SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 10, No. 4, pp. 1435{1460

A Theoretical Framework of the Scaled Gaussian Stochastic Process in Prediction and Calibration

Mengyang Gu[y], Fangzheng Xie[z], and Long Wang[x]

Abstract. Model calibration or data inversion is one of the fundamental tasks in uncertainty quantification. In this work, we study the theoretical properties of the scaled Gaussian stochastic process (S-GaSP) for modeling the discrepancy between reality and the imperfect mathematical model. We establish an explicit connection between the Gaussian stochastic process (GaSP) and S-GaSP through the orthogonal series representation. The predictive mean estimator in the S-GaSP calibration model converges to reality at the same rate as the GaSP with suitable choices of the regularization and scaling parameters. We also show that the calibrated mathematical model in the S-GaSP calibration converges to the one that minimizes the L_2 loss between reality and the mathematical model, whereas the GaSP model with other, widely used covariance functions does not have this property. Numerical examples confirm the excellent finite sample performance of our approaches.

Key words. model misspecification, Bayesian prior, scaled Gaussian stochastic process prior, convergence, interpretability, orthogonal series representation

MSC codes. 62A01, 62F15, 62M20, 62P30

DOI. 10.1137/21M1409949

1. Introduction.

Mathematical models are developed by scientists and engineers based on their expert knowledge of reproducing physical reality. With the rapid development of computational techniques in recent years, many mathematical models have been implemented in computer codes; these are often referred to as computer models or simulators.

Some parameters of mathematical models are unknown or unobservable in experiments. For example, the Kilauea volcano recently had one of its biggest eruptions in 2018. The location and volume of the magma chamber, as well as the magma supply and storage rate of this volcano, however, are unobservable. Field data, such as satellite interferograms and GPS measurements of the ground deformation, were used to estimate these parameters [1, 2]. Using field observations to estimate parameters in a mathematical model and to identify the possible discrepancy between the mathematical model and the reality is widely known as model calibration or data inversion.

For any p-dimensional observable input x 2 X, denote $y^F(x)$ 2 R as the field observation and $f^M(x;)$ 2 R as a mathematical model with q-dimensional unobservable calibration

Received by the editors April 12, 2021; accepted for publication (in revised form) April 26, 2022; published electronically October 31, 2022.

https://doi.org/10.1137/21M1409949

Funding: This research was supported by National Science Foundation grant DMS-2053423.

[y] Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA 93106-3110 USA (mengyang@pstat.ucsb.edu).

[z] Department of Statistics, Indiana University, Bloomington, Bloomington, IN 47408 USA (fxie@iu.edu).

[x] Department of Applied Mathematical and Statistics, Johns Hopkins University, Baltimore, MD 21218 USA (long.wang@jhu.edu).

parameters 2 . Furthermore, let $y^R(x) = E[y^F(x)]$ represent the reality. A routinely used framework for calibrating the imperfect mathematical model is [15, 4]

(1) $$y^F(x) = f^M(x;) + (x) + ;$$

where is the noise and () is a discrepancy function between the reality and the mathematical model. Since the mathematical model is often developed by experts, we assume that the mean and trend of the observations are already included in the mathematical model.

The discrepancy function was modeled as a Gaussian stochastic process (GaSP), in [15] and the framework has been widely studied in recent years [8, 14, 18]. This approach is referred to as GaSP calibration. Define the prediction error by the L_2 loss Z

(2) $$\text{prediction error}(;) = {}_{x \, 2 \, X} [y^R(x) \quad \hat{y}^R_{;}(x)]^2 dx;$$

where $\hat{y}^R_{;}(x)$ is the prediction (e.g., predictive mean) of the reality, and the subscript (;) implies the dependence on both the mathematical model and the discrepancy function. Since both the mathematical model and the discrepancy function are jointly estimated in GaSP calibration, the prediction error was found to be smaller when compared with the prediction error from the mathematical model or nonparametric regression alone [15].

It was shown in follow-up studies, however, that the calibrated mathematical model in GaSP calibration can be far away from the reality, which results in an identifiability problem between the calibration parameters and the discrepancy function [3, 19, 26]. This is because the variability of the observations in GaSP calibration with some frequently used kernels, such as the Matern kernel, can be explained mostly by the estimated discrepancy function instead of the calibrated mathematical model.

Note that the true value of the parameter in (1) is not well defined due to the inclusion of an unknown function . However, one can evaluate the distance between the reality and the calibrated computer model via some frequently used loss functions. A few recent studies measure the goodness of calibration in terms of the L_2 loss between the calibrated mathematical model and reality [25, 26, 29]. Thus, the calibration error may be defined by

(3) $$\text{calibration error}() = Z_{x \, 2 \, X} [y^R(x) \quad f^M(x;)]^2 dx:$$

These studies seek to find the L_2 minimizer of that minimizes the calibration error, i.e., $_{L_2}$:= argmin_2 calibrationerror(). To estimate the L_2 minimizer, a few two-step ap-proaches were developed. In [25], for instance, the reality is first estimated through a nonparametric regression model without the assistance of the mathematical model. The calibration parameters are then estimated by minimizing the L_2 loss between the mathematical model and the estimator of the reality. For some complex applications, however, it is crucial to jointly estimate the reality and the calibration parameters, as the mathematical model developed based on the experts' knowledge can be very helpful for predicting the reality.

A recent model for the discrepancy function, called the scaled Gaussian stochastic process (S-GaSP) [12], was constructed such that both the prediction error and the calibration error are small. We call this approach the S-GaSP calibration. Note that the calibration error can be written as $R_{x \, 2 \, X} (x)^2 dx$. In S-GaSP, the prior distribution of the calibration error has more probability mass near small values, reflecting one's preference for smaller values of the L_2 norm of the discrepancy function. In contrast to the two-step approaches that seek L_2 minimizers, both GaSP and S-GaSP define a sampling model of the discrepancy, such that the uncertainty of the parameters can be obtained exactly. Though the computational strategy of the S-GaSP calibration was studied in [12], theoretical properties of this process are yet to be examined.

In this work, we study the theoretical properties of S-GaSP. The contribution of this article contains three parts. First, we establish the explicit connection between GaSP and S-GaSP through the orthogonal representation of the process, which describes the reproducing kernel Hilbert space associated with an S-GaSP. Second, we show that the maximum likelihood estimator of the calibration parameter and the predictive mean in GaSP can be obtained by minimizing a squared error loss with a penalty term depending on the L_2 norm and the reproducing kernel Hilbert space norm. Third, we show that the predictive mean from S-GaSP converges to the reality at the same rate as the one from GaSP with suitable choices of the regularization and scaling parameters. Last but not least, using the same regularization and scaling parameters, the calibration parameters in S-GaSP converge to L_2 , whereas GaSP calibration with other, widely used kernels does not enjoy this property. To the best of our knowledge, this work represents the first effort in the literature showing that the joint estimation of the calibration parameters and the model discrepancy by S-GaSP (essentially a GaSP with a transformed kernel) allows these two convergence properties to hold at the same time, bridging the gap between the joint estimation procedure and other two-step estimation methods.

Although the two convergence properties discussed in this work can be achieved using the aforementioned two-step approaches [25, 29], finite sample studies suggest that jointly estimating the calibration parameters and the discrepancy as in S-GaSP calibration leads to high predictive accuracy. Additionally, since the sampling model is fully specified, uncertainties of the parameters in S-GaSP calibration can be naturally assessed through posterior distribu-tions via a Bayesian approach. Numerical results in section 5 suggest the maximum likelihood estimator (MLE) and Bayesian estimation are often similar, yet Bayesian estimation is more robust than numerical optimization.

The rest of this paper is organized as follows. In section 2, we introduce S-GaSP along with the orthogonal series representation and the joint estimation in calibration. Two convergence properties are discussed in section 3. In section 4, we introduce the discretized S-GaSP along with the parameter estimation under the Frequentist framework and Bayesian framework. A comparison between S-GaSP calibration and other alternatives is discussed in section 5 with numerical evidence provided in section 6. We conclude this work in section 7. The proofs of the theoretical results are given in the supplementary material (supplement.pdf [local/web 442KB]). We implement model calibration approaches in the RobustCalibration R package available on the Comprehensive R Archive Network (CRAN) [10], which allows for both MLE and Bayesian estimation by posterior sampling. The parallel partial Gaussian stochastic process emulator for computer models with scalar-valued or vectorized output [11] can be called as a surrogate model from the package when the computer model is expensive. The source code of the examples in this article is available at https://github.com/UncertaintyQuantification/SGaSP-Theory.

2. The scaled Gaussian stochastic process.

Denote $()$ GaSP$(0; {}^2K(;))$ with variance ^{2} and correlation function $K(;)$ such that, for any inputs $fx_ig_{i=1}^n$, the marginal distribution $((x_1); : : : ; (x_n))^T$ follows a multivariate normal distribution with covariance $\mathrm{Cov}((x_i); (x_j)) = {}^2K(x_i; x_j)$. In order to have the mathematical model explain more variability, the S-GaSP prior introduced in [12] places more probability mass on a smaller random L_2 distance between the mathematical model and reality, as this measure is often used to quantify how well a mathematical model fits the reality. The S-GaSP calibration model is defined as the following hierarchical model :

$$(4)\quad \begin{aligned} & y^F(x) = f^M(x;) + {}_z(x) + ; \\ & {}_z(x) = \quad (x) \ j \ \int_{2X} {}^2()d = Z \quad ; \\ & () \ \ \mathrm{GaSP}(0; {}^2K(;)); \\ & Z \ \ p_Z(); \qquad N(0; {}_0); \ {}^2 \end{aligned}$$

where, conditional on all parameters, the default choice of $p_Z()$ is defined as

$$(5)\quad p_Z(z) = \frac{g_Z(z)p\ (Z = z)}{\int_0^1 g_Z(t)p(Z = t)dt};$$

with $g_Z(z)$ being a nonincreasing scaling function and $p(Z = z)$ being the density of Z at z induced by a GaSP with mean 0 and covariance ${}^2K(;)$.

We call ${}_z()$ in (4) the scaled Gaussian stochastic process (S-GaSP). Given $Z = z$, S-GaSP becomes a GaSP constrained at the space $\int_{x2X} {}^2(x)dx = z$. Note that Z represents the L_2 distance between the reality and mathematical model. Given that $g()$ is a nondecreasing function, the measure for Z induced by S-GaSP has a larger prior probability mass near 0 than the one by GaSP, reflecting one's belief that the L_2 loss between the mathematical model and reality should be small.

It is easy to see that when $g_Z()$ is a constant function, S-GaSP reduces to GaSP without any constraint. Conditioning on all parameters, we assume

$$(6)\quad g_Z(z) = \frac{}{2^{2z}}\exp \quad \frac{z}{2} ; \frac{z}{2}$$

with a scaling parameter ${}_z$. We select $p_Z()$ in (5) and $g_Z()$ in (6) for computational reasons, as any marginal distribution of ${}_z$ still follows a multivariate normal distribution [12, Lemma 2.3]. Other scaling functions may also be used, but we do not pursue them in this study.

2.1. Orthogonal series representation and marginal distribution.

Based on Karhunen{ Loeve theorem, GaSP with a stationary kernel admits the representation for any $x \ 2 \ X$;

$$(7)\quad (x) = \sum_{k=1}^{1} p_{\ k} Z_k \, {}_k(x) ;$$

where $Z_k \overset{i:i:d}{\ } N(0; 1)$, and ${}_k$ and ${}_k()$ are the kth eigenvalue and eigenfunction of the kernel $K(;)$, respectively. The S-GaSP can also be represented as an orthogonal series as shown below.

Lemma 2.1 (Karhunen{Loeve expansion for the S-GaSP). Assume $p_Z()$ and $g_Z()$ are defined in (5) and (6), respectively. For any x 2 X, the S-GaSP defined in (4) can be written as

$$ {}_z(x) = \sum_{k=1}^{1} \sqrt{\frac{{}_k}{1+{}_{z\,k}}}\, Z_k\, {}_k(x); $$

where $Z_k \overset{i.i.d.}{} N(0;1)$, and $_k$ and $_k()$ are the k th eigenvalue and eigenfunction of the kernel $K(;)$, respectively.

The covariance function of S-GaSP can also be decomposed as an infinite orthogonal series, which is an immediate consequence of the fact that S-GaSP is indeed a GaSP with a transformed kernel (see Lemma 2.3 in [12] and Lemma 2.1).

Corollary 2.2. Assume $p_Z()$ and $g_Z()$ are defined in (5) and (6), respectively. The S-GaSP defined in (4) follows a multivariate normal distribution

$$ [{}_z(x_1);:::;{}_z(x_n) \mid {}^2 R_z] \; MN(0;{}^2 R_z); $$

where the $(i;j)$ entry of R_z is

$$ K_z(x_i;x_j) = \sum_{k=1}^{1} \frac{{}_k}{1+{}_{z\,k}} (x_i)(x_j): \tag{8} $$

Corollary 2.2 implies that the ith eigenvalue of the kernel function $K_z(;)$ in S-GaSP is $_{z;k} := {}_k=(1+{}_{z\,k})$; and the kth eigenfunction $_k()$ is the same as the one in GaSP. The form (8) does not give an explicit expression for the kernel in S-GaSP. Instead of truncating the series, one may discretize the integral $\int_{x2X} {}^2(x)dx$, which leads to an explicit expression of the covariance matrix, discussed in section 4.

To see the difference between GaSP prior and S-GaSP prior, we generate $n = 200$ equally spaced inputs x_i 2 [0; 1] for $i = 1;:::;200$. The covariance matrix of these inputs is assumed to follow a unit variance Matern kernel with roughness parameter 5=2, and the $(i;j)$ term is

$$ K(d_{i;j}) = \left(1+\frac{\sqrt{5}\,d_{i;j}}{}+\frac{5d_{i;j}^2}{3\,{}^2}\right)\exp\left(\frac{\sqrt{5}\,d_{i;j}}{}\right); \tag{9} $$

where $d_{i;j} = |x_i \;\; x_j|$ for $i = 1;:::;n$ and $j = 1;:::;n$. The empirical eigenvalues are graphed as red symbols in the left panel of Figure 1, and the empirical CDF of the L_2 loss is shown in the right panel of Figure 1. First, when the correlation is large (i.e., large range parameter), the eigenvalues are typically more widely separated (left panel), and the prior probability of large L_2 loss $\int {}^2(x)dx$ is large, even if the variance is assumed to be the same (right panel). Large correlation in GaSP is sometimes needed for accurate predictions; however, it simultaneously means the prior probability mass of the L_2 loss (between the reality and computer model) is large, leading to an identifiability problem. In comparison, the induced eigenvalues by the S-GaSP are graphed as blue symbols in the left panel of Figure 1, where the large eigenvalues are truncated yet the small eigenvalues remain almost the same.

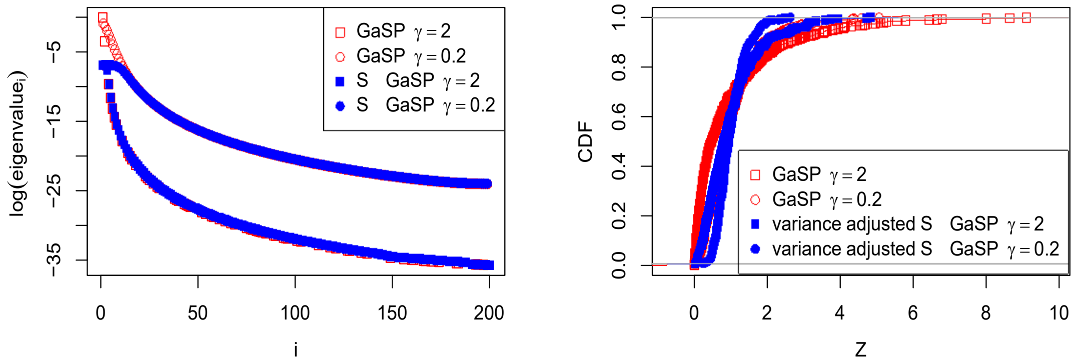

Figure 1. Eigenvalues and the cumulative distribution function (CDF) of the L_2 loss between GaSP and S-GaSP. The covariance is computed based on n = 200 inputs equally spaced from [0;1]. In the left panel, the red symbols give the logarithm of approximated eigenvalues $\tilde{}_i=n$ by the Matern kernel in (9) with two range parameters, where $\tilde{}_i$ denotes the empirical eigenvalue of covariance matrix ^2R, with the $(i;j)$th term being $R_{i;j}$ = $K(d_{i;j})$ from the Matern kernel in (9) and ^{2} = 1. The scaled eigenvalues $\tilde{}_i = \frac{\tilde{}_i=n}{n+ {}_z}$ by S-GaSP, with ${}_z$ = 10^{3} are graphed by the blue symbols, for i = $1;:::;n$. In the right panel, the empirical CDF of approximated L_2 loss Z = $\int {}^2(x)dx$ of GaSP and S-GaSP is shown. Each curve is computed based on 500 simulations, each containing n = 200 observations. The variance parameter ^{2} in S-GaSP is adjusted such that the summation of the eigenvalues is the same as the GaSP model with the Matern kernel.

Since we leave the variance as a free parameter estimated by the data, the spread of eigenvalues in S-GaSP is less extreme than that in GaSP. In the right panel of Figure 1, we plot the empirical distribution of the L_2 loss by S-GaSP, and the variance is adjusted such that the summation of the eigenvalues between GaSP and S-GaSP is the same. The S-GaSP has less prior mass on the large L_2 loss, reducing the identiability problem caused by large correlation.

The following Corollary 2.3 provides a decomposition of the L_2 loss Z in S-GaSP, which follows from Lemma 2.1 in [12] and Corollary 2.2.

Corollary 2.3. Assume the same conditions in Lemma 2.1 hold. The distribution of Z = $\int_{x \in X} {}^2(x)dx$ induced by S-GaSP follows

$$Z \quad {}^2 \sum_{k=1}^{1} \frac{{}^2 k}{1 + {}_z k} \; {}_k(1);$$

where $f_k(1)g_{k=1}^{1}$ are independent chi-squared random variables with one degree of freedom.

Denote H and H_z as the reproducing kernel Hilbert space attached to GaSP with kernel $K(;)$ and S-GaSP with kernel $K_z(;)$, respectively. Let the native norm associated with $K(;)$ and $K_z(;)$ be $h;i_H$ and $h;i_{H_z}$, respectively. We conclude this subsection by the explicit connection between the inner product of GaSP and that of S-GaSP.

Lemma 2.4. Assume $p_Z()$ and $g_Z()$ are dened as in (5) and (6), respectively. Let $h()$ = $\sum_{i=1}^{1} h_i {}_i()$ and $g()$ = $\sum_{i=1}^{1} g_i {}_i()$ be the elements in H. It holds that

$$hh;gi_{H_z} = hh;gi_H + {}_z hh;gi_{L_2(X)}:$$

2.2. Joint estimation in the S-GaSP calibration. With the specification of $p_Z()$ in (5) and $g_Z()$ in (6), after marginalizing out $_z = [_z(x_1); : : : ; _z(x_n)]$, the marginal distribution of the field observations in (4) follows a multivariate normal distribution

$$(10)\qquad [y^F \; j \; ; {}^2; \; ;_{0z}] \quad MN(f^M; {}^2((n) \; {}^1 R_z + I_n));$$

with the regularization parameter $_0 := {}^2 = (n^2)$ and the $(i; j)$ entry of R_z as defined in (8).

Denote $L_z()$ as the likelihood for in (10). The joint estimator of $(; _z())$ can be written as a penalized kernel ridge regression (KRR) estimator [28], where both the reproducing kernel Hilbert space (RKHS) norm and the L_2 norm of the discrepancy function are penalized simultaneously.

Lemma 2.5. The maximum likelihood estimator $\hat{}_{;_z;n} := \text{argmax}_2 L_z()$ and posterior mean $_{;_z;n}() :\triangleq E[_z() \; j \; y^F ; \hat{}_{;_z;n}; ; _z]$ are the same as the following penalized KRR estimator:

$$(11)\qquad \hat{}_{;n}; \; \hat{}_{; ;n}() \;=\; \underset{()2H; \; 2}{\text{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} (y^F(x_i) \quad f^M(x_i;) \quad (x_i))^2 + kk^2_{H_z}$$

with $kk^2_{H_z} = kk^2_H + _z kk^2_{L_2(X)}$.

The connection between KRR and the posterior mean in a GaSP regression model is well known. Here we establish an analogous connection in the S-GaSP calibration. In Lemma 2.5, both the L_2 norm and the native norm of the discrepancy function are penalized in S-GaSP calibration. When the discrepancy function is modeled as a GaSP, however, the L_2 norm of the discrepancy function is not a direct penalty (see the supplementary material supplement.pdf [local/web 442KB]). This property of the S-GaSP calibration is the key to guaranteeing that, under some regularity conditions, the estimated calibration parameters obtained from the joint estimation procedure (11) converges to the L_2 minimizer. A more detailed discussion is provided in section 3.

3. Convergence properties of the S-GaSP calibration.

We discuss two convergence properties of the S-GaSP calibration in this section. First, the predictive mean estimator of the reality converges to the truth at the optimal rate with suitable choices of the regularization and scaling parameters. Second, the estimated calibration parameters by the S-GaSP calibration converge to $_{L_2}$ when sample size increases. These two properties are obtained by jointly estimating the discrepancy function and calibration parameters in (11).

3.1. Convergence to the reality. Let us first consider the nonparametric regression,

$$(12)\qquad y(x_i) = f_0(x_i) + {}_i; \quad {}_i \overset{i:i:d:}{\sim} N(0; {}_0^2); \qquad i = 1; : : : ; n;$$

where $f_0()$ denotes underlying truth. Here we place a zero-mean S-GaSP prior on the unknown truth with the default choices of $p_Z()$ and $g_Z()$ as in (5) and (6), respectively. This is a special case where the mathematical model is zero, i.e., $f^M(x;) = 0$, and we will soon extend it to the general case when the mathematical model is not zero. For illustrative purposes, we follow [25] to assume that $x_1; : : : ; x_n$ are independently sampled from $\text{Unif}([0; 1]^p)$.

Assume the underlying truth $f_0() := E_y[y()]$ resides in the p-dimensional Sobolev space,

$$(13)\qquad W_2^m(X) = \left\{ f() = \sum_{k=1}^{1} f_k {}_k() \; 2 \; L_2(X) : \sum_{k=1}^{1} k^{2m=p} f_k^2 < 1 \right\};$$

with smoothness $m >$ p=2; and $f_k()g^1_{k=1}$ being a sequence of the orthonormal basis of $L_2(X)$. For any integer vector $k = (k_1; : : : ; k_p)^T$ and a function $f(x_1; : : : ; x_p) : X ! R$, denote by D^k the mixed partial derivative operator $D^k f() := @^{jkj}f()=@^{k_1}x_1 : : : @^{k_p}x_p$ with $jkj = P^p_{i=1} k_i$. For any function in $W_2^m(X)$, we have $kD^k f()k_{L_2(X)} < 1$ for any $jkj < m$.

Recall that $= _0=(n\ )$ in (10). By Lemma 2.5, the posterior mean estimator of $f()$ with a S-GaSP prior is equivalent to the KRR estimator

$$(14)\qquad \hat{f}_{;_z;n} = \underset{f\,2\,H}{\operatorname{argmin}} \; " \frac{1}{n} X^n_{i=1} (y(x_i) \quad f(x_i))^2 + kfk^2_H + {}_z kfk^2_{L_2(X)} \# :$$

Recall that $f_k g^1_{k=1}$ and $f_k g^1_{k=1}$ are eigenvalues and eigenfunctions of the kernel $K(;)$ associated with H, respectively. For all k, we assume the eigenvalues satisfy

$$(15)\qquad ck^{\;2m=p} \;\; _k \;\; Ck^{\;2m=p}$$

for some constants c and $C > 0$. For all $k\ 2\ N^+$ and $x\ 2\ X$, we assume the eigenfunctions are bounded uniformly,

$$(16)\qquad \sup_{x\,2\,X} \sup j\ _k(x)j \;\; C\ ;_{k1}$$

where $C > 0$ is a constant depending on the kernel $K(;)$.

We are now ready to state the convergence rate of the S-GaSP for the nonparametric regression model in (12).

Theorem 3.1. Assume the eigenvalues and eigenfunctions of $K(;)$ satisfy (15) and (16), respectively. Further assume $f_0\ 2\ W_2^m(X)$; and denote $:= (2m \quad p)^2=f2m(2m + p)g$. Consider the nonparametric regression model (12). For sufficiently large n, and any > 2 and $C\ 2\ (0;1)$, with probability at least $1 \quad \exp f\ (\quad 2)=3g \quad \exp \quad n^C$, we have

$$\hat{} \quad f_0 k_{L_2(X)} \;\; 2 \; h p \; 2 \; kf_0k_{L_2(X)} + kf_0k_H \; + C_{K\,0} \quad \frac{p}{n} \; i \; kf_{;z;n\,\frac{}{m}}^{\;\;2m+p}$$

and

$$k\hat{f}_{;z;n} \quad f_0k_H \;\; 2 \; h p \; 2 \; kf_0k_{L_2(X)} + kf_0k_H \; + C_{K\,0} \quad p \;\; i$$

by choosing $= n^{\;2m=(2m+p)}$ and ${}_z = \;\; ^{1=2}$, where the constant C_K depends on $K(;)$.

Our proof for Theorem 3.1 stems from [31], where the convergence rate was proved in a nonparametric regression model under the supremum norm. The proof of the convergence rate in Theorem 3.1 has two main differences in comparison to the proof of convergence of the nonparametric regression in [31]. First, ${}_z$ can go to infinity in Theorem 3.1, and consequently $jjjj_{H_z}$ is unbounded, whereas $jjjj_H$ was assumed bounded in proving the convergence of a nonparametric regression approach in [31]. Second, we generalize the proof to multivariate inputs. These two conditions make the proof more challenging. To solve this problem, we substantially modify the tools (e.g., see Lemma SM3.2 in the supplementary material) to prove Theorem 3.1.

The conditions in Theorem 3.1 can be relaxed in various ways. First, from the proof of Theorem 3.1, it is easy to see that if $= O(n^{2m=(2m+p)})$ and $_z \; O(^{1=2})$, the estimator still converges to the truth in the L_2 distance with the same rate $O(n^{m=(2m+p)})$. Second, the design can be generalized to other space filling designs. Additionally, although the stationarity of the process is often assumed for computational purposes, it is not required in Theorem 3.1. Note that the regularity parameter and kernel parameters are held fixed in Theorem 3.1. We discuss the estimation of and the parameters in the kernel function in section 4.1.

We are ready to discuss the convergence of estimating the reality in the calibration. The estimator for the reality in the S-GaSP calibration model is defined as

$$\hat{y}^R_{;z;n}(x) := f^M(x; \hat{}_{;z;n}) + \hat{}_{;z;n}(x) \tag{17}$$

for any $x \; 2 \; X$, where $(\hat{}_{;z;n}; \hat{}_{;z;n})$ is the estimator of the penalized KRR obtained by minimizing the loss in (11). The following Corollary 3.2 gives the convergence rate of the S-GaSP calibration model in predicting the reality. Similar to the extension for Theorem 3.1, the conditions in Corollary 3.2 can be relaxed by letting $= O(n^{2m=(2m+p)})$ and $_z \; O(^{1=2})$ to obtain the same convergence rate.

Corollary 3.2. Assume $y^R() \quad f^M(;) \; 2 \; W^m_2(X)$ for any 2 and $\sup_2 ky^R() \quad f^M(;)k_H < 1$. Let the eigenvalues and eigenfunctions of $K(;)$ satisfy (15) and (16), respectively. For sufficiently large n, any $> 2;$ and $C \; 2 \; (0; 1)$, with probability at least $1 \quad \exp f \; (\quad 2)=3g \quad \exp(\; n^C)$, we have

$$k\hat{y}^R_{;z;n}() \quad y^R()k_{L_2(X)} \; 2 \; {}^p 2 \; \bar{} \sup_2 ky^R() \quad f^M(;)k_{L_2(X)} + \sup ky^R() \quad f^M(;)k_H \quad + C_{K0} \; {}^p \; n^{\frac{}{2m+p} \frac{m}{2}}$$

by choosing $= n^{2m=(2m+p)}$ and $_z = \; ^{1=2}$, where $= (2m \quad p)^2=(2m(2m+p))$ and C_K is a constant only depending on the kernel.

In the example below, we illustrate the convergence using the function studied in [31], where $y^R()$ lies in the Sobolev space $W^m_2(X)$ with $m = 3$ and $X = [0; 1]$ when $J \; ! \; 1$.

Example 1. Let the reality be $y^R(x) = 2 \sum^J_{j=1} j \; ^6 \cos(5(j \quad 0{:}5)x) \sin(5j)$; and consider $y^F(x) = y^R(x) + \;$, where $N(0; 0{:}05^2)$ independently. Let $f^M(x;) = \;$. The goal is to predict $y^R(x)$ at $x \; 2 \; [0; 1]$ and estimate .

As a motivating example, we let $K(;)$ follow the Matern kernel in (9) with the range parameter $= 1$, as the reproducing kernel Hilbert space attached to the GaSP with this kernel is equal to Sobolev space $W^3_2(X)$. We test 50 configurations with the number of observations $n \; 2 \; [\exp(5); \exp(10)]$, and the design points $fx_i g^n_{i=1}$ are equally spaced in $[0; 1]$. In each configuration, $N = 100$ simulation replicates are implemented, and we let $J = 100$ in each simulation. We then compute the average root of the mean squared error as follows:

$$\text{AvgRMSE}_{f^M+} = \frac{1}{N} \sum^N_{i=1} \sqrt{\frac{1}{n} \sum^n_{j=1} (\hat{y}^R_i(x_j) \quad y^R(x_j))^2}; \tag{18}$$

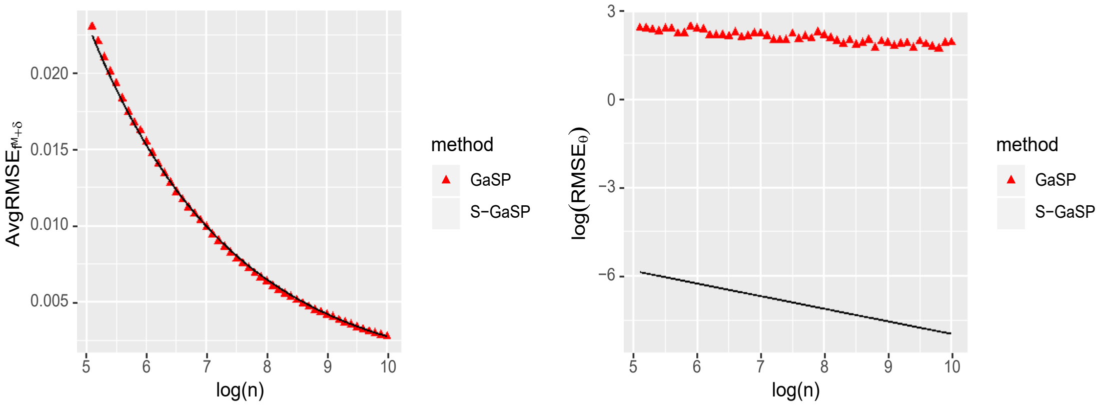

Figure 2. Prediction and calibration by the GaSP and discretized S-GaSP calibration models for Example 1. In the left panel, the AvgRMSE$_{f^M+}$ of the GaSP calibration and that of the discretized S-GaSP calibration are graphed as the red triangles and blue dots, respectively; the black curve is n $^{m=(2m+p)}=5$, representing the upper bound by Corollary 3.2 (up to a constant). In the right panel, the natural logarithm of the RMSE of the GaSP calibration and that of the discretized S-GaSP calibration are graphed as the red triangles and blue circles, respectively; the black line is log(n $^{m=(2m+p)}=40$), the upper bound from Theorem 3.3 (up to a constant). = n $^{2m=(2m+p)}$ 10 ^{4} and $_z$ = $^{1=2}$ are assumed.

where $\hat{y}^R_i(x_j)$ is an estimator of the reality at x_j for j = 1;:::;n. The subscript f^M + indicates that both the calibrated mathematical model and the discrepancy can be used for prediction.

For both GaSP and S-GaSP calibration, the joint estimator, i.e., the posterior mean of the reality and the MLE of the calibration parameters discussed in Lemma 2.5, is implemented for each experiment at each conguration. In the left panel of Figure 2, the posterior mean estimator of the reality in both GaSP and S-GaSP calibration converges to the reality at the same rate, which matches the theoretical upper bound from Corollary 3.2. Here, for computational purposes we graph the results of discretized S-GaSP calibration, which replaces the integral $\int_{x \in X}$ $^2(x)dx$ in the S-GaSP model in (4) by (1=n) $\sum_{i=1}^{n}$ $^2(x_i)$ in Figure 2. The discretized S-GaSP calibration is discussed in section 4.

To evaluate whether the calibrated mathematical model (here only a mean parameter) ts the data, we use the root of the mean squared error between the estimator of the calibration parameters and the L_2 minimizer, i.e., RMSE = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{}_i \quad _{L_2})^2}$, where $\hat{}_i$ is an estimator of in the ith experiment.

Although GaSP and S-GaSP perform equally well in predictions for Example 1, and the estimator of the calibration parameter in the discretized S-GaSP calibration converges to the L_2 minimizer, that in the GaSP calibration does not converge to $_{L_2}$, shown in the right panel of Figure 2. This problem is caused by the dierence between the RKHS norm and the L_2 norm. As illustrated in Lemma 2.5, both the RKHS norm and L_2 norm of the discrepancy function are penalized in the S-GaSP calibration model, whereas the GaSP calibration model does not penalize the L_2 norm of the discrepancy function. In section 3.2, we further show that under some regularity conditions, the calibrated parameters in the S-GaSP calibration do converge to the L_2 minimizer with the same choices of regularization parameter and scaling parameter as in Corollary 3.2.

3.2. Convergence of calibration parameters. We first list some regularity conditions for the convergence of calibration parameters. These conditions are also assumed in other studies for computer model calibration [25].

A1: $_{L_2}$ is the unique L_2 minimizer and an interior point of .

A2: The Hessian matrix $\int \frac{@^2(y^R(x) \quad f^M(x;))^2}{@@^T}dx$ is invertible in a neighborhood of $_{L_2}$.

A3: For all $j = 1;:::;q$, it holds that $\sup_2 \left\| \frac{@f^M(;)}{@_j} \right\|_H < 1$.

A4: The function class $fy^R() \quad f^M(;) : 2 g$ is Donsker. A5: $\sup_2 ky^R() \quad f^M(;)k_H < 1$.

A6: The eigenvalues and eigenfunctions of $K(;)$ satisfy (15) and (16), respectively.

Assumptions A1{A3 are regularity conditions of $_{L_2}$ and the mathematical model $y^M(;)$ around $_{L_2}$. Assumption A1 states that the L_2 loss minimizer is uniquely defined. If this condition is violated, reparameterization of calibration parameters may be required to reduce the intrinsic dimension of the parameter space. Assumptions A2{A3 ensure the Hessian matrix of the squared residuals is non-singular and the derivatives of computer models are bounded in the RKHS norm. The Donsker property in Assumption A4 is a standard requirement and ensures that the covering number of the function space for the discrepancy does not increase too fast [16]. Assumptions A5{A6 ensure the residuals between the reality and computer model are well-behaved such that the KRR estimator $\hat{}_{z;}$ converges to $y^R() \quad f^M(;)$ uniformly for each 2 in terms of the L_2 loss.

Below, Theorem 3.3 guarantees the convergence of calibration parameters. As the calibration parameters and discrepancy function are estimated jointly in our approach, we extend the tools for proving the convergence of the two-step calibration approach [25] to prove Theorem 3.3. The detailed proof is given in the supplementary material (supplement.pdf [local/web 442KB]).

Theorem 3.3. Under assumptions A1{A6, the estimated parameters in (11) satisfy

$$\hat{}_{;z;n} = _{L_2} + O_p(n^{\frac{m}{2m+p}});$$

provided that $= O(n^{2m=(2m+p)})$ and $_z = O(^{1=2})$.

Note that $= O(n^{2m=(2m+p)})$ and $_z = O(^{1=2})$ also guarantee that the posterior mean estimator in the S-GaSP calibration converges to the reality at the rate $O_p(n^{m=(2m+p)})$ in terms of the L_2 loss. We briefly compare Theorem 3.3 with some existing results. The convergence rate of the calibration parameter is slower than $O(n^{1=2})$ obtained in the two-step approach in both [25] and its Bayesian counterpart [30]. In [25] the authors established the $\sqrt{n}$-consistency of the L_2-calibration estimator, whereas a semi-parametric Bernstein-von Mises (BvM) theorem was developed in [30], providing a Bayesian analogy of the $\sqrt{n}$-consistency and the semi-parametric efficiency of the L_2 calibration method. In these approaches, the model of interest is a semiparametric model in the sense that the calibration parameter can be viewed as a functional of the unknown infinite- dimensional regression function y^F by considering the functional below

$$(y^F) = \mathrm{argmin} k y^F(x) \quad f^M(x;) k^2_{L_2(}$$

$$):$$

In contrast, the S-GaSP model does not incorporate the above explicit constraint directly. This is different from the work of [30], in which the authors formulated the prior specification by first assigning a Gaussian process prior model on y^F and then established the semiparametric BvM theorem of the posterior distribution of the induced posterior on (y^F). In the current framework, such an exact constraint is not directly imposed on the relation between y^F and . Due to the doubly-penalized least squares approach, given a realization of the discrepancy function (and hence, that of y^F), is not directly determined by (y^F) but follows a distribution concentrated around (y^F) a posteriori instead. Therefore, the semi-parametric BvM theorem no longer applies here. In addition, this extra layer of randomness results in a slower convergence rate than the parametric rate that can be obtained in a sufficiently regular semiparametric model [6].

Though the $O(n \; ^{1=2})$ rate may be obtained by choosing $= O(n \; ^{2m=(2m+p)})$ and $_z = O(\; ^{1=2})$, we should be aware that, however, the L_2 minimizer is not the true calibration parameter but the one that minimizes the L_2 distance between the calibrated mathematical model and reality. In practice, the residuals between reality and the calibrated mathematical model with the L_2 minimizer may behave like noises, making them hard to be estimated by a nonprametric regression model alone. In comparison, the joint estimation of the discrepancy function and calibration parameters was found to have a smaller predictive error in numerical examples.

The calibrated parameters of the GaSP calibration, on the other hand, typically do not converge to the L_2 minimizer. Let $\frac{@f^M(;)}{@_j}\hat{} := \frac{@f^M(;)}{@_j} j_{= \hat{}}$. A key difference between the GaSP and the S-GaSP calibrations is stated in the following Corollary 3.4, which is an immediate consequence of the proof of Theorem 3.3.

Corollary 3.4. Under assumptions A1{A6, the estimator for the calibration parameters in the S-GaSP calibration in (11) satisfies

$$\frac{1}{z} \, _{;z;n}(); \; \frac{@f^M(;_{;z;n})}{@_j} \hat{} \; _H + \; _{;z;n}(); \; \frac{@f^M(;_{;z;n})}{@_j} \hat{} \; _{L_2(X)} = 0:$$

Further assuming the mathematical model is differentiable at $_{;n}$, where $\hat{}_{;n}$ is the estimator of via replacement of the norm $jj \; jj_{H_z}$ by $jj \; jj_H$ in (11), we see that the estimator of the calibration parameters in the GaSP calibration satisfies

$$_{;n}(); \; \frac{@f^M(;_{;n})}{@_j} \hat{} \; _H = 0$$

for any $_j$, $j = 1; : : : ; q$.

To ensure the convergence of an estimator $\hat{}$ to the L_2 minimizer, one typical requirement is that $h_{L_2}(); \frac{@f^M(;)}{@_j} j \hat{} \; i_{L_2(X)} = o_p(1)$. It is easy to see that the S-GaSP satisfies this condition with $1=_z = o(1)$. However, because of the difference between the RKHS norm and the L_2 norm, the estimated parameters $\hat{}_{;n}$ in the GaSP calibration model can be far away from the L_2 minimizer. As a result, the calibrated mathematical model may not fit the data in the GaSP calibration model, as found in previous studies [25, 29]. In Example 1, the estimated parameters in the discretized S-GaSP calibration converge to the L_2 minimizer when the sample size increases, whereas the estimated parameters in GaSP calibration with an unscaled kernel function do not converge, as shown in the right panel of Figure 2.

4. Discretized scaled Gaussian stochastic process.

We address the computational issue in the S-GaSP calibration in this section. Instead of truncating the kernel function in (8) by the first several terms, we select N_C distinct points to discretize the input space $[0; 1]^p$ and replace $\int_{2X} ()^2 d$ by $(1=N_C) \sum_{i=1}^{N_C} (x_i^C)^2$ in the S-GaSP model in (4).

Here we let the discretization points be the observed inputs, i.e., $x_i^C = x_i$ for $i = 1; : : : ; N_C$ and $N_C = n$. The discretized S-GaSP is to replace $_z$ in (4) by

$$(19) \qquad {}_{z_d}(x) = \left\{ (x) \; j \; \frac{1}{n} \sum_{i=1}^{n} (x_i)^2 = Z_d \right\} ;$$

with density $p_{Z_d}()$ as defined in (5). After marginalizing out Z_d, it follows from Lemma 2.4 in [12] that $_{z_d}()$ is a zero-mean GaSP with the covariance function

$$(20) \qquad {}^2 K_{z_d}(x_a; x_b) = {}^2 (K(x_a; x_b) \quad r^T(x_a) \tilde{R}^{\;1} r(x_b))$$

for any $x_a; x_b$ 2 X, where $\tilde{R} := R + n I_n =_z$. Denote the ith largest eigenvalues of R and $\tilde{R}$ by ~ and ~ $_{;i\cdot z}$ It is easy to see ~ $_{;i}=n$ = ~ $=_i fn(1 + {}_z$~ $=n)g$, which coincides with the shrinkage of eigenvalues of a nondiscretized S-GaSP in (8), by using ~ $=n$ as an empirical approximation to $_i$.

Recall = $^2=(_0 n^2)$. We have the following predictive distribution of the discretized S-GaSP calibration model.

Theorem 4.1. Assume $_{z_d}()$ in (19) with $p_{Z_d}()$ and $g_{Z_d}()$ defined as in (5) and (6), respectively. The predictive distribution of the field data at any x 2 X by the discretized S-GaSP calibration model in (19) is a multivariate normal distribution

$$y^F(x) \; j \; y^F ; ; {}^2 ; ;_0 \quad MN(\hat{}_{z_d}(x); {}^2((n) \;\; {}^1 K_{z_d}(x; x) + 1));$$

where $\hat{}_{z_d}(x) = f^M(x;) + \frac{r^T(x)}{1+_z} \left(R + \frac{n}{1+_z} I_n \right)^{1} \left(y^F \quad f^M \right)$; and $K_{z_d}(x; x) = K(x; x)$ $r^T(x) \left(I_n + \left(R + \frac{n}{1+_z} I_n \right)^{1} \frac{n}{(1+_z)_z} \right) R \;\; \tilde{}^{1} r(x)$, with $r(x) = (K(x; x_1); : : : ; K(x; x_n))^T$; and $\tilde{R} = R + \frac{n}{_z} I_n$ with the $(i; j)$ entry of R being $K(x_i; x_j)$.

Theorem 4.1 indicates that the predictive mean of the reality in the discretized S-GaSP calibration model shrinks the posterior mean towards the mean function. When $_z = 0$, the shrinkage is zero, and the discretized S-GaSP becomes the GaSP.

Interestingly, when the observations contain no noise, the predictive mean and variance of the field data from the GaSP calibration model and the discretized S-GaSP calibration model are exactly the same, as stated in the following Lemma 4.2.

Lemma 4.2. Assume the conditions in Theorem 4.1 hold. If $^2_0$ = 0, the predictive distribution of the field data at any x 2 X by the discretized S-GaSP model in (19) is a multivariate normal distribution with the predictive mean and variance as follows:

$$E[y^F(x)\; j\; y^F; ; ; _z] = f^M(x;) + r^T(x) R \;\; ^1(y^F \qquad f^M);$$
$$V[y^F(x)\; j\; y^F; ; ; _z] = \; ^2 \;\; K(x; x) \qquad r^T(x) R \;\; ^1 r(x) \;\; ;$$

where $r(x) = (K(x; x_1); : : : ; K(x; x_n))^T$; and the $(i; j)$ entry of R is $K(x_i; x_j)$.

The discretized S-GaSP with the discretization points on the observed input has the same computational complexity as GaSP, as computing the inverse and determinant of the covariance matrix in the likelihood function takes $O(n^3)$ operations, where n is the number of field observations. The computational complexity between different calibration approaches will be compared in section 5.2.

4.1. Parameter estimation. We discuss the computational issue and the parameter estimation in this section. All the approaches are implemented in the RobustCalibration R package available on CRAN [10]. First, some of the mathematical models are numerical solutions of partial differential equations implemented as computer code, which is computationally expensive to run. In these cases, one often uses a statistical emulator to approximate the computer model based on a set of model runs [21, 23]. The GaSP emulator from the RobustGaSP R package is used to emulate the computer model for both scalar-value and vector-valued output when it is expensive to run in the RobustCalibration R package. One unique feature of the RobustCalibration R package is that the parallel partial Gaussian processes [11] from the RobustGaSP R package can be called to emulate computer models with massive numbers of coordinates.

We next discuss the estimation of the regularization parameters and the kernel parameters, which were held fixed in some studies. In practice, estimating these parameters rather than fixing them can improve the predictive performance. For any $x_a = (x_{a1}; : : : ; x_{ap})^T$ and $x_b = (x_{b1}; : : : ; x_{bp})^T$, the kernel is often assumed to have a product form [15],

$$K(x_a; x_b) = \prod_{i=1}^{p} K_i(d_i); \tag{21}$$

where $d_i = jx_{ai} \quad x_{bi}j$ for $i = 1; : : : ; p$, and $K_i()$ is a one-dimensional kernel function. One widely used kernel function is the Matern kernel [13]. When the roughness parameter is a half-integer, the Matern kernel has an explicit expression. For example, the Matern kernel with roughness parameter 5=2 is given in (9). A good feature of the Matern kernel is that the sample path of a GaSP is d_i 1e times differentiable, where $_i$ is the roughness parameter.

Denote by $= (_1; : : : ; _p)^T$ the unknown range parameters in the covariance. The parameters in the discretized S-GaSP calibration model are the calibration parameters , the variance parameter of the noise $^2_0$, the regularization parameter $= \;^2 = (n \;^2_0)$, the scaling parameter $_z$ and range parameters . The MLE of the variance of noise follows $\hat{}^2_{0;MLE} = S^2_{z_d}$ where $S^2_{z_d} = (y^F \quad f^M)^T \tilde{R}^{\;1}_{z_d}(y^F \quad f^M)$ with $\tilde{R}_{z_d} = \quad R \;\;^1 + \;_z I_n = n \qquad ^1 + n I_n$; and $R^{\;1}_{z_d} = \;_z = (ng) + \;(R + \;n I_n = g) \;\;^1 = g^2$ with $g = \;_z + 1$. Marginalizing out $_z()$ and plugging $\hat{}^2_{0;MLE}$ into the likelihood of the discretized S-GaSP model in (19), one has the profile likelihood

$$L_{Z_d}(;;;_z) / \ |R_{z_d}|^{1=2}(S^2_{z_d})^{n=2}: \tag{22}$$

One may use an MLE approach to numerically maximize the profile likelihood in (22) to estimate the parameters. Note that $_z$ reflects one's tolerance of how well a mathematical model should predict the reality without the discrepancy function, and thus this parameter may be chosen based on the experts' knowledge. Because of the conditions discussed in Theorem 3.3, $_z$ may be fixed to be proportional to $^{1=2}$ or be related to the sample size. For all numerical examples, the scaling parameter of the S-GaSP is $_z = (c_z jjjj)^{1=2}$, where by default $c_z = 1$, $=$ $^2=(^2n)$; and $= (\tilde{}_1; :::; \tilde{}_p)^T$, with $\tilde{}_i$ being the normalized range parameter (normalized by the length of each coordinate of the input variable). This choice is also implemented as a default choice in the RobustCalibration package, and users can specify this parameter as well. The inclusion of the range parameters is due to the confounding issue between the range parameter and the variance parameter of the process, whereas the ratio of these parameters can typically be estimated consistently from the data [32].

In the RobustCalibration package, we implement both the MLE and the Bayesian method for estimating the parameters and making predictions. For the MLE, the low-storage quasi-Newton method was used for optimization [17]. For the Bayesian method, we assume that the prior distribution is $(;;;^2) /\ ()(;_0)=^2$; with $=$ $^2=^2_0$ being the nugget parameter. Here $(;)$ is chosen as the joint robust prior for the kernel parameters [9], and $()$ may be specified by the user to reflect the experts' knowledge. We assume a uniform distribution of $()$ in the numerical examples for demonstration purposes. In contrast to the MLE, the uncertainty in parameter estimation can be obtained by the posterior samples.

5. Comparison between different calibration approaches.

We compare a few calibration approaches in this section. One of the most popular frameworks for computer model calibration is the GaSP calibration approach [15], which models the discrepancy in (1) as a GaSP. The mathematical model and discrepancy function are jointly estimated under the Bayesian framework. The S-GaSP approach is an extended version of GaSP calibration and places more prior probability mass of the L_2 norm of the discrepancy near zero. Consequently, the calibrated mathematical models in the S-GaSP calibration fit the data better than the ones in the GaSP calibration.

The S-GaSP calibration approach was inspired by a few pioneering approaches seeking to find the L_2 minimizer of the calibration parameters [19, 25, 29]. The orthogonal Gaussian process proposed in [19] constrains the derivatives of the random L_2 norm of the discrepancy to be zeros, equivalently giving more prior probability mass of calibration parameters at the stationary points of the calibration parameters in terms of the L_2 loss. The S-GaSP model explores another transformation that avoids putting more prior probability mass at the local maxima of the L_2 loss of the discrepancy, and it has a closed-form likelihood function. The L_2 calibration was proposed in [25], where the reality is first estimated by a KRR nonpara-metric regression, and the calibration parameters are then estimated by minimizing the L_2 loss between the estimated reality and mathematical model. The least squares (LS) calibra-tion is proposed in [29], where the calibration parameters are estimated by first minimizing the squared error between the mathematical model and observations, and a nonparamet-ric approach is then applied to estimate the residuals between the reality and mathematical model.

5.1. Numerical comparison of predictive accuracy by the calibrated computer model and discrepancy. We use the following example to illustrate that jointly estimating the discrepancy function and calibration parameters can be helpful for predicting the reality.

Example 2. Let $y^F(x) = y^R(x) +$, where $N(0;0:05^2)$ independently and $y^R(x) = g_1(x)+g_2(x)$, with $g_1(x) = \sum_{j=1}^{1} j \; ^1 \cos(5(j \;\; 0:5)x)\sin(5j)$ and $g_2(x) = \sum_{j=1}^{1} j \; ^6 \cos(5(j \;\; 0:5)x)\sin(5j)$. Let the mathematical model be $f^M(x) = g_1$. The goal is to predict $y^R()$ and estimate . For any x, the first 100 terms in $g_1(x)$ and $g_2(x)$ were used in the computation.

We compare the GaSP, S-GaSP, L_2; and LS calibration approaches for Example 2. For both the GaSP and S-GaSP approaches, the calibration parameter, variance, and kernel parameters are estimated by the MLE. For the two-step approaches, a GaSP model is used as the nonparametric regression model in the first step of L_2 calibration and in the second step of the LS calibration. We use the Matern kernel function in (9) as the kernel function $K(;)$ for all methods.

In the left panel of Figure 3, we found that the predictive errors by the GaSP and S-GaSP calibrations are considerably smaller than those of the LS and L_2 calibration approaches. This is because the reality contains $g_1(x)$, which is difficult to predict by a nonparametric regression approach alone. The mathematical model specified herein, however, can explain this term with calibration parameter close to 1. The estimated calibrated parameter in both GaSP and S-GaSP is indeed close to 1 in all experiments, which leads to better predictive performance.

The L_2 minimizer of the calibration parameters in this example is around 1.775. Note that the estimated calibration parameter S-GaSP calibration may not converge to the L_2 minimizer when sample size increases. This is because $jjy^R() \quad f^M(;)jj_H$ is unbounded when $= 1$, which violates assumption A5. The calibrated computer model with calibration

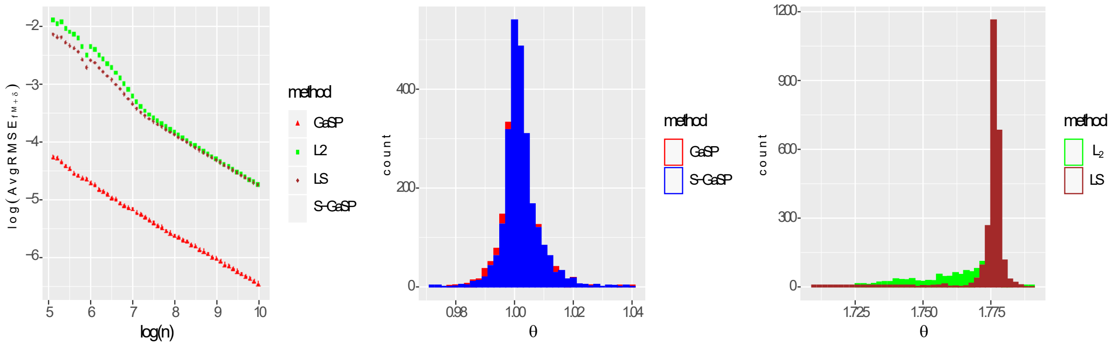

Figure 3. Comparison of different approaches in Example 2. In the left panel, the logarithm of the $AvgRMSE_{f^M+}$ for four calibration approaches is graphed at the logarithm of different sample sizes. The histogram of the estimated calibration parameter of each experiment for different approaches is given in the middle panel and the right panel.

parameter being around 1 improves the predictive accuracy and interpretation, as the residual discrepancy term is a smooth term that is easy to predict.

The GaSP and S-GaSP calibration approaches are not always more accurate in predicting the reality than the two-step approaches. Indeed, for Example 5, to be discussed in section 5, the prediction in the L_2 calibration approach is the best among all methods, as the observations can be easily predicted through a nonparametric regression without the mathematical model. When the reality is complicated, the joint estimation strategy implemented in both GaSP and S-GaSP calibration approaches seems to have smaller predictive errors, which will be illustrated by a few more numerical examples.

As the sampling model of the observations is well defined in both GaSP calibration and S-GaSP calibration, a Bayesian approach can be implemented, and the uncertainty of the parameters can be naturally obtained by their posterior distributions. For the L_2 calibration, the asymptotic distribution of the estimator of the calibration parameter may be used to approximately quantify the uncertainty in parameter estimation [25]. A bootstrap approach is developed to assess the uncertainty of parameters for the LS calibration approach [29].

5.2. Comparing computational complexity from different calibration approaches. The computational cost of computer model calibration may come from two parts of this method: evaluating the computer model, and computing the likelihood or loss function in calibration. First, for a slow computer model, the GaSP emulator discussed in section 4.1 can be used to approximate the output from the computer model. Second, the computational complexity of all the methods (L_2, LS, GaSP, and S-GaSP calibrations) discussed in section 5.1 is $O(n^3)$ for computing the likelihood function or loss function in general, where n is the number of field observations.

The MLE and the two-step approaches are faster than the Bayesian estimation by posterior samples, as they requires fewer iterations. However, there are two drawbacks to MLE and the two-step approaches. First, numerical optimization of the parameters could be unstable and could depend on the initial guesses when the numbers of calibration parameters and kernel parameters are large. Second, the uncertainty of the parameters can be obtained naturally by a Bayesian method but not straightforwardly by the MLE and the two-step approaches.

Both the MLE and posterior sampling approaches are implemented in the RobustCalibration package [10]. In the following, we will show results of both posterior sampling and MLE by the GaSP, S-GaSP, and no-discrepancy calibration approaches. The estimation by L_2 calibration and LS calibration will also be included for comparison.

6. Numerical study.

In this section, we compare the numerical performance of several methods based on the prediction error and the calibration error in (2) and (3), respectively, both evaluated with respect to the L_2 loss. The prediction error is our primary consideration, because the out-of-sample prediction for the reality examines how well we can reproduce the reality. The calibration describes how well the calibrated mathematical model fits reality in terms of the L_2 loss [25]. The parameters in a mathematical model often have scientific interpretation, whereas a nonlinear discrepancy function might not be interpretable. Thus the discrepancy function is not used for prediction in the second criterion.

For all examples, we compute $\mathrm{AvgRMSE}_{f^M+}$ in (18) and $\mathrm{AvgRMSE}_{f^M}$ through replacing prediction by the calibrated computer model output in (18). For assessing the uncertainty in predictions, we also compute the average length of the predictive interval and the proportion of the observations covered by the 95% predictive interval defined as follows:

$$L_{CI}(95\%) = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{i=1}^{n} \text{length}\{CI_{ij}(95\%)\},$$

$$P_{CI}(95\%) = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{i=1}^{n} 1\{y^R_j(x_i) \in CI_{ij}(95\%)\},$$

where $CI_{ij}(95\%)$ is the 95% predictive interval; N and n are the total number of experiments and of held-out test data in each experiment, respectively. An efficient method should have small $\mathrm{AvgRMSE}_{f^M+}$ and $\mathrm{AvgRMSE}_{f^M}$, short predictive interval, and $P_{CI}(95\%)$ close to the nominal 95%. For the real examples, we replace the test reality output by the held-out observations for out-of-sample predictions.

For GaSP, S-GaSP, and the no-discrepancy calibration (where the discrepancy is zero), the results by both posterior sampling and MLE are obtained by the RobustCalibration R package available from CRAN [10]. Five initial values were used to numerically optimize the calibration parameters and the kernel parameters in the MLE approach. Furthermore, we also implemented the L_2 calibration and LS calibration. The GaSP regression using the RobustGaSP R package is used to estimate the reality in the first step of the L_2 calibration and estimate the residual (between the calibrated mathematical model and the reality) in the second step of the LS calibration. The kernel function $K(\cdot,\cdot)$ is assumed to be the Matern covariance function (9) in all methods for demonstration purposes.

6.1. Simulated example. *Example* 3. Let $y^F(x) = y^R(x) + $, where $x = (x_1, x_2) \in [0, 1]^2$, $y^R(x) = \sin(0.2x_1)x_2 + \sin(2x_1)x_2 + 1$, and $_i \overset{iid}{\sim} N(0, 0.1^2)$. The mathematical model is $f^M(x,) = \sin(_1 x_1)x_2 + _2$ with $_1 \in [0, 10]$ and $_2 \in \mathbb{R}$.

We first consider a simulated study in Example 3 and test two configurations, where the sample sizes of the observations are taken to be $n = 25$ and $n = 50$, respectively. For each configuration, we test $N = 100$ experiments with $n = 2500$ held-out reality points, where the inputs are equally spaced in each interval. The input variable in each experiment is generated from the maximin Latin hypercube design [23]. Here we call the deSolve R package [24] to numerically solve the differential equation at each sampled or iterative parameter value. Emulating the vectorized output based on a GaSP emulator built in the RobustCalibration package is faster than directly calling the numerical solver for this example. It is shown in [10] that the calibration results based on the numerical solver and the emulator are quite similar for this example.

The predictive errors by different approaches are given in Table 1. First, we found that the $\mathrm{AvgRMSE}_{f^M+}$ of all methods that include a model of discrepancy is better than $\mathrm{AvgRMSE}_{f^M}$, as the mathematical model is imperfect compared to the reality. Second, the GaSP and S-GaSP calibration approaches perform better than the two-step LS and L_2 calibration methods in terms of $\mathrm{AvgRMSE}_{f^M+}$. Although around 95% of the held-out test data

Table 1

Predictive performance by different methods for Example 3. We show performance of the GaSP, S-GaSP, and no-discrepancy (N-D) calibration approaches based on posterior sampling (PS) and MLE in estimating parameters. Also included are the L_2 and LS calibration approaches. The smallest error is highlighted in bold.

n = 25	GaSP	GaSP	S-GaSP	S-GaSP	N-D	N-D	L_2	LS
	PS	MLE	PS	MLE	PS	MLE		
AvgRMSE_{f^M+}	.0556	.0686	.0558	.0656	/	/	.102	.0702
AvgRMSE_{f^M}	.143	.139	.131	.134	.130	.131	.131	.131
$L_{CI}(95\%)$	.206	.207	.209	.207	0.186	/	.320	.213
$P_{CI}(95\%)$	.931	.851	.932	.840	0.415	/	.871	.847
n = 50	GaSP	GaSP	S-GaSP	S-GaSP	N-D	N-D	L_2	LS
	PS	MLE	PS	MLE	PS	MLE		
AvgRMSE_{f^M+}	.0404	.0439	.0402	.0418	/	/	.0655	.0437
AvgRMSE_{f^M}	.144	.140	.130	.133	.128	.128	.128	.128
$L_{CI}(95\%)$	.153	.151	.151	.137	0.124	/	.245	.154
$P_{CI}(95\%)$	.938	.912	.935	.900	0.269	/	.949	.918

are covered by the 95% predictive interval in all methods, the average lengths of the predic-tive intervals by the GaSP and S-GaSP calibrations are shorter than those of the two-step approaches.

Second, the predictive error of the MLE approach by GaSP and S-GaSP is a little worse than that of the Bayesian approach by posterior sampling when the sample size is small, as the numerical optimization is less stable. Furthermore, as the uncertainty of calibration and the kernel parameters were not taken into account in the MLE, the proportion of the held-out data covered in the 95% interval ($P_{CI}(95\%)$) is typically much smaller than 95%.

For the L_2 calibration approach, the mathematical model is not used in the first step, and the parameters in the mathematical model are estimated in the second step to minimize the L_2 loss. Here, the high-requency term $(\sin(2x_1)x_2)$ makes the prediction in the first step by nonparametric regression less accurate than the joint estimation of the calibration parameters and the discrepancy function. For the LS approach, the residuals between the fitted mathematical model and the reality can be more difficult to predict by a GaSP model than a joint model of the mathematical model and discrepancy, as the residuals contain more changes of monotonicity locally than the reality.

The estimated calibration parameters of Example 3 are graphed in Figure 4. In the left panel, the estimate of ${}_1$ using the LS and L_2 methods is close to the L_2 minimizer (graphed as the solid line), whereas the estimate of ${}_1$ using the GaSP and S-GaSP is, in fact, closer to 2. This is because the model complexity is naturally built into the calibration: an estimated ${}_1$ that is close to 2 makes the prediction better, since the high-frequency term can be approximately explained by the calibrated mathematical model. The estimate of ${}_1$ being close to 2 is important to have better predictive performance, as shown in Table 1. The estimate of ${}_2$ is close to the L_2 minimizer by the S-GaSP, L_2; and LS calibration approaches, whereas GaSP calibration leads to an estimated ${}_2$ that has a large variability. Consequently, the GaSP calibration approach produces the larger calibration error (AvgRMSE_{f^M}) compared to other calibration methods, as illustrated in Table 1.

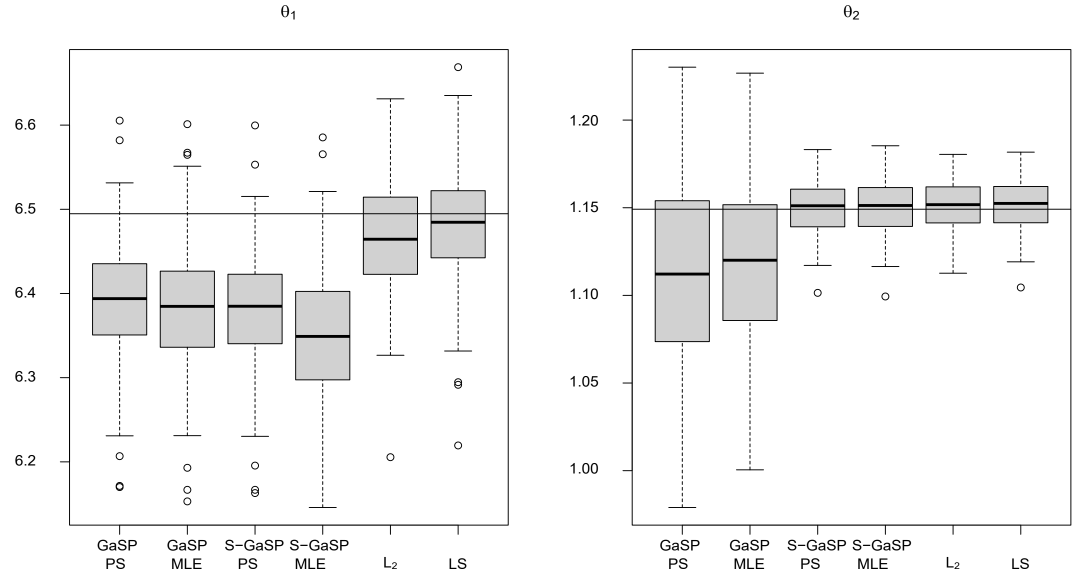

Figure 4. Boxplots of the posterior mean of calibration parameters from the GaSP with posterior sampling (PS), MLE, and from the S-GaSP, L_2, and LS calibration approaches for each experiment in Example 3. The no-discrepancy calibration with both PS and MLE are almost the same as L_2 and LS calibration, so they are not shown here. The solid lines are the L_2 minimizer, which is around 6:48 and 1:15 for $_1$ and $_2$, respectively.

Example 3 indicates that the calibrated mathematical model that minimizes the L_2 loss between the reality and mathematical model might not always be the optimal choice for predictions. In this example, when $_1$ is estimated to be close to 2, the predictive accuracy can be improved significantly. The other parameter, $_2$, is a mean parameter. The two types of errors are both small when $_2$ is estimated to be close to the L_2 minimize and when $_1$ is estimated to be close to 2, which is achieved in the S-GaSP calibration.

6.2. Chemical system interaction.

Example 4. Consider the system interaction between two chemical substances y_1 and y_2:

$$\underline{y}_1(t) = 10^{1}\ {}^{3}y_1(t);$$
$$\underline{y}_2(t) = 10^{1}\ {}^{3}y_1(t) \qquad 10^{2}\ {}^{3}y_2(t);$$

where 2 repeated observations of the second chemical substance are measured at each of the 6 time points $t =$ 10; 20; 40; 80; 160; 320 in [7]. The goal is to estimate $_1$ and $_2$ and to predict the values of the chemical substances across time.

We consider $(_1; _2)$ 2 $[0:5; 1:5]^2$. As the number of observations is limited, we first perform a leave-one-out comparison by holding out two repeated observations for prediction at each time point. We replace the reality in each criterion by the held-out observations to test each approach. The predictive performance of the leave-one-out comparison for Example 4 is given in Table 2.

First, the predictive errors ($\mathrm{AvgRMSE}_{f^M+}$) of GaSP or S-GaSP with posterior sampling are smaller than those of the MLE approach, as the Bayesian method is more stable, in

Table 2
Predictive performance by dierent methods for Example 4. The smallest error is highlighted in bold.

	GaSP, PS	GaSP, MLE	S-GaSP, PS	S-GaSP, MLE
$\mathrm{AvgRMSE}_{f^M+}$	5.19	6.47	5.39	6.50
$\mathrm{AvgRMSE}_{f^M}$	6.10	7.18	5.70	6.44
$L_{CI}(95\%)$	26.5	18.1	29.2	18.9
$P_{CI}(95\%)$	91:7%	83:3%	91:7%	83:3%
	N-D, PS	N-D, MLE	L_2	LS
$\mathrm{AvgRMSE}_{f^M+}$	/	/	10.2	6.39
$\mathrm{AvgRMSE}_{f^M}$	5.84	6.22	6.95	6.22
$L_{CI}(95\%)$	23.7	15.5	52.5	23.7
$P_{CI}(95\%)$	91:7%	75%	91:7%	100%

particular in a small sample scenario. The L_2 approach is less accurate in prediction in this example (shown in the upper panels in Figure 5), as predicting the reality by nonparametric regression can be inaccurate due to the small number of observations. On the other hand, the GaSP calibration with either posterior sampling or MLE has a large calibration error, based on predictions by the calibrated mathematical model alone. The L_2 calibration and no-discrepancy calibration have small $\mathrm{AvgRMSE}_{f^M}$. The S-GaSP calibration has a relatively small predictive error under both predictive criteria.

The upper panels in Figure 5 give posterior samples of the calibration parameters and MLE by GaSP, S-GaSP, and no-discrepancy calibration for Example 4. The posterior samples by S-GaSP and no-discrepancy calibration are very similar, and the MLEs by these two methods are close. This is reasonable, as the calibrated computer model ts the observations reasonably well, shown in the lower left panel. The posterior samples and MLE by GaSP are slightly dierent from those by S-GaSP or no-discrepancy calibration. Consequently, the calibrated computer model by GaSP underestimates the output, shown in the lower left panel in Figure 5.

6.3. Ion channel experiments.

In this example, we consider calibrating the mathematical model for the sodium ion channels using real observations from the whole cell voltage clamp experiments [20].

Example 5. The data sets consist of 19 observations of normalized current needed to maintain the membrane potential at 35mV over time [19]. Denote by the input variable x the natural logarithm of time. The mathematical model has the following expression:

$$f^M(x;\ ) = e_1^T \exp[\exp(x)A(\ )]e_4;$$

where e_i is a column vector with 1 at the ith element and 0 for the rest of the components, the rst exp is the matrix exponential, = $(_1; _2; _3)_T$, and

$$A(\ ) = \begin{pmatrix} {}_{2}\,{}_{3} & {}_{1} & 0 & 0 \\ {}_{2} & {}_{1}\,{}_{2} & {}_{1} & 0 \\ 0 & {}_{2} & {}_{2}\,{}_{1} & {}_{1} \\ 0 & 0 & {}_{2} & {}_{1} \end{pmatrix}:$$

The ranges of the calibration parameters considered here are $_i$ 2 [0; 10] for i = 1; 2; 3.

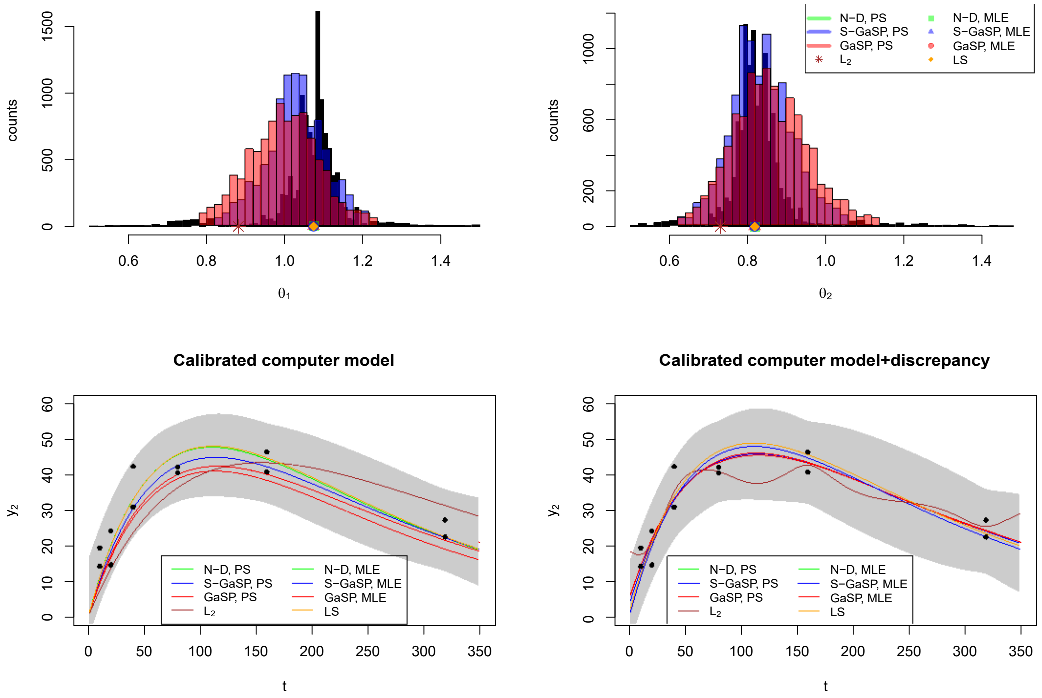

Figure 5. In the upper panels, histograms are posterior samples of two calibration parameters from the GaSP, S-GaSP, and no-discrepancy calibration approaches. Point estimates of MLE, L_2, and LS are plotted on the x-coordinate. In the lower left panel, calibrated computer models by different approaches were graphed, and the shared area is the 95% posterior interval from GaSP calibration. In the lower right panel, the predictive distribution of the summation of calibrated computer models and discrepancy is plotted, and the shaded area is the 95% posterior predictive interval from S-GaSP calibration. The black circles are observations.

We first consider leave-one-out comparisons for the ion channel experiment. The predictive errors by different approaches are given in Table 3. Here, since we have a moderately large number of observations for the one-dimensional input variable, GaSP regression by the RobustGaSP R package has high accuracy in predicting the reality even without the computer model. Including the mathematical model does not improves the prediction accuracy in this example. Therefore, the predictive error by the L_2 calibration is the smallest, as the GaSP regression without the mathematical model is used for predicting the reality in the first step. Between posterior sampling and the MLE approach, the posterior sampling typically has a smaller predictive error, and the predictive interval covers more held-out test points, which are consistent with previous examples.

The posterior samples and estimation of the reality by different calibration approaches are graphed in Figure 6. For better visualization, we reduce the posterior samples by 10 and only graph one-tenth of the total posterior samples. In this example, the posterior samples of the S-GaSP are closer to the ones with the no-discrepancy calibration. For this reason, the calibrated computer model by the S-GaSP calibration (graphed as the dashed blue curve in the lower right panel) fits the observations well, whereas the calibrated computer

Table 3
Predictive performance by dierent methods for Example 5.

	GaSP, PS	GaSP, MLE	S-GaSP, PS	S-GaSP, MLE
AvgRMSE$_{fM+}$	1:77 10^{3}	1:97 10^{3}	1:52 10^{3}	1:36 10^{3}
AvgRMSE$_{fM}$	1:13 10^{2}	1:75 10^{2}	5:62 10^{3}	5:76 10^{3}
L$_{CI}$(95%)	5:64 10^{3}	3:56 10^{3}	9:58 10^{3}	7:10 10^{3}
P$_{CI}$(95%)	94:7%	84:2%	100%	73:6%
	N-D, PS	N-D, MLE	L$_2$	LS
AvgRMSE$_{fM+}$	/	/	8:26 10^{4}	2:17 10 3
AvgRMSE$_{fM}$	6:45 10^{3}	6:51 10^{3}	4:83 10^{3}	6:49 10^{3}
L$_{CI}$(95%)	2:58 10^{2}	2:24 10^{2}	9:45 10^{3}	5:80 10^{3}
P$_{CI}$(95%)	84:2%	84:2%	84:2%	84:2%

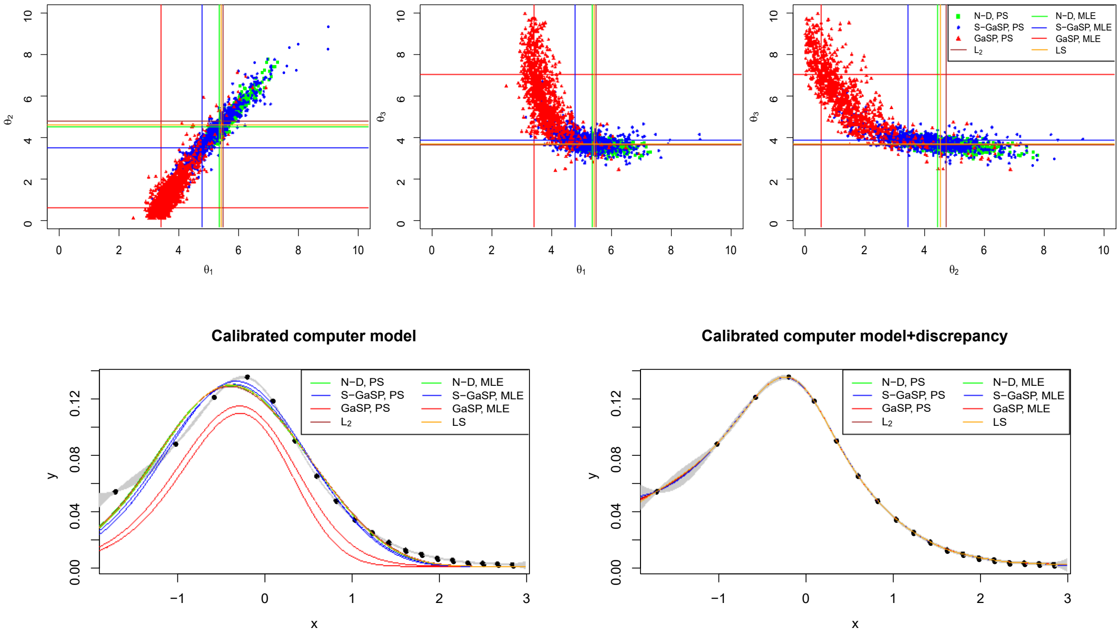

Figure 6. The posterior samples from no-discrepancy, S-GaSP, and GaSP calibrations are given in the upper panels. The crossings of each colored line are MLEs and L$_2$ and LS calibration. The t by the calibrated computer model by dierent approaches is plotted in the lower left panel, and the predictive mean by combining the calibrated computer model and discrepancy is plotted in the lower right panel. The black dots are observa-tions. The 95% predictive intervals by GaSP and S-GaSP with posterior sampling (PS) are plotted in the lower left and lower right panels, respectively.

model by the GaSP calibration (graphed as the dashed red curve in the lower right panel) under-estimates the values of the observations. This identiability issue of the GaSP calibration in this example was also reported in [19]. In contrast, the calibrated mathematical model ts the observations reasonably well in the S-GaSP calibration. Finally, when using the calibrated computer model and discrepancy function, all methods fit the observations well as shown in the lower right panel in Figure 6.

7. Concluding remarks.

We have introduced a theoretical framework of a scaled Gaussian stochastic process (S-GaSP) for calibration and prediction. We showed that under certain routinely used assumptions, the predictive mean of the S-GaSP calibration model converges to the reality as fast as the GaSP calibration with some suitable choices of the regularization parameter and scaling parameter. The estimated calibration parameters in the S-GaSP calibration converge to the L_2 minimizer with the same choices of the regularization parameter and scaling parameter, whereas those in the GaSP calibration typically do not converge to the L_2 minimizer. The results rely on the orthogonal series representation of the processes studied in this work. The GaSP, S-GaSP, and no-discrepancy calibration approaches were implemented in the RobustCalibration R package, where the parameters can be estimated by both posterior sampling and MLE. Furthermore, for computationally expensive computer models, a GaSP emulator for scalar-valued and vectorized output can be called in RobustCalibration for accelerating the computation.

The numerical studies indicate that jointly estimating the calibration parameters and the discrepancy function in GaSP and S-GaSP calibration can improve the predictive accuracy if the reality is too complicated to be predicted precisely by a nonparametric regression model alone. The mathematical model, which typically contains some information about the trend and the shape of the reality function, can be helpful in predictions. We also empirically found that the calibrated mathematical model that minimizes the L_2 distance between the reality and the mathematical model may not always be the best for reducing the predictive error. This is because the residuals may behave like noises, which could be difficult to estimate accurately by nonparametric models. The S-GaSP calibration gives predictions that are at least as accurate as the GaSP calibration, and the calibrated mathematical model by the S-GaSP calibration fits the real observations more closely than the ones by the GaSP calibration in almost all examples.

We outline a few extensions of the S-GaSP model below. First, from the theoretical perspective, we did not study the convergence of the discretized S-GaSP, whereas the numerical studies indicate that the convergence rate from the discretized S-GaSP is the same as from the S-GaSP. Second, we illustrated that the S-GaSP calibration can be implemented in both Frequentist and Bayesian ways. The contraction rate of the S-GaSP under the Bayesian framework, however, was not studied. The studies of the contraction rate of the GaSP regression may be extended to achieve this goal [5], and the adaptive approach with respect to the smoothness level of latent function in [27] may be extended in model calibration. Third, we fix the scaling parameter in the S-GaSP calibration as a function of the sample size, whereas historical information may be used to develop a reasonable prior for this parameter [22]. Furthermore, the S-GaSP calibration framework may be extended to include a model of correlated noises, as the field observations may contain time series and images [2].

Acknowledgment. We thank the editor, associate editor, and three referees for their great suggestions, which substantially improved this manuscript.

REFERENCES

[1] K. R. Anderson, I. A. Johanson, M. R. Patrick, M. Gu, P. Segall, M. P. Poland, E. K. Montgomery-Brown, and A. Miklius, Magma reservoir failure and the onset of caldera collapse at Klauea volcano in 2018, Science, 366 (2019), eaaz182.

[2] K. R. Anderson and M. P. Poland, Abundant carbon in the mantle beneath Hawai'i, Nat. Geosci., 10 (2017), pp. 704{708.

[3] P. D. Arendt, D. W. Apley, and W. Chen, Quantification of model uncertainty: Calibration, model discrepancy, and identifiability, J. Mech. Des., 134 (2012), 100908.

[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] A. Bhattacharya, D. Pati, and D. Dunson, Anisotropic function estimation using multi-bandwidth Gaussian processes, Ann. Stat., 42 (2014), pp. 352{381.

[6] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov, Efficient and Adaptive Estimation for Semiparametric Models, Vol. 4, Johns Hopkins University Press, Baltimore, 1993.

[7] G. Box and G. Coutie, Application of digital computers in the exploration of functional relationships, Proc. IEE-Part B Radio Electron. Eng., 103 (1956), pp. 100{107.

[8] M. Goldstein and J. Rougier, Probabilistic formulations for transferring inferences from mathematical models to physical systems, SIAM J. Sci. Comput., 26 (2004), pp. 467{487, https://doi.org/10.1137/S106482750342670X.

[9] M. Gu, Jointly robust prior for Gaussian stochastic process in emulation, calibration and variable selection, Bayesian Anal., 14 (2019), pp. 857{885.

[10] M. Gu, Robust Calibration: Robust Calibration of Computer Models in R, preprint, https://arxiv.org/abs/2201.01476, 2022.

[11] M. Gu and J. O. Berger, Parallel partial Gaussian process emulation for computer models with massive output, Ann Appl. Stat., 10 (2016), pp. 1317{1347.

[12] M. Gu and L. Wang, Scaled Gaussian stochastic process for computer model calibration and prediction, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1555{1583, https://doi.org/10.1137/17M1159890.

[13] M. S. Handcock and M. L. Stein, A Bayesian analysis of Kriging, Technometrics, 35 (1993), pp. 403{410.

[14] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, Computer model calibration using high-dimensional output, J. Am. Stat. Assoc., 103 (2008), pp. 570{583.

[15] M. C. Kennedy and A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B, 63 (2001), pp. 425{464.

[16] M. R. Kosorok, Introduction to Empirical Processes and Semiparametric Inference, Springer, 2008.

[17] D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Program., 45 (1989), pp. 503{528.

[18] R. Paulo, G. Garca-Donato, and J. Palomo, Calibration of computer models with multivariate output, Comput. Stat. Data Anal., 56 (2012), pp. 3959{3974.

[19] M. Plumlee, Bayesian calibration of inexact computer models, J. Am. Stat. Assoc., 112 (2017), pp. 1274{1285.

[20] M. Plumlee, V. R. Joseph and H. Yang, Calibrating functional parameters in the ion channel models of cardiac cells, J. Am. Stat. Assoc., 111 (2016), pp. 500{509.

[21] J. Sacks, W. J. Welch, T. J. Mitchell, and P. H. Wynn, Design and analysis of computer experiments, Statist. Sci., 4 (1989), pp. 409{435.

[22] J. M. Salter, D. B. Williamson, J. Scinocca, and V. Kharin, Uncertainty quantification for com-puter models with spatial output using calibration-optimal bases, J. Am. Stat. Assoc., (2019), pp. 1{24.

[23] T. J. Santner, B. J. Williams, and W. I. Notz, The Design and Analysis of Computer Experiments, Springer-Verlag, 2003.

[24] K. Soetaert, T. Petzoldt, and R. W. Setzer, Package deSolve: Solving initial value differential equations in R, J. Stat. Softw., 33 (2010), pp. 1{25.

[25] R. Tuo and C. J. Wu, Efficient calibration for imperfect computer models, Ann. Stat., 43 (2015), pp. 2331{2352.

[26] R. Tuo and C. F. J. Wu, A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 767{795, https://doi.org/10.1137/151005841.

[27] A. W. Van der Vaart and J. H. Van Zanten, Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth, Ann. Stat., 37 (2009), pp. 2655{2675.

[28] G. Wahba, Spline Models for Observational Data, CBMS-NSF Reg. Conf. Ser. Appl. Math. 59, SIAM, 1990.

[29] R. K. Wong, C. B. Storlie, and T. Lee, A frequentist approach to computer model calibration, J. R. Stat. Soc. Ser. B , 79 (2017), pp. 635{648.

[30] F. Xie and Y. Xu, Bayesian projected calibration of computer models, J. Am. Stat. Assoc., 116 (2021), pp. 1965{1982.

[31] Y. Yang, A. Bhattacharya, and D. Pati, Frequentist coverage and sup-norm convergence rate in Gaussian process regression, preprint, https://arxiv.org/abs/1708.04753, 2017.

[32] H. Zhang, Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, J. Am. Stat. Assoc., 99 (2004), pp. 250{261.

Supplementary Materials for A theoretical framework of the scaled Gaussian stochastic process

Mengyang Gu, Fangzheng Xie and Long Wang

All the formulas in this supplementary materials are cross-referenced in the main body of the article. We first give a brief introduction of the Gaussian stochastic process model and reproducing kernel Hilbert space in Section S2. The proof for Section 2 is given in Section S3. The proof for Theorem 3.1 and two auxiliary lemmas are provided in Section S4. Section S5 encloses the proof for Theorem 3.3 and provides additional results regarding the convergence of S-GaSP calibration when kernel parameters are estimated.

S1. Background: Gaussian stochastic process. All the formulas in this supplementary materials are cross-referenced in the main body of the article. We first give a brief introduction of the Gaussian stochastic process model and reproducing kernel Hilbert space in Section S2. The proof for Section 2 is given in Section S3. The proof for Theorem 3.1 and two auxiliary lemmas are provided in Section S4. Section S5 encloses the proof for Theorem 3.3 and provides additional results regarding the convergence of S-GaSP calibration when kernel parameters are estimated.

S2. Background: Gaussian stochastic process. Assume the mean and trend of the reality are properly modeled in the mathematical model. Consider to model the unknown discrepancy function in the calibration model (1) via a real-valued zero-mean Gaussian stochastic process () on a p-dimensional input domain X ,

$$\text{(S1)} \qquad () \;\; \text{GaSP}(0; {}^2K(;));$$

where ^{2} is a variance parameter and $K(x_a; x_b)$ is the correlation for any $x_a; x_b$ 2 X , parameterized by a kernel function. For simplicity, we assume X = $[0; 1]^p$ in this work.

For any $fx_1; :::; x_ng$, the outputs $((x_1); :::; (x_n))^T$ follow a multivariate normal distribution

$$\text{(S2)} \qquad [(x_1); :::; (x_n) \; j \; {}^2; R] \;\; MN(0; {}^2R);$$

where the $(i; j)$ entry of R is $K(x_i; x_j)$. Some frequently used kernel functions include the power exponential kernel and the Matern kernel. We defer the issue of estimating the parameters in the kernel function in Section 4 and assume $K(;)$ is known for now.

The reproducing kernel Hilbert space (RKHS), denoted as H, attached to the Gaussian stochastic process GaSP$(0; {}^2K(;))$, is the completion of the space of all functions

$$x \; ! \; \mathop{X}^{k}_{i=1} w_i K(x_i; x); \quad w_1; :::; w_k \; 2 \; R; \; x_1; :::; x_k; x \; 2 \; X; \; k \; 2 \; N;$$

with the inner product

$${}^{*}\mathop{X}^{k}_{i=1} w_i K(x_i;); \; \mathop{X}^{m}_{j=1} w_j K(x_j;) {}^{+}_{H} = \mathop{X}^{k}_{i=1}\mathop{X}^{m}_{j=1} w_i w_j K(x_i; x_j):$$

For any function $f() \; 2 \; H$, denote $kfk_H = {}^{p}\overline{hf; fi_H}$ the RKHS norm or the native norm. Because the evaluation maps in RKHS are bounded linear, it follows from the Riesz representation theorem that for each $x \; 2 \; X$ and $f() \; 2 \; H$, one has $f(x) = hf(); K(; x)i_H$.

Denote $L_2(X)$ the space of square-integrable functions $f : X \; ! \; R$ with ${}^{R}{}_{x 2 X} f^2(x)dx < 1$. We denote $hf; gi_{L_2(X)} := {}^{R}{}_{x 2 X} f(x)g(x)dx$ the usual inner product in $L_2(X)$. By Mercer's theorem, there exists an orthonormal sequence of continuous eigenfunctions $f_k g^{1}_{k=1}$ with a sequence of non-increasing and non-negative eigenvalues $f_k g^{1}_{k=1}$ such that

$$K(x_a; x_b) = \mathop{X}^{1}_{k=1} {}_{k}{}_{k}(x_a){}_{k}(x_b); \tag{S3}$$

for any $x_a; x_b \; 2 \; X$.

The RKHS H contains all functions $f() = {}^{P}{}^{1}_{k=1} f_k {}_k() \; 2 \; L_2(X)$ with $f_k = hf; {}_k i_{L_2(X)}$ and ${}^{P}{}^{1}_{k=1} f_k^2 = {}_k < 1$. For any $g() = {}^{P}{}^{1}_{k=1} g_k {}_k() \; 2 \; H$ and $f()$, the inner product can be represented as $hf; gi_H = {}^{P}{}^{1}_{k=1} f_k g_k = {}_k$. For more properties of the RKHS, we refer to Chapter 1 of [11] and Chapter 11 of [2].

S2.1. The equivalence between the maximum likelihood estimator and the kernel ridge regression estimator in calibration. Assume one has a set of observations $y^F := y^F(x_1); :::; y^F(x_n) {}^{T}$ and mathematical model outputs $f^M := (f^M(x_1;); :::; f^M(x_n;))^T$, where $= ({}_1; :::; {}_q)^T \; 2 \; R^q$ is a q-dimensional vector of the calibration parameters. Denote the regularization parameter $:= {}^2=(n^2)$. For the$_0$ calibration model (1) with modeled as a GaSP in (S1), the marginal distribution of y^F follows a multivariate normal 2 after marginalizing out ,

(S4) $[y^F \ j \ ; {}^2;]_0 \ MN(f^M; {}^2((n) \ {}^1_0R + I_n)):$

Let L() be the likelihood for in (S4) given the other parameters in the model. For any given , the maximum likelihood estimator (MLE) of is denoted as

(S5) $\hat{}_{;n} := \underset{2}{\text{argmax}} \, L():$

Conditioning on the observations, $\hat{}_{;n}$ and , the predictive mean of the discrepancy function at any x 2 X has the following expression

(S6) $\hat{}_{;n}(x) := E[(x) \ j \ y^F ; {}_{;n}; \hat{]} = r^T(x)(R + nI_n) \ {}^1y^F \quad f^M_{\hat{}\,;n}$

with $r(x) = (K(x_1; x); :::; K(x_n; x))^T$ and I_n being the n-dimensional identity matrix.

It is well-known that the predictive mean in (S6) can be written as the estimator for the kernel ridge regression (KRR). In the following lemma, we show that $(\hat{}_{;n}; {}_{;n}\hat{()})$ is equivalent to the KRR estimator.

Lemma S1. The maximum likelihood estimator$\hat{}_{;n}$ dened in (S5) and predictive mean estimator $\hat{}_{;n}()$ dened in (S6) can be expressed as the estimator of the kernel ridge regression as follows

$$(\hat{}_{;n}; {}_{;n}\hat{()}) = \text{argmin} \ `_{;n}(;); {}_{2;()2H}$$

where

(S7) $`_{;n}(;) = \frac{1}{n}\overset{n}{X}(y^F(x_i) \quad f^M(x_i;) \quad (x_i))^2 + kk^2 : {}_{i=1}{}_H$

Proof of Lemma S1. By the representer lemma [6, 11], for any 2 and x 2 X, one has

(S8) $\hat{}_{;n;}(x) = \overset{n}{X} w_i()K(x_i; x): {}_{i=1}$

Denote $w = (w_1(); :::; w_n())^T$. Since $hK(x_i;); K(x_j;)i_H = K(x_i; x_j)$, (S7) becomes to find and w that minimize

(S9) $$\frac{1}{n}(y^F \quad f^M \quad Rw)^T(y^F \quad f^M \quad Rw) + w^T Rw:$$

For any , solving the minimization for (S9) with regard to w gives

(S10) $$\hat{w} = (R + nI_n)\ ^1(y^F \quad f^M):$$

Then plugging $\hat{w}$ into (S9), based on the Woodbury matrix identity, one has

$$\begin{aligned}
&\frac{1}{n}(y^F \quad f^M \quad R\hat{w})^T(y^F \quad f^M \quad R\hat{w}) + \hat{w}^T R\hat{w} \\
&= \frac{1}{n}(y^F \quad f^M)^T[(I_n \quad R(R + nI_n)\ ^1)^T(I_n \quad R(R + nI_n)\ ^1)] \\
&\quad + (y^F \quad f^M)^T(R + nI_n)\ ^1R(R + nI_n)\ ^1(y^F \quad f^M) \\
&= (y^F \quad f^M)^T(R + nI_n)\ ^1(y^F \quad f^M); && \text{(S11)}
\end{aligned}$$

which shows that the minimizer of on right-hand side of (S11) is the same as the MLE of in (S5). Finally, plugging the estimator $\hat{}_{;n}$ into (S9), the result follows from the KRR estimator of () in (S8) with the weights in (S10). ■

Although modeling the discrepancy function by the GaSP typically improves the prediction accuracy of the reality, the penalty term of (S7) only contains kk_H to control the complexity of the discrepancy. As the RKHS norm is not equivalent to the L_2 norm, the calibrated computer model could deviate a lot from the best performed mathematical model in terms of the L_2 loss [8]. In Section 2, we introduce the scaled Gaussian stochastic process that predicts the reality as accurately as the GaSP with the aid of the mathematical model, but has more prior mass on the small L_2 distance between the reality and mathematical model. As a consequence, the KRR estimator of the new model penalizes both kk_H and $kk_{L_2(X)}$ simultaneously.

S3. Proof for Section 2.

Proof of Lemma 2.1. By Karhunen-Loeve expansion, we have

$$(x) = \sum_{i=1}^{1} p\overline{\ _i} Z_i\ _i(x)$$

with $Z_i \overset{i:i:d}{} N(0;1)$. Denote $W_k = \sum_{i=k+1}^{1}\ _i Z_i^2$ for any $k \ 2\ N^+$. From the definition of 4

$Z = \overset{R}{}_{x2X}\ {}^{2}(x)dx$ and $\overset{R}{}_{x2X}\ {}^{2}(x)_{i}dx = 1$ for any $i\ 2\ N^{+}$, it is straightforward to see that

$$\text{(S12)} \qquad Z = {}^{2}({}_{1}Z^{2}\ +_{1} + {}_{k}Z_{k} + W_{k}^{2}):$$

In the following expressions, we are conditioning on all parameters and they are dropped for simplicity. From the construction of

$$_{z}(x) = \quad (x)\ j \quad \int_{x2X} {}^{2}(x)dx = Z$$

and $Z \quad p_{Z}()$, the joint density of $(Z_{1}; :::; Z_{k}; W_{k})$ in the S-GaSP can be expressed as

$$\begin{aligned}
&p_{z}(Z_{1} = z_{1}; :::; Z_{k} = z_{k}; W_{k} = w_{k})\\
&= \int^{1} p\,(Z_{1} = z_{1}; :::; Z_{k} = z_{k}; W_{k} = w_{k}\ j\ Z = z)\,p_{Z}(Z = z)dz\\
&/ \int_{0}^{1} \frac{p\ (Z = z; Z_{1} = z_{1}; :::; Z_{k} = z_{k}; W_{k} = w_{k})}{p\ (Z = z\ )}\,g_{Z}(Z = z)p(Z = z)dz\\
&/ \ p(Z_{1} = z_{1}; :::; Z_{k} = z_{k})p(W_{k} = w_{k})\\
&\qquad \int_{0}^{1} p(Z = z\ j\ Z_{1} = z_{1}; :::; Z_{k} = z_{k}; W_{k} = w_{k})\exp \quad \frac{z_{z}}{2\ \ 2} \quad dz\\
&/ \ p(Z_{1} = z_{1}; :::; Z_{k} = z_{k})p(W_{k} = w_{k})\\
&\qquad \int_{0}^{1} 1\,z = {}^{2}({}_{1}z^{2} + {}_{1}^{+}\ {}_{k}z^{2} + w_{k})\,{}_{k}\exp \quad z^{z} \quad \frac{}{2\ \ 2} \quad dz\\
&/ \ \exp \ \ ! \ \frac{1}{2}\sum_{i=1}^{k} z_{i}^{2} \quad p(W_{k} = w_{k})\exp \ \ " \ \frac{z}{2}\sum_{i=1}^{k} {}_{i}z^{2}{}_{i} + w_{k} \ \ !\,\#\\
&= \ \prod_{i=1}^{k} \exp \ \ (\ \frac{1}{2}(1 + {}_{z\,i})z_{i} \quad {}^{2} \ \ ) \quad p(W_{k} = w_{k})\exp(\ \ {}_{z}w_{k}=2);
\end{aligned}$$

where 1 in the fourth step is a Dirac delta function.

After integrating out W_{k}, it is clear that Z_{i}'s are independently distributed as $N(0; 1=(1 + {}_{z\,i}))$ under the measure induced by the S-GaSP. Since k is arbitrary, we have

$$_{z}(x) = \sum_{i=1}^{1} \sqrt{\frac{i}{1 + {}_{z\,i}}}\,Z_{i\ i}(x)$$

with $Z_{i} \overset{i:i:d:}{} N(0; 1)$, from which the proof is complete. ■

Proof of Lemma 2.4. First note that for any $x_{a}; x_{b}\ 2\ X$, we have $K(x_{a}; x_{b}) = \overset{P}{}\,{}_{i=1}^{1}\ {}_{i\,i}(x_{a})_{i}(x_{b})$ 5 and $K_z(x_a; x_b) = \sum_{i=1}^{1} {}_{z;i}{}_i(x_a)_i(x_b)$ with ${}_{z;i} = {}_i=(1+{}_z{}_i)$. For $h() = \sum_{i=1} h_i{}_i() \; 2^1 H$ and $g() = \sum_{i=1}^{1} g_i{}_i() \; 2 \; H$, one has

$$hh; gi_{H_z} = \sum_{i=1}^{1} \frac{1}{{}_{z;i}} h_i g_i = \sum_{i=1}^{1} \frac{1}{{}_i} h_i g_i + {}_z \sum_{i=1}^{1} h_i g_i = hh; gi_H + {}_z hh; gi_{L_2}:$$ ■

Proof of Lemma 2.5. We show below that for any 2 and any $x \; 2 \; X$, one has

$$\hat{}_{;z;n}(x) = \sum_{i=1}^{n} w_{z;i}() K_z(x_i; x): \tag{S13}$$

For any $() \; 2 \; H$, decomposing it into the linear combination of the basis $f_z K_z(x_i;)g_{i=1}^n$ and the orthogonal complement $v()$ gives

$$() = \sum_{i=1}^{n} \tilde{w}_{z;i}()_z K_z(x_i;) + v();$$

where $hv(); {}_z K_z(; x_i)i_{H_z} = 0$ for $i = 1; :::; n$.

To evaluate $()$ at x_j for any $j = 1; :::; n$, we have

$$\begin{aligned} (x_j) &= \left\langle \sum_{i=1}^{n} \tilde{w}_{z;i}()_z K_z(x_i;) + v(); K_z(x_j;) \right\rangle_{H_z} \\ &= \sum_{i=1}^{n} \tilde{w}_{z;i}()_z K_z(x_i; x_j); \end{aligned}$$

which is independent from $v()$. Hence the first term on right-hand side of (11) is also independent from $v()$. For the second term on right-hand side of (11), since $v()$ is orthogonal to $fK_z(x_i;)g_{i=1}^n$, plugging in the decomposition of $()$, we have

$$\begin{aligned} kk_{H_z}^2 &= (k \sum_{i=1}^{n} \tilde{w}_{z;i}()_z K_z(x_i;)k_{H_z} + kvk_{H_z})^2 \\ & k \sum_{i=1}^{n} \tilde{w}_{z;i}()_z K_z(x_i;)k_{H_z}^2 : \end{aligned}$$

Thus choosing $v() = 0$ does not change the first term on right-hand side of (11), but also minimizes the second term on right-hand side of (11). Letting $w_{z;i}() = \tilde{w}_{z;i}()_z$, we have proved (S13). The rest of the proof can be derived similarly as the proof for Lemma S1, so it is omitted here. ■

S4. Proof for Section 3.1. We prove Theorem 3.1 in this Section. Two auxiliary lemmas used for the proof of Theorem 3.1 are given after the proof.

Proof for Theorem 3.1. Define a new inner product on H as

$$\text{(S14)} \qquad hf; gi = (1 + {}^{p})hf; gi_{L_2(X)} + hf; gi_H$$

Let $f = {}^{P}{}^{1}_{k=1} f_{k\,k}$ and $g = {}^{P}{}^{1}_{k=1} g_{k\,k}$ be elements in H. Then

$$hf; gi = (1 + {}^{p}) \overset{1}{\underset{k=1}{X}} f_k g_k + \overset{1}{\underset{k=1}{X}} \frac{{}_k f_k g}{{}_k} \overset{1}{\underset{k=1}{X}} 1 + {}^{p} + \frac{f_k g_k}{{}_k} :$$

By letting ${}_k$ by ${}_k{}^{1} = 1 + {}^{p} + ={}_k$, we can define a new reproducing kernel

$$\text{(S15)} \qquad K(x; x^0) = \overset{1}{X} {}_{k\,k}(x)_k(x^0)_{k=1}$$

Since $c^{1}k^{2m=p} \;{}_k\; C\;{}^{1}k^{2m=p}$ and $j_i()j < C$ for some positive constants c, C and C, bounding the sums by integrals, we have

$$\begin{aligned} \sup_{x;x^C} K(x; x^0) \; C^2 \overset{1}{\underset{k=1}{X}} \frac{1}{1 + ck^{2m=p}} \; C \; {}^{2}\overset{1}{\underset{k=1}{X}} \underset{1}{\overset{k}{Z}} \frac{1}{1 + cx^{2m=p}} dx \; \frac{1}{2 \quad p=2m} \\ = C \; {}^{p=2m}c \underset{1}{Z} \frac{(c)^{p=2m}}{1 + f(c)^{p=2m}xg^{2m=p} =} dx \\ C^2 c \;{}^{p=2m} \;{}^{p=2m} \underset{0}{\overset{0}{Z}}{}_{1} \frac{1}{1 + x^{2m=p}} dx: \end{aligned}$$

Thus

$$\text{(S16)} \qquad \sup K \; (x; x^0) \; C^2 \;{}_K{}^{p=(2m)}; {}_{x;x^0}$$

for some constant C_K depending on K. Define the following linear operators F : H ! H and P : H ! H via

$$(Fg)(x) = \underset{X}{Z} g(x^0)K(x; x^0)dx^0; \qquad \text{and} \quad (Pg)(x) = g(x) \quad (Fg)(x);$$

Clearly, we have

$$\langle f, Fg\rangle = \sum_{k=1}^{1} \langle f, {}_k\rangle_{L_2(X)}\langle g, {}_k\rangle_{L_2(X)} = \langle f, g\rangle_{L_2(X)};$$

$$\langle f, Pg\rangle = \langle f, g\rangle \qquad \langle f, Fg\rangle = p\,\langle f, g\rangle_{L_2(X)} + \langle f, g\rangle_H : \tag{S17}$$

Denote the loss function

$$\ell_n(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i \quad f(x_i))^2 + \underline{p}\,\|f\|^2_{L_2(X)} + \|f\|^2_H ;$$

and the estimator $\hat{f}_n := \mathrm{argmin}_{f\in H}\,\ell_n(f)$. Let $D\ell_n(f) : H \to H$ be the Frechet derivative of ℓ_n evaluated at f. Clearly, for any $g \in H$,

$$\begin{aligned} D\ell_n(f)g &= \frac{2}{n}\sum_{i=1}^{n}(f(x_i) \quad y_i)\langle K(x_i,), g()\rangle + 2\langle Pf, g\rangle \\ &= \left\langle \frac{2}{n}\sum_{i=1}^{n}(f(x_i) \quad y_i)K(x_i,) + 2(Pf)(), g() \right\rangle_* : \end{aligned} \tag{S18}$$

It follows that $D\ell_n(\hat{f}_n)g = 0$ for all $g \in H$, and hence, $S_n(\hat{f}_n)() = 0$, where

$$S_n(f)() = \frac{1}{n}\sum_{i=1}^{n}(y_i \quad f(x_i))K(x_i,) \qquad (Pf)():$$

Define $S(f)() = E_{y;x}(S_n(f)())$. Then

$$\begin{aligned} S(f)() &= \int_{x\in X}(f_0(x) \quad f(x))K(x,)dx \qquad (Pf)() \\ &= (F(f_0 \quad f))() \quad (Pf)() = (Ff_0)() \qquad f(); \end{aligned}$$

and therefore, $S(Ff_0)() = 0$. Let $f = \hat{f}_n \quad Ff_0$. By the definitions of S_n and S, we have

n o
S n (f n^) S (f n^) ( ) f S n (F f 0) S (F f 0) g ( )
n o n o
= S n (f n^) S n (F f 0) ( ) S (f n) ^ S (F f 0) ( )
= 1/n X n n o n o
i = 1 F f 0 (x i) f n^(x i) K (x i ;) + P (F f 0) ( ) f n ( ) ^
+ f^n ( ) (F f 0) ( )
= 1/n X n
i = 1 f (x i) K (x i ;) (P f) ( ) + (f) ( )
= 1/n X n f (x i) K (x i ;) + E x f f (x) K (x ;) g : i = 1

On the other hand, S n (f n^) ( ) = S (F f 0) ( ) = 0 and S (f n) ( ) = ^(F f 0) ( ) f n ( ) = ^ f ( ). Therefore,

n o
S n (f n^) S (f n^) ( ) f S n (F f 0) S (F f 0) g ( ) = f ( ) S n (F f 0) ( ):

Dene the event

( ) A n (t) =
n X
g (x i) K (x i ;) 1/n E x f g (x) K (x ;) g < t k g k for all g 2 H :
i = 1

Applying Lemma S2 on g()=kgk,

P x f A n (t) g 1 2 exp p(6m p)=(4m 2) n t 2
K

for some constant K > 0. The deviation threshold t will be specied later, and from now we consider data points (x i ; y i) n i = 1 over the event A n (t).

Over the event A n (t), we have

n o
S n (f n^) S (f n^) ( ) f S n (F f 0) S (F f 0) g ( )
= 1/n X n f (x i) K (x i ;) + E x f f (x) K (x ;) g i = 1
= f ( ) S n (F f 0) ( );

implying that

$$kf \quad S_n(Ff_0)k = \quad \frac{1}{n}\sum_{i=1}^{n} f(x_i)K(x_i;) \quad E_x\, ff(x)K(x;)g$$

$$tkfk:$$

Now we proceed to bound $kS_n(Ff_0)k$. Write

$$kS_n(Ff_0)k \qquad = \frac{1}{n}\sum_{i=1}^{n} \quad fy_i \quad Ff_0(x_i)gK(x_i;) \quad (PFf_0)()$$

$$\frac{1}{n}\sum_{i=1}^{n} ff_0(x_i) \quad Ff_0(x_i)gK(x_i;) \quad fF(f_0 \quad Ff_0)g()$$

$$+ \frac{1}{n}\sum_{i=1}^{n} {}_iK(x_i;)$$

$$= \frac{1}{n}\sum_{i=1}^{n} ff_0(x_i) \quad Ff_0(x_i)gK(x_i;) \quad E_x[ff_0(x) \quad Ff_0(x)gK(x;)]$$

$$+ \frac{1}{n}\sum_{i=1}^{n} {}_iK(x_i;)$$

$$tkf_0 \quad Ff_0k + \frac{1}{n}\sum_{i=1}^{n} {}_iK(x_i;) \quad ;$$

where the last inequality is due to the construction of the event $A_n(t)$. To bound the second term of the preceding display, we let $= [K(x_i; x_j)]_{n\,n}$ and $= [_1; : : : ; _n]^T$. By the Hanson-Wright inequality [7], for all $x > 0$, we have

$$P_x \; {}^T \; {}^2 \quad {}_0 \; tr() + 2\sqrt{tr()^2}x + 2kk_F x^2 \qquad e^{\;x^2}:$$

Since by the Cauchy-Schwarz inequality,

$$tr() = \sum_{i=1}^{n} kK(x_i;)k^2 = \sum_{i=1}^{n} K(x_i;x_i) \quad C^2 n^{p=(2m)}_{K};$$

$$tr(^2) \quad \sum_{i=1}^{n}\sum_{j=1}^{n} kK(x_i;)k_{L_2(X)}kK(x_j;)k_{L_2(X)} \quad C^4 n^2 {}^{p=m}_{K};$$

$$kk_F = \sqrt{tr()^2} \; C^2 n^{p=(2m)}_{K};$$

it follows that

$$\text{(S19)} \qquad tr() + 2\sqrt{tr()x^2} + 2kk_Fx^2 \quad C^2 n^{p=(2m)}_{K}(1 + 2x + 2x^2):$$

Set the event B_n to be

$$B_n = \left\{ \frac{1}{n}\sum_{i=1}^{n} e_iK(x_i;) < {}_0C_K n^{1=2} {}^{p=(4m)1=2} \right\};$$

where $= 2 + 3x^2$. Since $1 + 2x + 2x^2 \; 2 + 3x^2 =$, by taking $x = \sqrt{(- 2)=3}$, we have $P(B_n) \; 1 \; exp(\; (\; 2)=3)$ for any > 2. Putting all pieces obtained above together, we have

$$\begin{aligned} kfk \;\; kf \quad & S_n(Ff_0)k + kS_n(Ff_0)k \\ & tkfk + tkf_0 \quad Ff_0k + {}_0C_K n^{1=2} {}^{p=(4m)1=2} \quad \text{(S20)} \\ & = tkfk + tkPf_0k + {}_0C_K n^{1=2} {}^{p=(4m)1=2}; \end{aligned}$$

over the event $A_n(t) \setminus B_n$. Now take $= n^{2m=(2m+p)}$. Choose any $C \; 2 \; (0;1)$ and let $t = \sqrt{{}_K=(n^{(1 \; C)}\log(2))}$. Then, for suciently large n,

$$P_x fA_n(t)g \;\; 1 \quad 2\exp \; \frac{nt^2}{1} \quad \exp \left\{ n^C \right\};$$

$${}_K$$

where $= (2m \quad p)^2=(2m(2m + p))$, and therefore,

$$kfk \;\; kPf_0k + 2{}_0C_K n^{m=(2m+p)1=2};$$

with probability at least

$$
\begin{aligned}
PfA_n(t) \backslash B_n)g &= 1 \quad PfA_n^c(t) [\ B_n^c g \\
& 1 \quad PfA_n^c(t)g \quad P(B_n^c) = 1 \quad \exp f \quad (\quad 2)=3g \quad \exp f \quad n^C g
\end{aligned}
$$

for suciently large n. Observe that

$$
\begin{aligned}
kPf_0k^2 &= \sum_{k=1}^{1} (1 \quad {}_k)hf_0;{}_k i_{L_2(X)}{}_k() \quad = \quad {}^2 \sum_{k=1}^{1} \frac{(1 \quad {}_k)^2}{k} hf_0;{}_k i^2_{L_2(X)} \\
&= \sum_{k=1}^{1} \frac{(\overset{p}{} \quad + \ = \ )_k^2}{1 + \ p\overline{} + \ =_k} hf_0;{}_k i_{L_2(X)} \ {}^2 \quad \sum_{k=1} 2 + \ 2(1 = {}_k)^2 \ \frac{hf_0;{}_k i_{L_2(X)} \ 1}{1 + \ = \quad k} \quad 1 \quad 2 \\
& 2 \quad \sum_{k=1} hf_0;{}_k i^2_{L_2(X)} + \ 2 \quad \sum_{k=1} \frac{1}{k} hf_0;{}_k i^2_{L_2(X)} \\
&= 2kf_0k^2_{L_2(X)} + 2kf_0k_H \ {}^2 \\
& 2n^{\ 2m=(2m+p)} \ kf_0k_{L_2(X)} + kf_0k_H \ {}^2 :
\end{aligned}
$$

Hence, we proceed to compute

$$
\begin{aligned}
k\hat{f}_n \quad f_0k_{L_2(X)} \ & kf_n\hat{} \quad f_0k \\
& kf_n\hat{} \quad Ff_0k + kFf_0 \quad f_0k \\
&= kfk + kPf_0k \\
& 2 \quad \overset{p}{2}(kf_0k_{L_2(X)} + kf_0k_H) + 2_0C_K{}^{1=2} \quad n \quad {}^{m=(2m+p)}
\end{aligned}
$$

with probability at least 1 $\exp f \quad (\quad 2)=3g \quad \exp \quad n^C$ for suciently large n. The bound for $kf_{;z;n\wedge} \qquad f_0k_H$ follows immediately by the denition of jj jj, completing the proof. ■

The following the Lemma S1 is Theorem 3.6 in [5], which is needed for the proof of Lemma S2.

Lemma S1. Let $(X_j)_{j=0}^{1}$ be a sequence of random elements in a Hilbert space H with norm $k \ k_H$. Suppose that $(X_j)_{j=0}^{1}$ forms a martingale in the sense that $E(X_j \ j \ X_0; : : : ; X_j \ {}_1) = X_j$ a.s., and that the dierence sequence $(D_j)_{j=1}^{1} = (X_j \quad X_j \ {}_1)_{j=1}^{1}$ satises $kD_jk^2_H \ b_j$ a.s.^{2} and $\sum_{j=1}^{1} b_j^2 \ b^2$. Then for any t 0,

$$
P \ \sup_{j1} kX_jk_H \quad t \quad 2\exp \quad \frac{t^2}{2b^2} \ ! \ :
$$

The following maximum inequality for functional empirical processes in the Sobolev space $W_2^m(X;1)$, which generalizes Lemma 5.1 in [12] to multivariate functions, is of fundamental importance to the proof of Theorem 3.1.

Lemma S2. Denote $W_2^m(X;1) = ff \; 2 \; W_2^m(X) : kfk \quad 1g$. Suppose $x_1;:::;x_n$ are independently and uniformly drawn from X. Then there exists some constant $_K$ depending on the kernel K, such that for any $t > 0$,

$$P_x \quad \sup_{(X;1)^m} \quad \frac{1}{n}\overset{n}{\underset{i=1}{X}}[g(x_i)K(x_i;)] \qquad E_x\, fg(x)K(x;)g] \quad t \quad {}_{g2W_2} \; !$$
$$2\exp \left(\frac{{}^{d(6m \;\; d)=(4m^2)}nt^2}{K} \right) :$$

Proof of Lemma S2. We follow the argument used in the proof of Lemma 6.1 in [13]. Denote

$$fZ_n(g)g() = \frac{1}{n}\overset{n}{\underset{i=1}{X}}[g(x_i)K(x_i;)] \qquad E_x fg(x)K(x;)g]:$$

Fix g, $h \; 2 \; H$, n, and , consider the following sequence of martingale $(X_j)_{j=0}^{1}$ in H:

$$X_j = \begin{cases} 0; & \text{if } j = 0; \\ jfZ_j(g) \qquad Z_j(h)g; & \text{if } j = 1;:::;n \\ X_n; & \text{if } j \quad n+1: \end{cases}$$

Clearly, for $j = 1;:::;n$,

$$(X_j \quad X_{j \;\; 1})() = fg(x_j) \quad h(x_j)gK(x_j;) \qquad E_x[fg(x_j) \quad h(x_j)gK(x_j;)]$$

and $X_j \quad X_{j \;\; 1} = 0$ for $j \quad n+1$. Observe that

$$kK(x_j;)k = \sqrt{hK(x_j;);K(x_j;)i} = \sqrt{K(x_j;x_j)} \;\; C_K \;\; {}^{d=(4m)}$$

with probability one. Therefore, with probability one, we have

$$kX_j \quad X_{j \;\; 1}k^2 \;\; 4C^2{}_K{}^{d=(2m)}kg \qquad hk^2_{L_1}$$

for $j = 1;:::;n$, and hence, we invoke the bounded dierence inequality for martingales in

Banach space (Lemma S1) to derive

$$P(kZ_n(g) \quad Z_n(h)k \quad t) = P(knfZ_n(g) \qquad Z_n(h)gk \quad nt)$$

$$! \qquad P \quad \sup_{j1} kjfZ_j(g) \quad Z_j(h)gk \quad nt$$

$$2\exp \left(\frac{nt^2}{8C_K \quad 2^{d=(2m)}kg \quad hk_{L_1} \quad 2} \right) :$$

Applying Lemma 8.1 in [3], we obtain the following bound

$$\text{(S21)} \qquad kkZ_n(g) \quad Z_n(h)kk_{\ 2} \quad \frac{24C_K \ 2^{d=(4m)}}{p \overline{\quad} p \overline{n}} kg \qquad hk_{L_1};$$

where k k $_2$ is the Orlicz norm associated with $_2(s) = \exp(s^2)$ 1.

Now let = flog(3=2)g$^{1=2}$ and set (x) = $_2$(x). Clearly, (1) = 1=2, and (x)(y) (xy) for any x; y 1. Applying Lemma 8.2 in [3], the Orlicz norm of the maximum of nitely many random variables can be bounded by the maximum of these Orlicz norms as follows:

$$\max_{1ik} (i) \qquad = \max_{1ik} i \quad 2 \ ^1(k) \max_{1ik} k_i k = \qquad \frac{2}{\ } \ _2 \ ^1(k) \max_{1ik} k_i k \ _2; \ ^2$$

namely,

$$\text{(S22)} \qquad \max_{1ik} \ _2 \ ^1(k) \ \frac{2}{\ } \max_{1ik} k_i k \ _2;$$

where f$_i$g$_{i=1}^k$ are nitely many random variables.

Next we apply the \chaining" argument. Let " > 0 be some constant to be determined later. Construct a sequence of function classes $(G_j)_{j=0}^1$ in H(1) satisfying the following conditions:

(i) For any G_j and any $h_j; g_j$ 2 G_j, $kh_j \quad g_jk_{L_1}$ "=2^j, and G_j is maximal in the sense that for any g_j 2 G_j, there exists some h_j 2 G_j such that $kh_j \quad g_jk$ < "=2^j.

(ii) For any G_{j+1}, and any g_{j+1} 2 G_{j+1}, select a unique element g_j 2 G_j such that $kg_{j+1} \quad g_jk_{L_1}$ "=2^j. Thus, there exists a nite sequence $(g_0; g_1; : : : ; g_{j+1})$ such that $kg_i \qquad g_{i+1}k_{L_1}$ "=2^i for i = 0; : : : ; j, and g_i 2 G_i.

Therefore, for any $g_{j+1}; h_{j+1}$ 2 G_{j+1} with $kg_{j+1} \quad h_{j+1}k_{L_1}$ ", there exists two sequences

$(g_i)_{i=0}^{j+1}$, $(h_i)_{i=0}^{j+1}$, such that $g_i; h_i$ 2 G_i, maxfkg_i g_{i+1}k$_{L_1}$; kh_i h_{i+1}k$_{L_1}$g "=2^i, and that

$$\mathrm{k}g_0 \quad h_0\mathrm{k}_{L_1} \quad \overset{j}{\underset{i=0}{X}}(\mathrm{k}g_i \quad g_{i+1}\mathrm{k}_{L_1} + \mathrm{k}h_i \quad h_{i+1}\mathrm{k}_{L_1}) + \mathrm{k}h_{j+1} \quad g_{j+1}\mathrm{k}_{L_1}$$

$$2 \overset{j}{X} \frac{"}{2^i} + " \quad 5"; \; i=0$$

and hence, by (S21) one has

$$\text{(S23)} \qquad \mathrm{kk}Z_n(g_0) \quad Z_n(h_0)\mathrm{kk}_2 \quad \frac{5\sqrt{24}C_K^2}{\sqrt{n}} \; p=(4m)": $$

We also notice that G_j $H(1)$ ff 2 H : kfk$_H$ $^{1=2}$g, and therefore, the cardinality of G_j can be bounded by the metric entropy of ff 2 H : kfk$_H$ $^{1=2}$g, which is known in the literature [1]:

$$\log jG_jj \;\; \log N_{[]}\; "=2^j; \mathrm{ff}\; 2\; H : \mathrm{k}f\mathrm{k}_H \quad {}^{1=2}g; \mathrm{k}\;\mathrm{k}_{L_1} \quad c_0 \; \frac{2^j}{"}^{p=m} ; \; p=(2m)$$

where c_0 is some absolute constant.

Now suppose g, h are arbitrary functions in $H(1)$ such that kg hk$_{L_1}$ "=2. For any j 2, there exists $g_j; h_j$ 2 G_j such that

$$\mathrm{maxfk}g_j \quad g\mathrm{k}_{L_1}; \mathrm{k}h_j \quad h\mathrm{k}_{L_1}g \;\; "=2^j;$$

and hence, $\|g_j - h_j\|_{L_1} \le \epsilon$. Therefore, for any $j \ge 2$,

$$\begin{aligned}
&\sup_{g,h\in W_2^m(X,1),\|g-h\|_{L_1}\le\epsilon} \|Z_n(g) - Z_n(h)\|_{\psi_2} \\
&\le \sup_{g,h\in W_2^m(X,1),\|g-h\|_{L_1}\le\epsilon} \big\{\|Z_n(g) - Z_n(g_j)\|_{\psi_2} + \|Z_n(g_j) - Z_n(h_j)\|_{\psi_2} \\
&\quad + \|Z_n(h_j) - Z_n(h)\|_{\psi_2}\big\} \\
&\le \frac{2\sqrt{24}C_K^2}{\sqrt{n}} \max\{\|g - g_j\|_{L_1}, \|h - h_j\|_{L_1}\}^{d/(4m)} \\
&\quad + \sup_{g,h\in W_2^m(X,1),\|g-h\|_{L_1}\le\epsilon} \|Z_n(g_j) - Z_n(h_j)\|_{\psi_2} \\
&\le \frac{2\sqrt{24}C_K^2}{\sqrt{n}} \left(\frac{2^{m+1}\epsilon}{2^j}\right)^{d/(4m)} + \max_{g_j,h_j\in G_j,\|g_j-h_j\|_{L_1}\le\epsilon} \|Z_n(g_j) - Z_n(h_j)\|_{\psi_2}.
\end{aligned}$$

We focus on the second term of the preceding display. Fix $j \ge 2$, for any $g_j, h_j \in G_j$, consider the finite sequences $(g_0, g_1, \ldots, g_j)$ and $(h_0, h_1, \ldots, h_j)$ such that $g_i, h_i \in G_i$ and $\|g_i - g_{i+1}\|_{L_1} \le \epsilon/2^i$, $i = 1, \ldots, j-1$. Invoking the inequality (S22), we have

$$\begin{aligned}
&\max_{g_j,h_j\in G_j,\|g_j-h_j\|_{L_1}\le\epsilon} \|Z_n(g_j) - Z_n(h_j)\|_{\psi_2} \\
&\le \max_{g_j,h_j\in G_j,\|g_j-h_j\|_{L_1}\le\epsilon} \|\{Z_n(g_j) - Z_n(h_j)\} - \{Z_n(g_0) - Z_n(h_0)\}\|_{\psi_2} \\
&\quad + \max_{g_0,h_0\in G_0,\|g_j-h_j\|_{L_1}\le\epsilon} \|Z_n(g_0) - Z_n(h_0)\|_{\psi_2} \\
&\le \max_{g_j,h_j\in G_j,\|g_j-h_j\|_{L_1}\le\epsilon} \|\{Z_n(g_j) - Z_n(h_j)\} - \{Z_n(g_0) - Z_n(h_0)\}\|_{\psi_2} \\
&\quad + 2\sqrt{\log(1 + |G_0\times G_0|)} \max_{(g_0,h_0)\in G_0\times G_0,\|g_j-h_j\|_{L_1}\le\epsilon} \big\|\|Z_n(g_0) - Z_n(h_0)\|\big\|_{\psi_2}.
\end{aligned}$$

Clearly, the second term can be bounded by inequality (S23):

$$\begin{aligned}
&2\sqrt{\log(1 + |G_0\times G_0|)} \max_{(g_0,h_0)\in G_0\times G_0,\|g_j-h_j\|_{L_1}\le\epsilon} \big\|\|Z_n(g_0) - Z_n(h_0)\|\big\|_{\psi_2} \\
&\le \left(\frac{10\sqrt{24}C_K^2}{\sqrt{n}}\right) \epsilon^{p/(4m)} \sqrt{\log\left(1 + \exp\left(2c_0^{p/(2m)}\epsilon^{-p/m}\right)\right)};
\end{aligned}$$

since

$$|G_0 \times G_0| = |G_0|^2 \le \exp(2\log N_{[]}(\epsilon;\{f:\|f\|_\infty\le 1\};\|\cdot\|_{L_1})) \le \exp(2c_0 {}^{p/(2m)}\epsilon^{-p/m}); \tag{S24}$$

it suffices to bound the first term. Write

$$\begin{aligned}
&\max_{g_j,h_j\in G_j;\|g_j-h_j\|_{L_1}\le\epsilon}\|\{Z_n(g_j)-Z_n(h_j)\}-\{Z_n(g_0)-Z_n(h_0)\}\|_2\\
&\le 2\sum_{i=0}^{j-1}\max_{(g_i,g_{i+1})\in G_i\times G_{i+1};\|g_i-g_{i+1}\|_{L_1}\le\epsilon/2^i}\|Z_n(g_{i+1})-Z_n(g_i)\|_2\\
&\le 2\sum_{i=0}^{j-1}2\sqrt{\log(1+|G_i||G_{i+1}|)}\max_{(g_i,g_{i+1})\in G_i\times G_{i+1};\|g_i-g_{i+1}\|_{L_1}\le\epsilon/2^i}\|\|Z_n(g_{i+1})-Z_n(g_i)\|\|_2\\
&\le \frac{4\sqrt{24}C_K^2 {}^{p/(4m)}}{\sqrt{n}}\sum_{i=0}^{j-1}\sqrt{\log\left[1+\exp\left\{2c_0 {}^{p/(2m)}(\epsilon/2^i)^{-p/m}\right\}\right]}\,\epsilon/2^i;
\end{aligned}$$

where inequalities (S21) and (S22) are applied. Bounding the sum by integral, we have

$$\begin{aligned}
&\sum_{i=0}^{j-1}\sqrt{\log\left[1+\exp\left\{2c_0 {}^{p/(2m)}(\epsilon/2^i)^{-p/m}\right\}\right]}\,\epsilon/2^i\\
&\le\sum_{i=0}^{j-1}\int_{\epsilon/2^{i+1}}^{\epsilon/2^i}\sqrt{\log\{1+\exp(2c_0 {}^{p/(2m)}x^{-p/m})\}}\,dx\\
&\le\int_0^\epsilon\sqrt{\log\{1+\exp(2c_0 {}^{p/(2m)}x^{-p/m})\}}\,dx.
\end{aligned}$$

Putting all pieces above together, we obtain the following bound:

$$\begin{aligned}
&\sup_{g,h\in W_2^m(\mathcal{X};1);\|g-h\|_{L_1}\le\epsilon}\|Z_n(g)-Z_n(h)\|_2\\
&\lesssim\frac{{}^{p/(4m)}}{\sqrt{n}}\frac{\epsilon}{2^j}+\int_0^\epsilon\sqrt{\log\{1+\exp(2c_0 {}^{p/(2m)}x^{-p/m})\}}\,dx\\
&\quad+\epsilon\sqrt{\log\{1+\exp(2c_0 {}^{p/(2m)}\epsilon^{-p/m})\}}\ .
\end{aligned}$$

By taking j ! 1 , we can let the first term in the squared bracket tend to 0, and hence,

$$\sup_{g;h2W_2^m(X;1);\,kg\ \ hk_{L_1}"} kZ_n(g)\quad Z_n(h)k$$

$$. \frac{p=(4m)}{\sqrt{n}} \int_0^{"} \sqrt{\log f1 + \exp(2c_0\ ^{p=(2m)}x\ \ ^{p=m})g}\,dx \qquad 2$$

$$+\ "\ \sqrt{\log f1 + \exp(2c_0\ ^{p=}(^{2m)"}\ \ ^{p=m})g}$$

$$. \frac{p=(4m)}{\sqrt{n}} \int^{"} \sqrt{\log f1 + \exp(2c_0\ ^{p=(2m)}x\ \ ^{p=m})g}\,dx\text{: } 0$$

Now we take h = 0, which implies $Z_n(h) = 0$ by the construction of Z_n. Furthermore, by the property of reproducing kernel K and the Cauchy-Schwarz inequality,

$$kg\quad hk_{L_1}\quad \sup_{x2X} jg(x)j = \sup_{x2X} jhg(); K\ (x;)i\ j$$

$$\sup_{x2X} kgk\quad \sqrt{hK\ (x;); K\ (x;)i}\quad C_K\ \ ^{p=(4m)}:$$

Taking " = C_K $^{p=(4m)}$, we obtain

$$\sup_{g2W_2^m(X;1)} kZ_n(g)k$$

$$\sup_{g2W_2^m(X;1);\,kgk_{L^1}"} kZ_n(g)k \qquad 2$$

$$\sup_{g;h2W_2^m(X;1);\,kg\ \ hk_{L_1}"} kZ_n(g)\quad Z_n(h)k \qquad 2$$

$$.\ n\ ^{1=2}\ ^{p=(4m)} \int_0 \sqrt{\log f1 + \exp(2c_0\ ^{p=(2m)}x\ \ ^{p=m})}\,dx$$

$$.\ n\ ^{1=2}\ \ ^{p(6m\ \ p)=(8m^2)}:$$

Hence, invoking Lemma 8.1 in [3], we finally obtain

$$P\left(\sup_{g2W_2^m(X;1)} kZ_n(g)k > t\right)\quad 2\exp\left(\ 2\ \frac{nt_2^2}{p\ 6m\ \ p)=(4m\ \ )_K}\right);$$

for some absolute constant $_K$ depending on K only, completing the proof. ■

S5. Proof for Section 3.2. Denote $\hat{}_z := {}_{;z;n}$, $\hat{}_z() := \hat{}_{;z;n}()$ and $`_z(;) := `_{;z;n}(;)$ in (11).

We need the following Corollary S1 and Lemma S2 to prove theorem 3.3. Corollary S1 is a direct consequence of Theorem 3.1. We repeatedly use the fact that for any $f() \; 2 \; L_2(X)$, there exists a constant C such that $kfk_{L_2(X)} \; Ckfk_H$ in the following proof.

Corollary S1. Denote $\hat{}_{z;} = \operatorname{argmin}_{2H} `_z(;)$ for each $\; 2 \;$. Under the Assumptions A1 to A6, for suciently large n and any $\; > \; 2$ and $C \; 2 \; (0;1)$, with probability at least $1 \quad \exp f \; (\quad 2)=3g \qquad \exp f \; n^C g$, one has

$$\sup_2 k\hat{}_{z;}() \qquad (y^R() \quad f^M(;))k_{L_2(X)}$$
$$2^p 2 \sup_2 ky^R() \qquad f^M(;)k_{L_2(X)}$$
$$+ \sup_2 ky^R() \quad f^M(;)k_H \qquad + C_{K0}{}^{1=2} \quad n^{\quad 2m+p\frac{m}{};2}$$

and

$$\sup_2 k\hat{}_{z;}()k_H \;\; (2^p 2 + 1) \sup_2 k(y^R() \qquad f^M(;))k_H$$
$$+ 2^p 2 \sup_2 ky^R() \quad f^M(;)k_{L_2(X)} + 2 \quad \overline{p} 2C_{K0}{}^{1=2}$$

by choosing $= n^{\quad 2m=(2m+p)}$ and ${}_z = {}^{\quad 1=2}$, where C_K is a constant depending on the kernel $K(;)$.

Lemma S2. Under assumptions A1 to A6,

(i) it holds that

$$\sup_2 \frac{1}{n}\sum_{i=1}^n (y^R(x_i) \quad f^M(x_i;) \qquad \hat{}_{z;}(x_i))^2$$
$$\int_{x2X} (y^R(x) \qquad f^M(x;) \qquad {}_{;}(x))^2 dx = o_p(n^{\quad 1=2});$$

and

$$\sup_2 \frac{1}{n}\sum_{i=1}^n (y^R(x_i) \quad f^M(x_i;) \qquad \hat{}_{z;}(x_i))_i = o_p(n^{\quad 1=2});$$

(ii) for any $j = 1; :::; q$, one has

$$\frac{1}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i; \hat{}_z) \quad \hat{}_z(x_i))\frac{@f^M(x_i; \hat{}_z)}{@_j}$$

$$= \int_{x2X}(y^R(x) \quad f^M(x; \hat{}_z) \quad \hat{}_z(x))\frac{@f^M(x; \hat{}_z)}{@_j}dx + o_p(n^{1=2}):$$

Proof. Denote

$$W_2^m(H; B) := \left\{ f() = \sum_{j=1}^{1} f_j() \; 2 \; L_2(X) : \sum_{j=1}^{1} {}_j^{2m=p} f_j^2 \quad B^2 \right\};$$

and

$$s_i^2(;) := (y^R(x_i) \quad f^M(x_i;) \quad (x_i))^2;$$

$$u_i(;) := (y^R(x_i) \quad f^M(x_i;) \quad (x_i))_i;$$

$$r_i^2(;) := (y^R(x_i) \quad f^M(x_i;) \quad (x_i))\frac{@f \; @^M(x;)}{_j}$$

for $(;) \; 2 \quad W_2^m(X; B)$ and some $B > 0$ that will be specified later. Define the empirical processes

$$s^2(;) := \frac{1}{\overline{p} n}\sum_{i=1}^{n} f s_i^2(;) \qquad E_{x_i}[s_i^2(;)]g;$$

$$u(;) := \frac{1}{\overline{p} n}\sum_{i=1}^{n} f u_i(;) \qquad E_{x_i; i}[u_i(;)]g;$$

$$r(;) := \frac{1}{\overline{p} n}\sum_{i=1}^{n} r_i(;) \qquad E_{x_i}[r_i(;)]g;$$

where

$$E_{x_i}[s_i^2(;)] = \int_{x2X}(y^R(x) \quad f^M(x;) \quad (x))^2 dx;$$

$$E_{x_i; i}[u_i(;)] = \int_{x2X}(y^R(x) \quad f^M(x;) \quad (x))_i dx = 0;$$

$$E_{x_i}[r_i(;)] = \int_{x2X}(y^R(x) \quad f^M(x;) \quad (x))\frac{@f^M(x;)}{@_j}dx:$$

By Assumptions A3 and A4, the function classes f@f^M(;)=@$_j$: 2 g and F = fy^R() f^M(;); 2 g are Donsker. Note that, by denition, W^m(X ; B) is also Donsker. Since both W$_2$ (X ; B) and F are uniformly bounded, the function classes

$$f(y^R() \quad f^M(;) \quad ())^2 : 2 ; 2 W^m(X ; B)g; \qquad \text{and}$$

$$(y^R() \quad f^M(;) \quad ()) \frac{@f \quad ;M}{@( ) \quad j} : 2 ; 2 W^m(X ; B) \quad 2$$

are also Donsker classes. Furthermore, letting f$_{;}$(; x) = (y^R(x) f^M(x;) (x)), observe that for any ($_1$; $_1$) and ($_2$; $_2$), the distance

$$E_0 (f^1_{;1} \quad f^2_{;2})^2 \quad ^{1=2}$$
$$= {}_0 kf^M(;_1) \quad {}_1() \quad f^M(;_2) + {}_2()k_{L_2(X)}$$
$${}_0 \; kf^M(;_1) \quad f^M(;_2)k_{L_2(X)} + k_1() \quad {}_2()k_{L_2(X)}$$

can be bounded by the L$_2$(X)-distance of functions in ff^M(;) : 2 g and () 2 W^m(X ; B). In addition, by Assumption A4 ff^M(;) : 2 g and W^m(X ; B) are Donsker classes, it follows that the function class

$$ff_{;} \; 2 \; C(R \;\; X) : 2 ; 2 W_2^m(X ; B)g$$

is also Donsker, since its metric entropy can be upper bounded by those of ff^M(;) : 2 g and W$_2^m$(X ; B). By Theorem 2.4 in [4], for any t$_1$ > 0 and any B > 0, there exists t$_2$; t$_2^0$; t$_2^{00}$ > 0 such that

$$\text{(S25)} \qquad \limsup_{n\,!\,1} P \quad \sup_{kk_H B;\, 2;\, ky^R() \;\; f^M(;) \;\; ()k_{L_2(X)} t_2} \quad j\hat{S}(;)j > t_1 \quad ! \quad < t_1;$$

$$\text{(S26)} \qquad \limsup_{n\,!\,1} P \quad \sup_{kk_H B;\, 2;\, ky^R() \;\; f^M(;) \;\; ()k_{L_2(X)} t_2^0} \quad j(;)j > t_1 \quad ! \quad < t_1;$$

$$\text{(S27)} \qquad \limsup_{n\,!\,1} P \quad \sup_{kk_H B;\, 2;\, ky^R() \;\; f^M(;) \;\; ()k_{L_2(X)} t_2^{00}} \quad j(;)j > t_1 \quad ! \quad < t_1:$$

Note that by Corollary S1, sup$_2$ k$_{z;}$$\hat{k}_H$ is asymptotically tight, and therefore for any " > 0, there exists B$_0$ > 0 and some integer N 2 N$_+$, both depending on , such that P(sup$_2$ k$_{z;}$$\hat{k}_H$ > B$_0$) "=3 for all n > N. Now take B = B$_0$, t$_1$ = =3. Then we can choose t$_2$ to be a value that satises (S25), t$_2^0$ satisfying (S26), and t$_2^{00}$ satisfying (S27). By

Corollary S1 and Assumption A5, sup $k_{z;}\hat{(})\ (y^{R}()\ \ f^{M}(;))k_{L_2(X)} = O_P(n^{\ 2m=(2m+p)})$, and hence there exists $t_3 > 0$, depending on and n, such that for all $n > N$, it holds that

$$P \quad \sup_{2} ky^{R}() \quad f^{M}(;) \quad \hat{}_{z;}()k_{L_2(X)} \ \ t_3 \qquad < \ "=3:$$

Without loss of generality, we may require t_3 minf$t_2; t2 ; t^{0}_{0}g_2$ by taking suciently large n. Then for suciently large n, we obtain

$$P \quad \sup_{2} j\hat{s}^{2}(;\ _{z;})j\hat{} > \ "$$

$$P\ @ \quad \underset{k_{z;}k_H \hat{B_0};\ 2;\ ky^{R}()\ \ f^{M}(;)\ \ _{z;}()k_{L_2(X)}t_2}{\sup} \quad j\hat{s}^{2}(;\ _{z;})j\hat{} > \ t_1 A \ \ \sup \hat{}$$

$$+ \ P \quad \sup_{2} ky^{R}() \quad f^{M}(;) \quad \hat{}_{z;}()k_{L_2(X)} > \ t_2$$

$$+ \ P \quad \sup_{2} k_{z;}\hat{k}_H > \ B_0$$

$$< \ "=3 + \ "=3 + \ "=3 = \ ";$$

and similarly,

$$P \quad \sup_{2} j(;\ _{z;})j\hat{} > \ " \qquad < \qquad \text{and} \quad P \quad \sup_{2} j\text{u}(;\ _{z;})j\hat{} > \ " \qquad < \ :$$

Therefore,

$$\frac{1}{p\overline{n}} \sup_{2} j\$(;\ _{z;})j\hat{}$$

$$= \ \sup_{2} \ \frac{1}{n}\sum_{i=1}^{n} (y^{R}(x_i) \quad f^{M}(x_i;) \quad \hat{}_{z;}(x_i))^2$$

$$\int_{x\,2\,X} (y^{R}(x) \qquad f^{M}(x;) \quad \hat{}_{z;}(x))^2 dx = \ o_p(n^{\ 1=2});$$

and

$$\frac{1}{p\overline{n}} \sup_{2} j\text{u}(;\ _{z;})j\hat{} = \ \sup_{2} \quad \frac{1}{n}\sum_{i=1}^{n} (y^{R}(x_i) \quad f^{M}(x_i;) \quad \hat{}_{z;}(x_i))_i \ = \ o_p(n^{\ 1=2});$$

completing the proof of (i). The proof of (ii) can be completed by observing that

$$\frac{1}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i;\hat{}_z) \quad \hat{}_z(x_i))\frac{@f^M(x_i;\hat{}_z)}{@_j}$$

$$\int_{x\in X}(y^R(x) \quad f^M(x;\hat{}_z) \quad \hat{}_z(x))\frac{@f^M(x;\hat{}_z)}{@_j}dx$$

$$= \frac{1}{\sqrt{n}}|(_z;\hat{} \;\hat{}_{z;\,z})_j| \; \frac{1}{\sqrt{n}}\sup_2 |(;_{z;})_j| = o_p(n^{1=2}):$$

■

Proof for Theorem 3.3. Without loss of generality, it suces to prove the case when $_z =$ $^{1=2}$. For the general case when $_z = O($ $^{1=2})$, the proof follows similarly. We rst show $_z \, !^p$ $\hat{}_{L_2}$. By the denition of $_z$, $_{L_2}$, and the theory of M-estimators (see, Theorem 5.7 in [10]), it suces to show that $^{1=2}(`_z(;_{z;})$ $\hat{}$ $^2_0) \, !^p \; \|y^R() \quad f^M(;)\|^2_{L_2(X)}$ uniformly for each 2. Note that

$$`_z(\hat{}_{z;}();)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i;) \quad \hat{}_{z;}(x_i))^2 + \frac{1}{n}\sum_{i=1}^{n}\text{ilon}^2_i$$

$$+ \frac{2}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i;) \quad \hat{}_{z;}(x_i))_i + \|_{z;}\|^2_{\hat{H}} + {}^p\|_{z;}\|_{L_2(X)}\hat{}_{i=1} \quad {}^2$$

$$:= A_n + B_n + C_n + D_n + E_n:$$

For A_n, by Lemma S2 (i) and Corollary (S1), one has

(S28) $$\sup_2 \frac{1}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i;) \quad \hat{}_{z;}(x_i))^2 = o_p(n^{1=2})$$

Since $E[B_n] = {}^2_0$ and $V[B_n] = O(n^{1})$, Chebyshev's inequality implies $(1=n)\sum_{i=1}^{n}\text{ilon}^2_i = {}^2_0 + O_p(n^{1=2})$ for B_n. For C_n, Lemma S2 (i) guarantees that

$$\sup_2 \frac{2}{n}\sum_{i=1}^{n}(y^R(x_i) \quad f^M(x_i;) \quad \hat{}_{z;}(x_i))_i = o_p(n^{1=2})$$

Since $= O(n^{2m=(2m+p)})$, by the asymptotic tightness of $\sup \|_{z;}\|_{\hat{H}}$ (Corollary S1), one has $\sup_2 k_{z;}\hat{} k_H = o_p^2(n^{1=2})$. By putting the above all pieces together, we obtain

(S29) $$\sup_{2} {}^{1=2}(`_z(_{z;}(\hat{)};) \qquad {}^{2})_0 \quad k_{\hat{z};}k_{L_2}{}^2_{(X)} = O_p((n)^{1=2}):$$

For any , by the Cauchy-Schwarz inequality, one has

$$k_{z;}\hat{}k_{L_2}{}^2_{(X)} \qquad ky^R() \quad f^M(;)k_{L_2}{}^2_{(X)}$$

$$k(_{z;}(\hat{)} \qquad (y^R() \quad f^M(;))k_{L_2(X)}k_{z;}(\hat{)} + y^R() \qquad f^M(;)k_{L_2(X)}$$

Recall that

$$\sup_2 k(\hat{_{z;}}() \qquad (y^R() \quad f^M(;))k_{L_2(X)} = O_p(n^{m=(2m+d)})$$

by Corollary S1 and Assumption A4. Using Assumptions A4 and the asymptotic tightness of $\sup k_{z;}\hat{}k_H$ (Corollary S1), one has

$$k\hat{_{z;}}() + y^R() \qquad f^M(;)k_{L_2(X)}$$

$$k_{z;}(\hat{)}k_{L_2(X)} + \sup_2 ky^R() \qquad f^M(;)k_{L_2(X)}$$

$$Ck_{z;}()\hat{}k_H + \sup_2 ky^R() \qquad f^M(;)k_H = O_p(1):$$

Thus

$$\sup_2 k_{z;}\hat{}k^2{}_{2(X)} \qquad ky^R() \quad f^M(;)k^2{}_{2(X)} = O_p(n^{m=(2m+d)});$$

and hence,

$$\sup_2 {}^{1=2}(`_z(;_{z;}) \ \hat{} \qquad {}^{2})_0 \quad ky^R() \quad f^M(;)k^2{}_{2(X)} = o_p(1);$$

from which we conclude $_z\hat{}\,!^p\ _{L_2}$.

Next we derive the convergence rate of $_z\hat{}$. Apply the Frechet derivative on $`_z$ with regard to () and the partial derivative on $`_z$ with regard to $_j$, j = 1;:::;q. For any g() 2 H, $_z$ $\hat{}$

and $\hat{}_z$ satisfy

$$0 = \frac{2}{n} X^n (y_i^F \quad f(x_i;\hat{}_z) \quad \hat{}_z(x_i))g(x_i) + 2h_z(\hat{}); g()i_H \; {}_{i=1}$$

$$+ 2p\, \underline{h_z}(\hat{}); g()i_{L_2(X)};$$

(S30)

$$0 = \frac{2}{n} \underset{i=1}{X^n} (y_i^F \quad f(x_i;\hat{}_z) \quad \hat{}_z(x_i)) \frac{@f^M(x_i;\hat{}_z)}{@_j}:$$

(S31)

Choosing $g() = \frac{@f^M(;\hat{}_z)}{@_j}$ and plugging (S31) into (S30), one has

(S32) $$p \left\langle \hat{}_z(); \frac{@^M(;\hat{}_z)}{f @_j} \right\rangle_H + \left\langle \hat{}_z(); \frac{@^M(;\hat{}_z)}{f @_j} \right\rangle_{L_2(X)} = 0:$$

Substituting (S31) into (S32) and by Lemma S2 (ii), we have

$$0 = \frac{1}{n} \underset{i=1}{X^n} (y_i^F \quad f(x_i;\hat{}_z) \quad \hat{}_z(x_i)) \frac{@f^M(x_i;\hat{}_z)}{@_j}$$

$$= \int (y^R(x) \quad f^M(x;\hat{}_z)) \frac{@f^M(x;\hat{}_z)}{_j} dx + \left\langle \hat{}_z(); \frac{@f^M(;\hat{}_z)}{@_j} \right\rangle_{L_2(X)}$$

$$\frac{1}{n} \underset{i=1}{X^n} {}_i \frac{@f^M(x_i;\hat{}_z)}{@_j} + o_p(n^{1=2})$$

$$= \int \frac{@(y^R(x) \quad f^M(x;\hat{}_z))^2}{@_j} dx \quad p \left\langle \hat{}_z(); \frac{@f^M(;\hat{}_z)}{@_j} \right\rangle_H$$

$$\frac{1}{n} \underset{i=1}{X^n} {}_i \frac{@f^M(x_i;\hat{}_z)}{_j} + o_p(n^{1=2}):$$

Applying Taylor expansion to the rst term on the right-hand side at $_{L_2}$, for any $j = 1;:::;q$, we obtain

$$\left(\int \frac{@^2(y^R(x) \quad f^M(x;\tilde{}_z))^2}{@_j@} dx \right)^T (\hat{}_z \quad _{L_2})$$

$$= \int \frac{@^2(y^R(x) \quad f^M(x;_{L_2}))^2}{@_j@} dx + o_p(1)\right)^T (_z \quad \hat{} \quad _{L_2})$$

$$= p \left\langle \hat{}_z(); \frac{@f^M(;_z)}{@_j} \right\rangle_H + \frac{1}{n} \underset{i=1}{X^n} {}_i \frac{@f^M(x_i;_z)}{@_j} + o_p(n^{1=2});$$

(S33)

where $\tilde{}_z$ lies within the q dimensional rectangle between $_{L_2}$ and $\hat{}_z$. Observe that Corollary S1 and assumption A3 imply

$$\overset{*}{}{}_{z};\hat{} \frac{@f \;\; {}^M(;{}_z)}{@ \;\; {}_j} \overset{+}{\hat{}}_H \; k_z k_H \; \frac{@f\hat{}^M(;{}_{L_2})}{@_j} \; {}_H + o_p(1) = O_p(1):$$

Now we consider the second term. Dene the empirical process

$$G_n() = \frac{1}{p\overline{}n} \overset{n}{\underset{i}{X}} \; \frac{@f^M(x_i;)}{@_j} \quad {}_i \; \frac{@f^M(x_i;{}_{L_2})}{@_j} \; {}_{i=1}:$$

and denote

$$f(;x) = \frac{@f \;\; {}^M(x;)}{@ \;\;\; {}_j} \quad \frac{@f^M(x;{}_{L_2})}{@_j}:$$

Since

$$E_{ilon;x}[f_1(;x) \qquad f_2(;x)]^2$$

$$= \overset{ilon;}{E}{}_x \left[{}^2 \frac{@f^M(x; \;)}{@_j} {}_1 \quad \frac{@f^M(x; \;)}{@_j} {}_2 \right]^{2}$$

$$= \; {}^2_0 \; \frac{@f^M(x; \;)}{@_j} {}_1 \quad \frac{@f^M(x; \;)}{@_j} {}_2 \; {}^2_{L_2(X)} \quad ;$$

therefore the function class $ff(;x) \; 2 \; C(R \;\; X): \; 2 \; g$ is Donsker by Assumption A3, and hence, $G_n()$ converges weakly to a tight Gaussian stochastic process, denoted by $G()$. W.l.o.g., we may take $G()$ a version that has uniformly continuous sample paths (see Chapter 6 in [9]). Since $G_n({}_{L_2}) = 0$ for all n, it follows that $G({}_{L_2}) = 0$. By the consistency of $\hat{}_z$ and the continuous mapping theorem [10], $G_n(\hat{}_z) = G({}_{L_2}) + o_p(1) = o_p(1)$. Therefore,

$$\frac{1}{n} \overset{n}{\underset{i}{X}} \; \frac{@f^M(x_i;{}_z)}{@_j}\hat{} = \frac{1}{p\overline{}n} G_n(\hat{}_z) + \frac{1}{n} \overset{n}{\underset{i=1}{X}} {}_i \; \frac{@f^M(x_i;{}_{L_2})}{@_j} = O_p(n^{\;\;1=2}) \; {}_{i=1}$$

To sum up,

$$\int \frac{@^2(y^R(x) \qquad f^M(x;{}_{L_2}))^2}{@@^T} dx + o_p(1)(\hat{}_z \qquad {}_{L_2})$$

$$= O_p(n^{\;\;m=(2m+p)}) + O_p(n^{\;\;1=2}) + o_p(n^{\;\;1=2}) = O_p(n^{\;\;m=(2m+p)});$$

completing the proof. ■

S6. Proof and additional results for Section 4. The identities in the Lemma S1 are used repeatedly in the proof of the Theorem 4.1 and Lemma 4.2.

Lemma S1. Denote $^2 R_{z_d}$ the covariance matrix of $(_{z_d}(x_1); :::; _{z_d}(x_n))^T$, where the $(i; j)$ entry being $^2 K_{z_d}(x_i; x_j)$ dened in (20). Denote $r_{z_d}(x) = (K_{z_d}(x; x_1); :::; K_{z_d}(x; x_n))^T$ for any x 2 X. One has the following identities

$$R_{z_d}^{\;1} = R^{\;1} + \frac{z I_n}{n}; \tag{S34}$$

$$r^T_{z_d}(x) = \frac{n}{z} r^T(x) \tilde{R}^{\;1} = r^T(x) R^{\;1} R_{z_d}; \tag{S35}$$

for any x 2 X.

Proof. By the denitions of R_{z_d} and the Woodbury Identity, one has

$$R_{z_d} = R \quad R \tilde{R}^{\;1} R = R \quad I_n \quad I_n + \frac{z}{n} R^{\;1} \; ^{1\,!}$$

$$R \quad \frac{R^{z}}{n} + I_n \quad ^{1} = R^{\;1} + \frac{I_n z}{n} \quad ^{1\,=};$$

from which (S34) follows.

Equation (S35) can be shown similarly by noting $r^T_{z_d}(x) = r^T(x) \quad r^T(x) \tilde{R}^{\;1} R$ and the Woodbury Identity. ■

Proof of Theorem 4.1. The predictive mean is as follows

$$\begin{aligned} \hat{}_z(x) &= E[y^F(x) \; j \; y^F ; ; {}^2; ;_d] \\ &= f^M(x;) + r_{z_d}(x)^T (R_{z_d} + n I_n) \;^{1} (y^F \qquad f^M) \\ &= f^M(x;) + r(x)^T R^{\;1} R_{z_d} (R_{z_d} + n I_n) \;^{1} (y^F \qquad f^M) \\ &= f^M(x;) + r(x)^T R^{\;1} \quad I_n + n \quad R^{\;1} + \frac{z I_n}{n} \;^{1\,!} \quad (y^F \quad f^M) \\ &= f^M(x;) + \frac{r(x)^T}{(1 + z)} \quad R^{\;1} + \frac{n \; I_n}{1 + z} \;^{1} \quad (y^F \quad f^M); \end{aligned}$$

where the last two equalities follow from (S35) and (S34), respectively.

The predictive variance can be obtained using (S35) and (S34) as follows

$$\begin{aligned} K_z(x; x) = {} & K_{z_d}(x; x) \quad r^T_{z_d}(x) (R_{z_d} + n I_n) \;^{1} r_{z_d}(x) \\ = {} & K(x; x) \quad r^T(x) \tilde{R}^{\;1} r(x) \\ & (1 + z) \;^{1} r(x)^T \quad R^{\;1} + \frac{n \; I_n}{1 + z} \;^{1} \quad \frac{n}{z} \tilde{R}^{\;1} r(x) \end{aligned}$$ ■

from which the result follows.

Proof of Lemma 4.2. When $_0^2 = 0$, the predictive mean is as follows

$$\begin{aligned} E[y^F(x) \, j \, y^F ; ; ^2 ; ;_\sigma] &= f^M(x;) + r_{z_d}(x)^T R \;\; ^1(y^F_{\;\; z_d} \qquad f^M) \\ &= f^M(x;) + r(x)^T R \;\; ^1 R_{z_d} R_{z_d} \, (^1\!y^F \qquad f^M) \\ &= f^M(x;) + r(x)^T R \;\; ^1(y^F \quad f^M): \end{aligned}$$

The predictive variance can be obtained similarly. ■

REFERENCES

[1] D. E. Edmunds and H. Triebel, Function spaces, entropy numbers, differential operators, vol. 120, Cambridge University Press, 2008.

[2] S. Ghosal and A. Van der Vaart, Fundamentals of nonparametric Bayesian inference, vol. 44, Cambridge University Press, 2017.

[3] M. R. Kosorok, Introduction to empirical processes and semiparametric inference., Springer, 2008.

[4] E. Mammen and S. Van de Geer, Penalized quasi-likelihood estimation in partial linear models, The Annals of Statistics, (1997), pp. 1014{1035.

[5] I. Pinelis, Optimum bounds for the distributions of martingales in banach spaces, The Annals of Probability, (1994), pp. 1679{1706.

[6] C. E. Rasmussen, Gaussian processes for machine learning, MIT Press, 2006.

[7] M. Rudelson, R. Vershynin, et al., Hanson-wright inequality and sub-gaussian concentration, Electronic Communications in Probability, 18 (2013).

[8] R. Tuo and C. J. Wu, Efficient calibration for imperfect computer models, Annals of Statistics, 43 (2015), pp. 2331{2352.

[9] S. A. Van de Geer, Empirical Processes in M-estimation, vol. 6, Cambridge university press, 2000.

[10] A. W. Van der Vaart, Asymptotic statistics, vol. 3, Cambridge university press, 2000.

[11] G. Wahba, Spline models for observational data, vol. 59, SIAM, 1990.

[12] Y. Yang, A. Bhattacharya, and D. Pati, Frequentist coverage and sup-norm convergence rate in Gaussian process regression, arXiv preprint arXiv:1708.04753, (2017).

[13] Y. Yang, Z. Shang, and G. Cheng, Non-asymptotic theory for nonparametric testing, arXiv preprint arXiv:1702.01330, (2017).