Calibration of Inexact Computer Models with Heteroscedastic Errors*

Chih-Li Sung †, Beau David Barber ‡, and Berkley J. Walker §

Abstract. Computer models are commonly used to represent a wide range of real systems, but they often involve some unknown parameters. Estimating the parameters by collecting experimental data becomes essential in many scientific fields, ranging from engineering to biology. However, most of the existing methods are developed under the assumption that the experimental data contains homoscedastic measurement errors. Motivated by an experiment of plant relative growth rates where replicates are available, we propose a new calibration method for inexact computer models with heteroscedastic measurement errors. Asymptotic properties of the parameter estimators are derived which can be used to quantify the uncertainty of the estimates, and a goodness-of-fit test is developed to detect the presence of heteroscedasticity. Numerical examples and empirical studies demonstrate that the proposed method not only yields accurate parameter estimation, but also provides accurate predictions for physical data in the presence of both heteroscedasticity and model misspecification. An R package for the proposed methodology is provided in an open repository.

Key words. Gaussian process; Input-dependent noise; Plant biology; Replication; Uncertainty quantification.

AMS subject classifications. 00A20

1. Introduction.

Computer models, which use mathematical representations to simulate real systems, have been widely adopted to understand a real-world feature, phenomenon or event. The applications of computer models range from economics to the physical and biological sciences. For instance, a computer model based on mass action reaction mechanisms is developed in [12] to simulate the carbon flow in plant cell wall synthesis networks and to understand the patterns of carbon flow and their regulation, which underpin plant growth and development. A computer model often contains some unknown parameters that represent certain inherent attributes of the underlying systems but cannot be directly controlled or measurable in its physical experiment, which are called *calibration parameters* in the literature [48, 20]. When its physical experiment is available, these parameters are used to *calibrate* the computer model such that the model simulations agree with characteristics observed in the physical experiment. This process is called *calibration*, and it is of great importance for computer modelers because it not only improves the model prediction, but the estimated value of the calibration parameters also provides some scientific insight which can help modelers better understand the system. For example, the parameters in the cell adhesion study of [52] include kinetic rates and their estimated values provide the information of molecular interactions in the biological system.

This paper is motivated by the preponderance of computer models used to interpret bi-

*Submitted to the editors on May 5, 2021.

Funding: The authors recognize funding by the Division of Chemical Sciences, Geosciences and Biosciences, Office of Basic Energy Sciences of the U.S. Department of Energy Grant DE-FG02-91ER20021 (B.W.).

†Department of Statistics and Probability, Michigan State University (sungchih@msu.edu).

‡Agricultural and Biological Engineering, University of Illinois at Urbana-Champaign (bdbarbe2@illinois.edu).

§Plant Research Laboratory/Plant Biology, Michigan State University (berkley@msu.edu).

ological data in plant biology. For example, the underlying biochemistry facilitating carbon fixation by plants can be determined by calibrating a computer model with rates of photosynthesis measured as a function of light intensity or carbon dioxide concentration [49, 56]. Computer models are also used to quantify plant metabolic fluxes [37]. Oftentimes gathering and interpreting data from plant science experiments face the same challenges when used in calibration approaches: (1) models are inexact or imperfect due to simplifications or incomplete understanding of the system, and (2) data contain heteroscedastic variance due to limitations in measurement approaches and variability in plant development. A real problem in plant biology involving inexact models and heteroscedastic errors will be illustrated in Section 5. Apart from plant biology, many calibration problems in various fields, such as engineering and astronomy, also face these two challenges. See, for example, [36] where a sinusoidal model is used to estimate the period of a single periodic variable star but cannot perfectly represent the light curve shape. In these problems, the weighted least-squares (WLS) estimator is typically used to estimate the calibration parameters. The WLS estimator, however, can converge to an undesirable value of the parameters when the computer model is inexact and the error is input-dependent. More details can be found in [36].

In this paper, *we develop a new calibration framework for inexact computer models with replicated experiments potentially having input-dependent errors*, where a new statistical model is introduced that enables to estimate the calibration parameters and produce predictions in the face of heteroscedasticity. Moreover, the theoretical properties, including the asymptotic distribution of the estimator and the goodness-of-fit test to detect the presence of heteroscedasticity, are developed. Although there has been much work on calibration problems for inexact computer models in the statistics literature (e.g., [31, 54, 44, 26]), these methods are developed under a homoscedastic assumption, which may in turn lead to faulty inferences in the presence of heteroscedasticity. On the other hand, recent study by [36] proposes an adaptive estimator which accounts for heteroscedasticity and has lower asymptotic variance than ordinary least-square and WLS estimators. This method, however, is limited to a linear computer model and requires the assumption that the variances are independent of input variables.

It is worth noting that the focus of this paper is on input-dependent noises in the *physical experiments*, and the computer simulations can be either deterministic or stochastic, while recent studies, such as [1], [6] and [2], develop heteroscedastic models for tackling the heteroscedasticity inherent in stochastic computer simulations. The extension is non-trivial, as an accurate estimation for calibration parameters needs to account for the model inadequacy of computer models, while the methods therein focus on accurate predictions for expensive computer simulations. In addition, the theoretical properties of the parameter estimator are of great interest which can be used to quantify the uncertainty of the estimates, but have not been systematically studied.

The remainder of this paper is organized as follows. In Section 2, a heteroscedastic model is introduced, and the estimation procedure for the calibration parameters and the hyperparameters in the model is developed. The asymptotic distribution of the parameter estimator and the goodness-of-fit of the heteroscedastic model are presented in Section 3. Synthetic examples are illustrated in Section 4. The proposed framework is applied to the case study of plant relative growth rates in Section 5. Concluding remarks are given in Section 6. Estimation details, mathematical proofs, an `R` [46] package, `HetCalibrate` [51], and the `R` code for

implementation are provided in Supplemental Materials.

2. Heteroscedastic Modeling for Calibration Problems under Replication. Suppose that N observations are collected from the physical experiments, denoted by $y_1, \ldots, y_N$, and their corresponding inputs are $x_1, \ldots, x_N$, where $x_i \in \chi \subseteq \mathbb{R}^d$. In the case of replication, we further denote $\bar{x}_i$, $i = 1, \ldots, n$ as the n distinct input locations, where $n < N$, and $y_i^{(j)}$ as the j-th output out of $a_i \geq 1$ replicates at the distinct location $\bar{x}_i$, and denote its sample mean, $\sum_{j=1}^{a_i} y_i^{(j)}/a_i$, as $\bar{y}_i$. Furthermore, denote $f(x, \theta)$ as the computer model output, which is a function of the input $x \in \chi \subseteq \mathbb{R}^d$ and the calibration parameter $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^q$. When computer simulations are computationally demanding, such as the high-fidelity simulation in [38], a common approach is to run a computer experiment with various inputs and build a cheaper *emulator* for the actual computer simulations, for which Gaussian process modeling is often used [47, 48, 20]. Then, the calibration problem for inexact computer models with heteroscedastic errors can be represented as follows,

$$y(x_i) = \zeta(x_i) + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.1}$$

where $\zeta(x_i) = f(x_i, \theta) + b_\theta(x_i)$, and ϵ_i is the measurement or some stochastic error from the real system and independently follows an normal distribution with zero mean and variance $\mathbb{V}[\epsilon_i] = r(x_i)$, i.e., $\epsilon_i \sim \mathcal{N}(0, r(x_i))$. The function $\zeta(\cdot)$ is the *true process* of the real system, and the function $b_\theta(\cdot)$ is the *discrepancy* (or *bias*) between the true process and the computer model. The inclusion of the discrepancy term in the model is necessary because the computer models are often considered *inexact* or *imperfect*, meaning that even with an optimal calibration parameter, the computer model does not perfectly match the true process, which is referred to as *model inadequacy* in the literature [31]. Note that the dependence of $b_\theta(\cdot)$ on θ is often suppressed in the literature, but it is included here for clarity. It is also worth noting that when $r(x_i)$ is assumed to be constant, this model is a special case of [31]. When $f(x, \theta)$ is a constant mean or a linear function and $b_\theta(\cdot)$ follows a Gaussian process model that is independent of θ, the heteroscedastic model is closely related to the models of [1] and [6], where their primary objective is to emulate stochastic simulations, whereas our focus here is on estimation and inference of the calibration parameters as well as statistical correction of model predictions.

The inexact computer models were first discussed by [31] which model the model discrepancy as a Gaussian process (GP) model, and this method has been widely used in many applications (e.g., [27, 26, 58, 23]). Similar to [31], we assume that the distribution of $b_\theta(\cdot)$ is represented by a GP with zero mean and a positive-definite covariance function c, so that $b_\theta(x_1), \ldots, b_\theta(x_N)$ is an N-dimensional multivariate normal distribution with zero mean and covariance matrix $(c(x_i, x_j))_{1 \leq i \leq j \leq N}$. A scale $\nu > 0$ is commonly separated from a kernel function, $c(x_i, x_j) = \nu k(x_i, x_j; \boldsymbol{\varphi})$, where $\boldsymbol{\varphi}$ are hyperparameters of the kernel. That is,

$$\mathbf{b}_\theta \sim \mathcal{N}_N(\mathbf{0}_N, \nu \mathbf{K}_N),$$

where $\mathbf{b}_\theta := (b_\theta(x_1), \ldots, b_\theta(x_N))^T$, $\mathbf{0}_N$ is a zero vector of length N, and $\mathbf{K}_N$ is an $N \times N$ matrix with ij elements $k(x_i, x_j; \boldsymbol{\varphi})$. The dependency of $\boldsymbol{\varphi}$ will be suppressed in the rest of the paper for notational simplicity. Typical choices of the kernel function are Gaussian or

Matérn kernels which are independent of θ. However, recent studies (e.g., [21, 54, 55, 44]) indicate that these choices of kernel functions may lead to unreasonable calibration parameter estimation due to *unidentifiability* of the calibration parameters. Therefore, here we consider an *orthogonal kernel function* [44] to avoid the identifiability issue, which will be introduced in Section 2.2.

Thus, given the noise function $r(x)$, the observations $\mathbf{Y}_N = (y_1, \ldots, y_N)$ follow a multivariate normal distribution,

$$\mathbf{Y}_N \sim \mathcal{N}_N(\mathbf{f}(\theta), \nu(\mathbf{K}_N + \mathbf{\Lambda}_N)),$$

where $\mathbf{f}(\theta) = (f(x_1, \theta), \ldots, f(x_N, \theta))^T$, and $\mathbf{\Lambda}_N$ is an $N \times N$ diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_N$, where $\lambda_i = r(x_i)/\nu$. Based on the properties of conditional multivariate normal distributions, the predictive distribution of $y(x)$ at a new input setting x, $y(x)|\mathbf{Y}_N$, is a normal distribution, $\mathcal{N}(\mu(x), \sigma^2(x))$, where

$$\mu(x) = f(x, \theta) + \mathbf{k}(x)^T(\mathbf{K}_N + \mathbf{\Lambda}_N)^{-1}(\mathbf{Y}_N - \mathbf{f}(\theta))$$

and

$$\sigma^2(x) = \nu k(x, x) + r(x) - \nu \mathbf{k}(x)^T(\mathbf{K}_N + \mathbf{\Lambda}_N)^{-1}\mathbf{k}(x), \tag{2.2}$$

where $\mathbf{k}(x) = (k(x, x_1), \ldots, k(x, x_N))^T$.

The main difference of the model from the others in the literature lies in the the heteroscedastic error ϵ_i whose variance $r(x)$ is non-constant, while typical homoscedastic cases consider a constant variance, $r(x_i) = \tau^2$. This assumption makes the calibration problem more challenging, because the noise function $r(x)$ is unknown which needs to be estimated. A straightforward and sensible estimate is the sample variance of the replicates at each distinct location, that is, $\hat{r}(\bar{x}_i) = \sum_{j=1}^{a_i}(y_i^{(j)} - \bar{y}_i)^2/(a_i - 1)$, The estimate $\hat{r}(\bar{x}_i)$ was also used in [1] where they fit a GP model on the pairs $(\bar{x}_i, \hat{r}(\bar{x}_i))$. This estimate, however, often requires a minimal number of replicates. For example, [1] recommends $a_i \geq 10$ replicates for fitting a stochastic GP while [59] recommends $a_i \geq 5$. This is impractical in many applications because physical data from a real system is often time-consuming or too costly to collect. In addition, the predictive variance (2.2) still requires the value of $r(x)$ at the new input setting x but physical data at this input setting is not directly available for estimation. To this end, we employ a latent log-variance GP similar to [6] to model the noise function $r(\cdot)$, which does not require a minimal number of the replicates and is computationally efficient under replication.

2.1. Latent Variable Process for Modeling $r(\cdot)$. Since $r(\cdot) = \nu\lambda(\cdot)$, we instead model $\lambda(\cdot)$ and then $r(\cdot)$ can be obtained by multiplying the scale ν. Denote $\mathbf{\Lambda}_n = (\lambda(\bar{x}_1), \ldots, \lambda(\bar{x}_n))$ and $\mathbf{A}_n = \text{diag}(a_1, \ldots, a_n)$. Similar to [19] and [6], we model $\log\lambda_1, \ldots, \log\lambda_n$ and assume that their quantities are obtained via the predictive mean of a regularizing GP on new latent variables, $\delta_1, \ldots, \delta_n$. That is, assume $\delta_i = \delta_0(\bar{x}_i) + \eta_i$, where $\eta_i \overset{ind.}{\sim} \mathcal{N}(0, g\nu_{(g)}/a_i)$ and $\delta_0(\cdot)$ is a GP with zero mean and a positive-definite covariance function, $\nu_{(g)}k_{(g)}(\cdot, \cdot)$, where $\nu_{(g)}$ is a positive scale and $k_{(g)}$ is a kernel function that contains some hyperparameters, which are denoted by $\boldsymbol{\phi}$. Typical kernels such as Gaussian or Matérn kernels can be used for $k_{(g)}$. Thus,

denote $\boldsymbol{\Delta}_n = (\delta_1, \ldots, \delta_n)$, it follows that

$$\boldsymbol{\Delta}_n \sim \mathcal{N}_n(0, \nu_{(g)}(\mathbf{K}_{(g)} + g\mathbf{A}_n^{-1})),$$

and assume that $\log \lambda_i$ is the predictive mean, which gives

$$\log \boldsymbol{\Lambda}_n = \mathbf{K}_{(g)} \left(\mathbf{K}_{(g)} + g\mathbf{A}_n^{-1}\right)^{-1} \boldsymbol{\Delta}_n,$$

where $\mathbf{K}_{(g)} = \left(k_{(g)}(\bar{x}_i, \bar{x}_j)\right)_{1 \le i,j \le n}$. The form of $\log \boldsymbol{\Lambda}_n$ can be viewed as a smoother of the latent δ_i's, which are unknown and treated as additional parameters that will be estimated in Section 2.3 along with $\boldsymbol{\phi}$ and nugget g. The predictive value of $\log \lambda(x)$ at an new input x can then be obtained by $\log \lambda(x) = \mathbf{k}_{(g)}(x)^T \left(\mathbf{K}_{(g)} + g\mathbf{A}_n^{-1}\right)^{-1} \boldsymbol{\Delta}_n$, where $\mathbf{k}_{(g)}(x) = (k_{(g)}(x, \bar{x}_1), \ldots, k_{(g)}(x, \bar{x}_n))^T$.

2.2. Orthogonal Gaussian Process for Modeling $b_\theta(\cdot)$. Although the GP modeling of [31] (referred to as KO in the rest of the paper) for $b_\theta(\cdot)$ has been widely used, recent studies have raised concerns about its identifiability issue of the calibration parameters [35, 3, 4, 23, 21, 54, 55, 44, 60]. In particular, [54, 55, 60] define the true calibration parameter as

$$\theta^* = \arg\min_{\theta \in \Theta} \|\zeta(\cdot) - f(\cdot, \theta)\|^2_{L_2(\chi)}, \tag{2.3}$$

which minimizes the L_2 distance between the true process and the computer model, where $\|g\|_{L_2(\chi)} = \left(\int_\chi g(x)^2 \mathrm{d}x\right)^{1/2}$, and [54, 55] show that the KO estimator is asymptotically inconsistent with the true parameter, which could lead to poor approximations to physical systems. [44] further points out that the GP modeling of the bias $b_\theta(\cdot)$ should depend on θ, but the one in KO does not. Therefore, [44] provides an alternative GP modeling by *orthogonalizing* the model bias to avoid mixing the GP and the definition of the parameter, the idea of which is to create an *orthogonal kernel function* satisfying the condition

$$\int_\chi \frac{\partial}{\partial \theta} f(\xi, \theta) b_\theta(\xi) \mathrm{d}\xi,$$

which is shown to be a necessary condition to minimize the L_2 distance in (2.3). Note that while other modeling approaches may be considered here, such as [53] and [61], the orthogonal GP can be naturally adopted in a maximum likelihood or Bayesian framework, as the latent variable process $\delta(\cdot)$ is also a (regularizing) GP. More details regarding the potential alternatives will be discussed in Section 6.

An orthogonal kernel function has the form of

$$k(x_i, x_j) = k_0(x_i, x_j) - h_\theta(x_i)^T H_\theta^{-1} h_\theta(x_j), \tag{2.4}$$

where $k_0(\cdot, \cdot)$ is any valid kernel function on $\chi \times \chi$ and is independent of θ, such as Gaussian or Matérn kernels,

$$h_\theta(x) = \int_\chi \frac{\partial}{\partial \theta} f(\xi, \theta) k_0(x, \xi) \mathrm{d}\xi$$

and

$$H_\theta = \int_\chi \int_\chi \frac{\partial}{\partial\theta} f(\xi_1, \theta) \left(\frac{\partial}{\partial\theta} f(\xi_2, \theta) \right)^T k_0(\xi_1, \xi_2) \mathrm{d}\xi_1 \mathrm{d}\xi_2.$$

Note that the orthogonal kernel function $k(\cdot, \cdot)$ is dependent of θ but here it is suppressed for notational simplicity.

In practice, the integrals in h_θ and H_θ can be difficult to solve. This can be addressed by the stochastic average approximation, such as Monte Carlo integration [11]. For example, one can draw m uniform samples, $\xi_1, \ldots, \xi_m$, and then approximate $h_\theta(x)$ by

$$h_\theta(x) \approx \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial\theta} f(\xi_i, \theta) k_0(x, \xi_i).$$

When the gradient $\partial f(x, \theta)/\partial\theta$ is not available, it can be approximated by numerical methods, such as Richardson's extrapolation [34, 16]. When the evaluations of the computer model, $f(x, \theta)$, at any input pair $(x, \theta) \in \chi \times \Theta$ are computationally too demanding, a GP emulator is often used [47, 48], and the predictive distribution of the emulator can be taken as a fixed probabilistic definition of $f(\cdot, \cdot)$. Hence, the definition of orthogonal kernel function can be modified accordingly. We refer more details to [44].

2.3. Parameter Estimation. The estimation procedure for the model parameters herein is based on maximum likelihood estimation. This procedure is developed along the lines described in [6], which develop computationally efficient inference and prediction for a heteroscedastic GP when replication is present. The model parameters include the calibration parameters θ, the hyperparameters of the two GPs, $\boldsymbol{\varphi}, \boldsymbol{\phi}, g$, and the latent variables $\delta_1, \ldots, \delta_n$. Conditional on the parameters $\theta, \boldsymbol{\varphi}, \boldsymbol{\phi}, g, \delta_1, \ldots, \delta_n$, the scales ν and $\nu_{(g)}$ both have plug-in MLEs: $\hat{\nu} = N^{-1} \left(\mathbf{Y}_N - \mathbf{f}(\theta) \right)^T \left(\mathbf{K}_N + \boldsymbol{\Lambda}_N \right)^{-1} \left(\mathbf{Y}_N - \mathbf{f}(\theta) \right)$ and $\hat{\nu}_{(g)} = n^{-1} \boldsymbol{\Delta}_n^T \left(\mathbf{K}_{(g)} + g\mathbf{A}_n^{-1} \right)^{-1} \boldsymbol{\Delta}_n$. The log-likelihood conditional on $\hat{\nu}$ and $\hat{\nu}_{(g)}$ is then

$$\begin{aligned} \log L = &- \frac{N}{2} \log 2\pi - \frac{N}{2} \log \hat{\nu} - \frac{1}{2} \log \left| \mathbf{K}_N + \boldsymbol{\Lambda}_N \right| - \frac{N}{2} \\ &- \frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\nu}_{(g)} - \frac{1}{2} \log \left| \mathbf{K}_{(g)} + g\mathbf{A}_n^{-1} \right| - \frac{n}{2}, \end{aligned} \tag{2.5}$$

where the top line above is the mean-field component and the bottom line is the variance-field component. While optimizing the log-likelihood can be computationally demanding when N is large, because the inverse and determinant of $\mathbf{K}_N + \boldsymbol{\Lambda}_N$ requires $O(N^3)$ computations, the computation complexity can be efficiently reduced from $O(N^3)$ to $O(n^3 + N)$ by the Woodbury identity [25], which essentially only depends on the number of the distinct input locations. Moreover, since the gradient of the log-likelihood is available in a closed form, the parameters can be efficiently estimated by maximizing the log-likelihood via a Newton-like optimization algorithm, such as the quasi-Newton optimization [10]. We leave the computational details to Supplemental Materials SM1. An `R` package, `HetCalibrate` [51], is available in an open repository for the parameter estimation, which is via modifications to the source code of `R` package `hetGP` [5].

3. Inference and Goodness-of-fit. In this section, the asymptotic distribution for the maximum likelihood estimators is studied in Section 3.1, which is crucial for calibration problems because the inference of the calibration parameter is often of great interest, and a goodness-of-fit statistic is introduced in Section 3.2 to detect the presence of heteroscedasticity. These theoretical results will be applied to the case study in Section 5.

3.1. Asymptotic Distribution for the Estimator of the Calibration Parameter. Denote all the model parameters as $\boldsymbol{\omega} = (\theta, \boldsymbol{\psi}, \nu, \boldsymbol{\phi}, g, \nu_{(g)}, \delta_1, \ldots, \delta_n)$ and their estimators as $\hat{\boldsymbol{\omega}}_N$. Asymptotic results are presented here to show that $\hat{\boldsymbol{\omega}}_N$ is asymptotically normally distributed as both n and N become sufficiently large. The regularity conditions and proofs are given in Supplemental Materials SM2.

Theorem 3.1. *Under the regularity conditions in Supplemental Materials SM2, the maximum likelihood estimators $\hat{\boldsymbol{\omega}}_N$ are asymptotically consistent and normal as $n \to \infty$ and $N \to \infty$,*

$$\mathbf{B}_N(\boldsymbol{\omega})^{1/2}(\hat{\boldsymbol{\omega}}_N - \boldsymbol{\omega}) \xrightarrow{d} \mathcal{N}(0, I_m),$$

where I_m is the $m \times m$ identity matrix, m is the size of the vector $\boldsymbol{\omega}$, and $\mathbf{B}_N(\boldsymbol{\omega})$ is the information matrix whose closed-form expression is provided in Supplemental Materials SM2.

Hence, by the theorem, an approximate $(1-\alpha)\times 100\%$ confidence region of the calibration parameter θ can be constructed as

$$\left\{\theta \in \Theta \subseteq \mathbb{R}^q | \, (\mathbf{U}\hat{\boldsymbol{\omega}}_N - \theta)^T \left(\mathbf{U}\mathbf{B}_N(\hat{\boldsymbol{\omega}}_N)^{-1}\mathbf{U}^T\right)^{-1} (\mathbf{U}\hat{\boldsymbol{\omega}}_N - \theta) \leq \chi^2_{q,1-\alpha}\right\},$$

where $\mathbf{U}$ is a $q \times m$ matrix composed of the first through the q-th row of the m-dimensional identity matrix, and $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-squared distribution with q degrees of freedom.

3.2. Goodness of Fit: Heteroscedasticity Test. As the main assumption of the proposed model is the heteroscedastic assumption, to avoid overparameterization, it is essential to develop a hypothesis test to detect the presence of heteroscedasticity. There are a variety of test procedures proposed in the literature to detect heteroscedasticity. See, for example, [28, 24, 18, 32, 41]. However, these procedures are developed under the assumption that the specification of the regression function, or the computer model in our context, is correct. These test procedures may falsely indicate the presence of heteroscedasticity if the computer model is incorrect. One exception is the heteroscedastic test of [33], which is robust to the regression function misspecification. This test, however, is limited to detect a linear specification of measurement errors.

Given the model proposed herein, we can provide a more flexible heteroscedastic test for an inexact computer model. In particular, in the proposed model we have $r(x) = \nu\lambda(x)$ and $\log\lambda(x) = \mathbf{k}_{(g)}(x)^T \left(\mathbf{K}_{(g)} + g\mathbf{A}_n^{-1}\right)^{-1} \boldsymbol{\Delta}_n$. Then, under homoscedasticity, the variance function $r(x)$ is constant over x, which implies that all of the latent variables, $\delta_1, \ldots, \delta_n$, are equal to zero simultaneously. Therefore, a testable hypothesis to detect heteroscedasticity is

$$H_0 : \delta_1 = \cdots = \delta_n = 0 \quad \text{v.s.} \quad H_1 : \text{ at least one } \delta_i \text{ is non-zero.}$$

Based on the asymptotic results of Theorem 3.1, a test statistic for the null hypothesis is given in the next theorem. The proof can be done by Slutsky's theorem.

Theorem 3.2. *Under the regularity conditions in Theorem 3.1, the heteroscedastic test statistic*

$$(\mathbf{H}\hat{\boldsymbol{\omega}}_N)^T \left(\mathbf{H}\mathbf{B}_N(\boldsymbol{\omega})^{-1}\mathbf{H}^T\right)^{-1} (\mathbf{H}\hat{\boldsymbol{\omega}}_N) \xrightarrow{d} \chi^2_n$$

for n and N sufficiently large under the null hypothesis, where $\mathbf{H}$ is an $n \times m$ matrix composed of the last n rows of the m-dimensional identity matrix.

In a finite sample application, $\mathbf{B}_N(\boldsymbol{\omega})$ can be estimated by $\mathbf{B}_N(\hat{\boldsymbol{\omega}}_N)$, which can be shown to be consistent under the regularity conditions. Thus, with Theorem 3.2, an approximate level-α test is given by rejecting H_0 when

$$(\mathbf{H}\hat{\boldsymbol{\omega}}_N)^T \left(\mathbf{H}\mathbf{B}_N(\hat{\boldsymbol{\omega}}_N)^{-1}\mathbf{H}^T\right)^{-1} (\mathbf{H}\hat{\boldsymbol{\omega}}_N) \geq \chi^2_{n,1-\alpha}.$$

4. Numerical Study.

In this section, numerical experiments are conducted to examine the calibration performance of the proposed method, including one computer model with one calibration parameter and one with three calibration parameters. In the implementation of the proposed method, Matérn kernels with the smoothness parameter 5/2 are chosen for $k_{(g)}$, which have the form

$$k_{(g)}(x,y) = \left(1 + \frac{\sqrt{5}}{\phi}\|x-y\| + \frac{5}{3\phi^2}\|x-y\|^2\right) \exp\left(-\frac{\sqrt{5}}{\phi}\|x-y\|\right),$$

and the Matérn kernels with hyperparameter $\boldsymbol{\varphi}$ are chosen for k_0 in (2.4) to derive the orthogonal kernel k.

4.1. Example with One Calibration Parameter.

We consider an example adapted from [54]. Assume that the input x is uniformly distributed on $[0, 2\pi]$, the true process is $\zeta(x) = \exp(x/10)\sin x$, and the observations are given by $y_i = \zeta(x_i) + \epsilon_i$, where ϵ_i is independently normally distributed with zero mean and the variance $r(x_i) = (0.01 + 0.2(x_i - \pi)^2)^2$. Suppose that the computer output is given by the function $f(x,\theta) = \zeta(x) - \sqrt{\theta^2 - \theta + 1}(\sin\theta x + \cos\theta x)$. There does not exist a real number θ such that $f(\cdot,\theta) = \zeta(\cdot)$ because $\sqrt{\theta^2 - \theta + 1}(\sin\theta x + \cos\theta x)$ is always positive for any θ. Thus, this computer model is *inexact* because even with the optimal setting θ^*, there still exists discrepancy between $f(\cdot,\theta^*)$ and $\zeta(\cdot)$. The true parameter θ^* can be calculated as in (2.3), which is $\theta^* \approx -0.1789$.

In this numerical study, eight distinct input locations are selected with equal space in $[0, 2\pi]$, and 5 replicates are generated at each distinct location, that is, $a_1 = \ldots = a_8 = 5$. Figure 1 demonstrates the simulated data, in which three different methods are performed, which are: (left) the WLS estimator; (middle) the homoscedastic modeling, which is the frequentist version of the KO approach; (right) our proposed method. The calibration parameter estimates are -0.2784, 0.2674, and -0.1727, respectively. In this example, our proposed method provides a more accurate parameter estimate (the true parameter is $\theta^* \approx -0.1789$), which also can be seen from the upper panels, where the computer model with the estimate is closer to the true process than other two methods in the sense of L_2 distances. Figure 1 also shows that the WLS yields inaccurate predictions for physical data, and the KO approach

suggests unreasonably wide prediction intervals due to the constant variance assumption. On the other hand, the proposed method not only provides a more accurate parameter estimate, but it also provides more accurate predictions as well as more reasonable prediction intervals by recovering the variance process.

We conduct the simulation 100 times to examine the performance of our proposed method (labeled HetOGP), in comparison with the weighted least-squares (labeled WLS), homoscedastic modeling with a Matérn kernel (labeled HomGP), which is the frequentist version of the KO approach, homoscedastic modeling with an orthogonal kernel (labeled HomOGP), which is the frequentist version of the calibration approach in [44], and heteroscedastic modeling with a Matérn kernel (labeled HetGP), which is close to the model in [6]. Three main metrics are used for the comparison: (i) estimation bias, $\hat{\theta} - \theta^*$; (ii) root mean squared errors (RMSEs) based on 101 test equal-spaced locations in $[0, 1]$, $\left(\sum_{i=1}^{101}(\zeta(x_i) - \hat{y}(x_i))^2/101\right)^{1/2}$, where $\hat{y}(x)$ is the prediction mean for the input x; (iii) predictive score, which is a scoring rule provided by Equation (27) of [17] that combines prediction means and variances.

Figure 2 shows the results for the five methods based on the 100 simulations. The predictive score of WLS is not available because the prediction variances at unobserved input locations are not available for the WLS. First, from the left panel, it can be seen that our proposed method (HetOGP) outperforms the other methods in terms of the calibration pa-

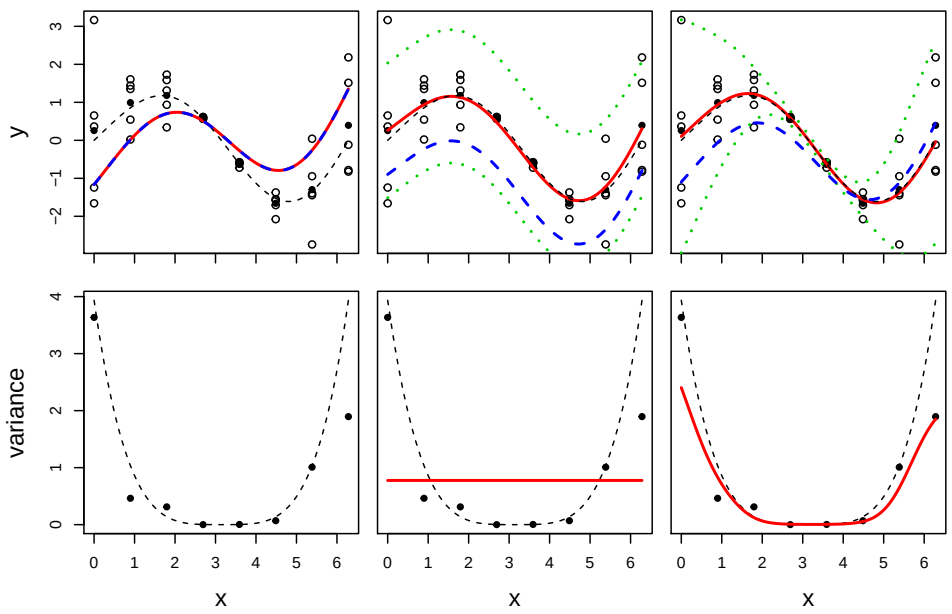

Figure 1. *Illustration of three methods: (left) WLS; (middle) KO; (right) the proposed method. Upper panels represent the replicates as open circles with the averaged observation $\bar{y}_i$ in filled circles at each distinct input location, the true process as a black dashed line, the computer model $f(\cdot, \hat{\theta})$ as a blue dashed line, and the prediction mean curve as a red solid line, with 95% prediction intervals in green dotted lines. Lower panels represent the sample variance $\hat{r}(\bar{x}_i)$ as black points, the true variance process as a dashed line, and the fitted variance process as a red solid line.*

rameter estimation. The estimates of WLS are very biased when the computer model is inexact. HomOGP gives relatively unbiased estimates, but the variation of the estimates is larger than HetOGP. From the middle and right panels, it shows that heteroscedastic modeling-based methods (HetGP and HetOGP) are better than WLS and homoscedastic modeling-based methods (HomGP and HomOGP) in terms of prediction performation. The reason of the poor predictive scores of HomGP and HomOGP is that, as shown in Figure 1, the homoscedastic modeling yields unreasonably wide prediction intervals when heteroscedasticity is present. HetGP results in superior prediction accuracy and prediction scores, but the estimates are quite off from the true parameter. On the other hand, our proposed method (HetOGP), which models the discrepancy function using an orthogonal GP to avoid the identifiability issue, provides more accurate parameter estimates. We also construct the 95% confidence intervals for the calibration parameter based on the result of Theorem 3.1, and 92 out of the 100 simulations cover the true parameter, which is close to the nominal coverage 95%. In terms of the computational cost, all the methods here are implemented within 3 seconds, on a laptop with 2.6 GHz CPU and 16 GB of RAM. Based on the estimation and prediction results, it suggests that our proposed method is more appropriate for the calibration problem in the face of heteroscedasticity and inexact computer models, which provides more accurate parameter estimates along with high prediction accuracy and prediction scores.

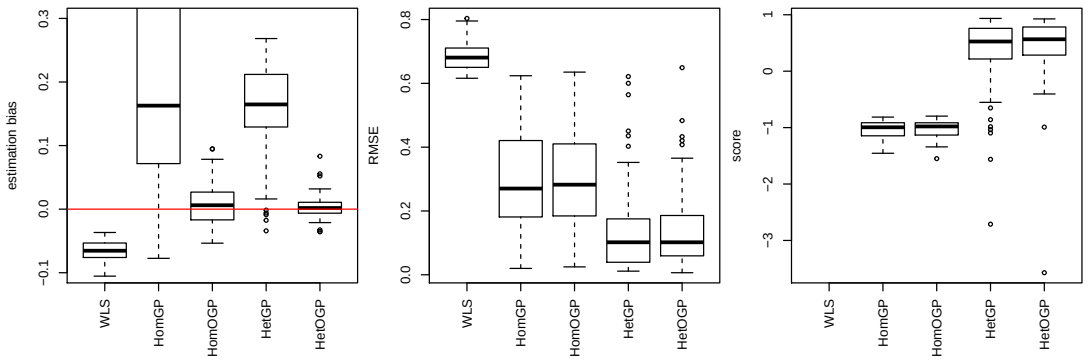

Figure 2. *The comparison of estimation and prediction performance. The left panel represents the estimation bias of the calibration parameter, with the red horizontal line indicating zero bias. The middle panel shows their root mean squared errors, and the right panel represents their predictive scores.*

To further examine the estimation performance, three additional variance processes are considered in the example: (i) a constant variance, $r(x) = 1$; (ii) variance with two weak peaks at $\pi/2$ and $3\pi/2$, $r(x) = 3(\exp\left(-3(x-\pi/2)^2\right) + \exp\left(-3(x-3\pi/2)^2\right)) + 0.01$; (iii) variance with one strong peak at $\pi/2$, $r(x) = 6\exp\left(-6(x-\pi/2)^2\right) + 0.01$. These variance functions are presented in the top panels of Figure 3 and 100 simulations are similarly conducted for each of the processes. The parameter estimation results are summarized in the bottom panels of Figure 3. It can be seen that when the variance is constant, HomOGP and HetOGP perform equally well, which is expected because the heteroscedastic modeling does not benefit from the homoscedastic case. When the variance process has two weak peaks as in (ii), our proposed method improves the estimation accuracy by accounting for the heteroscedasticity, which provides more accurate estimates with a lower variance than others. With a stronger peak as

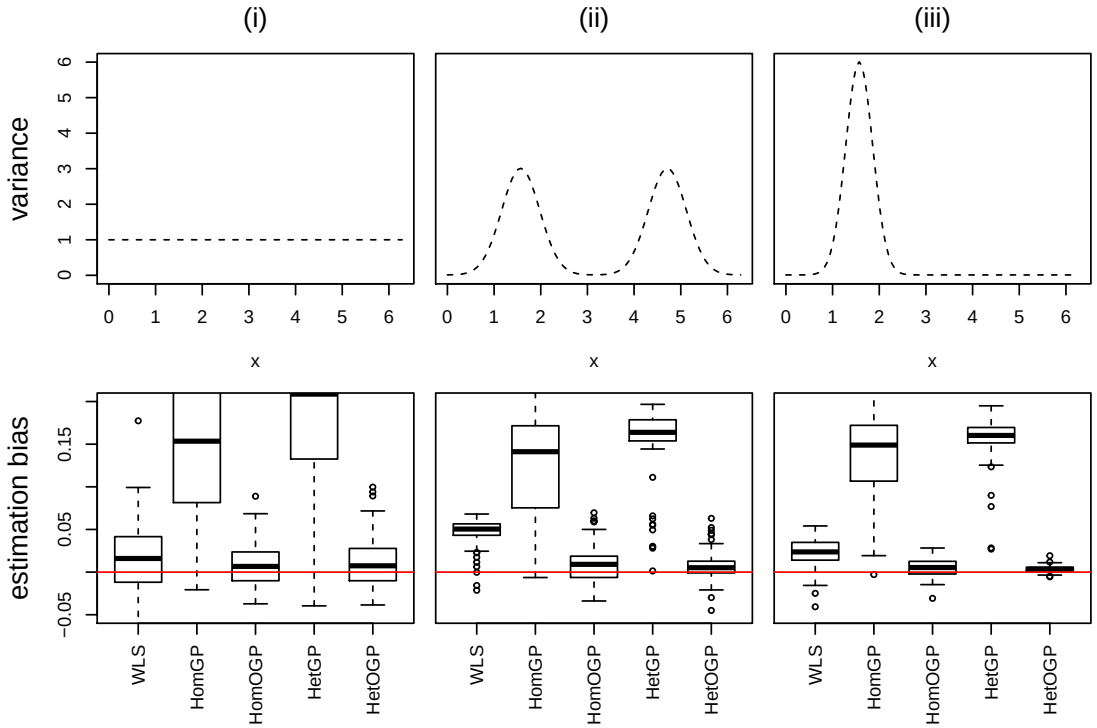

Figure 3. *True variance process (top) and the resulting estimation bias (bottom).*

in (iii), the benefit of heteroscedastic modeling is even more apparent.

We further use these three variance processes to examine the heteroscedastic test developed in Section 3.2, where the empirical powers are computed as the proportion of the time during the 1000 simulations that the null hypothesis is rejected with type I error 5%. The empirical powers with four different sample sizes are summarized in Table 1. The empirical powers of the constant variance (i) are close to the desired power, 5%, because the null hypothesis is true. The powers of (ii) and (iii) are both higher than 0.97 in this finite-sample setting, showing that the test can detect the presence of heteroscedasticity with high degree of power.

n	a	(i)	(ii)	(iii)
8	5	0.035	0.971	0.995
	10	0.019	1.000	0.999
16	10	0.021	1.000	1.000
	20	0.024	1.000	1.000

Table 1

Empirical powers of the heteroscedastic tests, where a is the number of replicates of each of the n distinct inputs, which gives the total sample size $N = na$.

4.2. Example with Three Calibration Parameters. In this subsection, we consider a calibration problem where two input variables and three calibration parameters are involved in a computer model. Assume that the input $x \in \mathbb{R}^2$ is uniformly distributed on $[0,1]^2$, the true process is $\zeta(x) = 4x_1 + x_1 \sin(5x_2)$, and the observations are given by $y_i = \zeta(x_i) + \epsilon_i$, where ϵ_i is independently normally distributed with zero mean and the variance $r(x_i) = 0.01 \exp(-10 \sin(x_1\pi) \cos(x_2\pi))$. Suppose that the computer output is given by the function $f(x, \theta) = \theta_1 + \theta_2 x_1 + \theta_3 x_2$, where $\theta = (\theta_1, \theta_2, \theta_3) \in \Theta = [0,1]^3$. By minimizing the L_2 distance as in (2.3), we have $\theta^* \approx (0.50, 4.14, -1.00)$.

Similar to the previous subsection, we conduct the simulation 100 times, where each simulation uses a two-dimensional Latin Hypercube sample (LHS, [39]) of size 30 on the unit cube for designing the input x. In this study we consider three different numbers of replicates, $\{2, 5, 10\}$, for each distinct input setting x, leading to $N \in \{60, 150, 300\}$. The estimation results are summarized in Figure 4, which shows the boxplots of estimation bias arranged by the numbers of replicates (three groups of five from left to right) for each calibration parameter. The results show that the proposed method (HetOGP) provides more accurate estimates than the other four methods for each of the three calibration parameters. HomGP and HomOGP provide relatively unbiased estimates, but the estimate variances are much larger than HetOGP. The estimates of the proposed method are more accurate with lower variances by the increase of the replicates, which agrees with the asymptotic result in Theorem 3.1.

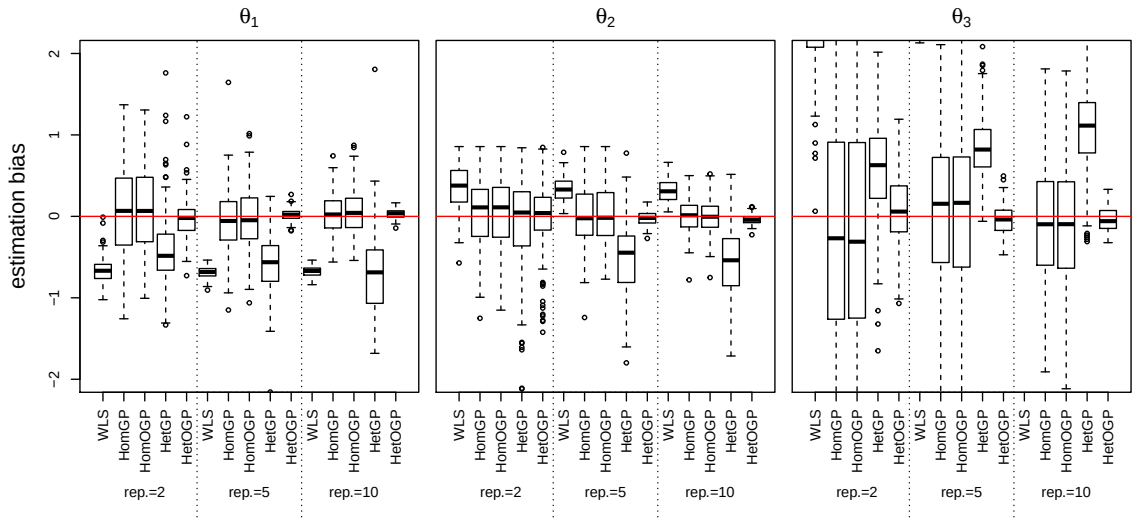

Figure 4. *The estimation bias of (left) θ_1; (middle) θ_2; (right) θ_3, with the red horizontal line indicating zero bias. Results with 2, 5 and 10 replicates at each input location are arranged in three groups of five along the x-axis in each panel.*

5. Case Study: Estimation of Plant Growth Rate.

5.1. Plant Growth Experiment. Plant relative growth rate [7, 29, 30] plays an important role to study the performance of plant productivity as related to environmental stress and

disturbance regimes. To calculate the relative growth rate of plants, a relatively simple, yet mechanistically accurate, computer model is commonly used. This model expresses plant biomass after x days, $S(x)$, in the form of an equation:

$$\frac{dS(x)}{dx} = \theta S(x), \tag{5.1}$$

where θ is a constant and defined as the relative growth rate. This differential equation has the solution $S(x) = S(0)\exp(\theta x)$. In this study the initial plant biomass is set $S(0) = 1$. The experimental data are observations from plant growth experiments, where the plant biomass is quantified by imaging the photosynthetically active areas of the tissues using a overhead camera system that measured fluorescence emitted from photosynthetically active tissues [40]. To account for technical and biological variation, multiple replicates are conducted in the experiment.

The three groups of plants differ by the presence or absence of certain genes involved in photorespiration. The different plant groups tested lacked distinct steps involved in photorespiration, either a critical enzymatic interconversion step (*glyk*) [8], or a transporter (*plgg*) [43] which can be circumvented via other transport mechanisms [57, 50]. These groups were compared to wild type plants (WT), which have a fully functioning photorespiratory pathway.

5.2. Calibration Results. We leverage the statistical developments to investigate the plant relative growth rates for the plants grown under ambient and high CO_2 concentrations, with the three plant groups: *glyk*, *plgg*, and WT. The experimental data consists of the total projected areas of the plants at 8 distinct time points, as the input variable $x \in [0, 20]$, and for each distinct time point, three to five replicates are measured. The computer model is as described in (5.1), which shares the same input variable x (time) and has a calibration parameter, the relative growth rate, $\theta \in [0, 1]$. The calibration results using the experimental data under ambient and high CO_2 concentrations are presented in Figures 5 and 6, respectively. First, it can be seen that in the experimental data under both ambient and high CO_2 concentrations, the variances tend to increase as the time increases. By performing the heteroscedasticity test developed in Section 3.2, the p-values for the three groups are given in Table 2, which shows that all of the p-values are less than 0.05, indicating that heteroscedastic modeling is essential for this data. Our approach takes into account the heteroscedasticity and provides the fitted variance process (as the red lines in the middle panels), which in turn gives sensible prediction intervals (as the green dotted lines in the top panels). Moreover, bottom panels present the fitted discrepancy function, $b_{\hat{\theta}}(x)$, with 95% pointwise confidence intervals based on the orthogonal GP modeling in Section 2.2. The discrepancy functions show that the computer model is imperfect as expected, especially for the data under CO_2 ambient concentrations (Figure 5), which may suggest that a quadratic polynomial is needed in the computer model. Future studies of plant growth rates may require more complex models if debiasing the computer model is of interest. Alternatively, it is also possible that the measurement itself could be imperfect, that is, measurement errors have non-zero mean that are input-dependent. For example, leaves could be more likely to overlap during the middle stage of growth, which makes it difficult to measure the growth by overhead images. These results show that the proposed method not only gives reasonable prediction means and intervals for the experimental data, but also provides important insights via the model discrepancy.

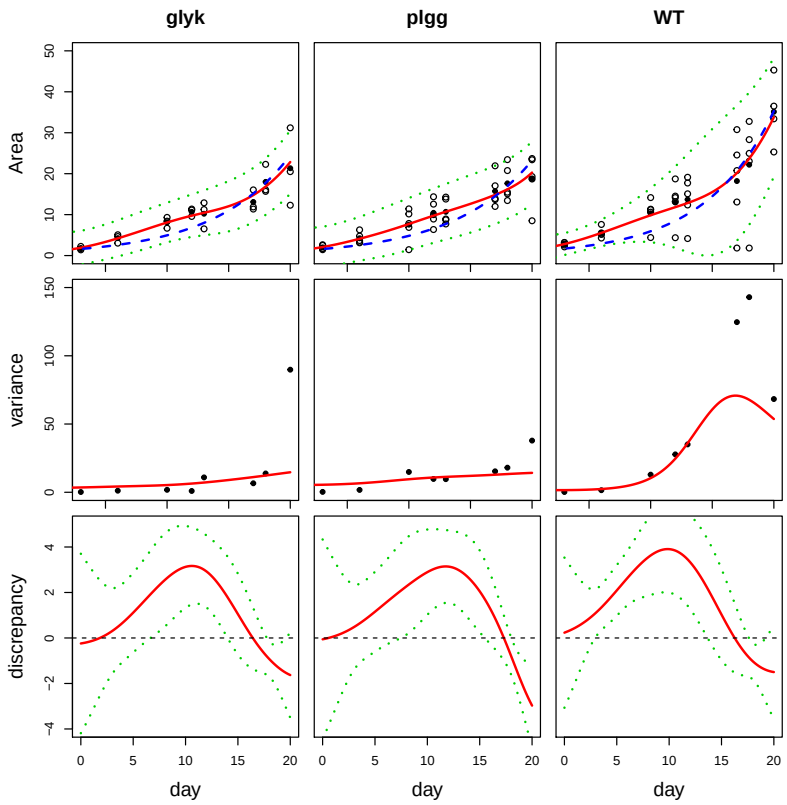

Figure 5. *Calibration results of the three plant groups under ambient CO_2: glyk (left), plgg (middle), and WY (right). Top panels represent the replicates as open circles with the averaged observation $\hat{y}_i$ in filled circles at each input location, the curve $f(x, \hat{\theta})$ as a blue dashed line, and the prediction mean curve as a red solid line, with 95% prediction intervals in green dotted lines. Middle panels represent the sample variance $\hat{r}(\bar{x}_i)$ as black points, and the fitted variance process as a red solid line. Lower panels represent the mean curve of the discrepancy function, with 95% pointwise confidence intervals in green dotted lines.*

CO_2 Concentration	Group	Relative Growth Rate		Het. Test
		Estimate	95% Confidence Interval	p-value
Ambient	*glyk*	0.1590	[0.1512, 0.1668]	0.0015
	plgg	0.1581	[0.1515, 0.1647]	0.0205
	WT	0.1780	[0.1697, 0.1864]	<0.0001
High	*glyk*	0.2128	[0.1947, 0.2308]	<0.0001
	plgg	0.2428	[0.2364, 0.2493]	<0.0001
	WT	0.2295	[0.2197, 0.2394]	<0.0001

Table 2

Estimated relative growth rates and p-values of heteroscedastic tests.

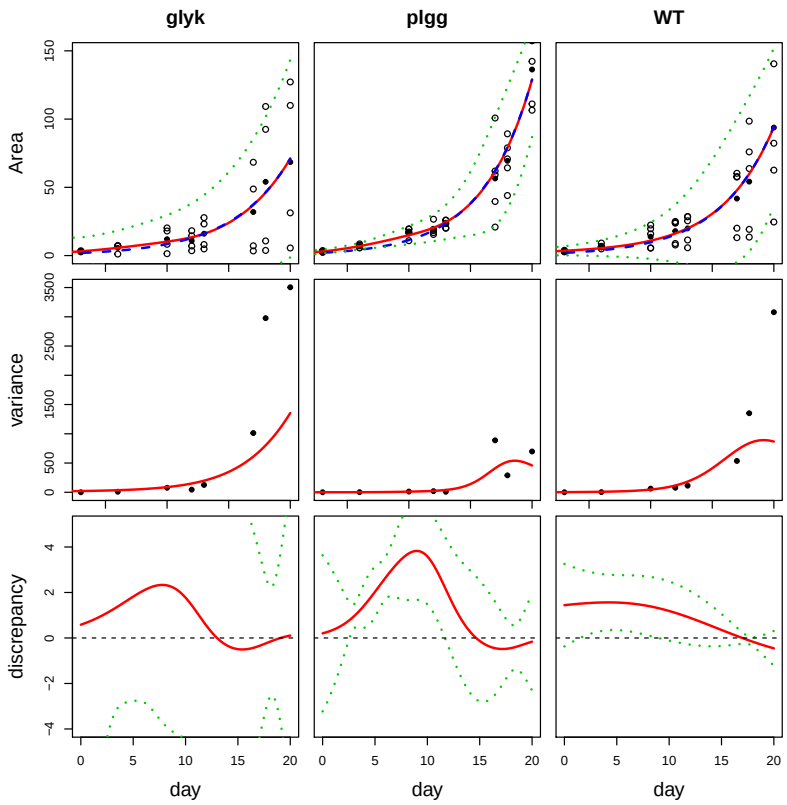

Figure 6. *Calibration results of the three plant groups under high CO_2: glyk (left), plgg (middle), and WY (right). Top panels represent the replicates as open circles with the averaged observation $\hat{y}_i$ in filled circles at each input location, the curve $f(x, \hat{\theta})$ as a blue dashed line, and the prediction mean curve as a red solid line, with 95% prediction intervals in green dotted lines. Middle panels represent the sample variance $\hat{r}(\bar{x}_i)$ as black points, and the fitted variance process as a red solid line. Lower panels represent the mean curve of the discrepancy function, with 95% pointwise confidence intervals in green dotted lines.*

The estimated relative growth rates, $\hat{\theta}$, are reported in Table 2, where the confidence intervals are constructed based on the asymptotic normality result in Theorem 3.1. First, we observe that relative growth rates under ambient CO_2 concentrations are slower than under high CO_2 concentrations across all plant groups. This is expected from the biological perspectives, because CO_2/O_2 under ambient CO_2 concentrations is low enough to drive high rates of photorespiration, which consumes energy and releases previously fixed carbon, decreasing growth. Secondly, the group *glyk* and *plgg* have slower relative growth rates than WT under ambient CO_2 concentrations, which is consistent with either a disruption in growth generally or specifically in photorespiration. These calibration parameter estimates and confidence intervals provide insight into the values of the relative growth rates of different plant groups, which are difficult to determine by physical experiments due to the limitation of the existing experimental techniques.

To examine our approach, we compare the performance with the least-squares approach

with logarithmic transformation on the response, which is commonly used to estimate growth rates in the literature [42], and a traditional method in the biological literature (see, for example, [13]), which uses least-squares estimates for each individual replicate and then takes the average of them. That is, $\hat{\theta}_j = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} (y_i^{(j)} - f(\bar{x}_i, \theta))^2$ and then obtain the estimate by $\hat{\theta} = \sum_{j=1}^{a} \hat{\theta}_j / a$, where a is the number of replicates. We use the experimental data under ambient CO_2 concentrations as an illustration, which is shown in Figure 7. Our approach appears to improve the calibration performance over other two methods, in the sense that the computer model outputs with our calibration parameter estimates, $f(x, \hat{\theta})$ in the blue dashed line, appear to be closer to the experimental data compared to the other two methods. In addition, it shows that the prediction intervals of the approach with logarithmic transformation (middle column) are much narrower than one would expect. This issue is resolved when the proposed heteroscedastic model is used.

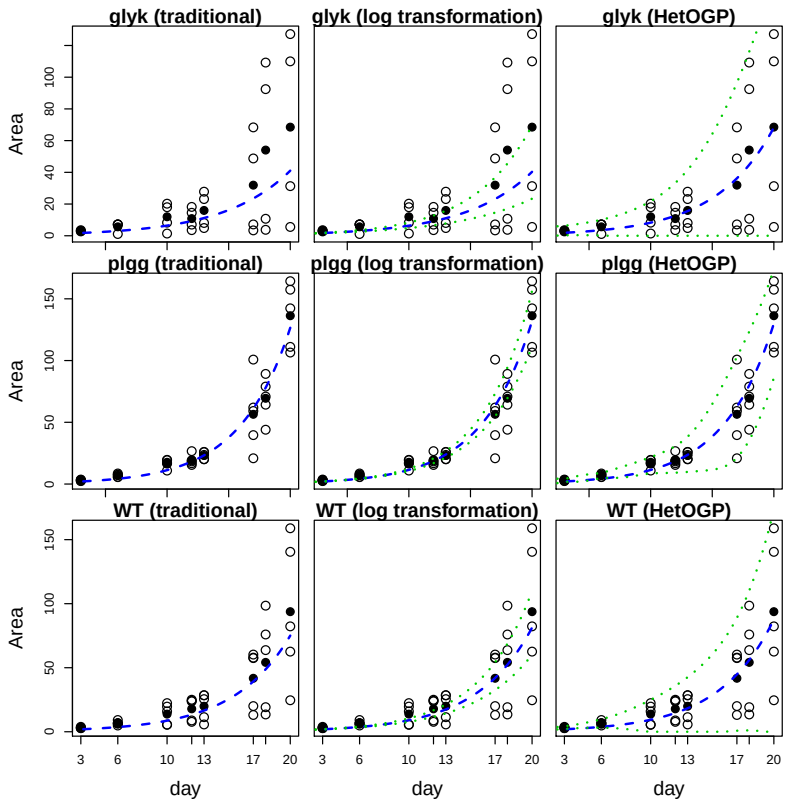

Figure 7. *Comparison of the traditional method (left), the logarithmic transformation (middle), and the proposed method (right) based on the experimental data under ambient CO_2, where the replicates are open circles with the averaged observation $\bar{y}_i$ in filled circles, the curve $f(x, \hat{\theta})$ is the blue dashed line, and 95% prediction intervals are the green dotted lines. The three rows represent the three plant groups: glyk (top), plgg (middle), and WT (bottom).*

6. Summary and Concluding Remarks. Calibration of computer models plays a crucial role in many scientific fields where computer models are essential to predict the reality. The

existing methods in the statistics literature, however, mainly focus on calibration with homoscedastic errors. Motivated by an experiment in plant biology, where the noise levels can vary dramatically across different input locations, we introduce a new calibration method to address the heteroscedasticity, where a latent variable process is used to model the error variance and an orthogonal Gaussian process is used to model the misspecification of a computer model. An `R` package is available for implementing the proposed method. We also study the asymptotic properties of the estimator and provide a goodness-of-fit statistic to detect the presence of heteroscedasticity. Our numerical studies demonstrate that when the errors are not homoscedastic, our proposed method not only successfully estimates calibration parameters accurately, but it also provides accurate predictions for a real system. The application to the plant relative growth rates illustrates that the proposed calibration method produces reasonable estimates of relative growth rates and uncertainty quantification for the physical experiments.

This work indicates several avenues for future research. First, in Theorem 3.1 we provide the asymptotic results of the maximum likelihood estimators for the parameters in the proposed model, which adopts the orthogonal GP of [44] for the model discrepancy. It is also worthwhile investigating whether the estimator $\hat{\theta}$ is theoretically consistent for the true calibration parameter θ^* as in (2.3), like the asymptotic results in [53] and [61]. As in the homoscedastic case, where the orthogonal GP modeling has been shown numerically to be consistent for θ^* [44, 61] and also conjectured to be theoretically consistent by [53], it is conceivable to conjecture that the estimator $\hat{\theta}$ is consistent for θ^* in the heteroscedastic case. This also could be verified by the numerical studies in Section 4. Alternatively, one may consider other modeling techniques for the discrepancy function that also address the identifiability issue of the calibration parameters, such as [22], [53], [61], and [15], which provide the potential to extend the proposed method with theoretical guarantees. However, the extension to heteroscedastic cases with the noise process in Section 2.1 is not straightforward. Second, instead of maximum likelihood estimation, Bayesian techniques can be naturally applied to the proposed method. Specifically, one could assign the priors of the calibration parameters as well as the hyperparameters in the model, and then draw samples from the posterior distribution using Markov chain Monte Carlo approaches, such as the Metropolis-Hastings sampler. With this Bayesian framework, the proposed method can also be naturally extended to the calibration problem with *functional* calibration parameters like in [9] and [45]. Last but not the least, it is worth developing the sampling schemes that satisfy the conditions for Theorems 3.1 and 3.2. A potential sampling scheme is, similar to [14], to choose distinct input locations whose minimum distance is sufficiently large. More details will be carefully investigated in the future work.

Supplemental Materials. Additional supporting materials can be found online, including the detailed proofs of Theorem 3.1, the detailed estimation procedure in Section 2.3, an `R` package `HetCalibrate` for implementing the proposed method, and the `R` code and data for reproducing the results in Sections 4 and 5.

REFERENCES

[1] B. Ankenman, B. L. Nelson, and J. Staum, *Stochastic kriging for simulation metamodeling*, Opera-

tions Research, 58 (2010), pp. 371–382.
[2] E. Baker, P. Barbillon, A. Fadikar, R. B. Gramacy, R. Herbei, D. Higdon, J. Huang, L. R. Johnson, P. Ma, A. Mondal, B. Pires, J. Sacks, and V. Sokolov, *Analyzing stochastic computer models: A review with opportunities*, arXiv preprint arXiv:2002.01321, (2020).
[3] M. Bayarri, J. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R. Parthasarathy, R. Paulo, J. Sacks, D. Walsh, et al., *Computer model validation with functional output*, The Annals of Statistics, 35 (2007), pp. 1874–1906.
[4] M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu, *A framework for validation of computer models*, Technometrics, 49 (2007), pp. 138–154.
[5] M. Binois and R. B. Gramacy, *hetGP: Heteroskedastic Gaussian Process Modeling and Design under Replication*, 2019, https://CRAN.R-project.org/package=hetGP. R package version 1.1.1.
[6] M. Binois, R. B. Gramacy, and M. Ludkovski, *Practical heteroscedastic Gaussian process modeling for large simulation experiments*, Journal of Computational and Graphical Statistics, 27 (2018), pp. 808–821.
[7] V. H. Blackman, *The compound interest law and plant growth*, Annals of Botany, 33 (1919), pp. 353–360.
[8] R. Boldt, C. Edner, Ü. Kolukisaoglu, M. Hagemann, W. Weckwerth, S. Wienkoop, K. Morgenthal, and H. Bauwe, *D-GLYCERATE 3-KINASE, the last unknown enzyme in the photorespiratory cycle in arabidopsis, belongs to a novel kinase family*, The Plant Cell, 17 (2005), pp. 2413–2420.
[9] D. A. Brown and S. Atamturktur, *Nonparametric functional calibration of computer models*, Statistica Sinica, 28 (2018), pp. 721–742.
[10] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM Journal on Scientific Computing, 16 (1995), pp. 1190–1208.
[11] R. E. Caflisch, *Monte Carlo and quasi-Monte Carlo methods*, Acta Numerica, 7 (1998), pp. 1–49.
[12] X. Chen, A. P. Alonso, and Y. Shachar-Hill, *Dynamic metabolic flux analysis of plant cell wall synthesis*, Metabolic Engineering, 18 (2013), pp. 78–85.
[13] A. B. Cousins, B. J. Walker, I. Pracharoenwattana, S. M. Smith, and M. R. Badger, *Peroxisomal hydroxypyruvate reductase is not essential for photorespiration in arabidopsis but its absence causes an increase in the stoichiometry of photorespiratory co2 release*, Photosynthesis Research, 108 (2011), pp. 91–100.
[14] N. Cressie and S. N. Lahiri, *Asymptotics for reml estimation of spatial covariance parameters*, Journal of Statistical Planning and Inference, 50 (1996), pp. 327–341.
[15] X. Dai and P. Chien, *Another look at statistical calibration: a non-asymptotic theory and prediction-oriented optimality*, arXiv preprint arXiv:1802.00021, (2018).
[16] B. Fornberg and D. M. Sloan, *A review of pseudospectral methods for solving partial differential equations*, Acta Numerica, 3 (1994), pp. 203–267.
[17] T. Gneiting and A. E. Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association, 102 (2007), pp. 359–378.
[18] L. G. Godfrey, *Testing for multiplicative heteroskedasticity*, Journal of Econometrics, 8 (1978), pp. 227–236.
[19] P. W. Goldberg, C. K. Williams, and C. M. Bishop, *Regression with input-dependent noise: A Gaussian process treatment*, in Advances in Neural Information Processing Systems, 1998, pp. 493–499.
[20] R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, CRC Press, 2020.
[21] R. B. Gramacy, D. Bingham, J. P. Holloway, M. J. Grosskopf, C. C. Kuranz, E. Rutter, M. Trantham, and R. P. Drake, *Calibrating a large computer experiment simulating radiative shock hydrodynamics*, The Annals of Applied Statistics, 9 (2015), pp. 1141–1168.
[22] M. Gu and L. Wang, *Scaled Gaussian stochastic process for computer model calibration and prediction*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1555–1583.
[23] G. Han, T. J. Santner, and J. J. Rawlinson, *Simultaneous determination of tuning and calibration parameters for computer experiments*, Technometrics, 51 (2009), pp. 464–474.
[24] A. C. Harvey, *Estimating regression models with multiplicative heteroscedasticity*, Econometrica, 44 (1976), pp. 461–465.
[25] D. A. Harville, *Matrix Algebra from a Statistician's Perspective*, New York: Springer-Verlag, 1998.

[26] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using high-dimensional output*, Journal of the American Statistical Association, 103 (2008), pp. 570–583.
[27] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne, *Combining field data and computer simulations for calibration and prediction*, SIAM Journal on Scientific Computing, 26 (2004), pp. 448–466.
[28] C. Hildreth and J. P. Houck, *Some estimators for a linear model with random coefficients*, Journal of the American Statistical Association, 63 (1968), pp. 584–595.
[29] R. Hunt, *Plant growth analysis: second derivatives and compounded second derivatives of splined plant growth curves*, Annals of Botany, 50 (1982), pp. 317–328.
[30] R. Hunt, *Plant Growth Curves. The Functional Approach to Plant Growth Analysis*, Edward Arnold Ltd., 1982.
[31] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society: Series B, 63 (2001), pp. 425–464.
[32] R. Koenker and G. Bassett Jr, *Robust tests for heteroscedasticity based on regression quantiles*, Econometrica, 50 (1982), pp. 43–61.
[33] B.-J. Lee, *A heteroskedasticity test robust to conditional mean misspecification*, Econometrica, 60 (1992), pp. 159–171.
[34] G. R. Lindfield and J. E. Penny, *Microcomputers in Numerical Analysis*, Halsted Press, 1989.
[35] J. Loeppky, D. Bingham, and W. Welch, *Computer model calibration or tuning in practice*, tech. report, University of British Columbia, Vancouver, BC, Canada, 2006.
[36] J. P. Long, *A note on parameter estimation for misspecified regression models with heteroskedastic errors*, Electronic Journal of Statistics, 11 (2017), pp. 1464–1490.
[37] F. Ma, L. J. Jazmin, J. D. Young, and D. K. Allen, *Isotopically nonstationary ^{13}C flux analysis of changes in arabidopsis thaliana leaf metabolism due to high light acclimation*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 16967–16972.
[38] S. Mak, C.-L. Sung, X. Wang, S.-T. Yeh, Y.-H. Chang, V. R. Joseph, V. Yang, and C. F. J. Wu, *An efficient surrogate model for emulation and physics extraction of large eddy simulations*, Journal of the American Statistical Association, 113 (2018), pp. 1443–1456.
[39] M. D. McKay, R. J. Beckman, and W. J. Conover, *Comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.
[40] E. H. Murchie and T. Lawson, *Chlorophyll fluorescence analysis: a guide to good practice and understanding some new applications*, Journal of Experimental Botany, 64 (2013), pp. 3983–3998.
[41] W. K. Newey and J. L. Powell, *Asymmetric least squares estimation and testing*, Econometrica, 55 (1987), pp. 819–847.
[42] C. T. Paine, T. R. Marthews, D. R. Vogt, D. Purves, M. Rees, A. Hector, and L. A. Turnbull, *How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists*, Methods in Ecology and Evolution, 3 (2012), pp. 245–256.
[43] T. R. Pick, A. Bräutigam, M. A. Schulz, T. Obata, A. R. Fernie, and A. P. M. Weber, *Plgg1, a plastidic glycolate glycerate transporter, is required for photorespiration and defines a unique class of metabolite transporters*, Proceedings of the National Acadamy of Sciences, 110 (2013), pp. 3185–3190.
[44] M. Plumlee, *Bayesian calibration of inexact computer models*, Journal of the American Statistical Association, 112 (2017), pp. 1274–1285.
[45] M. Plumlee, V. R. Joseph, and H. Yang, *Calibrating functional parameters in the ion channel models of cardiac cells*, Journal of the American Statistical Association, 111 (2016), pp. 500–509.
[46] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018, https://www.R-project.org/.
[47] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science, 4 (1989), pp. 409–423.
[48] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments (Second Edition)*, Springer New York, 2018.
[49] T. D. Sharkey, C. J. Bernacchi, G. D. Farquhar, and E. L. Singsaas, *Fitting photosynthetic carbon dioxide response curves for C_3 leaves*, Plant, Cell and Environment, 30 (2007), pp. 1035–1040.
[50] P. F. South, B. J. Walker, A. P. Cavanagh, V. Rolland, M. Badger, and D. R. Ort, *Bile*

acid sodium symporter bass6 can transport glycolate and is involved in photorespiratory metabolism in Arabidopsis thaliana, The Plant Cell, 29 (2017), pp. 808–823.

[51] C.-L. Sung, *HetCalibrate: Calibration of Inexact Computer Models with Heteroscedastic Errors*, 2020, https://github.com/ChihLi/HetCalibrate. R package version 0.1.

[52] C.-L. Sung, Y. Hung, W. Rittase, C. Zhu, and C. F. J. Wu, *Calibration for computer experiments with binary responses and application to cell adhesion study*, Journal of the American Statistical Association, 115 (2020), pp. 1664–1674.

[53] R. Tuo, *Adjustments to computer models via projected kernel calibration*, SIAM/ASA Journal on Uncertainty Quantification, 7 (2019), pp. 553–578.

[54] R. Tuo and C. F. J. Wu, *Efficient calibration for imperfect computer models*, The Annals of Statistics, 43 (2015), pp. 2331–2352.

[55] R. Tuo and C. F. J. Wu, *A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 767–795.

[56] S. von Caemmerer, *Steady-state models of photosynthesis*, Plant, Cell and Environment, 36 (2013), pp. 1617–1630.

[57] B. J. Walker, P. F. South, and D. R. Ort, *Physiological evidence for plasticity in glycolate/glycerate transport during photorespiration*, Photosynthesis research, 129 (2016), pp. 93–103.

[58] S. Wang, W. Chen, and K.-L. Tsui, *Bayesian validation of computer models*, Technometrics, 51 (2009), pp. 439–451.

[59] W. Wang and B. Haaland, *Controlling sources of inaccuracy in stochastic kriging*, Technometrics, 61 (2019), pp. 309–321.

[60] R. K. W. Wong, C. B. Storlie, and T. C. M. Lee, *A frequentist approach to computer model calibration*, Journal of the Royal Statistical Society: Series B, 79 (2017), pp. 635–648.

[61] F. Xie and Y. Xu, *Bayesian projected calibration of computer models*, Journal of the American Statistical Association, to appear (2020).