Submitted to Statistical Science

Supplementary Material for Distributed Bayesian Inference in Massive Spatial Data

Rajarshi Guhaniyogi[*]
Department of Statistics, Texas A & M University, College Station, Texas, U.S.A.

Cheng Li[†]
Department of Statistics and Data Science, National University of Singapore, Singapore

Terrance Savitsky[‡]
U.S. Bureau of Labor Statistics, Washington D.C., U.S.A.

Sanvesh Srivastava[§]
Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, U.S.A.

Abstract. We provide detailed technical proofs of the theorems and derive the posterior sampling algorithms described in Section 3 of the main manuscript, as well as an extension to the case when τ^2 is unknown and assigned a prior in Section 1.3. We include more detailed information on parameter estimation and convergence of Markov chains for the DISK posterior for the two simulation examples and real data analysis in Section 4 of the main manuscript.

Key words and phrases: Distributed Bayesian inference, Gaussian process, low-rank Gaussian process, modified predictive process, massive spatial data, Wasserstein distance, Wasserstein barycenter.

1. PROOF OF THEOREMS IN SECTION 3.4

Recall that the spatial regression model with a GP prior considered in Section 3.4 is

$$\begin{aligned} &y(\mathbf{s}_i) = w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \qquad \epsilon(\mathbf{s}_i) \sim N(0, \tau^2), \\ &w(\cdot) \sim \mathrm{GP}\{0, \lambda_n^{-1} C_{\boldsymbol{\alpha}}(\cdot,\cdot)\}, \qquad i = 1, \ldots, n. \end{aligned} \tag{1}$$

Writing this model for the n locations in $\mathcal{S}$ gives

$$\mathbf{y} = \mathbf{w}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \mid \mathcal{S} \sim N(\mathbf{0}, \tau^2 \mathbf{I}), \quad \mathbf{y} \mid \mathcal{S} \sim N(\mathbf{w}_0, \tau^2 \mathbf{I}), \tag{2}$$

where $\mathbf{w}_0 = \{w_0(\mathbf{s}_1), \ldots, w_0(\mathbf{s}_n)\}$ and $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)\}$ are the true value of the residual spatial surface and white noise realized at the locations in $\mathcal{S}$. We can

[*]rajguhaniyogi@tamu.edu
[†]stalic@nus.edu.sg
[‡]savitsky.terrance@bls.gov
[§]sanvesh-srivastava@uiowa.edu Corresponding author.

write the model in a similar format for each data subset. Let $\mathbf{s} \in \mathcal{D}$ be a location, $w_0(\mathbf{s})$ be the true value of the residual spatial surface, $\mathbb{E}_{\mathbf{s}^*}$, $\mathbb{E}_0$, $\mathbb{E}_{\mathcal{S}}$, $\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}$, and $\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}$ respectively be the expectations with respect to the distributions of $\mathbf{s}^*$, $(\mathcal{S}, \mathbf{y})$, $\mathcal{S}$, $\mathbf{y}$ given $\mathcal{S}$, and $(\mathbf{y}, \overline{w}(\mathbf{s}^*))$ given $\mathcal{S}, \mathbf{s}^*$.

In this section, we assume the Assumption A.5, so that the parameters $(\tau^2, \boldsymbol{\alpha})$ are fixed at their truth and the same across all subsets. In this case, if $\overline{w}(\mathbf{s}^*)$ is a random variable that follows the DISK posterior for estimating $w_0(\mathbf{s}^*)$, then conditional on τ^2 and $\boldsymbol{\alpha}$, $\overline{w}(\mathbf{s}^*)$ has the density $N(\overline{m}, \overline{v})$, where

$$\begin{aligned} \overline{m} &= \frac{1}{k}\sum_{j=1}^{k} \mathbf{c}_{j,*}^T (\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\,\mathbf{I})^{-1}\,\mathbf{y}_j, \\ \overline{v}^{1/2} &= \frac{1}{k}\sum_{j=1}^{k} v_j^{1/2},\ v_j = \lambda_n^{-1}\left\{c_{*,*} - \mathbf{c}_{j,*}^T (\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\,\mathbf{I})^{-1}\,\mathbf{c}_{j,*}\right\}, \end{aligned} \tag{3}$$

where $c_{*,*} = C_{\boldsymbol{\alpha}}(\mathbf{s}^*, \mathbf{s}^*)$, and $\mathbf{c}_{j,*}^T = \mathbf{c}_j^T(\mathbf{s}^*) = [C_{\boldsymbol{\alpha}}(\mathbf{s}_{j1}, \mathbf{s}^*), \ldots, C_{\boldsymbol{\alpha}}(\mathbf{s}_{jm}, \mathbf{s}^*)]$. In the proofs below, without confusion, we use the notation $\mathbf{c}_{j,*}$ and $\mathbf{c}_j(\mathbf{s}^*)$ interchangeably.

The Bayes L_2-risk in estimating w_0 using the DISK posterior is defined as

$$\begin{aligned} &\mathbb{E}_0\,\mathbb{E}_{\mathbf{s}^*}\left[\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2\right] \\ &\overset{(i)}{=} \mathbb{E}_{\mathcal{S}} \int_{\mathcal{D}} \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left[\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2\right] \mathbb{P}_{\mathbf{s}}(d\,\mathbf{s}^*), \end{aligned} \tag{4}$$

where (i) follows from Fubini's theorem. Using bias-variance decomposition,

$$\begin{aligned} &\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left[\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2\right] \\ &= \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left[\overline{w}(\mathbf{s}^*) - \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} + \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} - w_0(\mathbf{s}^*)\right]^2 \\ &= \left[\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} - w_0(\mathbf{s}^*)\right]^2 + \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left[\overline{w}(\mathbf{s}^*) - \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\}\right]^2 \\ &\equiv \mathrm{bias}^2_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} + \mathrm{var}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\}. \end{aligned}$$

If $\mathbf{c}_j^T(\cdot) = [\mathrm{cov}\{w(\cdot), w(\mathbf{s}_{j1})\}, \ldots, \mathrm{cov}\{w(\cdot), w(\mathbf{s}_{jm})\}] = \{C_{\boldsymbol{\alpha}}(\mathbf{s}_{j1}, \cdot), \ldots, C_{\boldsymbol{\alpha}}(\mathbf{s}_{jm}, \cdot)\}$, $\mathbf{c}^T(\cdot) = \{\mathbf{c}_1^T(\cdot), \ldots, \mathbf{c}_k^T(\cdot)\}$, $\mathbf{w}_{0j}^T = \{w_0(\mathbf{s}_{j1}), \ldots, w_0(\mathbf{s}_{jm})\}$, and $\mathbf{w}_0^T = \{\mathbf{w}_{01}^T, \ldots, \mathbf{w}_{0k}^T\}$, then the distribution of $\overline{w}(\mathbf{s}^*)$ in (3) implies that

$$\begin{aligned} \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} &= \frac{1}{k}\sum_{j=1}^{k} \mathbf{c}_j^T(\mathbf{s}^*)\left(\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k} I\right)^{-1} \mathbf{w}_{0j} \\ &= \mathbf{c}_*^T (k\,\mathbf{L} + \tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0, \\ \mathrm{var}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*)\} &= \mathrm{var}_{\mathbf{y}\,|\,\mathcal{S}}\left[\mathbb{E}\{\overline{w}(\mathbf{s}^*) \mid \mathbf{y}\}\right] + \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left[\mathrm{var}\{\overline{w}(\mathbf{s}^*) \mid \mathbf{y}\}\right] \\ &\overset{(i)}{=} \mathrm{var}_{\mathbf{y}\,|\,\mathcal{S}}\left[\frac{1}{k}\sum_{j=1}^{k} \mathbf{c}_{j,*}^T (\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\,\mathbf{I})^{-1}\,\mathbf{y}_j\right] + \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left[\overline{v}(\mathbf{s}^*)\right] \\ &= \tau^2\,\mathbf{c}^T(\mathbf{s}^*)(k\,\mathbf{L} + \tau^2\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}(\mathbf{s}^*) + \overline{v}(\mathbf{s}^*), \end{aligned} \tag{5}$$

where $\mathbf{L}$ is a block-diagonal matrix with $\mathbf{C}_{1,1}, \ldots, \mathbf{C}_{k,k}$ along the diagonal. The equality (i) holds due to the following reasons: (i) The true data follows $y(\mathbf{s}_{ji}) =$

$w_0(\mathbf{s}_{ji}) + \epsilon(\mathbf{s}_{ji})$ $(j = 1, \ldots, k$ and $i = 1, \ldots, m)$, where $\epsilon(\mathbf{s}_{ji})$'s are all independent with variance $\tau_0^2 = \tau^2$ by Assumption A.5; (ii) $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ conditional on $\mathcal{S}$ are jointly independent since they are a disjoint (random) partition of the full dataset by Assumption A.1, which implies that $\text{var}_{\mathbf{y}\,|\,\mathcal{S}}\left(\sum_{j=1}^k \mathbf{a}_j^T\,\mathbf{y}_j\right) = \tau^2 \sum_{j=1}^k \mathbf{a}_j^T\,\mathbf{a}_j$ for any vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^m$.

Therefore, the Bayes L_2-risk in (4) can be decomposed into three parts:

$$\begin{aligned}&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0 - w_0(\mathbf{s}^*)\}^2 + \tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\}\\ &\qquad + \mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\overline{v}(\mathbf{s}^*)\right\},\end{aligned} \tag{6}$$

which correspond to bias^2, var_{mean} and var_{DISK} in Theorem 3.1.

1.1 Proof of Theorem 3.1

The next three sections find upper bounds for each of the three terms in (6). The conclusion of Theorem 3.1 follows directly by combining the three upper bounds.

1.1.1 An upper bound for the squared bias Consider the squared-bias term in (6). For ease of presentation, assume that $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ are relabeled to

$$\{\mathbf{s}_{11}, \ldots, \mathbf{s}_{1m}, \ldots, \mathbf{s}_{k1}, \ldots, \mathbf{s}_{km}\}$$

corresponding to the k subsets. Define $\xi_{\mathbf{s}_{ji}}(\cdot) = C_{\boldsymbol{\alpha}}(\mathbf{s}_{ji}, \cdot)$,

$$\begin{aligned}\mathbf{w}_0^T &= (\langle w_0, \xi_{\mathbf{s}_{11}}\rangle_{\mathbb{H}}, \ldots, \langle w_0, \xi_{\mathbf{s}_{1m}}\rangle_{\mathbb{H}}, \ldots, \langle w_0, \xi_{\mathbf{s}_{k1}}\rangle_{\mathbb{H}}, \ldots, \langle w_0, \xi_{\mathbf{s}_{km}}\rangle_{\mathbb{H}})\\ &\equiv (\mathbf{w}_{01}^T, \ldots, \mathbf{w}_{0k}^T),\\ \mathbf{c}^T(\cdot) &= (\xi_{\mathbf{s}_{11}}, \ldots, \xi_{\mathbf{s}_{1m}}, \ldots, \xi_{\mathbf{s}_{k1}}, \ldots, \xi_{\mathbf{s}_{km}})\\ &= \{\mathbf{c}_1^T(\cdot), \ldots, \mathbf{c}_k^T(\cdot)\} \equiv (\mathbf{c}_1^T, \ldots, \mathbf{c}_k^T).\end{aligned} \tag{7}$$

The following lemma provides an upper bound on the squared bias of the DISK posterior.

Lemma 1.1 *If Assumptions A.1–A.5 in the main paper hold, then for some global constant $A > 0$,*

$$\begin{aligned}&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0 - w_0(\mathbf{s}^*)\}^2 \le \frac{8\tau^2\lambda_n}{n}\|w_0\|_{\mathbb{H}}^2\\ &\quad + \|w_0\|_{\mathbb{H}}^2 \inf_{d\in\mathbb{N}}\left[\frac{8n}{\tau^2\lambda_n}\rho^4\,\text{tr}(C_{\boldsymbol{\alpha}})\,\text{tr}(C_{\boldsymbol{\alpha}}^d) + \mu_1\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q\right].\end{aligned}$$

Proof Based on the term $\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0$ in (6), we define Δ_j $(j = 1, \ldots, k)$ and Δ as

$$\begin{aligned}\Delta_j(\cdot) &= \mathbf{y}_j^T(\mathbf{C}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\,\mathbf{I})^{-1}\,\mathbf{c}_j(\cdot) - w_0(\cdot) \equiv \tilde{w}_j(\cdot) - w_0(\cdot),\\ \Delta(\cdot) &= \mathbf{y}^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{c}(\cdot) - w_0(\cdot) = \frac{1}{k}\sum_{j=1}^k\{\tilde{w}_j(\cdot) - w_0(\cdot)\} = \frac{1}{k}\sum_{j=1}^k\Delta_j(\cdot),\end{aligned} \tag{8}$$

so that $\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta) = \mathbf{w}_0^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-1}\,\mathbf{c}(\cdot) - w_0(\cdot) = k^{-1}\sum_{j=1}^k \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)$ and $\mathbb{E}_{\mathcal{S}}\,\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta)\|_2^2$ yields the bias^2 term in (6). Jensen's inequality implies that $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta)\|_2^2 \le k^{-1}\sum_{j=1}^k \|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2$, so we only need to find upper bounds for $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2$ $(j = 1,\ldots,k)$.

We can recognize that the optimization problem below has $\tilde{w}_j(\cdot)$ defined in (8) as its solution,

$$\text{argmin}_{w\in\mathcal{H}} \sum_{i=1}^m \frac{\{w(\mathbf{s}_{ji}) - y(\mathbf{s}_{ji})\}^2}{2\tau^2/k} + \frac{1}{2}\lambda_n\|w\|_{\mathbb{H}}^2, \quad j = 1,\ldots,k. \tag{9}$$

Differentiating (9) and taking expectations with respect to $\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}$ implies that

$$\begin{aligned} &\sum_{i=1}^m \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left\{\tilde{w}_j(\mathbf{s}_{ji}) - y(\mathbf{s}_{ji})\right\}\,\xi_{\mathbf{s}_{ji}} + \frac{\tau^2\lambda_n}{k}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\tilde{w}_j) \\ &= \sum_{i=1}^m \langle \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j), \xi_{\mathbf{s}_{ji}}\rangle_{\mathbb{H}}\,\xi_{\mathbf{s}_{ji}} + \frac{\tau^2\lambda_n}{k}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\tilde{w}_j) = 0, \end{aligned} \tag{10}$$

where the last inequality follows because $y(\mathbf{s}_{ji}) = \langle w_0, \xi_{\mathbf{s}_{ji}}\rangle_{\mathbb{H}} + \epsilon(\mathbf{s}_{ji})$ and $\langle \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\epsilon), \xi_{\mathbf{s}_{ji}}\rangle_{\mathbb{H}} = \langle 0, \xi_{\mathbf{s}_{ji}}\rangle_{\mathbb{H}} = 0$. Using (8), $\Delta_j = \tilde{w}_j - w_0$, $\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\tilde{w}_j) = \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) + w_0$, and dividing by m in (10), we obtain that

$$\frac{1}{m}\sum_{i=1}^m \langle \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j), \xi_{\mathbf{s}_{ji}}\rangle_{\mathbb{H}}\,\xi_{\mathbf{s}_{ji}} + \frac{\tau^2\lambda_n}{km}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) = -\frac{\tau^2\lambda_n}{km} w_0. \tag{11}$$

If we define the jth sample covariance operator as $\hat{\Sigma}_j = \frac{1}{m}\sum_{j=1}^m \xi_{\mathbf{s}_{ji}} \otimes \xi_{\mathbf{s}_{ji}}$, then (11) reduces to

$$\begin{aligned} &\left(\hat{\Sigma}_j + \tfrac{\tau^2\lambda_n}{km}\,\mathbf{I}\right)\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) = -\frac{\tau^2\lambda_n}{km} w_0 \\ &\implies \|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_{\mathbb{H}} \le \|w_0\|_{\mathbb{H}}, \quad j = 1,\ldots,k, \end{aligned} \tag{12}$$

where the last inequality follows because $\hat{\Sigma}_j$ is a positive semi-definite matrix.

The rest of the proof finds an upper bound for $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2$. We now reduce this problem to a finite dimensional one indexed by a chosen $d \in \mathbb{N}$. Let $\boldsymbol{\delta}_j = (\delta_{j1},\ldots,\delta_{jd},\delta_{j(d+1)},\ldots,\delta_{j\infty}) \in L_2(\mathbb{N})$ such that

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) = \sum_{i=1}^\infty \delta_{ji}\varphi_i, \quad \delta_{ji} = \langle \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j), \varphi_i\rangle_{L^2(\mathbb{P})}, \\ &\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2 = \sum_{i=1}^\infty \delta_{ji}^2, \quad j = 1,\ldots,k. \end{aligned} \tag{13}$$

Define the vectors $\boldsymbol{\delta}_j^{\downarrow} = (\delta_{j1},\ldots,\delta_{jd})$ and $\boldsymbol{\delta}_j^{\uparrow} = (\delta_{j(d+1)},\ldots,\delta_{j\infty})$, so $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2 = \|\,\boldsymbol{\delta}_j^{\downarrow}\|_2^2 + \|\,\boldsymbol{\delta}_j^{\uparrow}\|_2^2$ and we upper bound $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2$ by separately upper bounding $\|\,\boldsymbol{\delta}_j^{\downarrow}\|_2^2$ and $\|\,\boldsymbol{\delta}_j^{\uparrow}\|_2^2$. Using the expansion $C_{\boldsymbol{\alpha}}(\mathbf{s},\mathbf{s}') = \sum_{j=1}^\infty \mu_j\varphi_j(\mathbf{s})\varphi_j(\mathbf{s}')$ for any $\mathbf{s},\mathbf{s}' \in \mathcal{D}$, we have the following upper bound for $\|\,\boldsymbol{\delta}_j^{\uparrow}\|_2^2$:

(14)
$$\| \boldsymbol{\delta}_j^{\uparrow} \|_2^2 = \frac{\mu_{d+1}}{\mu_{d+1}} \sum_{i=d+1}^{\infty} \delta_{ji}^2 \leq \mu_{d+1} \sum_{i=d+1}^{\infty} \frac{\delta_{ji}^2}{\mu_i} \overset{(i)}{\leq} \mu_{d+1} \| \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) \|_{\mathbb{H}}^2 \overset{(ii)}{\leq} \mu_{d+1} \|w_0\|_{\mathbb{H}}^2,$$

where (i) follows because $\| \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) \|_{\mathbb{H}}^2 = \sum_{i=1}^{\infty} \delta_{ji}^2/\mu_i$ and (ii) follows from (12).

We then derive an upper bound for $\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2$. Let $\mathbf{M} = \mathrm{diag}(\mu_1, \ldots, \mu_d) \in \mathbb{R}^{d \times d}$, $\Phi^j \in \mathbb{R}^{m \times d}$ be a matrix such that

$$\Phi_{ih}^j = \varphi_h(\mathbf{s}_{ji}), \quad i = 1, \ldots, m, \quad h = 1, \ldots, d, \quad j = 1, \ldots, k, \tag{15}$$

$w_0 = \sum_{i=1}^{\infty} \theta_i \varphi_i$, and the tail error vector $\mathbf{v}_j = (v_{j1}, \ldots, v_{jm})^T \in \mathbb{R}^m$ $(j = 1, \ldots, k)$ such that

$$v_{ji} = \sum_{h=d+1}^{\infty} \delta_{jh} \varphi_h(\mathbf{s}_{ji}), \quad i = 1, \ldots, m.$$

For any $g \in \{1, \ldots, d\}$, taking the $\mathbb{H}$-inner product with respect φ_g in (12) yields

$$\begin{aligned} &\left\langle \left(\frac{1}{m} \sum_{i=1}^{m} \xi_{\mathbf{s}_{ji}} \otimes \xi_{\mathbf{s}_{ji}} + \tfrac{\tau^2 \lambda_n}{km} \mathbf{I} \right) \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j), \varphi_g \right\rangle_{\mathbb{H}} \\ &= -\frac{\tau^2 \lambda_n}{km} \langle w_0, \varphi_g \rangle_{\mathbb{H}} = -\frac{\tau^2 \lambda_n}{km} \frac{\theta_g}{\mu_g}, \quad j = 1, \ldots, k. \end{aligned} \tag{16}$$

Expanding the left hand side in (16), we obtain that

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} \langle \varphi_g, \xi_{\mathbf{s}_{ji}} \rangle_{\mathbb{H}} \, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}} \left\{ \Delta_j(\mathbf{s}_{ji}) \right\} + \frac{\tau^2 \lambda_n}{km} \langle \varphi_g, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j) \rangle_{\mathbb{H}} \\ &= \frac{1}{m} \sum_{i=1}^{m} \varphi_g(\mathbf{s}_{ji}) \, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}} \left\{ \Delta_j(\mathbf{s}_{ji}) \right\} + \frac{\tau^2 \lambda_n}{km} \frac{\delta_{jg}}{\mu_g}. \end{aligned}$$

The term $\frac{1}{m} \sum_{i=1}^{m} \varphi_g(\mathbf{s}_{ji}) \, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}} \left\{ \Delta_j(\mathbf{s}_{ji}) \right\}$ on the right hand side is

$$\begin{aligned} &= \frac{1}{m} \sum_{i=1}^{m} \Phi_{ig}^j \sum_{h=1}^{d} \delta_{jh} \varphi_h(\mathbf{s}_{ji}) + \frac{1}{m} \sum_{i=1}^{m} \Phi_{ig}^j \sum_{h=d+1}^{\infty} \delta_{jh} \varphi_h(\mathbf{s}_{ji}) \\ &= \frac{1}{m} \sum_{h=1}^{d} \delta_{jh} \sum_{i=1}^{m} \Phi_{ig}^j \, \Phi_{ih}^j + \frac{1}{m} \sum_{i=1}^{m} \Phi_{ig}^j \, v_{ji} \\ &= \frac{1}{m} \sum_{h=1}^{d} \delta_{jh} \left(\Phi^{j^T} \Phi^j \right)_{gh} + \frac{1}{m} \sum_{i=1}^{m} \left(\Phi^{j^T} v_j \right)_g \\ &= \frac{1}{m} \left(\Phi^{j^T} \Phi^j \, \boldsymbol{\delta}^{\downarrow} \right)_g + \frac{1}{m} \left(\Phi^{j^T} \mathbf{v}_j \right)_g. \end{aligned} \tag{17}$$

Substitute (17) in (16) for $g = 1, \ldots, d$ to obtain that

$$\frac{1}{m} \Phi^{j^T} \Phi^j \, \boldsymbol{\delta}_j^{\downarrow} + \frac{1}{m} \Phi^{j^T} \mathbf{v}_j + \frac{\tau^2 \lambda_n}{km} \mathbf{M}^{-1} \, \boldsymbol{\delta}_j^{\downarrow} = -\frac{\tau^2 \lambda_n}{km} \mathbf{M}^{-1} \, \boldsymbol{\theta}^{\downarrow}$$

$$\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j + \frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1}\right)\boldsymbol{\delta}_j^{\downarrow} = -\frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} - \frac{1}{m}\,\Phi^{j^T}\,\mathbf{v}_j\,. \tag{18}$$

The proof is completed by showing that the right hand side expression in (18) gives an upper bound for $\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2$. Define $\mathbf{Q} = \left(\mathbf{I} + \frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1}\right)^{1/2}$, then

$$\begin{aligned}
&\frac{1}{m}\,\Phi^{j^T}\,\Phi^j + \frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1} = \mathbf{I} + \frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1} + \frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I} \\
&= \mathbf{Q}\left\{\mathbf{I} + \mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1}\right\}\mathbf{Q}
\end{aligned}$$

and using this in (18) gives

$$\left\{\mathbf{I} + \mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1}\right\}\mathbf{Q}\,\boldsymbol{\delta}_j^{\downarrow} = -\frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} - \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j\,. \tag{19}$$

Now we define the $\mathbb{P}$-measureable event

$$\mathcal{E}_1 = \left\{\left|\!\left|\!\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1}\right|\!\right|\!\right| \le 1/2\right\}, \tag{20}$$

where $|||\cdot|||$ is the matrix operator norm. We have that $\mathbf{I} + \mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1} \succeq (1/2)\,\mathbf{I}$ whenever $\mathcal{E}_1$ occurs. Furthermore, when $\mathcal{E}_1$ occurs, (19) implies that

$$\begin{aligned}
\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2 &\le \| \mathbf{Q}\,\boldsymbol{\delta}_j^{\downarrow} \|_2^2 \le 4\left\| \frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} + \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j \right\|_2^2 \\
&\le 8\left\| \frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} \right\|_2^2 + 8\left\| \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j \right\|_2^2,
\end{aligned}$$

where the last inequality follows because $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$.

Since $\mathcal{E}_1$ is $\mathbb{P}$-measureable, $\mathbb{E}_{\mathcal{S}}\left(\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2\right) = \mathbb{E}_{\mathcal{S}}\left\{\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2\,\mathbf{1}\left(\mathcal{E}_1\right)\right\} + \mathbb{E}_{\mathcal{S}}\left\{\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2\,\mathbf{1}\left(\mathcal{E}_1^c\right)\right\}$ and the previous display gives

$$\mathbb{E}_{\mathcal{S}}\left\{\| \boldsymbol{\delta}_j^{\downarrow} \|_2^2\,\mathbf{1}\left(\mathcal{E}_1\right)\right\} \le 8\left\| \frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} \right\|_2^2 + 8\,\mathbb{E}_{\mathcal{S}}\left\| \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j \right\|_2^2. \tag{21}$$

From Lemma 10 in Zhang et al. (2015), we have that under our assumptions A.1-A.5, there exists a universal constant $A > 0$ that does not depend on λ_n, n, τ^2, such that

$$\begin{aligned}
&\left\| \frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} \right\|_2^2 \le \frac{\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2, \\
&\mathbb{E}_{\mathcal{S}}\left\| \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j \right\|_2^2 \le \frac{km}{\tau^2\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)\|w_0\|_{\mathbb{H}}^2, \\
&\mathbb{P}\left(\mathcal{E}_1^c\right) \le \left\{ A\max\left(\sqrt{\max(q, \log d)}, \frac{\max(q, \log d)}{m^{1/2-1/q}}\right)\frac{\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}} \right\}^q \\
&\qquad = \left\{ \frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}} \right\}^q. \qquad (22)
\end{aligned}$$

Since $\mu_1 \geq \mu_2 \geq \ldots \geq 0$, the optimality condition in (12) implies that

$$\left\|\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\right\|_2^2 = \frac{\mu_1}{\mu_1}\sum_{i=1}^{\infty}\delta_{ji}\varphi_i \leq \mu_1\sum_{i=1}^{\infty}\frac{\delta_{ji}}{\mu_i}\varphi_i = \mu_1\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_{\mathbb{H}}^2 \leq \mu_1\|w_0\|_{\mathbb{H}}^2. \tag{23}$$

Using the shorthand (22) and (23), we obtain that

$$\mathbb{E}_{\mathcal{S}}\left\{\|\,\boldsymbol{\delta}_j^{\downarrow}\,\|_2^2\,\mathbf{1}\left(\mathcal{E}_1^c\right)\right\} \leq \mathbb{E}_{\mathcal{S}}\left\{\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2\,\mathbf{1}\left(\mathcal{E}_1^c\right)\right\} \leq \mathbb{P}(\mathcal{E}_1^c)\mu_1\|w_0\|_{\mathbb{H}}^2. \tag{24}$$

Combining (21) and (24) gives

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}(\|\,\boldsymbol{\delta}_j\,\|_2^2) \leq& \frac{8\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2 + \frac{8km}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\|w_0\|_{\mathbb{H}}^2\\ &+\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}}\right\}^q\mu_1\|w_0\|_{\mathbb{H}}^2.\end{aligned} \tag{25}$$

Finally, we use that $\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta)\|_2^2 \leq k^{-1}\sum_{j=1}^k\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta_j)\|_2^2 = k^{-1}\sum_{j=1}^k\|\,\boldsymbol{\delta}_j\,\|_2^2$ to obtain that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}(\|\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}(\Delta)\|_2^2) &\leq \frac{8\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2 + \frac{8km}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\|w_0\|_{\mathbb{H}}^2\\ &\quad+\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}}\right\}^q\mu_1\|w_0\|_{\mathbb{H}}^2\\ &= \frac{8\tau^2\lambda_n}{n}\|w_0\|_{\mathbb{H}}^2 + \|w_0\|_{\mathbb{H}}^2\left[\frac{8n}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d) + \mu_1\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q\right],\end{aligned} \tag{26}$$

where we have replaced km by n in the last equality. Taking the infimum over $d \in \mathbb{N}$ leads to the proof. ∎

1.1.2 An upper bound for the first variance term The following lemma provides an upper bound the first part of the variance term in (6).

Lemma 1.2 *If Assumptions A.1–A.5 in the main paper hold, then*

$$\begin{aligned}&\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\} \leq\\ &\left(\frac{2n}{k\lambda_n}+\frac{4\|w_0\|_{\mathbb{H}}^2}{k}\right)\inf_{d\in\mathbb{N}}\left[\mu_{d+1}+12\frac{n}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\\ &\left.+\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q\right]+\frac{12\tau^2\lambda_n}{kn}\|w_0\|_{\mathbb{H}}^2+12\frac{\tau^2\lambda_n}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right).\end{aligned}$$

Proof Continuing from (8), we start by finding an upper bound for $\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)\,|\,\mathcal{S}}\,\|\Delta_j\|_{\mathbb{H}}^2$, which is required later to upper bound $\mathbb{E}_0\,\|\Delta_j\|_{\mathbb{H}}^2$. From (8) we have

$$\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)\,|\,\mathcal{S}}\,\|\Delta_j\|_{\mathbb{H}}^2 \leq 2\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)\,|\,\mathcal{S}}\,\|\tilde{w}_j\|_{\mathbb{H}}^2 + 2\|w_0\|_{\mathbb{H}}^2. \tag{27}$$

An upper bound for $\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\tilde{w}_j\|_{\mathbb{H}}^2$ gives the desired bound. Using the objective in (9),

$$\begin{aligned}\frac{1}{2}\|\tilde{w}_j\|_{\mathbb{H}}^2 &\overset{(i)}{\leq} \sum_{i=1}^{m}\frac{\{\tilde{w}_j(\mathbf{s}_{ji})-y(\mathbf{s}_{ji})\}^2}{2\tau^2\lambda_n/k}+\frac{1}{2}\|\tilde{w}_j\|_{\mathbb{H}}^2\\ &\overset{(ii)}{\leq} \sum_{i=1}^{m}\frac{\{w_0(\mathbf{s}_{ji})-y(\mathbf{s}_{ji})\}^2}{2\tau^2\lambda_n/k}+\frac{1}{2}\|w_0\|_{\mathbb{H}}^2,\end{aligned} \tag{28}$$

where (i) follows because the term inside the summation is non-negative and (ii) follows because $\tilde{w}_j$ minimizes the objective. Since $w(\mathbf{s}_{ji})-y(\mathbf{s}_{ji})=-\epsilon(\mathbf{s}_{ji})$ and $\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\epsilon^2(\mathbf{s}_{ji})\}\leq\tau^2$ by Assumption A.2, (28) reduces to

$$\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\tilde{w}_j\|_{\mathbb{H}}^2 \leq \frac{k}{\tau^2\lambda_n}\sum_{i=1}^{m}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\{\epsilon(\mathbf{s}_{ji})\}^2+\|w_0\|_{\mathbb{H}}^2 \leq \frac{km}{\lambda_n}+\|w_0\|_{\mathbb{H}}^2. \tag{29}$$

Substituting (29) in (27) gives

$$\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\Delta_j\|_{\mathbb{H}}^2 \leq \frac{2km}{\lambda_n}+4\|w_0\|_{\mathbb{H}}^2. \tag{30}$$

First notice that

$$\begin{aligned}&\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\}\\ &=\frac{1}{k^2}\sum_{j=1}^{k}\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_{j*}^T\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\,\mathbf{I}\right)^{-2}\mathbf{c}_{j*}\right\}.\end{aligned} \tag{31}$$

and from (6) we have

$$\begin{aligned}&\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_{j*}^T\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\,\mathbf{I}\right)^{-2}\mathbf{c}_{j*}\right\}\\ &=\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathrm{var}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\mathbf{c}_{j*}^T\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\,\mathbf{I}\right)^{-1}\mathbf{y}_j\right\}\\ &\leq\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\mathbf{c}_{j*}^T\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\,\mathbf{I}\right)^{-1}\mathbf{y}_j-w_0(\mathbf{s}^*)\right\}^2\\ &=\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\Delta_j\|_2^2.\end{aligned} \tag{32}$$

Substituting (32) to (31) leads to

$$\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\}\leq\mathbb{E}_{\mathbf{s}^*}\left\{\frac{1}{k^2}\sum_{j=1}^{k}\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\Delta_j\|_2^2\right\}. \tag{33}$$

We then find an upper bound for $\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\Delta_j\|_2^2$ by following similar steps to the proof of Lemma 1.1. Let $\boldsymbol{\delta}_j\in L_2(\mathbb{N})$ be the expansion of Δ_j in the basis $\{\varphi_i\}_{i=1}^{\infty}$, so that $\Delta_j=\sum_{i=1}^{\infty}\delta_{ji}\varphi_i$ (the $\boldsymbol{\delta}_j$ sequence here is different from the one in the previous section). Similar to Section 1.1.1, choose a fixed $d\in\mathbb{N}$ and truncate Δ_j by defining $\Delta_j^{\downarrow}$, $\Delta_j^{\uparrow}$, $\boldsymbol{\delta}_j^{\downarrow}$, and $\boldsymbol{\delta}_j^{\uparrow}$ as

$$\Delta_j^{\downarrow}=\sum_{i=1}^{d}\delta_{ji}\varphi_i,\quad \Delta_j^{\uparrow}=\sum_{i=d+1}^{\infty}\delta_{ji}\varphi_i=\Delta_j-\Delta_j^{\downarrow},$$

$$\boldsymbol{\delta}_j^{\downarrow} = (\delta_{j1}, \ldots, \delta_{jd}), \quad \boldsymbol{\delta}_j^{\uparrow} = (\delta_{j(d+1)}, \ldots, \delta_{j\infty}).$$

The orthonormality of $\{\varphi_i\}_{i=1}^{\infty}$ implies that

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j\|_2^2 &= \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j^{\downarrow}\|_2^2 + \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{0|\,\mathcal{S}}\, \|\Delta_j^{\uparrow}\|_2^2 \\ &= \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\,\boldsymbol{\delta}_j^{\downarrow}\,\|_2^2 + \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\,\boldsymbol{\delta}_j^{\uparrow}\,\|_2^2. \end{aligned} \tag{34}$$

First, the upper bound for $\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\,\boldsymbol{\delta}_j^{\uparrow}\,\|_2^2$ follows from (14),

$$\begin{aligned} \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j^{\uparrow}\|_2^2 &= \sum_{i=d+1}^{\infty} \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}(\delta_{ji}^2) \\ &= \mu_{d+1} \sum_{i=d+1}^{\infty} \frac{\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}(\delta_{ji}^2)}{\mu_{d+1}} \leq \mu_{d+1} \sum_{i=d+1}^{\infty} \frac{\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}(\delta_{ji}^2)}{\mu_i} \\ &= \mu_{d+1}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j^{\uparrow}\|_{\mathbb{H}}^2 \leq \mu_{d+1}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j\|_{\mathbb{H}}^2, \end{aligned}$$

and using (30),

$$\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j^{\uparrow}\|_2^2 \leq \mu_{d+1} \left(\frac{2km}{\lambda_n} + 4\|w_0\|_{\mathbb{H}}^2 \right). \tag{35}$$

We now find an upper bound for $\mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\, \|\Delta_j^{\downarrow}\|_2^2$. Following Section 1.1.1, define the error vector $\mathbf{v}_j = (v_{j1}, \ldots, v_{jm})^T \in \mathbb{R}^m$ with $v_{ji} = \sum_{h=d+1}^{\infty} \delta_{ji}\varphi_h(\mathbf{s}_{ji})$ $(i = 1, \ldots, m)$, and $\mathbf{M} = \mathrm{diag}(\mu_1, \ldots, \mu_d)$. From (9) and (10), $\tilde{w}_j(\cdot)$ in (8) satisfies

$$\frac{1}{m} \sum_{i=1}^{m} \langle \xi_{\mathbf{s}_{ji}}, \tilde{w}_j - w_0 - \epsilon \rangle_{\mathbb{H}}\, \xi_{\mathbf{s}_{ji}} + \frac{\tau^2 \lambda_n}{km} \tilde{w}_j = 0. \tag{36}$$

For any $g \in \{1, \ldots, d\}$, taking the $\mathbb{H}$-inner product with respect φ_g in (36) to obtain that

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} \langle \xi_{\mathbf{s}_{ji}}, \Delta_j - \epsilon \rangle_{\mathbb{H}}\, \langle \xi_{\mathbf{s}_{ji}}, \varphi_g \rangle_{\mathbb{H}} + \frac{\tau^2 \lambda_n}{km} \langle \Delta_j + w_0, \varphi_g \rangle_{\mathbb{H}} = \\ &\frac{1}{m} \sum_{i=1}^{m} \{\Delta_j(\mathbf{s}_{ji}) - \epsilon(\mathbf{s}_{ji})\}\, \varphi_g(\mathbf{s}_{ji}) + \frac{\tau^2 \lambda_n}{km} \frac{\delta_{jg}}{\mu_g} + \frac{\tau^2 \lambda_n}{km} \frac{\theta_g}{\mu_g} = 0, \\ &\frac{1}{m} \sum_{i=1}^{m} \left\{ \sum_{h=1}^{d} \delta_{jh}\varphi_h(\mathbf{s}_{ji}) + \sum_{h=d+1}^{\infty} \delta_{jh}\varphi_h(\mathbf{s}_{ji}) - \epsilon(\mathbf{s}_{ji}) \right\} \varphi_g(\mathbf{s}_{ji}) + \frac{\tau^2 \lambda_n}{km} \frac{\delta_{jg}}{\mu_g} = -\frac{\tau^2 \lambda_n}{km} \frac{\theta_g}{\mu_g}, \\ &\frac{1}{m} \sum_{h=1}^{d} \left\{ \sum_{i=1}^{m} \varphi_h(\mathbf{s}_{ji}) \varphi_g(\mathbf{s}_{ji}) \right\} \delta_{jh} + \frac{1}{m} \sum_{i=1}^{m} \{v_{ji} - \epsilon(\mathbf{s}_{ji})\}\, \varphi_g(\mathbf{s}_{ji}) + \frac{\tau^2 \lambda_n}{km} \frac{\delta_{jg}}{\mu_g} = -\frac{\tau^2 \lambda_n}{km} \frac{\theta_g}{\mu_g}, \\ &\frac{1}{m} \left(\Phi^{j^T} \Phi^j\, \boldsymbol{\delta}_j^{\downarrow} \right)_g + \frac{1}{m} \left\{ \Phi^{j^T} (\mathbf{v}_j - \boldsymbol{\epsilon}_j) \right\}_g + \frac{\tau^2 \lambda_n}{km} (\mathbf{M}^{-1}\, \boldsymbol{\delta}_j^{\downarrow})_g = -\frac{\tau^2 \lambda_n}{km} (\mathbf{M}^{-1}\, \boldsymbol{\theta}^{\downarrow})_g. \end{aligned}$$

Writing this equation in the matrix form yields,

$$\left(\frac{1}{m}\, \Phi^{j^T}\, \Phi^j + \frac{\tau^2 \lambda_n}{km}\, \mathbf{M}^{-1} \right) \boldsymbol{\delta}_j^{\downarrow} = -\frac{\tau^2 \lambda_n}{km}\, \mathbf{M}^{-1}\, \boldsymbol{\theta}^{\downarrow} - \frac{1}{m}\, \Phi^{j^T}\, \mathbf{v}_j + \frac{1}{m}\, \Phi^{j^T}\, \boldsymbol{\epsilon}_j\,. \tag{37}$$

Following Section 1.1.1, by defining $\mathbf{Q} = (\mathbf{I} + \frac{\tau^2\lambda_n}{km}\,\mathbf{M}^{-1})^{1/2}$, (37) reduces to

$$\begin{aligned}&\left\{\mathbf{I} + \mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1}\right\}\mathbf{Q}\,\boldsymbol{\delta}_j^{\downarrow}\\ &= -\frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} - \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j + \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\boldsymbol{\epsilon}_j\,.\end{aligned} \tag{38}$$

On the event $\mathcal{E}_1$ defined as in (20), we have that $\mathbf{I} + \mathbf{Q}^{-1}\left(\frac{1}{m}\,\Phi^{j^T}\,\Phi^j - \mathbf{I}\right)\mathbf{Q}^{-1} \succeq (1/2)\,\mathbf{I}$. Furthermore, when $\mathcal{E}_1$ occurs, (38) implies that

$$\begin{aligned}\|\Delta_j^{\downarrow}\|_2^2 &\le \|\,\mathbf{Q}\,\boldsymbol{\delta}_j^{\downarrow}\,\|_2^2 \le 4\left\|-\frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow} - \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j + \frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\boldsymbol{\epsilon}_j\right\|_2^2\\ &\le 12\left\|\frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow}\right\|_2^2 + 12\left\|\frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j\right\|_2^2 + 12\left\|\frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\boldsymbol{\epsilon}_j\right\|_2^2,\end{aligned}$$

where the last inequality follows because $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$ for any $a, b, c \in \mathbb{R}$. Since $\mathcal{E}_1$ is $\mathbb{P}$-measureable,

$$\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left(\|\Delta_j^{\downarrow}\|_2^2\right) = \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\|\Delta_j^{\downarrow}\|_2^2\,\mathbf{1}\left(\mathcal{E}_1\right)\right\} + \mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\|\Delta_j^{\downarrow}\|_2^2\,\mathbf{1}\left(\mathcal{E}_1^c\right)\right\}.$$

If the event $\mathcal{E}_1$ occurs, then the upper bounds for the first term and the last two terms in the last inequality are given by Lemmas 10 and 7 of Zhang et al. (2015), respectively, and we have that

$$\begin{aligned}&\left\|\frac{\tau^2\lambda_n}{km}\,\mathbf{Q}^{-1}\,\mathbf{M}^{-1}\,\boldsymbol{\theta}^{\downarrow}\right\|_2^2 \le \frac{\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2,\\ &\mathbb{E}_{\mathcal{S}}\left\|\frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\mathbf{v}_j\right\|_2^2 \le \frac{km}{\tau^2\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)\left(\frac{2km}{\lambda_n} + 4\|w_0\|_{\mathbb{H}}^2\right),\\ &\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\|\frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\boldsymbol{\epsilon}_j\right\|_2^2\\ &\le \frac{1}{m^2}\sum_{h=1}^{d}\sum_{i=1}^{m}\frac{1}{1 + \frac{\tau^2\lambda_n}{km}\frac{1}{\mu_h}}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\varphi_h^2(\mathbf{s}_{ji})\epsilon^2(\mathbf{s}_{ji})\right\}.\end{aligned} \tag{39}$$

Since the error $\epsilon(\cdot)$ and $w(\cdot)$ are independent, by Assumption A.4,

$$\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\varphi_h^2(\mathbf{s}_{ji})\epsilon^2(\mathbf{s}_{ji})\right\} = \mathbb{E}_{\mathcal{S}}\left\{\varphi_h^2(\mathbf{s}_{ji})\right\}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\epsilon^2(\mathbf{s}_{ji})\right\} \le \tau^2,$$

and the last inequality in (39) simplifies to

$$\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\|\frac{1}{m}\,\mathbf{Q}^{-1}\,\Phi^{j^T}\,\boldsymbol{\epsilon}_j\right\|_2^2 \le \frac{\tau^2}{m}\sum_{h=1}^{d}\frac{1}{1 + \frac{\tau^2\lambda_n}{km}\frac{1}{\mu_h}} \le \frac{\tau^2}{m}\gamma\left(\frac{\tau^2\lambda_n}{km}\right).$$

Hence when the event $\mathcal{E}_1$ occurs,

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\|\Delta_j^{\downarrow}\|_2^2\,\mathbf{1}(\mathcal{E}_1)\right\} \le\\ &12\frac{\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2 + 12\frac{km}{\tau^2\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)\left(\frac{2km}{\lambda_n} + 4\|w_0\|_{\mathbb{H}}^2\right) + 12\frac{\tau^2}{m}\gamma\left(\frac{\tau^2\lambda_n}{km}\right).\end{aligned} \tag{40}$$

If the event $\mathcal{E}_1$ does not occur, then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\|\Delta_j^{\downarrow}\|_2^2\,\mathbf{1}(\mathcal{E}_1^c)\right\}\\
&\leq \mathbb{E}_{\mathcal{S}}\left\{\mathbf{1}(\mathcal{E}_1^c)\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\,\|\Delta_j^{\downarrow}\|_2^2\right\} \overset{(i)}{\leq} \mathbb{P}(\mathcal{E}_1^c)\left(\frac{2km}{\lambda_n}+4\|w_0\|_{\mathbb{H}}^2\right)\\
&\overset{(ii)}{=} \left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}}\right\}^q\left(\frac{2km}{\lambda_n}+4\|w_0\|_{\mathbb{H}}^2\right), \qquad (41)
\end{aligned}$$

where (i) follows from (30) and (ii) follows from (22). Substituting (40), (41), and (35) in (34) implies that

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\left\{\|\Delta_j\|_2^2\right\} \leq\; & 12\frac{\tau^2\lambda_n}{km}\|w_0\|_{\mathbb{H}}^2+12\frac{\tau^2}{m}\gamma\left(\frac{\tau^2\lambda_n}{km}\right)+\\
& \left[\mu_{d+1}+12\frac{km}{\tau^2\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)+\right.\\
& \left.\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{km})}{\sqrt{m}}\right\}^q\right]\left(\frac{2km}{\lambda_n}+4\|w_0\|_{\mathbb{H}}^2\right). \qquad (42)
\end{aligned}$$

Therefore, substituting (42) in (33) implies that

$$\begin{aligned}
&\tau^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\} \leq\\
&\left(\frac{2n}{k\lambda_n}+\frac{4\|w_0\|_{\mathbb{H}}^2}{k}\right)\left[\mu_{d+1}+12\frac{n}{\tau^2\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)+\right.\\
&\left.\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q\right]+\frac{12\tau^2\lambda_n}{kn}\|w_0\|_{\mathbb{H}}^2+12\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right). \qquad (43)
\end{aligned}$$

where we have replace km by n. Taking the infimum over $d\in\mathbb{N}$ leads to the proof. ■

1.1.3 An upper bound for the second variance term The following lemma provides an upper bound the second part of the variance term in (6).

Lemma 1.3 *If Assumptions A.1–A.5 in the main paper hold, then*

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\overline{v}(\mathbf{s}^*) \leq 3\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right)\\
&+\inf_{d\in\mathbb{N}}\left[\left\{\frac{4n}{\tau^2\lambda_n^2}\,\mathrm{tr}(C_{\boldsymbol{\alpha}})+\frac{1}{\lambda_n}\right\}\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)+\lambda_n^{-1}\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2}{n})}{\sqrt{m}}\right\}^q\right].
\end{aligned}$$

Proof First we have the following relation between $\overline{v}$ and the subset variance v_j:

$$\overline{v}(\mathbf{s}^*)=\left\{k^{-1}\sum_{j=1}^k v_j^{1/2}(\mathbf{s}^*)\right\}^2 \leq \frac{1}{k}\sum_{j=1}^k v_j(\mathbf{s}^*)$$

$$= \frac{1}{k}\sum_{j=1}^{k}\lambda_n^{-1}\left\{C_{\boldsymbol{\alpha}}(\mathbf{s}^*,\mathbf{s}^*) - \mathbf{c}_j^T(\mathbf{s}^*)\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\mathbf{c}_j(\mathbf{s}^*)\right\}. \tag{44}$$

Since $C_{\boldsymbol{\alpha}}(\mathbf{s},\mathbf{s}') = \sum_{i=1}^{\infty}\mu_i\varphi_i(\mathbf{s})\varphi_i(\mathbf{s}')$ for $\mathbf{s},\mathbf{s}'\in\mathcal{D}$, we have

$$C_{\boldsymbol{\alpha}}(\mathbf{s}^*,\mathbf{s}^*) = \sum_{a=1}^{\infty}\mu_a\varphi_a^2(\mathbf{s}^*), \quad \{\mathbf{c}_j(\mathbf{s}^*)\}_i = \sum_{a=1}^{\infty}\mu_a\varphi_a(\mathbf{s}_{ji})\varphi_a(\mathbf{s}^*), \quad i=1,\ldots,m.$$

These together with the orthogonality property of $\{\varphi_i\}_{i=1}^{\infty}$ imply that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\{v_j(\mathbf{s}^*)\} = \lambda_n^{-1}\sum_{a=1}^{\infty}\mu_a\,\mathbb{E}_{\mathbf{s}^*}\varphi_a^2(\mathbf{s}^*)\\
&-\lambda_n^{-1}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\mu_a\mu_b\left\{\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\right\}_{i'i''}\\
&\times\mathbb{E}_{\mathcal{S}}\left[\varphi_a(\mathbf{s}_{ji})\varphi_b(\mathbf{s}_{ji'})\,\mathbb{E}_{\mathbf{s}^*}\{\varphi_a(\mathbf{s}^*)\varphi_b(\mathbf{s}^*)\}\right]\\
&=\lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}}) - \lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{a=1}^{\infty}\mu_a^2\left\{\left(\mathbf{C}_{j,j}+\tfrac{\tau^2}{k}\mathbf{I}\right)^{-1}\right\}_{ii'}\varphi_a(\mathbf{s}_{ji})\varphi_a(\mathbf{s}_{ji'})\\
&=\lambda_n^{-1}\sum_{a=1}^{d}\mu_a - \lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\sum_{a=1}^{d}\mu_a^2\left[\sum_{i=1}^{m}\sum_{i'=1}^{m}\left\{\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\right\}_{ii'}\varphi_a(\mathbf{s}_{ji})\varphi_a(\mathbf{s}_{ji'})\right] +\\
&\lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d) - \lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\sum_{a=d+1}^{\infty}\mu_a^2\left[\sum_{i=1}^{m}\sum_{i'=1}^{m}\left\{\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\right\}_{i'i''}\varphi_a(\mathbf{s}_{ji})\varphi_a(\mathbf{s}_{ji'})\right]
\end{aligned}$$

$$\overset{(i)}{\leq}\lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\sum_{a=1}^{d}\left\{\mu_a - \mu_a^2\,\boldsymbol{\varphi}_a^{j^T}(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I})^{-1}\boldsymbol{\varphi}_a^j\right\} + \lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d), \tag{45}$$

where iath element of the matrix Φ^j (defined in the proof of Lemma 1.1) is $\varphi_a(\mathbf{s}_{ji})$, $\boldsymbol{\varphi}_a^j$ is the ath column of Φ^j, and (i) follows because $\left(\mathbf{C}_{j,j}+\frac{\tau^2\lambda_n}{k}\mathbf{I}\right)$ is a positive definite matrix and $\boldsymbol{\varphi}_a^{j^T}\left(\mathbf{C}_{j,j}+\frac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\boldsymbol{\varphi}_a^j \geq 0$.

Let $\mathbf{M} = \operatorname{diag}(\mu_1,\ldots,\mu_d)$ and $\mathbf{Q} = \left(\mathbf{I}+\frac{\tau^2\lambda_n}{km}\mathbf{M}^{-1}\right)^{1/2}$ as defined in the proofs of Lemmas 1.1 and 1.2. Define a $d\times d$ matrix $\mathbf{B} \equiv \mathbf{M} - \mathbf{M}\,\Phi^{j^T}\left(\mathbf{C}_{j,j}+\frac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\Phi^j\,\mathbf{M}$, so that from (45),

$$\begin{aligned}
&\operatorname{tr}(\mathbf{B}) = \sum_{a=1}^{d}\left\{\mu_a - \mu_a^2\,\boldsymbol{\varphi}_a^{j^T}\left(\mathbf{C}_{j,j}+\tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\boldsymbol{\varphi}_a^j\right\},\\
&\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\{v_j(\mathbf{s}^*)\} \leq \lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\operatorname{tr}(\mathbf{B}) + \lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d).
\end{aligned} \tag{46}$$

Let

$$\begin{aligned}
&\mathbf{C}_{j,j} = \Phi^j\,\mathbf{M}\,\Phi^{j^T} + \Phi^{j\uparrow}\,\mathbf{M}^{\uparrow}\,\Phi^{j\uparrow^T} \equiv \Phi^j\,\mathbf{M}\,\Phi^{j^T} + \mathbf{C}_{j,j}^{\uparrow},\\
&\mathbf{M}^{\uparrow} = \operatorname{diag}(\mu_{d+1},\ldots,\mu_{\infty}), \quad \Phi^{j\uparrow} = [\boldsymbol{\varphi}_{d+1}^j,\cdots,\boldsymbol{\varphi}_{\infty}^j],
\end{aligned}$$

then the Woodbury formula (Harville, 1997) and the definition of $\mathbf{Q}$ imply that

$$\begin{aligned}\mathbf{B} &= \left\{\mathbf{M}^{-1} + \Phi^{j^T}\left(\mathbf{C}^{\uparrow}_{j,j} + \tfrac{\tau^2\lambda_n}{k}\mathbf{I}\right)^{-1}\Phi^j\right\}^{-1} \\ &= \frac{\tau^2\lambda_n}{km}\left\{\mathbf{I} + \frac{\tau^2\lambda_n}{km}\mathbf{M}^{-1} + \frac{1}{m}\Phi^{j^T}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j} + \mathbf{I}\right)^{-1}\Phi^j - \mathbf{I}\right\}^{-1} \\ &= \frac{\tau^2\lambda_n}{km}\mathbf{Q}^{-2}\left[\mathbf{I} + \mathbf{Q}^{-1}\left\{\frac{1}{m}\Phi^{j^T}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j} + \mathbf{I}\right)^{-1}\Phi^j - \mathbf{I}\right\}\mathbf{Q}^{-1}\right]^{-1}. \qquad (47)\end{aligned}$$

Define the event $\mathcal{E}_2 = \left\{\frac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j} \preceq \frac{1}{4}\mathbf{I}\right\}$. Since the matrix $\mathbf{C}^{\uparrow}_{j,j}$ is nonnegative definite, we have the relation that

$$\left\{\mathrm{tr}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j}\right) \le \frac{1}{4}\right\} \subseteq \left\{\mathrm{s}_{\max}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j}\right) \le \frac{1}{4}\right\} \subseteq \mathcal{E}_2,$$

$\mathrm{s}_{\max}(\mathbf{A})$ is the maximum eigenvalue of the square matrix $\mathbf{A}$. Therefore, by Markov's inequality, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2^c) &\le \mathbb{P}\left\{\mathrm{tr}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j}\right) > \frac{1}{4}\right\} \le 4\,\mathbb{E}_{\mathcal{S}}\,\mathrm{tr}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j}\right) \\ &= \frac{4k}{\tau^2\lambda_n}\sum_{i=1}^{m}\sum_{a=d+1}^{\infty}\mu_a\,\mathbb{E}_{\mathcal{S}}\,\varphi_a^2(\mathbf{s}_{ji}) = \frac{4km}{\tau^2\lambda_n}\,\mathrm{tr}\left(C^d_{\boldsymbol{\alpha}}\right). \qquad (48)\end{aligned}$$

Now on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ (with $\mathcal{E}_1$ defined in (20)), we have that

$$\begin{aligned}&\mathbf{I} + \mathbf{Q}^{-1}\left\{\frac{1}{m}\Phi^{j^T}\left(\tfrac{k}{\tau^2\lambda_n}\mathbf{C}^{\uparrow}_{j,j} + \mathbf{I}\right)^{-1}\Phi^j - \mathbf{I}\right\}\mathbf{Q}^{-1} \\ &\overset{(i)}{\succeq} \mathbf{I} + \mathbf{Q}^{-1}\left\{\frac{1}{m}\Phi^{j^T}\left(\frac{1}{4}\mathbf{I} + \mathbf{I}\right)^{-1}\Phi^j - \mathbf{I}\right\}\mathbf{Q}^{-1} \\ &= \mathbf{I} - \frac{1}{5}\mathbf{Q}^{-2} + \frac{4}{5}\mathbf{Q}^{-1}\left\{\frac{1}{m}\Phi^{j^T}\Phi^j - \mathbf{I}\right\}\mathbf{Q}^{-1} \\ &\overset{(ii)}{\succeq} \mathbf{I} - \frac{1}{5}\mathbf{I} - \frac{4}{5}\cdot\frac{1}{2}\mathbf{I} = \frac{2}{5}\mathbf{I}, \qquad (49)\end{aligned}$$

where (i) follows on the event $\mathcal{E}_2$, and (ii) holds on the event $\mathcal{E}_1$ and from the fact $\mathbf{Q}^{-2} \preceq \mathbf{I}$.

Therefore, by combining (48), (49), and the upper bound for $\mathbb{P}(\mathcal{E}_1^c)$ given in (22) under our assumptions, we obtain that

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\,\mathrm{tr}(\mathbf{B}) \\ &\le \mathbb{E}_{\mathcal{S}}\left\{\mathrm{tr}(\mathbf{B})\,\mathbf{1}(\mathcal{E}_1\cap\mathcal{E}_2)\right\} + \mathbb{E}_{\mathcal{S}}\left[\mathrm{tr}(\mathbf{B})\left\{\mathbf{1}(\mathcal{E}_1^c) + \mathbf{1}(\mathcal{E}_2^c)\right\}\right] \\ &\overset{(i)}{\le} \frac{5}{2}\frac{\tau^2\lambda_n}{km}\,\mathrm{tr}\left(\mathbf{Q}^{-2}\right) + \mathrm{tr}(C_{\boldsymbol{\alpha}})\left\{\mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c)\right\} \\ &\overset{(ii)}{\le} 3\frac{\tau^2\lambda_n}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right) + \frac{4n}{\tau^2\lambda_n}\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C^d_{\boldsymbol{\alpha}}) \\ &\quad + \mathrm{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q, \qquad (50)\end{aligned}$$

where (i) follows from (49), and (ii) follows from (48), (22), and by replacing km with n.

(45), (47), and (50) together yield

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\overline{v}_j(\mathbf{s}^*)\right\} \\
&\le \lambda_n^{-1}\,\mathbb{E}_{\mathcal{S}}\,\mathrm{tr}(\mathbf{B}) + \lambda_n^{-1}\,\mathrm{tr}\left(C_{\boldsymbol{\alpha}}^d\right) \\
&\le 3\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right) + \left\{\frac{4n}{\tau^2\lambda_n^2}\,\mathrm{tr}(C_{\boldsymbol{\alpha}}) + \frac{1}{\lambda_n}\right\}\mathrm{tr}(C_{\boldsymbol{\alpha}}^d) \\
&\quad + \lambda_n^{-1}\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2}{n})}{\sqrt{m}}\right\}^q. \qquad (51)
\end{aligned}$$

Since the righthand side of (51) does not depend on j, a further upper bound for (44) is given by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\overline{v}(\mathbf{s}^*)\right\} \le \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\overline{v}_j(\mathbf{s}^*)\right\} \\
&\le 3\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right) + \left\{\frac{4n}{\tau^2\lambda_n^2}\,\mathrm{tr}(C_{\boldsymbol{\alpha}}) + \frac{1}{\lambda_n}\right\}\mathrm{tr}(C_{\boldsymbol{\alpha}}^d) \\
&\quad + \lambda_n^{-1}\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2}{n})}{\sqrt{m}}\right\}^q. \qquad (52)
\end{aligned}$$

Taking the infimum over $d \in \mathbb{N}$ leads to the proof. ■

1.2 Proof of Theorem 3.2

The proof of parts (i)–(iv) are as follows.

(i) Since d^* is a constant integer and $k = o(n)$, we can take m sufficiently large such that $n \ge m > \max(d^*, e^q)$. In the upper bounds of Theorem 3.1, we choose $d = n$ in every infimum to make the upper bounds larger. This implies that $\mathrm{tr}\left(C_{\boldsymbol{\alpha}}^d\right) = 0$, $\mu_{d+1} = 0$, and $b(m,d,q) \le \log n$. Also notice that in this case, $\gamma(a) \le d^*$ for any $a > 0$. Then, with $\lambda_n = 1$, Theorem 3.1 implies that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 \\
&\le \left(8\|w_0\|_{\mathbb{H}}^2 + 12k^{-1}\|w_0\|_{\mathbb{H}}^2 + 15d^*\right)\frac{\tau^2}{n} \\
&\quad + \left\{\mu_1\|w_0\|_{\mathbb{H}}^2 + \frac{2n}{k} + \frac{4\|w_0\|_{\mathbb{H}}^2}{k} + \mathrm{tr}(C_{\boldsymbol{\alpha}})\right\}\left(\frac{A\rho^2 d^*\log n}{\sqrt{n/k}}\right)^q \\
&\le O(n^{-1}) + \{1 + o(1)\}\frac{2\left(A\rho^2 d^*\log n\right)^q k^{r/2-1}}{n^{r/2-1}} \\
&= O(n^{-1}),
\end{aligned}$$

where the last equality follows from the condition on k.

(ii) In the upper bounds of Theorem 3.1, we choose $d = n^2$ in every infimum for sufficiently large n such that $\log d = 2\log n > q$. Then

$$\mu_{d+1} \le c_{1\mu}\exp\left(-c_{2\mu}n^{2\kappa}\right) = O(n^{-4}),$$

$$b(m,d,q) \le \max\left(\sqrt{\log d}, \frac{\log d}{m^{1/2-1/q}}\right) \le \log d \le 2\log n,$$

$$\begin{aligned}\operatorname{tr}\left(C_{\boldsymbol{\alpha}}^d\right) &= \sum_{i=n^2+1}^{\infty} \mu_i \le \sum_{i=n^2+1}^{\infty} c_{1\mu}\exp\left(-c_{2\mu}i^{\kappa}\right) \le c_{1\mu}\int_{n^2}^{\infty}\exp\left(-c_{2\mu}z^{\kappa}\right)dz \\ &= c_{1\mu}\int_{n^{2\kappa}}^{\infty}\frac{1}{\kappa}t^{\frac{1}{\kappa}-1}\exp\left(-c_{2\mu}t\right)dt, \end{aligned} \tag{53}$$

where in the last step, we use the change of variable $t = z^{\kappa}$. If $\kappa \ge 1$, then since $t \ge n^{2\kappa} \ge 1$, we have $t^{\frac{1}{\kappa}-1} \le 1$. If $0 < \kappa < 1$, then there exists a large $n_0 \in \mathbb{N}$ that depends on only $c_{2\mu}$ and κ, such that for all $n \ge n_0$ and $t \ge n^{2\kappa}$, we have $t^{\frac{1}{\kappa}-1} \le \exp(c_{2\mu}t/2)$. Therefore, in all cases,

$$\operatorname{tr}\left(C_{\boldsymbol{\alpha}}^d\right) \le \frac{c_{1\mu}}{\kappa}\int_{n^{2\kappa}}^{\infty}\exp\left(-c_{2\mu}t/2\right)dt = \frac{2c_{1\mu}}{c_{2\mu}\kappa}\exp\left(-c_{2\mu}n^{2\kappa}/2\right) = O(n^{-4}). \tag{54}$$

Let $d_1 = \left(\frac{2}{c_{2\mu}}\log n\right)^{1/\kappa}$. For sufficiently large n, with $\lambda_n \equiv 1$, $\gamma(\tau^2\lambda_n/n)$ can be bounded as

$$\begin{aligned}\gamma(\tau^2\lambda_n/n) = \gamma(\tau^2/n) &= \sum_{i=1}^{\infty}\frac{\mu_i}{\mu_i+\frac{\tau^2}{n}} = \sum_{i=1}^{\lfloor d_1\rfloor+1}\frac{\mu_i}{\mu_i+\frac{\tau^2}{n}} + \sum_{i=\lfloor d_1\rfloor+2}^{\infty}\frac{\mu_i}{\mu_i+\frac{\tau^2}{n}} \\ &\le d_1 + 1 + \frac{n}{\tau^2}\sum_{i=\lfloor d_1\rfloor+1}^{\infty} c_{1\mu}\exp\left(-c_{2\mu}i^{\kappa}\right) \\ &\le d_1 + 1 + \frac{n}{\tau^2}\int_{d_1}^{\infty} c_{1\mu}\exp\left(-c_{2\mu}z^{\kappa}\right)dz \\ &= d_1 + 1 + \frac{nc_{1\mu}}{\tau^2\kappa}\int_{d_1^{\kappa}}^{\infty} t^{\frac{1}{\kappa}-1}\exp\left(-c_{2\mu}t\right)dt \\ &\le d_1 + 1 + \frac{nc_{1\mu}}{\tau^2\kappa}\int_{d_1^{\kappa}}^{\infty}\exp\left(-c_{2\mu}t/2\right)dt \\ &= d_1 + 1 + \frac{nc_{1\mu}}{c_{2\mu}\tau^2\kappa}\exp\left(-c_{2\mu}d_1^{\kappa}/2\right) \\ &= \left(\frac{2}{c_{2\mu}}\log n\right)^{1/\kappa} + 1 + \frac{c_{1\mu}}{c_{2\mu}\tau^2\kappa} = O\left((\log n)^{1/\kappa}\right). \end{aligned} \tag{55}$$

Therefore, from (53), (54), (55), and the bounds in Theorem 3.1, we obtain that

$$\begin{aligned}&\mathbb{E}_{\mathbf{s}^*}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\mathcal{S}}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 \\ &\le O(n^{-1}) + 15\frac{\tau^2}{n}\gamma\left(\frac{\tau^2}{n}\right) + \{1+o(1)\}\frac{2n}{k}\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2}{n})}{\sqrt{m}}\right\}^q \\ &\le O(n^{-1}) + O\left((\log n)^{1/\kappa}/n\right) + O(1)\cdot\frac{n}{k}\left\{\frac{(\log n)^{1/\kappa}\cdot\log n}{\sqrt{n/k}}\right\}^q \\ &\le O\left((\log n)^{1/\kappa}/n\right) + O(1)\cdot\frac{k^{\frac{q}{2}-1}(\log n)^{\frac{q(1+\kappa)}{\kappa}}}{n^{\frac{q}{2}-1}} \\ &= O\left((\log n)^{1/\kappa}/n\right), \end{aligned}$$

where the last equality follows from the condition on k.

(iii) Let $\lambda_n = 1$. In the upper bounds of Theorem 3.1, we choose $d = \lfloor n^{3/(2\eta-1)} \rfloor$ in every infimum for sufficiently large n such that $\log d \geq \log\left(n^{\frac{3}{2\eta-1}} - 1\right) > q$. Then

$$\begin{aligned}
\mu_{d+1} &\leq c_\mu n^{-6\eta/(2\eta-1)} \leq c_\mu n^{-3}, \\
\operatorname{tr}\left(C_{\boldsymbol{\alpha}}^d\right) &= \sum_{i=d+1}^{\infty} \mu_i \leq \sum_{i=d+1}^{\infty} c_\mu i^{-2\eta} \leq c_\mu \int_d^\infty \frac{1}{z^{2\eta}} dz \\
&= \frac{c_\mu}{2\eta-1} d^{-(2\eta-1)} \leq \frac{c_\mu}{2\eta-1} n^{-3}, \\
b(m,d,q) &\leq \max\left(\sqrt{\log d}, \frac{\log d}{m^{1/2-1/q}}\right) \leq \log d \leq \frac{3}{2\eta-1} \log n. \qquad (56)
\end{aligned}$$

$\gamma(\tau^2\lambda_n/n) = \gamma(\tau^2/n)$ can be bounded as

$$\begin{aligned}
\gamma(\tau^2/n) &= \sum_{i=1}^{\infty} \frac{1}{1 + \frac{\tau^2}{n\mu_i}} \leq \sum_{i=1}^{\infty} \frac{1}{1 + \frac{\tau^2 i^{2\eta}}{c_\mu n}} \\
&\leq n^{1/(2\eta)} + 1 + \frac{c_\mu n}{\tau^2} \sum_{i=\lfloor n^{1/(2\eta)} \rfloor + 2}^{\infty} \frac{1}{i^{2\eta}} \\
&\leq n^{1/(2\eta)} + 1 + \frac{c_\mu n}{\tau^2} \int_{n^{1/(2\eta)}}^{\infty} \frac{1}{z^{2\eta}} dz \\
&= n^{1/(2\eta)} + 1 + \frac{c_\mu n}{\tau^2(2\eta-1) n^{(2\eta-1)/(2\eta)}} \leq \left(\frac{c_\mu}{\tau^2(2\eta-1)} + 1\right) n^{1/(2\eta)}. \qquad (57)
\end{aligned}$$

From (56), (57), and the bounds in Theorem 3.1, we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*} \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{y}, \overline{w}(\mathbf{s}^*)|\mathcal{S}} \{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 \\
&\leq O(n^{-1}) + 15\frac{\tau^2}{n}\gamma\left(\frac{\tau^2}{n}\right) + \{1 + o(1)\}\frac{2n}{k}\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2}{n})}{\sqrt{m}}\right\}^q \\
&\leq O(n^{-1}) + \frac{15\tau^2\left(2 + \frac{c_\mu}{\tau^2(2\eta-1)}\right) n^{1/(2\eta)}}{n} \\
&\quad + \{1 + o(1)\}\frac{2n}{k}\left\{\frac{3A\rho^2\left(2 + \frac{c_\mu}{\tau^2(2\eta-1)}\right) n^{1/(2\eta)} \log n}{(2\eta-1)\sqrt{n/k}}\right\}^q \\
&\leq O(n^{-1}) + O\left(n^{-\frac{2\eta-1}{2\eta}}\right) + O(1) \cdot \frac{k^{\frac{q}{2}-1}(\log n)^q}{n^{\frac{q}{2}-1-\frac{q}{2\eta}}} \\
&= O\left(n^{-\frac{2\eta-1}{2\eta}}\right),
\end{aligned}$$

where the last equality follows from the condition on k.

(iv) Now let $\lambda_n = c_1 n^{1/(2\eta+1)}$. In the upper bounds of Theorem 3.1, we choose $d = \lfloor n^{3/(2\eta-1)} \rfloor$ in every infimum for sufficiently large n, in the same way as in

Part (iii). Therefore, (56) still holds true. Furthermore, since $\lambda_n = c_1 n^{1/(2\eta+1)}$, we have that

$$\begin{aligned}
\gamma(\tau^2\lambda_n/n) &= \sum_{i=1}^{\infty}\left(1+\frac{\tau^2\lambda_n}{n\mu_i}\right)^{-1} \le \sum_{i=1}^{\infty}\left(1+\frac{\tau^2 c_1 i^{2\eta}}{c_\mu n^{2\eta/(2\eta+1)}}\right)^{-1} \\
&\le n^{1/(2\eta+1)} + 1 + \frac{c_\mu n^{\frac{2\eta}{2\eta+1}}}{\tau^2 c_1} \sum_{i=\lfloor n^{\frac{1}{2\eta+1}}\rfloor+2}^{\infty} \frac{1}{i^{2\eta}} \\
&\le n^{1/(2\eta+1)} + 1 + \frac{c_\mu n^{\frac{2\eta}{2\eta+1}}}{\tau^2 c_1} \int_{n^{\frac{1}{2\eta+1}}}^{\infty} \frac{1}{z^{2\eta}} dz \\
&= n^{1/(2\eta+1)} + 1 + \frac{c_\mu n^{\frac{2\eta}{2\eta+1}}}{\tau^2 c_1 (2\eta-1) n^{(2\eta-1)/(2\eta+1)}} \\
&\le \left(\frac{c_\mu}{\tau^2(2\eta-1)} + 1\right) n^{\frac{1}{2\eta+1}}. \qquad (58)
\end{aligned}$$

From (56), (58), and the bounds in Theorem 3.1, we obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y},\overline{w}(\mathbf{s}^*)|\,\mathcal{S}}\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\}^2 \\
&\le O(\lambda_n/n) + 15\frac{\tau^2}{n}\gamma\left(\frac{\tau^2\lambda_n}{n}\right) + \{1+o(1)\}\frac{2n}{k\lambda_n}\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\tau^2\lambda_n}{n})}{\sqrt{m}}\right\}^q \\
&\le O\left(n^{-\frac{2\eta}{2\eta+1}}\right) + 15\tau^2\left(\frac{c_\mu}{\tau^2(2\eta-1)}+1\right)n^{-\frac{2\eta}{2\eta+1}} \\
&\quad + \{1+o(1)\}\frac{2n^{\frac{2\eta}{2\eta+1}}}{k}\left\{\frac{3A\rho^2\left(\frac{c_\mu}{\tau^2(2\eta-1)}+1\right)n^{\frac{1}{2\eta+1}}\log n}{(2\eta-1)\sqrt{n/k}}\right\}^q \\
&\le O(n^{-1}) + O\left(n^{-\frac{2\eta}{2\eta+1}}\right) + O(1)\cdot\frac{k^{\frac{q}{2}-1}(\log n)^q}{n^{\frac{(2\eta-1)q}{2(2\eta+1)}-\frac{2\eta}{2\eta+1}}} \\
&= O\left(n^{-\frac{2\eta}{2\eta+1}}\right),
\end{aligned}$$

where the last equality follows from the condition on k.

1.3 Extension to Unknown τ^2

In this section, we extend the convergence rates of Bayes L_2-risk in Theorem 3.2 to the case where the covariance function is parameterized in a different way and is scaled by τ^2, such that τ^2 is unknown and assigned a prior distribution. We modify the GP prior on $w(\cdot)$ in Equation (11) of the main text to the following

$$\begin{aligned}
&y(\mathbf{s}_i) = w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \qquad \epsilon(\mathbf{s}_i) \sim N\left(0, \tau^2\right), \\
&w(\cdot) \sim \mathrm{GP}\{0, \lambda_n^{-1}\tau^2 C_{\boldsymbol{\alpha}}(\cdot,\cdot)\}; \qquad (59)
\end{aligned}$$

that is, $C_{\boldsymbol{\alpha}}$ is scaled with τ^2, the same as the error variance. This parameterization has also been used in the application of GP models before. We maintain the same eigen-decomposition of the kernel $C_{\boldsymbol{\alpha}_0}(\cdot,\cdot)$ and the Assumptions A.3 and A.4 as before. We assume that $\boldsymbol{\alpha}$ is still fixed at its truth $\boldsymbol{\alpha}_0$, but now impose a prior on τ^2.

A.5′ (Prior) For each of the k subsets, τ^2 is assigned a prior with a bounded support in $(0, \overline{\tau}^2]$ for some finite constants $\overline{\tau}^2 > 0$.

Let $\mathbb{E}_{\tau^2|\mathbf{y}}$ and $\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\tau^2,\mathbf{y},\mathbf{s}^*}$ be the expectations of $\{\tau_j^2 : j = 1, \ldots, k\}$ given $\mathbf{y}$, and $\overline{w}(\mathbf{s}^*)$ given $\mathbf{y}$, $\{\tau_j^2 : j = 1, \ldots, k\}$, and $\mathbf{s}^*$, respectively, where τ_j^2 is drawn from the posterior of τ^2 given $\mathbf{y}_j$ from the jth subset posterior. Then the Bayes L_2-risk of the DISK posterior for $\overline{w}(\cdot)$ can be written as

$$\mathbb{E}_{\mathbf{s}^*}\, \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\, \mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2,\mathbf{s}^*} \left\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\right\}^2. \tag{60}$$

Then, we have the following corollary when a prior distribution is imposed on τ^2.

Corollary 1.1 *If Assumptions A.1 – A.4 and A.5′ hold, then all the convergence rates in the four cases of Theorem 3.2 still hold true for the Bayes L_2-risk given in (62).*

Proof [Proof of Corollary 1.1] We proceed to prove a similar bound for the Bayes L_2 risk to Theorem 3.1 in the main paper under A.5′. By A.5′, we need to account for the randomness in the posterior of $p(\tau_j^2|\,\mathbf{y}_j)$ across $j = 1, \ldots, k$. Based on the model (59), we can see that conditional on the subset posterior draws of τ_j^2 from the subset posterior $p(\tau^2|\,\mathbf{y}_j)$ for $j = 1, \ldots, k$, the DISK posterior draw $\overline{w}(\mathbf{s}^*)$ follows the distribution $N(\overline{m}, \overline{v})$, with

$$\begin{aligned} \overline{m} &= \frac{1}{k}\sum_{j=1}^{k} \mathbf{c}_{j,*}^T \left\{\mathbf{C}_{j,j} + \tfrac{\lambda_n}{k}\,\mathbf{I}\right\}^{-1} \mathbf{y}_j, \\ \overline{v}^{1/2} &= \frac{1}{k}\sum_{j=1}^{k} v_j^{1/2}, \quad v_j = \frac{\tau_j^2}{\lambda_n}\left\{c_{*,*} - \mathbf{c}_{j,*}^T \left(\mathbf{C}_{j,j} + \tfrac{\lambda_n}{k}\,\mathbf{I}\right)^{-1} \mathbf{c}_{j,*}\right\}, \end{aligned} \tag{61}$$

where $\mathbf{c}_{j,*}, \mathbf{C}_{j,j}, c_{*,*}$ are defined similarly to those in (3) according to the base kernel $C_{\boldsymbol{\alpha}_0}$. Notice that $\overline{m}$ does not depend on τ_j^2 due to the rescaled kernel $\tau^2 C_{\boldsymbol{\alpha}_0}$ in (59).

Let $\mathbb{E}_{\mathbf{s}^*}$, $\mathbb{E}_0$, $\mathbb{E}_{\mathcal{S}}$, $\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}$, and $\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}$, $\mathbb{E}_{\tau^2|\,\mathbf{y}}$ respectively be the expectations with respect to the distributions of $\mathbf{s}^*$, $(\mathcal{S}, \mathbf{y})$, $\mathcal{S}$, $\mathbf{y}$ given $\mathcal{S}$, $\overline{w}(\mathbf{s}^*)$ given $\mathbf{y}$ and $\{\tau_j^2 : j = 1, \ldots, k\}$, and $\{\tau_j^2 : j = 1, \ldots, k\}$ given $\mathbf{y}$. Then based on A.5′, the Bayes L_2-risk of the DISK posterior for $\overline{w}(\cdot)$ can be written as

$$\mathbb{E}_{\mathbf{s}^*}\, \mathbb{E}_{\mathcal{S}}\, \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\, \mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\right\}^2. \tag{62}$$

To upper bound (62), we apply the law of total variance repeatedly to obtain that

$$\begin{aligned} & \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\, \mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\right\}^2 \\ = \; & \left[\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\, \mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*)\right\} - w_0(\mathbf{s}^*)\right]^2 + \mathrm{var}_{\mathbf{y},\tau^2,\overline{w}(\mathbf{s}^*)|\,\mathcal{S}} \left\{\overline{w}(\mathbf{s}^*)\right\} \\ = \; & \left[\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\, \mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*)\right\} - w_0(\mathbf{s}^*)\right]^2 \\ & + \mathrm{var}_{\mathbf{y}\,|\,\mathcal{S}} \left[\mathbb{E}_{\tau^2|\,\mathbf{y}}\, \mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*)\right\}\right] \\ & + \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}} \left(\mathrm{var}_{\tau^2|\,\mathbf{y}} \left[\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*)\right\}\right]\right) \\ & + \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}} \left(\mathbb{E}_{\tau^2|\,\mathbf{y}} \left[\mathrm{var}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2} \left\{\overline{w}(\mathbf{s}^*)\right\}\right]\right). \end{aligned} \tag{63}$$

Using (61), we can derive that

$$
\begin{aligned}
&\left[\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\,\mathbb{E}_{\tau^2|\,\mathbf{y}}\,\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}\left\{\overline{w}(\mathbf{s}^*)\right\} - w_0(\mathbf{s}^*)\right]^2 \\
&= \left[k^{-1}\sum_{j=1}^{k}\mathbf{c}_{j,*}^T\left\{\mathbf{C}_{j,j}+\tfrac{\lambda_n}{k}\,\mathbf{I}\right\}^{-1}\mathbf{w}_{0j} - w_0(\mathbf{s}^*)\right]^2, \\
&= \left\{\mathbf{c}_*^T(k\,\mathbf{L}+\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0 - w_0(\mathbf{s}^*)\right\}^2, \qquad (64)\\
&\quad \mathrm{var}_{\mathbf{y}\,|\,\mathcal{S}}\left[\mathbb{E}_{\tau^2|\,\mathbf{y}}\,\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}\left\{\overline{w}(\mathbf{s}^*)\right\}\right] \\
&= \mathrm{var}_{\mathbf{y}\,|\,\mathcal{S}}\left[k^{-1}\sum_{j=1}^{k}\mathbf{c}_{j,*}^T\left\{\mathbf{C}_{j,j}+\tfrac{\lambda_n}{k}\,\mathbf{I}\right\}^{-1}\mathbf{y}_j\right] \\
&= \tau_0^2\,\mathbf{c}^T(\mathbf{s}^*)(k\,\mathbf{L}+\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}(\mathbf{s}^*), \qquad (65)\\
&\quad \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left(\mathrm{var}_{\tau^2|\,\mathbf{y}}\left[\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}\left\{\overline{w}(\mathbf{s}^*)\right\}\right]\right) \\
&= \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left(\mathrm{var}_{\tau^2|\,\mathbf{y}}\left[\frac{1}{k}\sum_{j=1}^{k}\mathbf{c}_{j,*}^T\left\{\mathbf{C}_{j,j}+\tfrac{\lambda_n}{k}\,\mathbf{I}\right\}^{-1}\mathbf{y}_j\right]\right) = 0, \qquad (66)\\
&\quad \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left(\mathbb{E}_{\tau^2|\,\mathbf{y}}\left[\mathrm{var}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}\left\{\overline{w}(\mathbf{s}^*)\right\}\right]\right) = \mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\left\{\mathbb{E}_{\tau^2|\,\mathbf{y}}(\overline{v})\right\}, \qquad (67)
\end{aligned}
$$

where (64) and (67) follow from (61) and (5), (65) follows similarly to (5), and (66) is zero because $\overline{m}$ does not depend on τ_j^2 $(j = 1, \ldots, k)$. Next, we find upper bound for (64), (65), and (67), respectively.

First, we notice that (64) has the same expression as (5) by setting $\tau^2 = 1$ in (5). Therefore, the proof and the conclusion of Lemma 1.1 still works as before, by setting $\tau^2 = 1$, i.e.

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\{\mathbf{c}_*^T(k\,\mathbf{L}+\lambda_n\,\mathbf{I})^{-1}\,\mathbf{w}_0 - w_0(\mathbf{s}^*)\}^2 \le \frac{8\lambda_n}{n}\|w_0\|_{\mathbb{H}}^2 \\
&\quad + \|w_0\|_{\mathbb{H}}^2 \inf_{d\in\mathbb{N}}\left[\frac{8n}{\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d) + \mu_1\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\lambda_n}{n})}{\sqrt{m}}\right\}^q\right]. \qquad (68)
\end{aligned}
$$

Second, we notice that (65) differs from (5) only with the τ^2 outside replaced by the true error variance τ_0^2, and that $\tau^2 = 1$ in $(k\,\mathbf{L}+\tau^2\lambda_n\,\mathbf{I})^{-2}$. We carefully inspect and modify the proof of Lemma 1.2 to obtain that

$$
\begin{aligned}
&\tau_0^2\,\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{\mathbf{c}_*^T(k\,\mathbf{L}+\lambda_n\,\mathbf{I})^{-2}\,\mathbf{c}_*\right\} \le \\
&\left(\frac{2\tau_0^2 n}{k\lambda_n} + \frac{4\|w_0\|_{\mathbb{H}}^2}{k}\right)\inf_{d\in\mathbb{N}}\left[\mu_{d+1} + 12\frac{n}{\lambda_n}\rho^4\,\mathrm{tr}(C_{\boldsymbol{\alpha}})\,\mathrm{tr}(C_{\boldsymbol{\alpha}}^d)\right. \\
&\left. + \left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\lambda_n}{n})}{\sqrt{m}}\right\}^q\right] + \frac{12\lambda_n}{kn}\|w_0\|_{\mathbb{H}}^2 + 12\frac{\tau_0^2\lambda_n}{n}\gamma\left(\frac{\lambda_n}{n}\right). \qquad (69)
\end{aligned}
$$

Third, $\overline{v}$ (and v_j) in (61) differs from $\overline{v}$ (and v_j) in (3) only in that (61) has a τ_j^2 factor outside and it has $\tau^2 = 1$ inside $\left(\mathbf{C}_{j,j}+\frac{\tau^2\lambda_n}{k}\,\mathbf{I}\right)^{-1}$ in (3). Using the expression of v_j in (61) and the upper bound $\tau_j^2 \le \overline{\tau}^2$ in A.5′, we carefully inspect and modify the proof of Lemma 1.2 to obtain that

$$
\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\,\mathbb{E}_{\tau^2|\,\mathbf{y}}(\overline{v})
$$

$$
\begin{aligned}
&\leq \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\,\mathbb{E}_{\tau^2|\,\mathbf{y}}(v_j)\\
&\leq \frac{\overline{\tau}^2\lambda_n^{-1}}{k}\sum_{j=1}^{k} \mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\left\{c_{*,*} - \mathbf{c}_{j,*}^T\left(\mathbf{C}_{j,j}+\tfrac{\lambda_n}{k}\,\mathbf{I}\right)^{-1}\mathbf{c}_{j,*}\right\}\\
&\leq \overline{\tau}^2\left\{\frac{3}{n}\gamma\left(\frac{\lambda_n}{n}\right) + \inf_{d\in\mathbb{N}}\left[\left\{\frac{4n}{\lambda_n^2}\operatorname{tr}(C_{\boldsymbol{\alpha}}) + \frac{1}{\lambda_n}\right\}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\right.\\
&\qquad\left.\left. + \lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{1}{n})}{\sqrt{m}}\right\}^q\right]\right\}. \qquad (70)
\end{aligned}
$$

Now we can combine (62), (63), (64), (65), (66), (67), (68), (69), and (70) to obtain that

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{s}^*}\,\mathbb{E}_{\mathcal{S}}\,\mathbb{E}_{\mathbf{y}\,|\,\mathcal{S}}\,\mathbb{E}_{\tau^2|\,\mathbf{y}}\,\mathbb{E}_{\overline{w}(\mathbf{s}^*)|\,\mathbf{y},\tau^2}\left\{\overline{w}(\mathbf{s}^*) - w_0(\mathbf{s}^*)\right\}^2\\
&\leq \frac{8\lambda_n}{n}\|w_0\|_{\mathbb{H}}^2 + \|w_0\|_{\mathbb{H}}^2\inf_{d\in\mathbb{N}}\left[\frac{8n}{\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d) + \mu_1\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\lambda_n}{n})}{\sqrt{m}}\right\}^q\right]\\
&\quad + \left(\frac{2\tau_0^2 n}{k\lambda_n} + \frac{4\|w_0\|_{\mathbb{H}}^2}{k}\right)\inf_{d\in\mathbb{N}}\left[\mu_{d+1} + 12\frac{n}{\tau^2\lambda_n}\rho^4\operatorname{tr}(C_{\boldsymbol{\alpha}})\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\\
&\quad\left. + \left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{\lambda_n}{n})}{\sqrt{m}}\right\}^q\right] + \frac{12\lambda_n}{kn}\|w_0\|_{\mathbb{H}}^2 + 12\frac{\tau_0^2\lambda_n}{n}\gamma\left(\frac{\lambda_n}{n}\right)\\
&\quad + \overline{\tau}^2\left\{\frac{3}{n}\gamma\left(\frac{\lambda_n}{n}\right) + \inf_{d\in\mathbb{N}}\left[\left\{\frac{4n}{\lambda_n^2}\operatorname{tr}(C_{\boldsymbol{\alpha}}) + \frac{1}{\lambda_n}\right\}\operatorname{tr}(C_{\boldsymbol{\alpha}}^d)\right.\right.\\
&\qquad\left.\left. + \lambda_n^{-1}\operatorname{tr}(C_{\boldsymbol{\alpha}})\left\{\frac{Ab(m,d,q)\rho^2\gamma(\frac{1}{n})}{\sqrt{m}}\right\}^q\right]\right\}. \qquad (71)
\end{aligned}
$$

We notice that the upper bound in (71) differs from the upper bound in Theorem 3.1 only by some multiplicative constants in each term and inside the $\gamma(\cdot)$ functions. In the previous proof of Theorem 3.2, these constants will only change the multiplicative constants and do not affect the convergence rates of the Bayes L_2-risk of $\overline{w}(\cdot)$. As a result, the convergence rate results of Theorem 3.2 continue to hold for (71) under the various conditions specified in the different cases of Theorem 3.2. This completes the proof. ■

2. SAMPLING FROM THE SUBSET POSTERIOR DISTRIBUTIONS USING A FULL-RANK GP PRIOR

Recall the univariate spatial regression model for the data observed at the ith location in subset j using a GP prior is

$$
y(\mathbf{s}_{ji}) = \mathbf{x}(\mathbf{s}_{ji})^T\boldsymbol{\beta} + w(\mathbf{s}_{ji}) + \epsilon(\mathbf{s}_{ji}), \quad j = 1,\ldots,k, \quad i = 1,\ldots,m_j. \qquad (72)
$$

For the simulations and real data analysis, we assume that $C_{\boldsymbol{\alpha}}(\mathbf{s}_{ji},\mathbf{s}_{ji'}) = \sigma^2\rho(\mathbf{s}_{ji},\mathbf{s}_{ji'};\phi)$ and $D_{\boldsymbol{\alpha}}(\mathbf{s}_{ji},\mathbf{s}_{ji'}) = \mathbf{1}(i=i')\tau^2$, where σ^2,ϕ,τ^2 are positive scalars, $\rho(\cdot,\cdot)$ is a

known positive definite correlation function, and $\mathbf{1}(i = i') = 1$ if $i = i'$ and 0 otherwise. This implies that $\boldsymbol{\alpha} = (\sigma^2, \tau^2, \phi)$. The model in (72) is completed by putting priors on the unknown parameters. The priors distributions on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ have the following forms:

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}), \quad \sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \quad \tau^2 \sim \mathrm{IG}(a_\tau, b_\tau), \quad \phi \sim \mathrm{U}(a_\phi, b_\phi), \tag{73}$$

where $\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}, a_\sigma, b_\sigma, a_\tau, b_\tau, a_\phi$, and b_ϕ are constants, N represents the multivariate Gaussian distribution of appropriate dimension, $\mathrm{IG}(a, b)$ represents the Inverse-Gamma distribution with mean $a/(b+1)$ and variance $b/\{(a-1)^2(a-2)\}$ for $a > 2$, and $\mathrm{U}(a, b)$ represents the uniform distribution on the interval $[a, b]$. The spatial process $w(\cdot)$ is assigned a GP prior as

$$w(\cdot) \mid \sigma^2, \phi \sim \mathrm{GP}\{0, C_{\boldsymbol{\alpha}}(\cdot, \cdot)\}, \quad C_{\boldsymbol{\alpha}}(\cdot, \cdot) = \sigma^2 \rho(\cdot, \cdot; \phi). \tag{74}$$

The training data $\{\mathbf{x}(\mathbf{s}_{j1}), y(\mathbf{s}_{j1})\}, \ldots, \{\mathbf{x}(\mathbf{s}_{jm_j}), y(\mathbf{s}_{jm_j})\}$ are observed at the m_j spatial locations and $\mathcal{S}_j = \{\mathbf{s}_{j1}, \ldots, \mathbf{s}_{jm_j}\}$ contains the locations in subset j.

Consider the setup for predictions and inferences on subset j. Let $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_l^*\}$ be the set of locations such that $\mathcal{S}^* \cap \mathcal{S}_j = \emptyset$. If $\mathbf{w}_j^T = \{w(\mathbf{s}_{j1}), \ldots, w(\mathbf{s}_{jm_j})\}$ and $\boldsymbol{\epsilon}_j^T = \{\epsilon(\mathbf{s}_{j1}), \ldots, \epsilon(\mathbf{s}_{jm_j})\}$, then (72) implies that $\mathbf{w}_j$ apriori follows $N\{\mathbf{0}, \mathbf{C}_{j,j}(\boldsymbol{\alpha})\}$, where $\mathbf{C}_{j,j}(\boldsymbol{\alpha})$ is the block of $\mathbf{C}(\boldsymbol{\alpha})$ that corresponds to the locations in $\mathcal{S}_j$, and $\boldsymbol{\epsilon}_j$ follows $N(\mathbf{0}, \tau^2 \mathbf{I})$, where $\mathbf{I}$ is the identity matrix of appropriate dimension. Given the training data on subset j, our goal is to predict $\mathbf{y}_j^* = \{y(\mathbf{s}_1^*), \ldots, y(\mathbf{s}_l^*)\}$ and to perform posterior inference on $\mathbf{w}_j^* = \{w(\mathbf{s}_1), \ldots, w(\mathbf{s}_l)\}$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\alpha}_j$, where the subscript j denotes that the predictions and inferences condition only on subset j. Standard Markov chain Monte Carlo (MCMC) algorithms exist to achieve this goal (Banerjee et al., 2014), but conditioning only on subset j ignores the information contained in the other $(k-1)$ subsets, resulting in greater posterior uncertainty compared to the full data posterior distribution.

Stochastic approximation is an approach for proper uncertainty quantification that modifies the likelihood used for sampling from the subset posterior distributions for predictions and inferences. The likelihoods for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\mathbf{w}_j$ are raised to the power of k to compensate for the data in the other $(k-1)$ subsets, where we assume that $m_1 = \cdots = m_k = m$ and $k = n/m$. First, consider stochastic approximation for the likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Integrating out $\mathbf{w}_j$ in (72) gives

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \boldsymbol{\eta}_j, \quad \boldsymbol{\eta}_j \sim N\{\mathbf{0}, \mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}\}, \tag{75}$$

where $\mathbf{X}_j = [\mathbf{x}(\mathbf{s}_{j1}) : \cdots : \mathbf{x}(\mathbf{s}_{jm})]^T \in \mathbb{R}^{m \times p}$ is the design matrix for subset j. The likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ given $\mathbf{y}_j$, $\mathbf{X}_j$ after stochastic approximation is

$$\{l_j(\boldsymbol{\beta}, \boldsymbol{\alpha})\}^k = (2\pi)^{-mk/2} |\, \mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}\,|^{-k/2} e^{-\frac{k}{2}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})^T \{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}\}^{-1} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})}. \tag{76}$$

The prior distribution for $\boldsymbol{\beta}$ in (73), the pseudo likelihood in (76), and Bayes rule implies that the density of the jth subset posterior distribution for $\boldsymbol{\beta}$ given the rest is

$$\boldsymbol{\beta} \mid \text{rest} \propto e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})^T [k^{-1} \{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}\}]^{-1} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})} e^{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \Sigma_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})}.$$

This implies that the complete conditional distribution of $\boldsymbol{\beta}_j$ has density $N(\mathbf{m}_{j\boldsymbol{\beta}}, \mathbf{V}_{j\boldsymbol{\beta}})$, where

$$\begin{aligned}\mathbf{V}_{j\boldsymbol{\beta}} &= \left[k\,\mathbf{X}_j^T\{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2\,\mathbf{I}\}^{-1}\,\mathbf{X}_j + \Sigma_{\boldsymbol{\beta}}^{-1}\right]^{-1},\\ \mathbf{m}_{j\boldsymbol{\beta}} &= \mathbf{V}_{j\boldsymbol{\beta}}\left[k\,\mathbf{X}_j^T\left\{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2\,\mathbf{I}\right\}^{-1}\mathbf{y}_j + \Sigma_{\boldsymbol{\beta}}^{-1}\,\boldsymbol{\mu}_{\boldsymbol{\beta}}\right].\end{aligned} \tag{77}$$

If the density of the prior distribution for $\boldsymbol{\alpha}$ is assumed to be $\pi(\sigma^2)\pi(\tau^2)\pi(\phi)$, where the prior densities $\pi(\sigma^2)$, $\pi(\tau^2)$, and $\pi(\phi)$ are defined in (73), then the pseudo likelihood in (76), and Bayes rule implies that the density of the jth subset posterior distribution for $\boldsymbol{\alpha}$ given the rest is

$$\begin{aligned}\boldsymbol{\alpha} \mid \text{rest} \propto\ & |\,\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tau^2\,\mathbf{I}\,|^{-k/2} e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})^T\left[k^{-1}\left\{\mathbf{C}_{j,j}(\boldsymbol{\alpha})+\tau^2\,\mathbf{I}\right\}\right]^{-1}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})}\\ & \left(\sigma^2\right)^{-a_\sigma-1} e^{-b_\sigma/\sigma^2}\left(\tau^2\right)^{-a_\tau-1} e^{-b_\tau/\tau^2}(b_\phi - a_\phi)^{-1}.\end{aligned} \tag{78}$$

This density does not have a standard form, so we use a Metropolis-Hastings step with a normal random walk proposal and sample $\boldsymbol{\alpha}_j$ using the `metrop` function in the R package `mcmc` (R Development Core Team, 2017).

Second, we derive the posterior predictive distribution of $\mathbf{w}_j^*$ given the rest. The GP prior on $(\mathbf{w}_j, \mathbf{w}_j^*)$ implies that the density of $\mathbf{w}_j^*$ given $\mathbf{w}_j$ is

$$\mathbf{w}_j^* \mid \mathbf{w}_j \sim N\left\{\mathbf{C}_{*,j}(\boldsymbol{\alpha})\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{w}_j, \mathbf{C}_{*,*}(\boldsymbol{\alpha}) - \mathbf{C}_{*,j}(\boldsymbol{\alpha})\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{C}_{j,*}(\boldsymbol{\alpha})\right\}, \tag{79}$$

where $\text{cov}(\mathbf{w}_j^*, \mathbf{w}_j^*) = \mathbf{C}_{*,*}(\boldsymbol{\alpha})$, $\text{cov}(\mathbf{w}_j^*, \mathbf{w}_j) = \mathbf{C}_{*,j}(\boldsymbol{\alpha})$, and $\text{cov}(\mathbf{w}_j, \mathbf{w}_j^*) = \mathbf{C}_{j,*}(\boldsymbol{\alpha})$. Given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{y}_j$, and $\mathbf{X}_j$, (72) implies that the likelihood of $\mathbf{w}_j$ after stochastic approximation is

$$\{l_j(\mathbf{w}_j)\}^k = (2\pi)^{-mk/2}|\tau^2\,\mathbf{I}\,|^{-k/2} e^{-\frac{k}{2\tau^2}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{w}_j)^T(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{w}_j)}. \tag{80}$$

The GP prior on $\mathbf{w}_j$, the pseudo likelihood in (80), and Bayes rule implies that the density of the subset posterior distribution for $\mathbf{w}_j$ given the rest is

$$\mathbf{w}_j \mid \text{rest} \propto e^{-\frac{1}{2\tau^2/k}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{w}_j)^T(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta} - \mathbf{w}_j)}\, e^{-\frac{1}{2}\mathbf{w}_j^T\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{w}_j}.$$

This implies that the complete conditional distribution of $\mathbf{w}_j$ has density $N(\mathbf{m}_{\mathbf{w}_j}, \mathbf{V}_{\mathbf{w}_j})$, where

$$\mathbf{V}_{\mathbf{w}_j} = \left\{\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha}) + \tfrac{k}{\tau^2}\,\mathbf{I}\right\}^{-1}, \quad \mathbf{m}_{\mathbf{w}_j} = \frac{k}{\tau^2}\,\mathbf{V}_{\mathbf{w}_j}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}); \tag{81}$$

therefore, (79) and (81) imply that the complete conditional distribution of $\mathbf{w}_j^*$ has density $N(\mathbf{m}_{\mathbf{w}_j^*}, \mathbf{V}_{\mathbf{w}_j^*})$, where

$$\begin{aligned}\mathbf{m}_{\mathbf{w}_j^*} &= \mathbb{E}(\mathbf{w}_j^* \mid \text{rest}) = \mathbf{C}_{*,j}(\boldsymbol{\alpha})\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbb{E}(\mathbf{w}_j \mid \text{rest})\\ &= \mathbf{C}_{*,j}(\boldsymbol{\alpha})\left\{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) + \tfrac{\tau^2}{k}\,\mathbf{I}\right\}^{-1}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})\end{aligned} \tag{82}$$

and

$$\mathbf{V}_{\mathbf{w}_j^*} = \text{var}(\mathbf{w}_j^* \mid \text{rest}) = \mathbb{E}\left\{\text{var}(\mathbf{w}_j^* \mid \mathbf{w}_j) \mid \text{rest}\right\} + \text{var}\left\{\mathbb{E}(\mathbf{w}_j^* \mid \mathbf{w}_j) \mid \text{rest}\right\}$$

(83)
$$= \mathbf{C}_{*,*}(\boldsymbol{\alpha}) - \mathbf{C}_{*,j}(\boldsymbol{\alpha})\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{C}_{j,*}(\boldsymbol{\alpha}) + \mathbf{C}_{*,j}(\boldsymbol{\alpha})\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{V}_{\mathbf{w}_j}\,\mathbf{C}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{C}_{j,*}(\boldsymbol{\alpha}).$$

Finally, we derive the posterior predictive distribution of $\mathbf{y}_j^*$ given the rest. If $\boldsymbol{\beta}_j$, τ_j^2, $\mathbf{w}_j^*$ are the samples from the jth subset posterior distribution of $\boldsymbol{\beta}$, τ^2, and $\mathbf{w}^*$, then (72) implies that $\mathbf{y}_j^*$ given the rest is sampled as

$$\mathbf{y}_j^* = \mathbf{X}_j\,\boldsymbol{\beta}_j + \mathbf{w}_j^* + \boldsymbol{\epsilon}_j^*, \quad \boldsymbol{\epsilon}_j^* \sim N(\mathbf{0}, \tau_j^2\,\mathbf{I});$$

therefore, the complete conditional distribution of $\mathbf{y}_j^*$ is $N(\boldsymbol{\mu}_{\mathbf{y}_j^*}, \mathbf{V}_{\mathbf{y}_j^*})$, where

(84)
$$\boldsymbol{\mu}_{\mathbf{y}_j^*} = \mathbf{X}_j\,\boldsymbol{\beta}_j + \mathbf{w}_j^*, \quad \mathbf{V}_{\mathbf{y}_j^*} = \tau_j^2\,\mathbf{I}\,.$$

All full conditionals except that of $\boldsymbol{\alpha}$ are analytically tractable in terms of standard distributions in subset j ($j = 1, \ldots, k$). The Gibbs sampler with a Metropolis-Hastings step iterates between the following four steps until sufficient number of samples of $\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j, \mathbf{w}_j^*$, and $\mathbf{y}_j^*$ are drawn post convergence to the stationary distribution:

1. Sample $\boldsymbol{\beta}_j$ from $N(\boldsymbol{\mu}_{j\,\boldsymbol{\beta}}, \mathbf{V}_{j\,\boldsymbol{\beta}})$, where $\boldsymbol{\mu}_{j\,\boldsymbol{\beta}}$ and $\mathbf{V}_{j\,\boldsymbol{\beta}}$ are defined in (77).
2. Sample $\boldsymbol{\alpha}_j$ using the Metropolis-Hastings algorithm from the jth subset posterior density (up to constants) of $\boldsymbol{\alpha}_j$ in (78) with a normal random walk proposal.
3. Sample $\mathbf{w}_j^*$ from $N(\boldsymbol{\mu}_{\mathbf{w}_j^*}, \mathbf{V}_{\mathbf{w}_j^*})$, where $\boldsymbol{\mu}_{\mathbf{w}_j^*}$ and $\mathbf{V}_{\mathbf{w}_j^*}$ are defined in (82) and (83).
4. Sample $\mathbf{y}_j^*$ from $N(\boldsymbol{\mu}_{\mathbf{y}_j^*}, \mathbf{V}_{\mathbf{y}_j^*})$, where $\boldsymbol{\mu}_{\mathbf{y}_j^*}$ and $\mathbf{V}_{\mathbf{y}_j^*}$ are defined in (84).

3. SAMPLING FROM THE SUBSET POSTERIOR DISTRIBUTIONS USING A LOW-RANK GP PRIOR

For clarity, we focus on the modified predictive process (MPP) prior as a representative example of low-rank GP prior. The Gibbs sampling algorithm derived in this section is easily extended to other low-rank GP priors. Following the setup in Section 2, we assume that $C_{\boldsymbol{\alpha}}(\mathbf{s}_{ji}, \mathbf{s}_{ji'}) = \sigma^2\rho(\mathbf{s}_{ji}, \mathbf{s}_{ji'}; \phi)$ and $D_{\boldsymbol{\alpha}}(\mathbf{s}_{ji}, \mathbf{s}_{ji'}) = \mathbf{1}(i = i')\tau^2$, $\boldsymbol{\alpha} = (\sigma^2, \tau^2, \phi)$, the prior distributions on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ have the same forms as in (73), and $\mathcal{S}_j$ contains the locations in subset j. Following the previous section, we assume that $m_1 = \cdots = m_k = m$ and $k = n/m$. The only change in this section is that the spatial process $w(\cdot)$ in (72) is assigned a MPP prior derived from parent GP prior in (74). MPP projects the parent GP $w(\cdot)$ onto a subspace spanned by its realization over a set of r locations, $\mathcal{S}^{(0)} = \{\mathbf{s}_1^{(0)}, \ldots, \mathbf{s}_r^{(0)}\}$, known as the "knots", where no conditions are imposed on $\mathcal{S} \cap \mathcal{S}^{(0)}$. Let $\mathbf{c}(\cdot, \mathcal{S}^{(0)}) = \left\{C_{\boldsymbol{\alpha}}(\cdot, \mathbf{s}_1^{(0)}), \ldots, C_{\boldsymbol{\alpha}}(\cdot, \mathbf{s}_r^{(0)})\right\}^T$ and $\mathbf{w}^{(0)} = \left\{w(\mathbf{s}_1^{(0)}), \ldots, w(\mathbf{s}_r^{(0)})\right\}^T$ be $r \times 1$ vectors and $\mathbf{C}(\mathcal{S}^{(0)})$ be an $r \times r$ matrix whose (i, j)th entry is $C_{\boldsymbol{\alpha}}(\mathbf{s}_i^{(0)}, \mathbf{s}_j^{(0)})$. The MPP prior defines

(85)
$$\tilde{w}(\cdot) = \mathbf{c}^T(\cdot, \mathcal{S}^{(0)})\,\mathbf{C}(\mathcal{S}^{(0)})^{-1}\,\mathbf{w}^{(0)} + \tilde{\epsilon}(\cdot),$$

where the processes $\tilde{\epsilon}(\cdot)$ and $w(\cdot)$ are mutually independent and $\tilde{\epsilon}(\cdot)$ is a GP with mean 0, $\mathrm{cov}\{\tilde{\epsilon}(\mathbf{a}), \tilde{\epsilon}(\mathbf{b})\} = \delta(\mathbf{a})\,\mathbf{1}(\mathbf{a} = \mathbf{b})$ for any $\mathbf{a}, \mathbf{b} \in \mathcal{D}$, and

$$\delta(\mathbf{s}_{ji}) = C_{\boldsymbol{\alpha}}(\mathbf{s}_{ji}, \mathbf{s}_{ji}) - \mathbf{c}^T(\mathbf{s}_{ji}, \mathcal{S}^{(0)})\,\mathbf{C}(\mathcal{S}^{(0)})^{-1}\,\mathbf{c}(\mathbf{s}_{ji}, \mathcal{S}^{(0)}).$$

The process $\tilde{w}(\cdot)$ is a low-rank GP with mean 0 and

$$\text{cov}\{\tilde{w}(\mathbf{a}), \tilde{w}(\mathbf{b})\} = \mathbf{c}^T(\mathbf{a}, \mathcal{S}^{(0)})\, \mathbf{C}(\mathcal{S}^{(0)})^{-1}\, \mathbf{c}(\mathbf{b}, \mathcal{S}^{(0)}) + \delta(\mathbf{a})\, \mathbf{1}_{\mathbf{a}=\mathbf{b}}$$

for any $\mathbf{a}, \mathbf{b} \in \mathcal{D}$. If we replace $w(\cdot)$ by $\tilde{w}(\cdot)$ in (72), then

$$y(\mathbf{s}_{ji}) = \mathbf{x}(\mathbf{s}_{ji})^T\, \boldsymbol{\beta} + \tilde{w}(\mathbf{s}_{ji}) + \epsilon(\mathbf{s}_{ji}), \quad j = 1, \ldots, k, \quad i = 1, \ldots, m_j. \tag{86}$$

and our definition in (85) implies that $\tilde{w}(\cdot)$ is assigned a MPP prior (Finley et al., 2009).

We start by defining mean and covariance functions specific to univariate spatial regression using MPP. Let $\tilde{\mathbf{w}}_j = \{\tilde{w}(\mathbf{s}_{j1}), \ldots, \tilde{w}(\mathbf{s}_{jm})\}$ and $\tilde{\mathbf{w}}_j^* = \{\tilde{w}(\mathbf{s}_1), \ldots, \tilde{w}(\mathbf{s}_l)\}$. The MPP prior is identical to the FITC approximation in sparse approximate GP regression, so we use the FITC notations to simplify the description of posterior computations (Quiñonero-Candela and Rasmussen, 2005). Define $\mathbf{Q}_{j,j} = \mathbf{C}_{j,0}(\boldsymbol{\alpha})\, \mathbf{C}^{-1}(\mathcal{S}^{(0)})\, \mathbf{C}_{0,j}(\boldsymbol{\alpha})$, where $\text{cov}\{w(\mathbf{s}_{ja}), w(\mathbf{s}_b^{(0)})\} = \{\mathbf{C}_{j,0}(\boldsymbol{\alpha})\}_{a,b}$ $(a = 1, \ldots, m;$ $b = 1, \ldots, r)$ and $\mathbf{C}_{0,j}(\boldsymbol{\alpha}) = \mathbf{C}_{j,0}^T(\boldsymbol{\alpha})$. The density of $(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_j^*)$ under the GP prior implied by MPP is $N\{\mathbf{0}, \tilde{\mathbf{C}}(\boldsymbol{\alpha})\}$, where 2×2 block form of $\tilde{\mathbf{C}}(\boldsymbol{\alpha})$ is defined using

$$\begin{aligned}
\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) &= \mathbf{Q}_{j,j} + \text{diag}\{\mathbf{C}_{j,j}(\boldsymbol{\alpha}) - \mathbf{Q}_{j,j}\} = \text{cov}(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_j), \\
\tilde{\mathbf{C}}_{j,*}(\boldsymbol{\alpha}) &= \mathbf{Q}_{j,*} = \text{cov}(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_j^*), \\
\tilde{\mathbf{C}}_{*,*}(\boldsymbol{\alpha}) &= \mathbf{Q}_{*,*} + \text{diag}\{\mathbf{C}_{*,*}(\boldsymbol{\alpha}) - \mathbf{Q}_{*,*}\} = \text{cov}(\tilde{\mathbf{w}}_j^*, \tilde{\mathbf{w}}_j^*), \\
\tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha}) &= \mathbf{Q}_{*,j} = \text{cov}(\tilde{\mathbf{w}}_j^*, \tilde{\mathbf{w}}_j).
\end{aligned} \tag{87}$$

Stochastic approximation is implemented following Section 2. First, consider stochastic approximation for the likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Integrating out $\tilde{\mathbf{w}}_j$ in (86) gives

$$\mathbf{y}_j = \mathbf{X}_j\, \boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_j, \quad \tilde{\boldsymbol{\eta}}_j \sim N\{\mathbf{0}, \tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}\}. \tag{88}$$

The likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ given $\mathbf{y}_j$, $\mathbf{X}_j$ after stochastic approximation is

$$\{l_j(\boldsymbol{\beta}, \boldsymbol{\alpha})\}^k = (2\pi)^{-mk/2} |\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}|^{-k/2} e^{-\frac{k}{2}(\mathbf{y}_j - \mathbf{X}_j\, \boldsymbol{\beta})^T \{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}\}^{-1} (\mathbf{y}_j - \mathbf{X}_j\, \boldsymbol{\beta})}. \tag{89}$$

The prior distribution for $\boldsymbol{\beta}$ in (73), the pseudo likelihood in (89), and Bayes rule implies that the density of the jth subset posterior distribution for $\boldsymbol{\beta}$ given the rest is

$$\boldsymbol{\beta} \mid \text{rest} \propto e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j\, \boldsymbol{\beta})^T [k^{-1}\{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}\}]^{-1} (\mathbf{y}_j - \mathbf{X}_j\, \boldsymbol{\beta})}\, e^{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \Sigma_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})}.$$

This implies that the complete conditional distribution of $\boldsymbol{\beta}_j$ has density $N(\tilde{\mathbf{m}}_{j\,\boldsymbol{\beta}}, \tilde{\mathbf{V}}_{j\,\boldsymbol{\beta}})$, where

$$\begin{aligned}
\tilde{\mathbf{V}}_{j\,\boldsymbol{\beta}} &= \left[k\, \mathbf{X}_j^T \{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}\}^{-1}\, \mathbf{X}_j + \Sigma_{\boldsymbol{\beta}}^{-1}\right]^{-1}, \\
\tilde{\mathbf{m}}_{j\,\boldsymbol{\beta}} &= \tilde{\mathbf{V}}_{j\,\boldsymbol{\beta}} \left[k\, \mathbf{X}_j^T \left\{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2\, \mathbf{I}\right\}^{-1} \mathbf{y}_j + \Sigma_{\boldsymbol{\beta}}^{-1}\, \boldsymbol{\mu}_{\boldsymbol{\beta}}\right].
\end{aligned} \tag{90}$$

Following Section 2, the density of the jth subset posterior distribution for $\boldsymbol{\alpha}$ given the rest is

$$\begin{aligned}\boldsymbol{\alpha} \mid \text{rest} \propto & |\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}|^{-k/2} e^{-\frac{1}{2}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})^T [k^{-1}\{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tau^2 \mathbf{I}\}]^{-1} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})} \\ & \left(\sigma^2\right)^{-a_\sigma - 1} e^{-b_\sigma/\sigma^2} \left(\tau^2\right)^{-a_\tau - 1} e^{-b_\tau/\tau^2} (b_\phi - a_\phi)^{-1}. \end{aligned} \tag{91}$$

This density does not have a standard form, so we use a Metropolis-Hastings step with a normal random walk proposal and sample $\boldsymbol{\alpha}_j$ using the `metrop` function in the R package `mcmc`.

Second, we derive the posterior predictive distribution of $\tilde{\mathbf{w}}_j^*$ given the rest. The MPP prior on $(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_j^*)$ implies that the density of $\tilde{\mathbf{w}}_j^*$ given $\tilde{\mathbf{w}}_j$ is

$$\tilde{\mathbf{w}}_j^* \mid \tilde{\mathbf{w}}_j \sim N\left\{\tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\tilde{\mathbf{w}}_j, \tilde{\mathbf{C}}_{*,*}(\boldsymbol{\alpha}) - \tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,*}(\boldsymbol{\alpha})\right\}. \tag{92}$$

Given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{y}_j$, and $\mathbf{X}_j$, (86) implies that the likelihood of $\tilde{\mathbf{w}}_j$ after stochastic approximation is

$$\{l_j(\tilde{\mathbf{w}}_j)\}^k = (2\pi)^{-mk/2} |\tau^2 \mathbf{I}|^{-k/2} e^{-\frac{k}{2\tau^2}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta} - \tilde{\mathbf{w}}_j)^T (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta} - \tilde{\mathbf{w}}_j)}. \tag{93}$$

The MPP prior on $\tilde{\mathbf{w}}_j$, the pseudo likelihood in (93), and Bayes rule implies that the density of the subset posterior distribution for $\tilde{\mathbf{w}}_j$ given the rest is

$$\tilde{\mathbf{w}}_j \mid \text{rest} \propto e^{-\frac{1}{2\tau^2/k}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta} - \tilde{\mathbf{w}}_j)^T (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta} - \tilde{\mathbf{w}}_j)} e^{-\frac{1}{2}\tilde{\mathbf{w}}_j^T \tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\tilde{\mathbf{w}}_j}.$$

This implies that the complete conditional distribution of $\tilde{\mathbf{w}}_j$ has density $N(\mathbf{m}_{\tilde{\mathbf{w}}_j}, \mathbf{V}_{\tilde{\mathbf{w}}_j})$, where

$$\mathbf{V}_{\tilde{\mathbf{w}}_j} = \left\{\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha}) + \tfrac{k}{\tau^2}\mathbf{I}\right\}^{-1}, \quad \mathbf{m}_{\tilde{\mathbf{w}}_j} = \frac{k}{\tau^2}\mathbf{V}_{\tilde{\mathbf{w}}_j}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}); \tag{94}$$

therefore, (92) and (94) imply that the complete conditional distribution of $\tilde{\mathbf{w}}_j^*$ has density $N(\mathbf{m}_{\tilde{\mathbf{w}}_j^*}, \mathbf{V}_{\tilde{\mathbf{w}}_j^*})$, where

$$\begin{aligned}\mathbf{m}_{\tilde{\mathbf{w}}_j^*} &= \mathbb{E}(\tilde{\mathbf{w}}_j^* \mid \text{rest}) = \tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbb{E}(\tilde{\mathbf{w}}_j \mid \text{rest}) \\ &= \tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\left\{\tilde{\mathbf{C}}_{j,j}(\boldsymbol{\alpha}) + \tfrac{\tau^2}{k}\mathbf{I}\right\}^{-1}(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}) \end{aligned} \tag{95}$$

and

$$\begin{aligned}\mathbf{V}_{\tilde{\mathbf{w}}_j^*} &= \text{var}(\tilde{\mathbf{w}}_j^* \mid \text{rest}) = \mathbb{E}\left\{\text{var}(\tilde{\mathbf{w}}_j^* \mid \tilde{\mathbf{w}}_j) \mid \text{rest}\right\} + \text{var}\left\{\mathbb{E}(\tilde{\mathbf{w}}_j^* \mid \tilde{\mathbf{w}}_j) \mid \text{rest}\right\} \\ &= \tilde{\mathbf{C}}_{*,*}(\boldsymbol{\alpha}) - \tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,*}(\boldsymbol{\alpha}) + \tilde{\mathbf{C}}_{*,j}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\,\mathbf{V}_{\tilde{\mathbf{w}}_j}\,\tilde{\mathbf{C}}_{j,j}^{-1}(\boldsymbol{\alpha})\tilde{\mathbf{C}}_{j,*}(\boldsymbol{\alpha}). \end{aligned} \tag{96}$$

Finally, we derive the posterior predictive distribution of $\mathbf{y}_j^*$ given the rest. If $\boldsymbol{\beta}_j$, τ_j^2, $\tilde{\mathbf{w}}_j^*$ are the samples from the jth subset posterior distribution of $\boldsymbol{\beta}$, τ^2, and $\tilde{\mathbf{w}}^*$, then (86) implies that $\mathbf{y}_j^*$ given the rest is sampled as

$$\mathbf{y}_j^* = \mathbf{X}_j \boldsymbol{\beta}_j + \tilde{\mathbf{w}}_j^* + \boldsymbol{\epsilon}_j^*, \quad \boldsymbol{\epsilon}_j^* \sim N(\mathbf{0}, \tau_j^2 \mathbf{I});$$

therefore, the complete conditional distribution of $\mathbf{y}_j^*$ has density $N(\tilde{\boldsymbol{\mu}}_{\mathbf{y}_j^*}, \tilde{\mathbf{V}}_{\mathbf{y}_j^*})$, where

$$\tilde{\boldsymbol{\mu}}_{\mathbf{y}_j^*} = \mathbf{X}_j \boldsymbol{\beta}_j + \tilde{\mathbf{w}}_j^*, \quad \tilde{\mathbf{V}}_{\mathbf{y}_j^*} = \tau_j^2 \mathbf{I}. \tag{97}$$

All full conditionals except that of $\boldsymbol{\alpha}$ are analytically tractable in terms of standard distributions in subset j ($j = 1, \ldots, k$). The Gibbs sampler with a Metropolis-Hastings step iterates between the following four steps until sufficient number of samples of $\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j, \tilde{\mathbf{w}}_j^*$, and $\mathbf{y}_j^*$ are drawn post convergence to the stationary distribution:

1. Sample $\boldsymbol{\beta}_j$ from $N(\tilde{\boldsymbol{\mu}}_{j\boldsymbol{\beta}}, \tilde{\mathbf{V}}_{j\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\mu}}_{j\boldsymbol{\beta}}$ and $\tilde{\mathbf{V}}_{j\boldsymbol{\beta}}$ are defined in (90).
2. Sample $\boldsymbol{\alpha}_j$ using the Metropolis-Hastings algorithm from the jth subset posterior density (up to constants) of $\boldsymbol{\alpha}_j$ in (91) with a normal random walk proposal.
3. Sample $\tilde{\mathbf{w}}_j^*$ from $N(\boldsymbol{\mu}_{\tilde{\mathbf{w}}_j^*}, \mathbf{V}_{\tilde{\mathbf{w}}_j^*})$, where $\boldsymbol{\mu}_{\tilde{\mathbf{w}}_j^*}$ and $\mathbf{V}_{\tilde{\mathbf{w}}_j^*}$ are defined in (95) and (96).
4. Sample $\mathbf{y}_j^*$ from $N(\tilde{\boldsymbol{\mu}}_{\mathbf{y}_j^*}, \tilde{\mathbf{V}}_{\mathbf{y}_j^*})$, where $\tilde{\boldsymbol{\mu}}_{\mathbf{y}_j^*}$ and $\tilde{\mathbf{V}}_{\mathbf{y}_j^*}$ are defined in (97).

4. COMPARISONS BETWEEN DIVIDE-AND-CONQUER COMPETITORS

4.1 Setup

We compare the four competitors based on the divide-and-conquer technique. Extending Section 4 of the main manuscript, we compare the performance based on learning the process parameters, interpolating the unobserved spatial surface, and predicting the response at new locations. This section presents two simulation studies and one real data analysis. Recall that the first and second simulations (*Simulation 1*) generate data from a spatial linear model where the spatial processes are simulated from a GP and an analytic function with local features, respectively. The number of locations in the two simulations is moderately large with $n = 10,000$. Continuing from the main manuscript, our real data analysis is based on a large data subset of sea surface temperature data with $n = 1,00,000$ locations. For all the three simulations, the response at $(n+l)$ locations is modeled as

$$y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + w(\mathbf{s}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \tau^2), \quad \mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^2, \tag{98}$$

for $i = 1, \ldots, n + l$, where $\mathcal{D}$ is the spatial domain, $y(\mathbf{s}_i)$, $x(\mathbf{s}_i)$, $w(\mathbf{s}_i)$, and ϵ_i are the response, covariate, spatial process, and idiosyncratic error values at the location $\mathbf{s}_i$, β_0 is the intercept, β_1 models the covariate effect, and l is the number of new locations.

The three-step DISK, WASP, DPMC and CMC frameworks are applied using the low-rank MPP priors using the algorithm outlined in Section 3.3 of the main paper with two partitioning schemes. The first partitioning scheme randomly partitions the spatial locations in k groups. In the second partitioning scheme, we divide the spatial domain into sixteen square grid cells and randomly allocate locations in every grid cell into k groups.

We compare the quality of prediction and estimation of spatial surface at predictive locations $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_l^*\}$. If $w(\mathbf{s}_{i'}^*)$ and $y(\mathbf{s}_{i'}^*)$ are the value of the

spatial surface and response at $\mathbf{s}^*_{i'} \in \mathcal{S}^*$, then the estimation and prediction errors are defined as

$$(99)\quad \text{Est Err}^2 = \frac{1}{l}\sum_{i'=1}^{l}\{\hat{w}(\mathbf{s}^*_{i'}) - w(\mathbf{s}^*_{i'})\}^2, \quad \text{Pred Err}^2 = \frac{1}{l}\sum_{i'=1}^{l}\{\hat{y}(\mathbf{s}^*_{i'}) - y(\mathbf{s}^*_{i'})\}^2,$$

where $\hat{w}(\mathbf{s}^*_{i'})$ and $\hat{y}(\mathbf{s}^*_{i'})$ denote the estimates of $w(\mathbf{s}^*_{i'})$ and $y(\mathbf{s}^*_{i'})$ obtained using any distributed or non-distributed methods. For sampling-based methods, we set $\hat{w}(\mathbf{s}^*_{i'})$ and $\hat{y}(\mathbf{s}^*_{i'})$ to be the medians of posterior MCMC samples for $w(\mathbf{s}^*_{i'})$ and $y(\mathbf{s}^*_{i'})$, respectively, for $i' = 1, \ldots, l$. We also estimate the point-wise 95% credible or confidence intervals (CIs) of $w(\mathbf{s}^*_{i'})$ and predictive intervals (PIs) of $y(\mathbf{s}^*_{i'})$ for every $\mathbf{s}_{i'} \in \mathcal{S}^*$ and compare the CI and PI coverages and lengths for every method. Finally, we compare the performance of all the methods for parameter estimation using the posterior medians or point estimates and the 95% CIs.

4.2 Simulation 1: Spatial Linear Model Based On GP

This section compares DISK with its divide-and-conquer competitors under the two partitioning schemes and is a continuation of Section 4.2 of the main manuscript. The four divide-and-conquer methods, CMC DISK, WASP, and DPMC, have similar performance in parameter estimation (Tables 1, 2, and 3). The parameter estimates obtained using all these methods are close to the truth and estimation errors are also very similar. The 95% credible intervals of β_0, β_1, τ^2 in cover the true values and their lower and upper quantiles are very similar. All the four methods underestimate σ^2 and overestimate ϕ slightly. Both results are the impacts of parent MPP prior, which also shows a similar pattern for the two choices of r. We notice that the coverage of CMC is smaller than that of DISK, WASP, and DPMC. More importantly, the choice of r, k, or partitioning scheme has a minimal impact on parameter estimation in DISK, WASP, and DPMC.

The inferential and predictive performance of DISK, WASP, and DPMC are similar, but CMC shows significant differences (Table 4). There are minimal differences in the prediction and estimation errors of CMC, DISK, WASP, and DPMC. This indicates that the estimate of posterior medians are very similar in all the four methods; however, the pointwise coverage of CMC in prediction of the response and inference on the spatial surface is significantly smaller than the nominal value for every choice of r and k. On the other hand, DISK, WASP, and DPMC have nominal coverage in prediction and inference on the spatial surface. Furthermore, their CI and PI coverage values are robust to the choices of r, k, and partitioning scheme.

In summary, the DISK, WASP, and DPMC have similar inferential and predictive performance. While CMC's point estimates are close to those of DISK, WASP, and DPMC, its inferential and predictive performance is worse than its three competitors. The partitioning scheme, random or grid-based, has no impact on the performance of all the four divide-and-conquer methods.

4.3 Simulation 2: Spatial Linear Model Based On Analytic Spatial Surface

This section compares DISK with its divide-and-conquer competitors and is a continuation of Section 4.3 of the main manuscript. Our conclusions remain similar as those observed in the previous section. Specifically, CMC DISK, WASP, and DPMC have similar performance in parameter estimation (Tables 5, 6, and

TABLE 1

The errors in estimating the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1), \sigma^2, \phi, \tau^2$ in Simulation 1 for the divide-and-conquer methods under random and grid-based partitioning. The parameter estimates for the Bayesian methods $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples and their true values are $\boldsymbol{\beta}_0 = (1, 2)$, $\sigma_0^2 = 1, \phi_0 = 4$ and $\tau_0^2 = 0.1$. The entries in the table are averaged across 10 simulation replications.

	$\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|\|$	$\|\hat{\sigma}^2 - \sigma_0^2\|$	$\|\hat{\phi} - \phi_0\|$	$\|\hat{\tau}^2 - \tau_0^2\|$
	Random Partitioning			
CMC ($r = 200$, $k = 10$)	0.09	0.12	0.68	0.01
CMC ($r = 400$, $k = 10$)	0.09	0.12	0.75	0.01
CMC ($r = 200$, $k = 20$)	0.10	0.13	0.95	0.02
CMC ($r = 400$, $k = 20$)	0.10	0.13	0.82	0.02
DISK ($r = 200$, $k = 10$)	0.09	0.11	0.64	0.01
DISK ($r = 400$, $k = 10$)	0.09	0.11	0.64	0.01
DISK ($r = 200$, $k = 20$)	0.10	0.12	0.66	0.02
DISK ($r = 400$, $k = 20$)	0.10	0.12	0.66	0.02
WASP ($r = 200$, $k = 10$)	0.09	0.11	0.64	0.01
WASP ($r = 400$, $k = 10$)	0.09	0.11	0.63	0.01
WASP ($r = 200$, $k = 20$)	0.10	0.12	0.66	0.02
WASP ($r = 400$, $k = 20$)	0.10	0.12	0.66	0.02
DPMC ($r = 200$, $k = 10$)	0.09	0.11	0.64	0.01
DPMC ($r = 400$, $k = 10$)	0.09	0.11	0.63	0.01
DPMC ($r = 200$, $k = 20$)	0.10	0.12	0.66	0.02
DPMC ($r = 400$, $k = 20$)	0.10	0.12	0.66	0.02
	Grid-Based Partitioning			
CMC ($r = 200$, $k = 10$)	0.09	0.12	0.63	0.01
CMC ($r = 400$, $k = 10$)	0.09	0.12	0.65	0.01
CMC ($r = 200$, $k = 20$)	0.10	0.13	0.77	0.01
CMC ($r = 400$, $k = 20$)	0.10	0.13	0.83	0.01
DISK ($r = 200$, $k = 10$)	0.09	0.12	0.62	0.01
DISK ($r = 400$, $k = 10$)	0.09	0.12	0.62	0.01
DISK ($r = 200$, $k = 20$)	0.10	0.12	0.63	0.01
DISK ($r = 400$, $k = 20$)	0.10	0.12	0.64	0.01
WASP ($r = 200$, $k = 10$)	0.09	0.12	0.62	0.01
WASP ($r = 400$, $k = 10$)	0.09	0.12	0.62	0.01
WASP ($r = 200$, $k = 20$)	0.10	0.12	0.63	0.01
WASP ($r = 400$, $k = 20$)	0.10	0.12	0.64	0.01
DPMC ($r = 200$, $k = 10$)	0.09	0.12	0.62	0.01
DPMC ($r = 400$, $k = 10$)	0.09	0.12	0.62	0.01
DPMC ($r = 200$, $k = 20$)	0.10	0.12	0.63	0.01
DPMC ($r = 400$, $k = 20$)	0.10	0.12	0.64	0.01

7); however, the inferential and predictive performance of DISK, WASP, and DPMC are significantly better than those of CMC (Table 8). The partitioning scheme, random or grid-based, has no impact on the inferential and predictive performance of CMC, DISK, WASP, and DPMC. The results are also robust to the choices of k and r.

4.4 Real data analysis: Sea Surface Temperature data

This section is a continuation of Section 4.3 of the main manuscript and compares DISK with its divide-and-conquer competitors in analyzing the Sea Surface Temperature (SST) data. We have chosen random partitioning based on our conclusions in the previous two simulations. Our results for SST data analysis are also very similar to those in the previous two simulations. Specifically, CMC DISK, WASP, and DPMC have similar performance in parameter estimation, but signif-

TABLE 2

The estimates of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1), \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 1 for the divide-and-conquer methods under random partitioning. The parameter estimates for the Bayesian methods $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications.

	β_0	β_1	σ^2	ϕ	τ^2
Truth	1.00	2.00	1.00	4.00	0.10
	Parameter Estimates				
CMC ($r = 200$, $k = 10$)	1.00	2.00	0.91	4.38	0.10
CMC ($r = 400$, $k = 10$)	1.00	2.00	0.91	4.41	0.10
CMC ($r = 200$, $k = 20$)	1.00	2.00	0.90	4.55	0.10
CMC ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.46	0.10
DISK ($r = 200$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
DISK ($r = 400$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
DISK ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
DISK ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
WASP ($r = 200$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
WASP ($r = 400$, $k = 10$)	1.00	2.00	0.92	4.34	0.11
WASP ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
WASP ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
DPMC ($r = 200$, $k = 10$)	1.00	2.00	0.92	4.35	0.11
DPMC ($r = 400$, $k = 10$)	1.00	2.00	0.92	4.34	0.11
DPMC ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
DPMC ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.38	0.11
	95% Credible Intervals				
CMC ($r = 200$, $k = 10$)	(0.98, 1.02)	(2.00, 2.00)	(0.90, 0.93)	(4.28, 4.49)	(0.10, 0.11)
CMC ($r = 400$, $k = 10$)	(0.98, 1.02)	(2.00, 2.00)	(0.90, 0.93)	(4.31, 4.52)	(0.10, 0.11)
CMC ($r = 200$, $k = 20$)	(0.99, 1.01)	(2.00, 2.00)	(0.89, 0.92)	(4.49, 4.61)	(0.10, 0.10)
CMC ($r = 400$, $k = 20$)	(0.99, 1.01)	(2.00, 2.00)	(0.90, 0.92)	(4.40, 4.53)	(0.10, 0.10)
DISK ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
DISK ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
DISK ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.07, 4.67)	(0.09, 0.13)
DISK ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.07, 4.68)	(0.09, 0.13)
WASP ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
WASP ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.69)	(0.09, 0.12)
WASP ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.07, 4.67)	(0.09, 0.13)
WASP ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.07, 4.68)	(0.09, 0.13)
DPMC ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(4.00, 4.70)	(0.09, 0.12)
DPMC ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.99, 4.70)	(0.09, 0.12)
DPMC ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.06, 4.68)	(0.09, 0.13)
DPMC ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.98, 2.01)	(0.86, 0.96)	(4.06, 4.69)	(0.09, 0.13)

icant differences exist in their predictive performance (Table 9). CMC's predictive coverage is much smaller than the nominal value, which matches our conclusions in the previous two simulations. DISK outperforms WASP and DPMC in predictions in that its MSPE is the smallest among them. DISK also has better nominal predictive coverage than WASP and DPMC while having comparable 95% PI lengths. The results are also robust to the choices of r. We conclude that DISK performs better than its divide-and-conquer competitors in SST data analysis.

4.5 Computation time comparisons

We report the run-times of all the methods used in the simulated and real data analysis in Section 4 of the main paper. Since distributed methods partition the data into the same subset size and fit the same MPP model for subset posterior inference, the run times are identical for any method in Simulation 1 and 2. Thus, we only present run times for simulation and for the sea surface temperature data; see Tables 10 and 11 for the run-times in $\log_{10}$ seconds for Simulation

TABLE 3

The estimates of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1), \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 1 for the divide-and-conquer methods under grid-based partitioning. The parameter estimates for the Bayesian methods $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications.

	β_0	β_1	σ^2	ϕ	τ^2
Truth	1.00	2.00	1.00	4.00	0.10
	Parameter Estimates				
CMC ($r = 200$, $k = 10$)	1.00	2.00	0.91	4.37	0.10
CMC ($r = 400$, $k = 10$)	1.00	2.00	0.91	4.37	0.10
CMC ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.44	0.10
CMC ($r = 400$, $k = 20$)	1.00	2.00	0.90	4.48	0.10
DISK ($r = 200$, $k = 10$)	1.00	2.00	0.91	4.35	0.11
DISK ($r = 400$, $k = 10$)	1.00	2.00	0.91	4.35	0.11
DISK ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
DISK ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
WASP ($r = 200$, $k = 10$)	1.00	2.00	0.91	4.35	0.11
WASP ($r = 400$, $k = 10$)	1.00	2.00	0.91	4.34	0.11
WASP ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
WASP ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
DPMC ($r = 200$, $k = 10$)	1.00	2.00	0.91	4.35	0.11
DPMC ($r = 400$, $k = 10$)	1.00	2.00	0.91	4.34	0.11
DPMC ($r = 200$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
DPMC ($r = 400$, $k = 20$)	1.00	2.00	0.91	4.37	0.11
	95% Credible Intervals				
CMC ($r = 200$, $k = 10$)	(0.98, 1.02)	(2.00, 2.00)	(0.89, 0.93)	(4.27, 4.47)	(0.10, 0.11)
CMC ($r = 400$, $k = 10$)	(0.98, 1.02)	(2.00, 2.00)	(0.89, 0.93)	(4.27, 4.48)	(0.10, 0.11)
CMC ($r = 200$, $k = 20$)	(0.99, 1.01)	(2.00, 2.00)	(0.90, 0.92)	(4.38, 4.50)	(0.10, 0.11)
CMC ($r = 400$, $k = 20$)	(0.99, 1.01)	(2.00, 2.00)	(0.89, 0.92)	(4.42, 4.54)	(0.10, 0.11)
DISK ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.99, 4.69)	(0.09, 0.12)
DISK ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.99, 4.70)	(0.09, 0.12)
DISK ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.06, 4.67)	(0.09, 0.13)
DISK ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.07, 4.67)	(0.09, 0.13)
WASP ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.99, 4.69)	(0.10, 0.12)
WASP ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.98, 4.70)	(0.09, 0.12)
WASP ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.06, 4.67)	(0.09, 0.13)
WASP ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.07, 4.67)	(0.09, 0.13)
DPMC ($r = 200$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.98, 4.70)	(0.09, 0.12)
DPMC ($r = 400$, $k = 10$)	(0.92, 1.08)	(1.99, 2.01)	(0.86, 0.98)	(3.98, 4.71)	(0.09, 0.12)
DPMC ($r = 200$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.05, 4.69)	(0.09, 0.13)
DPMC ($r = 400$, $k = 20$)	(0.94, 1.06)	(1.99, 2.01)	(0.86, 0.96)	(4.07, 4.68)	(0.09, 0.13)

1 and in $\log_{10}$ hours for sea surface temperature data analysis, respectively. Similar to our observations in the performance comparisons, the run-times for the distributed methods are independent of the partitioning schemes. The run-times cannot be compared directly from the tables due to the differences in implementation. Specifically, distributed methods are implemented in R for all values of r and k, whereas most non-distributed methods are implemented in R and a higher-level language, such as C/C++ and Fortran.

The combination step in any distributed method requires a very small time compared to the time required for sampling on the subsets. For example, the time required for combination using the WASP is the largest among all the four distributed methods, but the maximum of WASP's combination time is only 8% of the maximum time required for sampling on the subsets. On an average, the combination steps of the other three methods require less than 1% of the time required for sampling on the subsets. This implies that run-times for all the four distributed methods in the two simulations are fairly similar (Table 10). In the real data analysis, the combination steps of the all the four distributed methods

TABLE 4

Inference on the values of spatial surface and response at the locations in $\mathcal{S}_*$ in Simulation 1 for the divide-and-conquer methods under random and grid-based partitioning. The estimation and prediction errors are defined in (99) and coverage and credible intervals are calculated pointwise for the locations in $\mathcal{S}_*$. The entries in the table are averaged over 10 simulation replications.

	Est Err	Pred Err	95% CI Coverage		95% CI Length	
	GP	Y	GP	Y	GP	Y
	Random Partitioning					
CMC ($r = 200$, $k = 10$)	0.56	0.64	0.38	0.39	0.81	0.81
CMC ($r = 400$, $k = 10$)	0.43	0.52	0.40	0.41	0.74	0.74
CMC ($r = 200$, $k = 20$)	0.58	0.67	0.27	0.28	0.57	0.57
CMC ($r = 400$, $k = 20$)	0.46	0.55	0.28	0.29	0.52	0.52
DISK ($r = 200$, $k = 10$)	0.55	0.64	0.97	0.97	3.20	3.45
DISK ($r = 400$, $k = 10$)	0.42	0.51	0.97	0.97	2.88	3.15
DISK ($r = 200$, $k = 20$)	0.58	0.67	0.97	0.97	3.25	3.51
DISK ($r = 400$, $k = 20$)	0.46	0.55	0.97	0.97	2.98	3.25
WASP ($r = 200$, $k = 10$)	0.55	0.64	0.96	0.96	3.25	3.25
WASP ($r = 400$, $k = 10$)	0.42	0.51	0.96	0.96	2.97	2.97
WASP ($r = 200$, $k = 20$)	0.58	0.67	0.96	0.96	3.30	3.30
WASP ($r = 400$, $k = 20$)	0.46	0.55	0.96	0.96	3.06	3.06
DPMC ($r = 200$, $k = 10$)	0.55	0.64	0.97	0.97	3.46	3.46
DPMC ($r = 400$, $k = 10$)	0.42	0.51	0.97	0.97	3.17	3.17
DPMC ($r = 200$, $k = 20$)	0.58	0.67	0.97	0.97	3.53	3.53
DPMC ($r = 400$, $k = 20$)	0.46	0.55	0.97	0.97	3.28	3.28
	Grid-Based Partitioning					
CMC ($r = 200$, $k = 10$)	0.75	0.80	0.38	0.39	0.81	0.81
CMC ($r = 400$, $k = 10$)	0.65	0.72	0.40	0.40	0.74	0.74
CMC ($r = 200$, $k = 20$)	0.76	0.82	0.27	0.28	0.57	0.57
CMC ($r = 400$, $k = 20$)	0.68	0.74	0.28	0.28	0.52	0.52
DISK ($r = 200$, $k = 10$)	0.75	0.80	0.97	0.97	3.45	3.45
DISK ($r = 400$, $k = 10$)	0.65	0.72	0.97	0.97	3.15	3.15
DISK ($r = 200$, $k = 20$)	0.76	0.82	0.97	0.97	3.51	3.51
DISK ($r = 400$, $k = 20$)	0.68	0.74	0.97	0.97	3.26	3.26
WASP ($r = 200$, $k = 10$)	0.74	0.80	0.96	0.96	3.25	3.25
WASP ($r = 400$, $k = 10$)	0.65	0.72	0.96	0.96	2.97	2.97
WASP ($r = 200$, $k = 20$)	0.76	0.82	0.96	0.95	3.30	3.30
WASP ($r = 400$, $k = 20$)	0.68	0.74	0.96	0.96	3.06	3.06
DPMC ($r = 200$, $k = 10$)	0.74	0.80	0.97	0.97	3.46	3.46
DPMC ($r = 400$, $k = 10$)	0.65	0.72	0.97	0.97	3.16	3.16
DPMC ($r = 200$, $k = 20$)	0.76	0.82	0.97	0.97	3.53	3.53
DPMC ($r = 400$, $k = 20$)	0.68	0.74	0.97	0.97	3.28	3.28

require less that 1% of the time required for sampling on the subsets, so all of them have nearly identical run-times (Table 11).

5. MARKOV CHAINS ON THE SUBSETS IN DISK

Any divide-and-conquer method runs modified Markov chain Monte Carlo algorithms in parallel on the subsets to obtain draws from the subset posterior distributions. In our context, we draw parameter and response values from the respective posterior distributions on every subset. There are no theoretical results that guarantee convergence of the Markov chain produced by the sampling algorithms to its stationary distribution in a spatial linear model with MPP prior. This further complicates the theoretical analysis of the Markov chain produced on the subsets in DISK, where the likelihood is modified. We are not aware any rigorous approach for comparing the Markov chains obtained from the subset and true posterior distributions. We use heuristics based on trace plots and auto correlation functions of the Markov chains for parameters, spatial surface, and predictive surface to judge "convergence" to the respective subset posterior dis-

TABLE 5

The errors in estimating the parameters β, τ^2 in Simulation 2 for the divide-and-conquer methods under random and grid-based partitioning. The parameter estimates for the Bayesian methods $\hat{\beta}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples and $\beta_0 = 1$ and $\tau_0^2 = 0.01$. The entries in the table are averaged across 10 simulation replications.

	$\|\hat{\beta} - \beta_0\|$	$\|\hat{\tau}^2 - \tau_0^2\|$
	Random Partitioning	
CMC ($r = 200$, $k = 10$)	0.03	0.00
CMC ($r = 400$, $k = 10$)	0.03	0.09
CMC ($r = 200$, $k = 20$)	1.41	0.09
CMC ($r = 400$, $k = 20$)	1.41	0.09
DISK ($r = 200$, $k = 10$)	0.18	0.04
DISK ($r = 400$, $k = 10$)	0.13	0.04
DISK ($r = 200$, $k = 20$)	0.18	0.04
DISK ($r = 400$, $k = 20$)	0.13	0.04
WASP($r = 200$, $k = 10$)	0.68	0.09
WASP ($r = 400$, $k = 10$)	0.68	0.09
WASP ($r = 200$, $k = 20$)	0.72	0.09
WASP ($r = 400$, $k = 20$)	0.72	0.09
DPMC ($r = 200$, $k = 10$)	0.68	0.09
DPMC ($r = 400$, $k = 10$)	0.68	0.09
DPMC ($r = 200$, $k = 20$)	0.72	0.09
DPMC ($r = 400$, $k = 20$)	0.72	0.09
	Grid-Based Partitioning	
CMC ($r = 200$, $k = 10$)	0.03	0.09
CMC ($r = 400$, $k = 10$)	0.03	0.09
CMC ($r = 200$, $k = 20$)	0.02	0.09
CMC ($r = 400$, $k = 20$)	0.02	0.09
DISK ($r = 200$, $k = 10$)	0.03	0.09
DISK ($r = 400$, $k = 10$)	0.03	0.09
DISK ($r = 200$, $k = 20$)	0.02	0.09
DISK ($r = 400$, $k = 20$)	0.02	0.09
WASP($r = 200$, $k = 10$)	0.03	0.09
WASP ($r = 400$, $k = 10$)	0.03	0.09
WASP ($r = 200$, $k = 20$)	0.02	0.09
WASP ($r = 400$, $k = 20$)	0.02	0.09
DPMC ($r = 200$, $k = 10$)	0.03	0.09
DPMC ($r = 400$, $k = 10$)	0.03	0.09
DPMC ($r = 200$, $k = 20$)	0.02	0.09
DPMC ($r = 400$, $k = 20$)	0.02	0.09

tributions.

Unfortunately, it is impractical to compare trace plots and auto correlation functions obtained using subset and true posterior distributions; therefore, we compare the effective sample sizes of Markov chains for the parameters, spatial surface, and response obtained on the subsets using an MPP prior relative to those obtained using the full data and the same MPP prior. The number of posterior samples in both cases is 1000, which are obtained from a Markov chain of 10000 draws after using a burn-in of 5000 and collecting every fifth sample. The `effectiveSize` command coda R package is used for estimating the effective sample sizes for every choice of k and r (Plummer et al., 2006). We compute the ratio of the effective sample sizes of the Markov chains produced on the subsets in DISK to those obtained using the MPP prior and the full data. For two- or higher-dimensional parameters, spatial surface, and predictive surface, we average the ratio of the effective sample sizes across all the dimensions. While there are no theoretical justifying the convergence of the Markov chain to the stationary distribution, we still assume so because MPP has been used extensively for analyzing spatial data. This heuristic shows that the Markov chains obtained using the data subsets and full data are "similar" in that their effective sample

TABLE 6

The estimates of parameters $\beta, \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 2 for the divide-and-conquer methods under random partitioning. The parameter estimates for the Bayesian methods $\hat{\beta}$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications

	β	σ^2	ϕ	τ^2
Truth	1.00	-	-	0.01
	Parameter Estimates			
CMC ($r = 200$, $k = 10$)	1.03	0.22	0.11	0.01
CMC ($r = 400$, $k = 10$)	1.03	0.22	0.11	0.01
CMC ($r = 200$, $k = 20$)	0.98	0.23	0.13	0.01
CMC ($r = 400$, $k = 20$)	0.98	0.23	0.13	0.01
DISK ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
DISK ($r = 400$, $k = 10$)	0.98	0.22	0.14	0.01
DISK ($r = 200$, $k = 20$)	1.03	0.21	0.12	0.01
DISK ($r = 400$, $k = 20$)	0.98	0.22	0.14	0.01
WASP ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
WASP ($r = 400$, $k = 10$)	1.03	0.21	0.12	0.01
WASP ($r = 200$, $k = 20$)	0.98	0.22	0.14	0.01
WASP ($r = 400$, $k = 20$)	0.98	0.22	0.14	0.01
DPMC ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
DPMC ($r = 400$, $k = 10$)	1.03	0.21	0.12	0.01
DPMC ($r = 200$, $k = 20$)	0.98	0.22	0.14	0.01
DPMC ($r = 400$, $k = 20$)	0.98	0.22	0.14	0.01
	95% Credible Intervals			
CMC($r = 200$, $k = 10$)	(0.96, 1.11)	(0.22, 0.23)	(0.11, 0.12)	(0.01, 0.01)
CMC ($r = 400$, $k = 10$)	(0.96, 1.11)	(0.22, 0.23)	(0.11, 0.12)	(0.01, 0.01)
CMC ($r = 200$, $k = 20$)	(0.94, 1.02)	(0.22, 0.24)	(0.13, 0.13)	(0.01, 0.01)
CMC ($r = 400$, $k = 20$)	(0.94, 1.02)	(0.22, 0.24)	(0.12, 0.13)	(0.01, 0.01)
DISK ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
DISK ($r = 400$, $k = 10$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
DISK ($r = 200$, $k = 20$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
DISK ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
WASP ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
WASP ($r = 400$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
WASP ($r = 200$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
WASP ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
DPMC ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.17, 0.25)	(0.10, 0.15)	(0.01, 0.01)
DPMC ($r = 400$, $k = 10$)	(0.80, 1.27)	(0.17, 0.25)	(0.10, 0.15)	(0.01, 0.01)
DPMC ($r = 200$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.11, 0.18)	(0.01, 0.01)
DPMC ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.11, 0.19)	(0.01, 0.01)

sizes are very close.

The effective sample sizes of the Markov chains for the parameters and the spatial surface and response at the locations in $\mathcal{S}^*$ are very similar to those obtained using the full data and the same MPP prior in Simulation 1 (Table 12). The effective sample sizes decrease with k in Simulation 2 slightly for the spatial surface and response at the locations in $\mathcal{S}^*$ (Table 13); however, this spatial surface is not simulated from a GP in this simulation, so the comparisons are less reliable. The partitioning scheme, random or grid-based, has a minimal impact on the effective sample sizes. The ratio of the effective sample sizes are equal for the β, spatial surface, and predictions in Simulation 1; however, there are differences in the effective sample sizes of the Markov chains for σ^2, ϕ, τ^2 in both simulations. These differences mainly arise due to the non-identifiability of the covariance function parameters. In most spatial applications, the main interest lies in inference and prediction, where the effective sample sizes on the subsets are very similar to their full data benchmarks; therefore, we conclude that the Markov chains produced on the subsets in DISK have similar properties as their full data

TABLE 7

The estimates of parameters $\beta, \sigma^2, \phi, \tau^2$ and their 95% marginal credible intervals (CIs) in Simulation 2 for the divide-and-conquer methods under grid-based partitioning. The parameter estimates for the Bayesian methods $\hat{\beta}$, $\hat{\sigma}^2, \hat{\phi}$, $\hat{\tau}^2$ are defined as the posterior medians of their respective MCMC samples. The parameter estimates and upper and lower quantiles of 95% CIs are averaged over 10 simulation replications

	β	σ^2	ϕ	τ^2
Truth	1.00	-	-	0.01
	Parameter Estimates			
CMC ($r = 200$, $k = 10$)	1.03	0.22	0.11	0.01
CMC ($r = 400$, $k = 10$)	1.03	0.22	0.11	0.01
CMC ($r = 200$, $k = 20$)	0.98	0.23	0.13	0.01
CMC ($r = 400$, $k = 20$)	0.98	0.23	0.13	0.01
DISK ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
DISK ($r = 400$, $k = 10$)	1.03	0.21	0.12	0.01
DISK ($r = 200$, $k = 20$)	0.98	0.22	0.14	0.01
DISK ($r = 400$, $k = 20$)	0.98	0.22	0.14	0.01
WASP ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
WASP ($r = 400$, $k = 10$)	1.03	0.21	0.12	0.01
WASP ($r = 200$, $k = 20$)	0.98	0.22	0.14	0.01
WASP ($r = 400$, $k = 20$)	0.99	0.22	0.14	0.01
DPMC ($r = 200$, $k = 10$)	1.03	0.21	0.12	0.01
DPMC ($r = 400$, $k = 10$)	1.03	0.21	0.12	0.01
DPMC ($r = 200$, $k = 20$)	0.98	0.22	0.14	0.01
DPMC ($r = 400$, $k = 20$)	0.99	0.22	0.14	0.01
	95% Credible Intervals			
CMC($r = 200$, $k = 10$)	(0.96, 1.10)	(0.21, 0.23)	(0.11, 0.12)	(0.01, 0.01)
CMC ($r = 400$, $k = 10$)	(0.95, 1.10)	(0.21, 0.23)	(0.11, 0.12)	(0.01, 0.01)
CMC ($r = 200$, $k = 20$)	(0.94, 1.02)	(0.22, 0.23)	(0.13, 0.14)	(0.01, 0.01)
CMC ($r = 400$, $k = 20$)	(0.94, 1.02)	(0.22, 0.24)	(0.13, 0.13)	(0.01, 0.01)
DISK ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
DISK ($r = 400$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
DISK ($r = 200$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
DISK ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
WASP ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
WASP ($r = 400$, $k = 10$)	(0.80, 1.27)	(0.18, 0.24)	(0.11, 0.14)	(0.01, 0.01)
WASP ($r = 200$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
WASP ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.12, 0.18)	(0.01, 0.01)
DPMC ($r = 200$, $k = 10$)	(0.80, 1.27)	(0.17, 0.24)	(0.10, 0.15)	(0.01, 0.01)
DPMC ($r = 400$, $k = 10$)	(0.80, 1.27)	(0.17, 0.25)	(0.10, 0.15)	(0.01, 0.01)
DPMC ($r = 200$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.11, 0.18)	(0.01, 0.01)
DPMC ($r = 400$, $k = 20$)	(0.82, 1.16)	(0.17, 0.26)	(0.11, 0.18)	(0.01, 0.01)

versions in Simulations 1 and 2 in terms of effective sample size comparisons.

REFERENCES

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis 53*(8), 2873–2884.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, Volume 1. Springer.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news 6*(1), 7–11.

Quiñonero-Candela, J. and C. E. Rasmussen (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research 6*(Dec), 1939–1959.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research 16*, 3299–3340.

TABLE 8

Inference on the values of spatial surface and response at the locations in $\mathcal{S}_*$ in Simulation 2 for the divide-and-conquer methods under random and grid-based partitioning. The estimation and prediction errors are defined in (99) and coverage and credible intervals are calculated pointwise for the locations in $\mathcal{S}_*$. The entries in the table are averaged over 10 simulation replications.

	Est Err	Pred Err	95% CI Coverage		95% CI Length	
	GP	Y	GP	Y	GP	Y
	Random Partitioning					
CMC ($r = 200$, $k = 10$)	0.56	0.64	0.38	0.39	0.10	0.10
CMC ($r = 400$, $k = 10$)	0.43	0.52	0.40	0.41	0.10	0.10
CMC ($r = 200$, $k = 20$)	0.58	0.67	0.27	0.28	0.07	0.07
CMC ($r = 400$, $k = 20$)	0.46	0.55	0.28	0.29	0.07	0.07
DISK ($r = 200$, $k = 10$)	0.00	0.01	1.00	0.97	0.54	0.45
DISK ($r = 400$, $k = 10$)	0.00	0.01	1.00	0.97	0.45	0.47
DISK ($r = 200$, $k = 20$)	0.00	0.01	1.00	0.97	0.52	0.43
DISK ($r = 400$, $k = 20$)	0.00	0.01	1.00	0.97	0.43	0.44
WASP ($r = 200$, $k = 10$)	0.55	0.64	0.96	0.96	0.42	0.42
WASP ($r = 400$, $k = 10$)	0.42	0.51	0.96	0.96	0.40	0.40
WASP ($r = 200$, $k = 20$)	0.58	0.67	0.96	0.96	0.43	0.43
WASP ($r = 400$, $k = 20$)	0.46	0.55	0.96	0.96	0.41	0.41
DPMC ($r = 200$, $k = 10$)	0.55	0.64	0.97	0.97	0.45	0.45
DPMC ($r = 400$, $k = 10$)	0.42	0.51	0.97	0.97	0.43	0.43
DPMC ($r = 200$, $k = 20$)	0.58	0.67	0.97	0.97	0.46	0.46
DPMC ($r = 400$, $k = 20$)	0.46	0.55	0.97	0.97	0.44	0.44
	Grid-Based Partitioning					
CMC ($r = 200$, $k = 10$)	0.05	0.10	0.80	0.38	0.10	0.10
CMC ($r = 400$, $k = 10$)	0.04	0.10	0.85	0.37	0.10	0.10
CMC ($r = 200$, $k = 20$)	0.03	0.10	0.71	0.28	0.07	0.07
CMC ($r = 400$, $k = 20$)	0.03	0.10	0.70	0.28	0.07	0.07
DISK ($r = 200$, $k = 10$)	0.04	0.10	1.00	0.97	0.45	0.45
DISK ($r = 400$, $k = 10$)	0.04	0.10	1.00	0.96	0.42	0.42
DISK ($r = 200$, $k = 20$)	0.03	0.10	1.00	0.97	0.46	0.46
DISK ($r = 400$, $k = 20$)	0.03	0.10	1.00	0.96	0.44	0.44
WASP ($r = 200$, $k = 10$)	0.04	0.10	1.00	0.95	0.42	0.42
WASP ($r = 400$, $k = 10$)	0.04	0.10	1.00	0.94	0.40	0.40
WASP ($r = 200$, $k = 20$)	0.03	0.10	1.00	0.96	0.43	0.43
WASP ($r = 400$, $k = 20$)	0.03	0.10	1.00	0.95	0.41	0.41
DPMC ($r = 200$, $k = 10$)	0.04	0.10	1.00	0.97	0.45	0.45
DPMC ($r = 400$, $k = 10$)	0.04	0.10	1.00	0.96	0.43	0.43
DPMC ($r = 200$, $k = 20$)	0.03	0.10	1.00	0.97	0.46	0.46
DPMC ($r = 400$, $k = 20$)	0.03	0.10	1.00	0.97	0.44	0.44

TABLE 9

Parametric inference and prediction in SST data using the divide-and-conquer methods and MPP-based modeling with $r = 400, 600$ knots on $k = 300$ subsets. For parametric inference posterior medians are provided along with the 95% credible intervals (CIs) in the parentheses. Similarly, mean squared prediction errors (MSPEs) along with length and coverage of 95% predictive intervals (PIs) are presented. The upper and lower quantiles of 95% CIs and PIs are averaged over 10 simulation replications.

	β_0	β_1	σ^2	ϕ	τ^2
	Parameter Estimate				
CMC ($r = 400$, $k = 300$)	32.37	-0.32	12.38	0.03	0.18
CMC ($r = 600$, $k = 300$)	32.36	-0.32	12.31	0.03	0.18
DISK ($r = 400$, $k = 300$)	32.33	-0.32	11.82	0.04	0.18
DISK ($r = 600$, $k = 300$)	32.33	-0.32	11.85	0.04	0.18
WASP ($r = 400$, $k = 300$)	32.33	-0.32	11.82	0.04	0.18
WASP ($r = 600$, $k = 300$)	32.33	-0.32	11.85	0.04	0.18
DPMC ($r = 400$, $k = 300$)	32.33	-0.32	11.82	0.04	0.18
DPMC ($r = 600$, $k = 300$)	32.33	-0.32	11.85	0.04	0.18
	95% Credible Interval				
CMC ($r = 400$, $k = 300$)	(32.33, 32.4)	(-0.32, -0.32)	(12.37, 12.39)	(0.0339, 0.0340)	(0.18, 0.18)
CMC ($r = 600$, $k = 300$)	(32.33, 32.4)	(-0.32, -0.32)	(12.3, 12.31)	(0.0342, 0.0343)	(0.18, 0.18)
DISK ($r = 400$, $k = 300$)	(31.72, 32.93)	(-0.33, -0.31)	(11.24, 12.43)	(0.0373, 0.0412)	(0.18, 0.19)
DISK ($r = 600$, $k = 300$)	(31.72, 32.93)	(-0.33, -0.31)	(11.25, 12.45)	(0.0372, 0.0413)	(0.18, 0.19)
WASP ($r = 400$, $k = 300$)	(31.72, 32.93)	(-0.33, -0.31)	(11.22, 12.46)	(0.0372, 0.0413)	(0.18, 0.19)
WASP ($r = 600$, $k = 300$)	(31.72, 32.93)	(-0.33, -0.31)	(11.24, 12.47)	(0.0372, 0.0413)	(0.18, 0.19)
DPMC ($r = 400$, $k = 300$)	(31.72, 32.94)	(-0.33, -0.31)	(11.09, 12.55)	(0.0369, 0.0416)	(0.18, 0.19)
DPMC ($r = 600$, $k = 300$)	(31.72, 32.94)	(-0.33, -0.31)	(11.14, 12.56)	(0.0368, 0.0416)	(0.18, 0.19)
	Predictions				
	MSPE	95% PI Coverage	95% PI Length		
CMC ($r = 400$, $k = 300$)	0.74	0.05	0.08		
CMC ($r = 600$, $k = 300$)	0.67	0.05	0.07		
DISK ($r = 400$, $k = 300$)	0.43	0.95	2.65		
DISK ($r = 600$, $k = 300$)	0.36	0.95	2.34		
WASP ($r = 400$, $k = 300$)	0.66	0.93	2.39		
WASP ($r = 600$, $k = 300$)	0.60	0.92	2.11		
DPMC ($r = 400$, $k = 300$)	0.66	0.95	2.67		
DPMC ($r = 600$, $k = 300$)	0.60	0.94	2.36		

TABLE 10

Run-time (in $\log_{10}$ seconds) of the non-distributed methods and distributed methods under the random and grid-based partitioning schemes in Simulations 1 and 2, where MPP prior is used on the subsets.

	INLA	LaGP	NNGP ($m = 10$)	NNGP ($m = 20$)	NNGP ($m = 30$) ($m = 30$)	LatticeKrig	GpGp	
	1.08	0.08	2.96	3.42	3.74	2.03	0.96	
	Vecchia ($m = 10$)	Vecchia ($m = 20$)	Vecchia ($m = 30$)	MPP ($r = 200$)	MPP ($r = 400$)			
	2.76	3.20	3.50	3.97	4.31			
	$k = 10$							
	CMC		DISK		WASP		DPMC	
	$r = 200$	$r = 400$	$r = 200$	$r = 400$	$r = 200$	$r = 400$	$r = 200$	$r = 400$
Random	3.18	3.18	3.18	3.18	3.20	3.20	3.18	3.18
Grid	3.18	3.18	3.18	3.18	3.20	3.20	3.18	3.18
	$k = 20$							
	CMC		DISK		WASP		DPMC	
	$r = 200$	$r = 400$	$r = 200$	$r = 400$	$r = 200$	$r = 400$	$r = 200$	$r = 400$
Random	3.17	3.17	3.17	3.17	3.20	3.20	3.17	3.17
Grid	3.17	3.17	3.17	3.17	3.20	3.20	3.17	3.17

TABLE 11

Run-time (in $\log_{10}$ hours) of laGP and the distributed methods in the sea surface temperature data analysis, where MPP prior is used on the subsets.

laGP	MPP, $r = 400$, $k = 300$				MPP, $r = 600$, $k = 300$			
	CMC	DISK	WASP	DPMC	CMC	DISK	WASP	DPMC
-1.32	1.67	1.67	1.67	1.67	1.69	1.69	1.69	1.69

TABLE 12

The ratio of effective sample sizes of the Markov chains produced on the subsets using the MPP prior and those obtained using the full data and the same MPP prior in Simulation 1 under random and grid-based partitioning. The effective sample sizes have been averaged over the parameter dimensions and over 10 simulation replications.

	β	σ^2	ϕ	τ^2	GP	Y
	Random Partitioning					
$k = 10$ and $r = 200$	0.99	0.35	3.24	0.53	1.00	1.00
$k = 20$ and $r = 200$	1.00	0.61	3.92	0.40	1.00	1.00
$k = 10$ and $r = 400$	1.0	0.93	2.53	0.57	1.00	1.00
$k = 20$ and $r = 400$	1.11	1.45	2.88	0.43	1.00	1.00
	Grid-Based Partitioning					
$k = 10$ and $r = 200$	1.00	0.34	3.37	0.55	1.00	1.00
$k = 20$ and $r = 200$	1.00	0.61	3.89	0.39	1.00	1.00
$k = 10$ and $r = 400$	1.11	1.00	2.44	0.55	1.00	1.00
$k = 20$ and $r = 400$	1.11	1.65	2.97	0.41	1.00	1.00

TABLE 13

The ratio of effective sample sizes of the Markov chains produced on the subsets using the MPP prior and those obtained using the full data and the same MPP prior in Simulation 2 under random and grid-based partitioning.The effective sample sizes have been averaged over the parameter dimensions and over 10 simulation replications.

	β	σ^2	ϕ	τ^2	GP	Y
	Random Partitioning					
$k = 10$ and $r = 200$	0.98	0.46	1.32	3.45	0.93	1.00
$k = 20$ and $r = 200$	0.69	0.20	1.25	3.30	0.79	1.00
$k = 10$ and $r = 400$	0.89	1.65	1.81	2.84	0.91	1.00
$k = 20$ and $r = 400$	0.57	1.05	1.94	2.60	0.73	1.00
	Grid-Based Partitioning					
$k = 10$ and $r = 200$	0.98	0.62	1.27	3.52	0.94	1.00
$k = 20$ and $r = 200$	0.65	0.24	1.33	3.32	0.79	1.00
$k = 10$ and $r = 400$	0.88	1.81	1.97	2.73	0.91	1.00
$k = 20$ and $r = 400$	0.63	0.94	2.05	2.70	0.77	1.00