"]])

  if (!is.finite(b0) || !is.finite(b1) || !is.finite(b2) || b2 <= 0 || b0 <= b1) return(NA_real_)

  -log((1 - 0.95) * b0 / (b0 - b1)) / b2
})

MIT.ci <- quantile(MIT.bt, c(0.025, 0.975), na.rm = TRUE)

# QUA bootstrap: vertex = -b1/(2*b2)
QUA.bt <- replicate(R, {
  db <- dat[sample.int(n, replace = TRUE), , drop = FALSE]

  fitb <- try(
    lm(response ~ input_level + I(input_level^2), data = db),
    silent = TRUE
  )

  if (inherits(fitb, "try-error")) return(NA_real_)

  b1 <- as.numeric(coef(fitb)[["input_level"]])
  b2 <- as.numeric(coef(fitb)[["I(input_level^2)"]])

  if (!is.finite(b1) || !is.finite(b2) || b2 == 0) return(NA_real_)

  -b1 / (2 * b2)
})


QUA.ci <- quantile(QUA.bt, c(0.025, 0.975), na.rm = TRUE)

# Results
tibble(
  Model    = c("LP","QP","MIT","QUA"),
  Estimate = c(LP.est, QP.est, MIT.est, QUA.est),
  CI_2.5   = c(LP.ci[1], QP.ci[1], MIT.ci[1], QUA.ci[1]),
  CI_97.5  = c(LP.ci[2], QP.ci[2], MIT.ci[2], QUA.ci[2])
)
```

```
# AIC weights and AIC-WB optimum distribution

# Compute AIC weights

AICs <- c(
  LP  = AIC(fm.LP),
  QP  = AIC(fm.QP),
  MIT = AIC(fm.MIT),
  QUA = AIC(fm.QUA)
)

dAIC <- AICs - min(AICs)
w <- exp(-0.5 * dAIC) / sum(exp(-0.5 * dAIC))
w


# AIC-WB point estimate and CI

theta_aic_hat <- sum(w * c(LP = LP.est, QP = QP.est, MIT = MIT.est, QUA = QUA.est), na.rm
= TRUE)

theta_aic_bt <- w["LP"]  * LP.bt +
              w["QP"]  * QP.bt +
              w["MIT"] * MIT.bt +
              w["QUA"] * QUA.bt

theta_aic_ci <- quantile(theta_aic_bt, c(0.025, 0.975), na.rm = TRUE)

theta_aic_hat
theta_aic_ci
```

R code for Bagging

```
# Bagging (random model per resample)

# Define candidate response functions

# Linear Plateau Model
lin_p <- function(x, a, b, xs){
  if_else(x < xs,
          a + (b*x),
          a + (b*xs))
}

# Quadratic Plateau Model
quad_p <- function(x, b0, b1, b2){
  xp  <- b1/(2*b2)
  x_2 <- x^2
  xp_2 <- xp^2

  if_else(x < xp,
          b0 + (b1*x) - (b2*x_2),
          b0 + (b1*xp) - (b2*xp_2))
}

# Mitscherlich Model
mits <- function(x, b0, b1, b2){
  b0 + (b1 - b0) * exp(-b2*x)
}



# Bagging procedure

df_to_boot <- data.frame(
  input_level = dat$input_level,
  response    = dat$response
)

set.seed(123)
K <- 5000

df_boot <- replicate(
  K,
  df_to_boot[sample(nrow(df_to_boot), replace = TRUE), ],
  simplify = FALSE
)
```

```
for (i in seq_len(K)) {
  df_boot[[i]] <- df_boot[[i]] %>%
    mutate(boot_id = i, .before = input_level) %>%
    as_tibble()
}


# Fit random models to each bootstrap resample

set.seed(123)

model_choices <- c("LP", "QP", "MIT", "QUA")
model_probs   <- c(0.25, 0.25, 0.25, 0.25)

bagg_models <- vector("list", K)

for (i in seq_len(K)) {

  mtofit <- sample(model_choices, size = 1, prob = model_probs)

  fitM <- tryCatch(
    {
      if (mtofit == "LP") {

        nlsLM(
          response ~ lin_p(x = input_level, a, b, xs),
          data = df_boot[[i]],
          start = list(
            a  = unname(coef(lm(response ~ input_level, data = df_boot[[i]]))[1]),
            b  = unname(coef(lm(response ~ input_level, data = df_boot[[i]]))[2]),
            xs = unname(quantile(df_boot[[i]]$input_level, 0.6, na.rm = TRUE))
          ),
          control = nls.lm.control(maxiter = 1024)
        )

      } else if (mtofit == "QP") {

        nlsLM(
          response ~ quad_p(x = input_level, b0, b1, b2),
          data = df_boot[[i]],
          start = list(
            b0 = unname(coef(lm(response ~ input_level, data = df_boot[[i]]))[1]),
            b1 = unname(coef(lm(response ~ input_level, data = df_boot[[i]]))[2]),
            b2 = unname(abs(coef(lm(response ~ input_level, data = df_boot[[i]]))[2]) /
                          (2*unname(quantile(df_boot[[i]]$input_level, 0.6, na.rm = TRUE)) + 1e-8))
```

```
      ),
      control = nls.lm.control(maxiter = 1024)
    )

  } else if (mtofit == "MIT") {

    nlsLM(
      response ~ mits(x = input_level, b0, b1, b2),
      data = df_boot[[i]],
      start = list(
        b0 = unname(max(df_boot[[i]]$response, na.rm = TRUE)),
        b1 = unname(min(df_boot[[i]]$response, na.rm = TRUE)),
        b2 = 1 / unname(max(df_boot[[i]]$input_level, na.rm = TRUE))
      ),
      control = nls.lm.control(maxiter = 1024)
    )

  } else {

    lm(response ~ input_level + I(input_level^2), data = df_boot[[i]])
  }
 },
 error = function(e) NA
)

bagg_models[[i]] <- tryCatch(
 {
  if (identical(fitM, NA)) return(NA)

  if (mtofit == "QUA") {

    theta <- unname(-coef(fitM)[["input_level"]] / (2 * coef(fitM)[["I(input_level^2)"]]))

    tibble(
      model = mtofit,
      par1  = as.numeric(coef(fitM)[["(Intercept)"]]),
      par2  = as.numeric(coef(fitM)[["input_level"]]),
      par3  = as.numeric(coef(fitM)[["I(input_level^2)"]]),
      theta = as.numeric(theta)
    )

  } else {

    cf <- coef(fitM)

    theta <- if (mtofit == "LP") {
```

```
        as.numeric(cf[["xs"]])

      } else if (mtofit == "QP") {

        b1 <- as.numeric(cf[["b1"]])
        b2 <- as.numeric(cf[["b2"]])
        if (is.finite(b1) && is.finite(b2) && b2 > 0) b1/(2*b2) else NA_real_

      } else {

        b0 <- as.numeric(cf[["b0"]])
        b1 <- as.numeric(cf[["b1"]])
        b2 <- as.numeric(cf[["b2"]])

        if (!is.finite(b0) || !is.finite(b1) || !is.finite(b2) || b2 <= 0 || b0 <= b1) {
          NA_real_
        } else {
          -log((1 - 0.95) * b0 / (b0 - b1)) / b2
        }
      }

      tibble(
        model = mtofit,
        par1  = as.numeric(cf[1]),
        par2  = as.numeric(cf[2]),
        par3  = ifelse(length(cf) >= 3, as.numeric(cf[3]), NA_real_),
        theta = as.numeric(theta)
      )
    }
  },
  error = function(e) NA
 )
}

bagg_models_NONA <- bagg_models[
 !sapply(bagg_models, function(x) is.null(x) || identical(x, NA) || !inherits(x, "data.frame"))
]

df_bagg <- bind_rows(bagg_models_NONA)


# Bagging results

# Summaries by model (median + 95% CI)
theta_by_model <- df_bagg %>%
```

```
  dplyr::filter(is.finite(theta)) %>%
  dplyr::group_by(model) %>%
  dplyr::summarise(
    n        = dplyr::n(),
    estimate = median(theta, na.rm = TRUE),
    CI_2.5   = unname(quantile(theta, 0.025, na.rm = TRUE)),
    CI_97.5  = unname(quantile(theta, 0.975, na.rm = TRUE)),
    .groups  = "drop"
  ) %>%
  dplyr::arrange(factor(model, levels = c("LP","QP","MIT","QUA")))

theta_by_model

# Overall Bagging estimate (median + 95% CI)
theta_bag     <- df_bagg$theta
theta_bag_hat <- median(theta_bag, na.rm = TRUE)
theta_bag_ci  <- quantile(theta_bag, c(0.025, 0.975), na.rm = TRUE)

theta_bag_hat
theta_bag_ci
```

R code for BMA

```
# Bayesian model averaging

# Candidate Bayesian models in Stan

# Linear-Plateau model
stan_model_LinPl <- "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real<lower=0> b0;
  real<lower=0> b1;
  real<lower=0> tau;
  real<lower=0> sigma;
}
model {
  vector[N] mu;
  for(i in 1:N) {
    if (x[i] < tau) mu[i] = b0 + b1 * x[i];
    else            mu[i] = b0 + b1 * tau;
  }
  target += normal_lpdf(y | mu, sigma);
  target += gamma_lpdf(b0  | 6, 6/7317.0);
  target += gamma_lpdf(b1  | 6, 6/33.3333);
  target += gamma_lpdf(tau | 6, 6/150.0);
  target += inv_gamma_lpdf(sigma | 6, 1000.0 * (6 - 1));
}
generated quantities {
  real OPT;
  OPT = tau;
}
"

# Quadratic-Plateau model
stan_model_QP <- "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real<lower=0> b0;
  real<lower=0> b1;
```

```
  real<lower=0> b2;
  real<lower=0> sigma;
}
model {
  vector[N] mu;
  for(i in 1:N) {
    real xp = b1 / (2 * b2);
    if (x[i] < xp) mu[i] = b0 + b1 * x[i] - b2 * x[i]^2;
    else           mu[i] = b0 + b1 * xp - b2 * xp^2;
  }
  target += normal_lpdf(y | mu, sigma);
  target += gamma_lpdf(b0 | 6, 6/6967.0);
  target += gamma_lpdf(b1 | 6, 6/50.45);
  target += gamma_lpdf(b2 | 6, 6/0.121);
  target += inv_gamma_lpdf(sigma | 6, 1000.0 * (6 - 1));
}
generated quantities {
  real OPT;
  OPT = fmax(b1 / (2 * b2), 0);
}
"

# Mitscherlich model
stan_model_MIT <- "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real<lower=0> b1;
  real<lower=b1> b0;
  real<lower=0> b2;
  real<lower=0> sigma;
}
model {
  target += normal_lpdf(y |  b0 + (b1 - b0) * exp(-b2 * x), sigma);
  target += gamma_lpdf(b0 | 6, 6/12544.0);
  target += gamma_lpdf(b1 | 6, 6/6817.0);
  target += gamma_lpdf(b2 | 6, 6/0.012);
  target += inv_gamma_lpdf(sigma | 6, 1000.0 * (6 - 1));
}
generated quantities {
  real OPT;
  OPT = -log((1 - 0.95) * b0 / (b0 - b1)) / b2;
}
```

```
"

# Quadratic model
stan_model_QD <- "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real<lower=0> b0;
  real<lower=0> b1;
  real<lower=0> b2;
  real<lower=0> sigma;
}
model {
  target += normal_lpdf(y | b0 + (b1 * x) - (b2 * x^2), sigma);
  target += gamma_lpdf(b0 | 6, 6/7214.0);
  target += gamma_lpdf(b1 | 6, 6/43.0);
  target += gamma_lpdf(b2 | 6, 6/0.09);
  target += inv_gamma_lpdf(sigma | 6, 1000.0 * (6 - 1));
}
generated quantities {
  real OPT;
  OPT = b1 / (2 * b2);
}
"

# Fit functions

bayes_model_LinPl <- function(x, y,
                          param_saved = c("b0","b1","tau","OPT","sigma"),
                          chains = 4, iter = 4000, warmup = 2000,
                          cores = 1, apt_del = 0.95, thin = 5, seed = 123){
  df.list <- list(x = x, y = y, N = length(y))
  stan(model_code = stan_model_LinPl,
       data = df.list,
       chains = chains, iter = iter, warmup = warmup,
       cores = cores, control = list(adapt_delta = apt_del),
       thin = thin, seed = seed, pars = param_saved)
}

bayes_model_QP <- function(x, y,
                       param_saved = c("b0","b1","b2","OPT","sigma"),
                       chains = 4, iter = 4000, warmup = 2000,
                       cores = 1, apt_del = 0.95, thin = 5, seed = 123){
```

```
  df.list <- list(x = x, y = y, N = length(y))
  stan(model_code = stan_model_QP,
       data = df.list,
       chains = chains, iter = iter, warmup = warmup,
       cores = cores, control = list(adapt_delta = apt_del),
       thin = thin, seed = seed, pars = param_saved)
}

bayes_model_MIT <- function(x, y,
                        param_saved = c("b0","b1","b2","OPT","sigma"),
                        chains = 4, iter = 4000, warmup = 2000,
                        cores = 1, apt_del = 0.95, thin = 5, seed = 123){
  df.list <- list(x = x, y = y, N = length(y))
  stan(model_code = stan_model_MIT,
       data = df.list,
       chains = chains, iter = iter, warmup = warmup,
       cores = cores, control = list(adapt_delta = apt_del),
       thin = thin, seed = seed, pars = param_saved)
}

bayes_model_QD <- function(x, y,
                       param_saved = c("b0","b1","b2","OPT","sigma"),
                       chains = 4, iter = 4000, warmup = 2000,
                       cores = 1, apt_del = 0.95, thin = 5, seed = 123){
  df.list <- list(x = x, y = y, N = length(y))
  stan(model_code = stan_model_QD,
       data = df.list,
       chains = chains, iter = iter, warmup = warmup,
       cores = cores, control = list(adapt_delta = apt_del),
       thin = thin, seed = seed, pars = param_saved)
}

# Fit all candidate models

bay.mod_LP  <- bayes_model_LinPl(x = dat$input_level, y = dat$response, cores = mc.cores)
bay.mod_QP  <- bayes_model_QP   (x = dat$input_level, y = dat$response, cores = mc.cores)
bay.mod_MIT <- bayes_model_MIT  (x = dat$input_level, y = dat$response, cores = mc.cores)
bay.mod_QD  <- bayes_model_QD   (x = dat$input_level, y = dat$response, cores = mc.cores)

# Posterior model probabilities via bridge sampling

LP_marglog  <- bridgesampling::bridge_sampler(bay.mod_LP,  silent = TRUE)
QP_marglog  <- bridgesampling::bridge_sampler(bay.mod_QP,  silent = TRUE)
MIT_marglog <- bridgesampling::bridge_sampler(bay.mod_MIT, silent = TRUE)
QD_marglog  <- bridgesampling::bridge_sampler(bay.mod_QD,  silent = TRUE)
```

```
m_prob <- bridgesampling::post_prob(
  LP_marglog, QP_marglog, MIT_marglog, QD_marglog,
  prior_prob = c(0.25, 0.25, 0.25, 0.25)
)

m_prob

# BMA posterior for the optimum

extract_OPT <- function(fit) rstan::extract(fit, pars = "OPT", permuted = TRUE)[["OPT"]]

draws_LP  <- extract_OPT(bay.mod_LP)
draws_QP  <- extract_OPT(bay.mod_QP)
draws_MIT <- extract_OPT(bay.mod_MIT)
draws_QD  <- extract_OPT(bay.mod_QD)

S <- min(length(draws_LP), length(draws_QP), length(draws_MIT), length(draws_QD))
draws_LP  <- draws_LP [seq_len(S)]
draws_QP  <- draws_QP [seq_len(S)]
draws_MIT <- draws_MIT[seq_len(S)]
draws_QD  <- draws_QD [seq_len(S)]

set.seed(123)
model_ids <- sample(
  c("LP","QP","MIT","QD"),
  size = S, replace = TRUE,
  prob = as.numeric(m_prob)
)

theta_bma <- numeric(S)
theta_bma[model_ids == "LP"]  <- sample(draws_LP,  sum(model_ids == "LP"),  replace =
TRUE)
theta_bma[model_ids == "QP"]  <- sample(draws_QP,  sum(model_ids == "QP"),  replace =
TRUE)
theta_bma[model_ids == "MIT"] <- sample(draws_MIT, sum(model_ids == "MIT"), replace =
TRUE)
theta_bma[model_ids == "QD"]  <- sample(draws_QD,  sum(model_ids == "QD"),  replace =
TRUE)

theta_bma_hat <- median(theta_bma, na.rm = TRUE)
theta_bma_ci  <- quantile(theta_bma, c(0.025, 0.975), na.rm = TRUE)

theta_bma_hat
theta_bma_ci
```