Calibration of imperfect geophysical models by multiple satellite interferograms with measurement bias

Mengyang Gu*, Kyle Anderson**, and Erika McPhillips*

* Department of Statistics and Applied Probability, UC Santa Barbara

** U. S. Geological Survey, Volcano Science Center

Abstract

Model calibration consists of using experimental or field data to estimate the unknown parameters of a mathematical model. The presence of model discrepancy and measurement bias in the data complicates this task. Satellite interferograms, for instance, are widely used for calibrating geophysical models in geological hazard quantification. In this work, we used satellite interferograms to relate ground deformation observations to the properties of the magma chamber at Kīlauea Volcano in Hawai'i. We derived closed-form marginal likelihoods and implemented posterior sampling procedures that simultaneously estimate the model discrepancy of physical models, and the measurement bias from the atmospheric error in satellite interferograms. We found that model calibration by aggregating multiple interferograms and downsampling the pixels in the interferograms can reduce the computation complexity compared to calibration approaches based on multiple data sets. The conditions that lead to no loss of information from data aggregation and downsampling are studied. Simulation illustrates that both discrepancy and measurement bias can be estimated, and real applications demonstrate that modeling both effects helps obtain a reliable estimation of a physical model's unobserved parameters and enhance its predictive accuracy. We implement the computational tools in the `RobustCalibration` package available on CRAN.

1 Introduction

Mathematical models are often used to describe various phenomena in science and engineering. To predict complex processes, one often first needs to estimate the unobserved parameters in the model using experimental observations or field data – a process generally known as model calibration. Denote the mathematical model by $f^M(\mathbf{x}, \boldsymbol{\theta})$, where $\mathbf{x}$ is a p_x-vector of observed input and $\boldsymbol{\theta}$ is a p_θ-vector of calibration parameters, both assumed to be real-valued. The superscript 'M' denotes the model. As the mathematical model may not represent reality perfectly, accurately estimating the difference between the mathematical model and reality can improve the predictive accuracy. In Kennedy and O'Hagan (2001), a Gaussian stochastic process (GaSP) defined on the observed input space, $\delta(\mathbf{x})$, was proposed to model the discrepancy between the mathematical model and reality. Modeling the discrepancy using a GaSP was subsequently examined in a number of other applications (Bayarri et al., 2007; Higdon et al., 2008; Arendt et al., 2012a).

In this work, we focus on spatially correlated patterns of measurement error from the data acquisition process, which we term *measurement bias*. We study interferometric synthetic aperture radar (InSAR) interferograms which, over the last 25 years, have made it possible to map deformation over broad swathes of the Earth's surface to sub-centimeter accuracy from space, revolutionizing scientists' understanding of Earth processes (Massonnet et al., 1995; Bürgmann et al., 2000; Pinel et al., 2014). InSAR interferograms are most often obtained using data from orbiting microwave-band radar satellites. By interfering two radar images of the surface taken from a satellite at different times, changes in the radar phase track temporal changes in the position of the Earth's surface along the oblique line-of-sight (LoS) vector between the satellite and ground. Because only fractional phase change can be measured directly, while the number of complete phase cycles between the satellite and ground is unknown, these images are wrapped by the radar's wavelength. Unwrapping an image by spatial integration of the phase gradient – relative to a point believed to be non-deforming or having zero ground displacement – yields relative LoS deformation in units of

distance change (Chen and Zebker, 2001).

Despite these advances, the interpretation of InSAR data is often greatly complicated by noise and bias. After removing the phase due to elevation of Earth's surface, the processed (or wrapped) phase for each pixel on the ground contains ground displacement, measurement bias, satellite orbital error, and look angle error. Measurement biases, in particular, are known to strongly affect many interferograms, caused most importantly by propagation delays due to atmospheric conditions, which yield spatially correlated noise which varies in time and space (e.g., Zebker et al., 1997; Hooper et al., 2007; Gong et al., 2016). Much work remains for mitigating and characterizing these uncertainties.

These observations motivate modeling the random measurement bias that can vary across different data sources due to environmental conditions or properties of devices, as well as a discrepancy function invariant across other data sources, to explain the difference between reality and model outcomes. The framework can be extended to integrate different types of observations, such as satellite radar interferograms, GPS, and tiltmeter observations (Anderson et al., 2019).

We utilize data from Kīlauea Volcano, one of the world's most active volcanoes. Kīlauea is located on the Island of Hawai‘i and erupted semi-continuously from 1983-2018. In 2018, a historically unprecedented rift eruption destroyed more than 700 homes and displaced thousands of residents (Neal et al., 2019). Figure 1 shows ground displacements at Kīlauea from October 2011 to May 2012, as the volcano's summit inflated due to magma storage (Anderson and Poland, 2016). This data was recorded from a satellite orbiting roughly north to south, which recorded LoS deformation along a vector oblique to the Earth's surface – roughly east to west and downward at an angle of 41 degrees. This image thus resolves a combination of predominantly vertical and east-west ground deformation. In Figure 1d, the image was subsampled by the quadtree algorithm for computational efficiency (Jonsson et al., 2002).

Of the sources of uncertainty in InSAR observations, the spatially correlated atmospheric

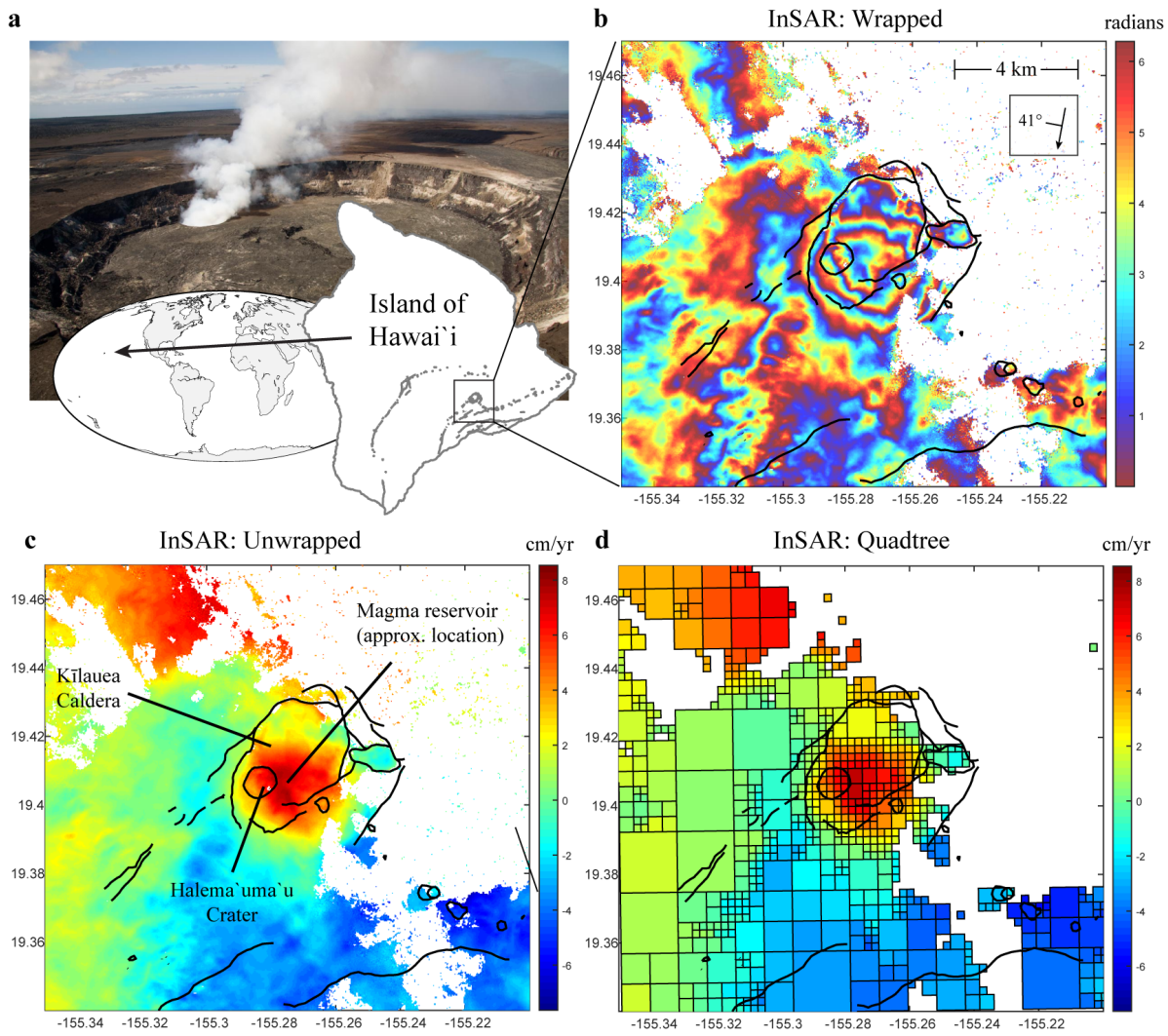

Figure 1: a) Overview maps showing the location of Kīlauea Volcano on the Island of Hawai'i, with background photo of Halema'uma'u Crater at the volcano's summit roughly as it appeared during the time of this study (USGS photo). (b) Wrapped InSAR interferogram from the COSMO-SkyMed satellite, spanning 20 Oct 2011 to 15 May 2012. The inset box shows the flight path of the satellite (arrow) and the downward look direction of the satellite at 41°. White areas have no data due to radar decorrelation. Number of data points is around 1.5×10^5. (c) Same data as in (b), but unwrapped. (d) Quadtree-processed interferogram. Thick black lines in panels b-d show cliffs and other important topographic features at the volcano; the large elliptical feature is Kīlauea Caldera.

term is usually the most important; spatial and temporal changes of just 20% in relative humidity can lead to errors of 10 cm in estimated ground deformation in some scenarios (Zebker et al., 1997). Figure 1b shows the wrapped InSAR phase at Kīlauea Volcano. A "bullseye" pattern near the center is due to real ground deformation, while most of the

remaining fringes are due to atmospheric conditions.

We highlight a few contributions of this study to address the main challenges involved in calibrating models using data from multiple sources. First, although discrepancy functions have been studied extensively for model calibration, discrepancy and measurement bias are rarely studied together. In this work, we simultaneously model both discrepancy and measurement bias functions and estimate their effects using multiple InSAR interferograms. This approach allows one to estimate geophysical model discrepancy and measurement bias due to distinct atmospheric conditions recorded in each InSAR interferogram. Furthermore, we derive the marginal likelihood and posterior distributions of the model parameters, which are useful for efficient posterior sampling. Third, InSAR interferograms can contain millions of pixels, so downsampling schemes such as the quadtree algorithm (Simons et al., 2002) are often applied, or multiple interferograms are averaged prior to modeling in order to reduce computational cost. However, the implications of modeling a single averaged interferogram rather than jointly modeling multiple individual interferograms has not been well-studied. Here we discuss the conditions under which these two approaches are equivalent and scenarios in which modeling individual data sets leads to more precise estimation. Simulated studies demonstrate these findings and confirm that multiple data sources can estimate the shared discrepancy function and source-dependent measurement bias. Finally, the new method has been implemented in the `RobustCalibration` package available on CRAN (Gu, 2022).

The rest of the paper is organized as follows. Section 2 introduces our approach that includes models of both the discrepancy function and measurement bias, as well as posterior sampling for Bayesian inference. Connections and differences between jointly modeling individual data sets and downsampled data, as well as different models of discrepancy functions, are also discussed. Simulated and real examples comparing several models are given in Section 3 and Section 4, respectively. We provide a short conclusion in Section 5. Lastly, supplementary materials contain extensive derivations of marginal likelihood functions, posterior distributions, theoretical results regarding the consistency of model

calibration, and numerical comparisons of different models of discrepancy functions and downsampling approaches. The code and data used in this article are publicly available: `https://github.com/UncertaintyQuantification/MultiCalibration`.

2 Model calibration by multiple sources of data

Let us consider the model of the lth source of real-valued field measurement, $y_l^F(\mathbf{x})$, with superscript 'F' meaning 'field,' to calibrate an imperfect model with the observable input $\mathbf{x} \in \mathbb{R}^{p_x}$ and unobservable calibration parameters $\boldsymbol{\theta} \in \mathbb{R}^{p_\theta}$,

$$y_l^F(\mathbf{x}) = f^M(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) + \delta_l(\mathbf{x}) + \mu_l + \epsilon_l(\mathbf{x}). \tag{1}$$

Here, μ_l, $\delta_l(\mathbf{x})$, and $\epsilon_l(\mathbf{x}) \sim \mathrm{N}(0, \sigma_{0l}^2)$ are the source-specific mean parameter, random measurement bias, and noise, respectively, for source $l = 1, ..., k$. $\delta(\mathbf{x})$ is a discrepancy term between reality and the computer model shared across data sources, which is independent of computer models and thus only depends on the observed input, as advocated in Kennedy and O'Hagan (2001). In our application, $\mathbf{x}$ is the spatial coordinates of an InSAR interferogram, $\boldsymbol{\theta}$ are physical model parameters listed in Table 1, and k is the number of InSAR interferograms. The satellite interferograms are taken at slightly different start and end dates to measure the ground deformation, and the true ground deformation is approximately the same in different interferograms. This motivates the inclusion of a discrepancy function $\delta(\mathbf{x})$ shared across all sources. On the other hand, the measurement bias $\delta_l(\mathbf{x})$ is distinct in each interferogram, as the atmospheric conditions were different when each of the interferograms were taken.

For a set of observable inputs $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, let us first assume that the marginal distributions of model discrepancy $\boldsymbol{\delta} = (\delta(\mathbf{x}_1), ..., \delta(\mathbf{x}_n))^T$ and measurement bias $\boldsymbol{\delta}_l = (\delta_l(\mathbf{x}_1), ..., \delta_l(\mathbf{x}_n))^T$ follow multivariate normal distributions: $\boldsymbol{\delta} \sim \mathrm{MN}(\mathbf{0}, \tau^2\mathbf{R})$ and $\boldsymbol{\delta}_l \sim \mathrm{MN}(\mathbf{0}, \sigma_l^2\mathbf{R}_l)$, respectively. The covariance matrix of the discrepancy is denoted as $\tau^2\mathbf{R}$, which contains the

variance parameter τ^2; the (i,j)th entry of the correlation matrix $\mathbf{R}$ is parameterized by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, while the (i,j)th entry of covariance matrix of the lth measurement bias is $\sigma_l^2 K_l(\mathbf{x}_i, \mathbf{x}_j)$. Here, σ_l^2 is the variance parameter for $l = 1, .., k$. We postpone the discussion of kernel functions and other models of discrepancy functions to Section 2.3.

In model (1), the physical reality, denoted as $y^R(\mathbf{x})$ at any coordinate $\mathbf{x}$, can be expressed as a summation of the mathematical model and discrepancy function, i.e., $y^R(\mathbf{x}) = f^M(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x})$, which follows the framework in Kennedy and O'Hagan (2001). The innovation in (1) is to explicitly model the measurement bias (spatially correlated pattern) contained in different sources of observations. Here we have two goals. The first goal is to estimate the calibration parameters, discrepancy function, and measurement bias. The second goal is to predict physical reality by combining the calibrated physical model with the discrepancy function.

InSAR measures ground displacements relative to a point assumed to be non-deforming, i.e., a spatial location assumed to have zero ground deformation, which introduces uncertainty. We therefore include an unknown mean parameter μ_l for each interferogram l in model (1) and estimate it using data. InSAR images may also contain long-wavelength "ramp" artifacts due to errors in satellite orbits, which may be corrected independently (for instance, using data from GPS sensors) (Simons and Rosen, 2007) or by estimation of linear or quadratic ramp parameters together with geophysical model parameters. For our case study, however, the geographic area of interest is relatively small, so we neglect these errors.

The closed-form marginal distributions and predictive distributions of the discrepancy, measurement bias, and the reality of model (1) are given in Section S2 of the supplementary materials. Since the computational complexity of each evaluation of the likelihood increases linearly to k, images are often averaged before modeling in geoscience studies to reduce computational cost. We first study the difference between these two ways of inference.

2.1 Model equivalence based on aggregated data and full data

First, let us consider the data with no correlated measurement bias $\delta_l(.)$, that is $y_l^F(\mathbf{x}) = y^R(\mathbf{x}) + \epsilon_l(\mathbf{x})$. Here, $y^R(\mathbf{x})$ is the unknown reality. The data can be modeled below

$$y_l^F(\mathbf{x}) = f^M(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) + \mu + \epsilon_l(\mathbf{x}), \tag{2}$$

where the independent noise follows $\epsilon_l(\mathbf{x}) \sim N(0, \sigma_0^2)$ for each $\mathbf{x}$, and $l = 1, ..., k$. We include the mean parameter μ here for scenarios where the mean of the reality is not directly modeled in the physical model f^M. Model (2) has been widely used in calibration of repeated experimental responses (Bayarri et al., 2007; Arendt et al., 2012b).

Denote $\bar{y}^F(\mathbf{x}) = \sum_{l=1}^k y_l^F(\mathbf{x})/k$ as the average value of the field data at the input $\mathbf{x}$. When (2) is assumed, the model of the aggregated data follows

$$\bar{y}^F(\mathbf{x}) = f^M(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}) + \mu + \bar{\epsilon}(\mathbf{x}), \tag{3}$$

where the noise independently follows $\bar{\epsilon}(\mathbf{x}) \sim N(0, \sigma_0^2/k)$ for each $\mathbf{x}$. In our real application, the InSAR interferograms are aligned on the same spatial coordinates $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Since the number of spatial coordinates is large, the uncertainty in aligning the spatial coordinates of InSAR interferograms is approximately negligible. Also, denote $\mathbf{y}_l^F = (y_l^F(\mathbf{x}_1), ..., y_l^F(\mathbf{x}_n))^T$ as the observations in source l and $\bar{\mathbf{y}}^F = (\sum_{l=1}^k y_l^F(\mathbf{x}_1)/k, ..., \sum_{l=1}^k y_l^F(\mathbf{x}_n)/k)^T$ as the aggregated data. In Lemma 1 below, we show the logarithm of the likelihood of the full data and the reduced data only differs by a constant relevant to the variance of the noise. The proof is given in Section S1 in the supplementary materials.

Lemma 1. *Integrating out* $\boldsymbol{\delta} \sim MN(\mathbf{0}, \tau^2\mathbf{R})$*, the natural logarithm of the marginal likelihood in model (2) follows*

$$\ell(\boldsymbol{\theta}, \mu, \sigma_0^2, \tau, \mathbf{R}) = c_{\sigma_0^2} + \bar{\ell}(\boldsymbol{\theta}, \mu, \sigma_0^2, \tau, \mathbf{R}), \tag{4}$$

where $c_{\sigma_0^2} = -\frac{n(k-1)}{2}\log(2\pi\sigma_0^2) - \frac{n}{2}\log(k) - \frac{\sum_{l=1}^k \sum_{i=1}^n (y_l^F(\mathbf{x}_i) - \bar{y}^F(\mathbf{x}_i))^2}{2\sigma_0^2}$ *and* $\bar{\ell}(\boldsymbol{\theta}, \mu, \sigma_0^2, \tau, \mathbf{R})$ *is the*

natural logarithm of the marginal likelihood model from (3):

$$\bar{\ell}(\boldsymbol{\theta}, \mu, \sigma_0^2, \tau, \mathbf{R}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\tilde{\boldsymbol{\Sigma}}\right| - \frac{(\bar{\mathbf{y}}^F - \mu\mathbf{1} - \mathbf{f}_{\boldsymbol{\theta}}^M)^T\tilde{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{y}}^F - \mu\mathbf{1} - \mathbf{f}_{\boldsymbol{\theta}}^M)}{2},$$

with $\tilde{\boldsymbol{\Sigma}} = \tau^2\mathbf{R} + \frac{\sigma_0^2}{k}\mathbf{I}_n$ *and* $\mathbf{f}_{\boldsymbol{\theta}}^M = (f_{\boldsymbol{\theta}}^M(\mathbf{x}_1, \boldsymbol{\theta}), ..., f_{\boldsymbol{\theta}}^M(\mathbf{x}_n, \boldsymbol{\theta}))^T$.

When the variance of the noise σ_0^2 is known, equation (4) implies that the estimation of calibration parameters, mean parameters, and discrepancy function based on the aggregated data and full data is the same, as the aggregated data $\bar{\mathbf{y}}^F$ is the sufficient statistics of these parameters (Casella and Berger, 2002). When σ_0^2 is unknown, the sufficient statistics are $\bar{\mathbf{y}}^F$ and s^2, where $s^2 = \sum_{l=1}^k \sum_{i=1}^n (y_l^F(\mathbf{x}_i) - \bar{y}^F(\mathbf{x}_i))^2$. This result was previously discussed in Bayarri et al. (2007). However, the efficiency of the estimators based on full and aggregated data was not compared. For instance, if the reality $y^R(\cdot)$ is a deterministic function, the usual unbiased estimator of σ_0^2 based on the full data is the sample variance $s^2/(n(k-1))$. Also, the variance when estimating σ_0^2 would be $2\sigma_0^4/(n(k-1))$. For aggregated data modeled in (3), even if the reality $y^R(\cdot)$ is known, the estimator of the variance σ_0^2 based on the sample variance is $k\sum_{i=1}^n(\bar{y}^F(\mathbf{x}_i) - y^R(\mathbf{x}_i))^2/(n-1)$, which has a variance of $2n\sigma_0^4/(n-1)^2$. This is larger than the sample variance $2\sigma_0^4/(n(k-1))$ based on full data. Since reality $y^R(\cdot)$ is unknown, the variance of the estimator of σ_0^2 based on the aggregated data typically becomes even larger. Thus, modeling aggregated data in (3) is less efficient in estimating σ_0^2, compared to modeling the full data in (2).

Second, for model (1) that contains the measurement bias, we have a similar result as Lemma 1. The logarithm of the likelihood of model (1) can be decomposed into two parts as in equation (4), where the first part contains individual data vectors $\mathbf{y}_1, ..., \mathbf{y}_k$, a weighted mean data vector $\bar{\mathbf{y}}_w = \sum_{l=1}^k(\mathbf{y}_l - \boldsymbol{\delta}_l - \mu_l\mathbf{1}_n)/(k\sqrt{\sigma_{0l}})$, and variances of the noises $\sigma_{01}^2, ...\sigma_{0k}^2$. The second part contains aggregated data $\bar{\mathbf{y}}_w$, calibration parameters $\boldsymbol{\theta}$, and covariance parameters. Thus, when the variances of the noises are known, parameter estimation based on the aggregated data $\bar{\mathbf{y}}_w$ and individual data sets is the same when the measurement bias is included. However, note that $\bar{\mathbf{y}}_w$ is different from the average $\bar{\mathbf{y}}$ and is generally not

observable. In our real application, this result indicates that using the averaged interferogram can lead to the loss of information if each interferogram contains distinct atmospheric errors, mean parameters, or unknown noise variance. Simulated studies in Section 3 further confirm this result.

2.2 Downsampling satellite interferograms

A single InSAR image is often composed of hundreds of thousands of pixels. Even for very simple geophysical models, the expense of computing deformation at all these points can be prohibitive, and subsampling techniques are typically employed. One approach is to uniformly sample a subset of pixels for calibration and prediction (Pritchard et al., 2002). As a result, posterior distributions of calibration parameters are often stable, with responses at only a few hundred pixels (Gu and Wang, 2018).

Another method of downsampling is the quadtree algorithm, in which one computes the average (or median) of groups of pixels ("boxes"), whose sizes are based on gradients in the image (Simons et al., 2002), the resolution of the forward model (Lohman and Simons, 2005), or both (Wang et al., 2014). The quadtree algorithm clusters the pixels in smaller boxes for regions with rapid changes in pixel values, while areas with less change are clustered in larger boxes. These algorithms have become widely used for modeling InSAR data (e.g., Montgomery-Brown et al., 2015; Anderson et al., 2019). The quadtree algorithm reduces the data from around a million pixels to a few hundred boxes, effectively reducing the number of observations by more than a thousand times. The quadtree algorithm may be considered a type of supervised data reduction method where the output values on the pixels are used for clustering. Other supervised algorithms for data reduction, such as those discussed in Joseph and Mak (2021), could also be useful.

Because the boxes in a quadtree-processed image, such as in Figure 1d, are computed from different numbers of pixels, in a calibration problem it is therefore important to consider the boxes' size (Simons et al., 2002; Lohman and Simons, 2005). However, this seems to have

been overlooked in many previous studies using quadtree-processed InSAR data for model calibration and prediction.

Suppose the lth quadtree-processed image is composed of J_l boxes, each box computed from $n_{j,l}$ pixels, for $j = 1, ..., J_l$ and $l = 1, ..., k$. Denote the lth quadtree-processed image by $\mathbf{y}_l^{F,Q} := \{y_{1,l}^{F,Q}, ..., y_{J_l,l}^{F,Q}\}$, where $y_{j,l}^{F,Q}$ is the average of the pixels of the jth quartree box for $j = 1, ..., J_l$. Let μ_l^Q, $f_{j,l}^{M,Q}(\boldsymbol{\theta})$, $\delta_{j,l}^Q$, and $\delta_{j,l}^Q$ denote the corresponding mean parameter, outputs of the mathematical model, discrepancy function, and measurement bias function evaluated at the centroid of the jth quadtree box at the lth quadtree-processed image. Consider the model of the quadtree-processed image l:

$$y_{j,l}^{F,Q} = f_{j,l}^{M,Q}(\boldsymbol{\theta}) + \mu_l^Q + \delta_{j,l}^Q + \delta_{j,l} + \epsilon_{j,l}^Q, \tag{5}$$

where $\epsilon_{j,l}^Q$ is a zero-mean Gaussian noise with variance $\frac{\sigma_0^2}{\omega_{j,l}}$ and $\omega_{j,l}$ is the weight for the jth image and lth source of data. Here we suppress the notation of spatial inputs as the point process data are compressed to areal data, and the correlation structure is defined between a finite set of areal units. Denote $S_{j,l}$ as the index set of pixels, where the pixels in this set belong to the jth quadtree box in the lth data source. The likelihood of model (5) of the quadtree-processed image is the same as that for model (1) of the original image if

$$\begin{aligned} &\sum_{i \in S_{j,l}} (y_l^F(\mathbf{x}_i) - f^M(\mathbf{x}_i, \boldsymbol{\theta}) - \delta(\mathbf{x}_i) - \delta_l(\mathbf{x}_i) - \mu_l)^2 \\ &\qquad = \omega_{j,l}(y_{j,l}^{F,Q} - f_{j,l}^{M,Q}(\boldsymbol{\theta}) - \delta_j^Q - \delta_{j,l}^Q - \mu_l^Q)^2. \end{aligned} \tag{6}$$

Other physical measurements may be useful for estimating discrepancy and measurement bias; however, in other cases, one may not know the discrepancy and measurement bias functions apriori. Some weighting schemes utilize the estimated covariance structure of the InSAR data (Lohman and Simons, 2005). In this work, we follow Simons et al. (2002) and Anderson and Poland (2016) by letting the weight of each quadtree box be proportional to the number of pixels in the box, i.e., $\omega_{j,l} \propto n_{j,l}$ for $j = 1, ..., J_l$. A quadtree box with a larger

size has a larger weight because it is averaged with more pixels.

2.3 Statistical models of discrepancy and measurement bias

We discuss specific discrepancy and measurement bias functions in this subsection. The discrepancy function is often modeled as a GaSP (Kennedy and O'Hagan, 2001):

$$\delta(\cdot) \sim \text{GaSP}(0, \tau^2 K(\cdot, \cdot)), \tag{7}$$

where $\tau^2 K(\cdot, \cdot)$ is a covariance function with variance τ^2. The identifiability issue, however, has been widely observed in modeling spatially correlated data, where the spatial random effect was confounded with a linear fixed effect, i.e., $f^M(\mathbf{x}, \boldsymbol{\theta})$ being a linear model of $\boldsymbol{\theta}$ (Reich et al., 2006; Hodges and Reich, 2010; Zhang, 2004). Wang et al. (2020), for example, show that the variance of generalized least squares estimator of the linear coefficients is bounded above zero under infill asymptotics. The non-identifiability of the calibration parameters was also recently observed when the discrepancy function was modeled by a GaSP (Arendt et al., 2012a; Tuo and Wu, 2015, 2016; Plumlee, 2017; Wong et al., 2017), where the calibrated physical models can be far away from reality in terms of L_2 distance.

Previous studies (Bayarri et al., 2007; Arendt et al., 2012b) suggest that repeated observations help identifiability as estimation accuracy of the variance of the data improves, which coincides with our discussion in Section 2.1. However, when the discrepancy is sampled from the true model, the MLE of the calibration parameter may not be consistent even if we have infinite repeated measurements, i.e., $k \to \infty$. An example is provided in Example S1 in the supplementary materials to illustrate this finding. This means that the repeated measurements are helpful in estimating the variance of the noise in the data, yet it cannot solve the identifiability issue.

When we model the discrepancy function by a GaSP with commonly used covariance functions, such as power exponential covariance or Matérn covariance (Rasmussen, 2006), the L_2

loss between the mathematical model and reality is $L_2(\boldsymbol{\theta}) = \int_{\mathbf{x}\in\mathcal{X}}(y^R(\mathbf{x}) - f^M(\mathbf{x},\boldsymbol{\theta}))^2 d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} \delta^2(\mathbf{x}) d\mathbf{x}$. $L_2(\boldsymbol{\theta})$ is a random variable whose measure is induced by the covariance function of the GaSP. The distribution of $L_2(\boldsymbol{\theta})$ can have a substantial probability mass at a large L_2 loss when the correlation in the data is large. In Gu and Wang (2018), this random L_2 loss is scaled to have more probability mass near zero. The construction of the discretized S-GaSP is summarized in S4 in the supplementary materials.

Starting from a GaSP model with any reasonable covariance function $\tau^2 K(\cdot,\cdot)$, and integrating out Z, the marginal distribution of the discretized GaSP $\boldsymbol{\delta}_z := (\delta_z(\mathbf{x}_1), ..., \delta_z(\mathbf{x}_n))^T$ follows a multivariate normal distribution with the following transformed covariance matrix:

$$\boldsymbol{\delta}_z \mid \tau, \mathbf{R}_z \sim \mathrm{MN}(\mathbf{0}, \tau^2 \mathbf{R}_z), \tag{8}$$

where $\mathbf{R}_z = \left(\mathbf{R}^{-1} + \frac{\lambda_z}{n}\mathbf{I}_n\right)^{-1}$ and the (i, j)th term of $\mathbf{R}$ is $K(\mathbf{x}_i, \mathbf{x}_j)$ with range parameters $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{p_x})^T$. Note that $\mathbf{R}_z$ is different from a covariance matrix with a nugget parameter.

A larger λ_z assigns more prior probability on the smaller sum of squares of the discrepancy function. Under some regularity conditions, a suitable choice of λ_z guarantees the predictive distribution in the S-GaSP model converges to the reality as fast as in the GaSP model. The estimation of the calibration parameters in the S-GaSP model minimizes the L_2 loss between the reality and mathematical model when the sample size goes to infinity. In numerical examples, we let $\lambda_z = C\sqrt{n}$, with $C = 100$, which guarantees two convergence properties under common regularity conditions (Gu et al., 2022).

Furthermore, the measurement biases are spatially correlated, which can be modeled as a spatial random effect via a GaSP,

$$\delta_l(\cdot) \sim \mathrm{GaSP}(0, \sigma_l^2 K_l(\cdot,\cdot)), \tag{9}$$

where $K_l(\cdot,\cdot)$ is the kernel function for $l = 1, ..., k$. For any $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, the marginal

distribution of $\boldsymbol{\delta}_l$ follows a multivariate normal distribution with covariance $\sigma_l^2\mathbf{R}_l$, for $l = 1, ..., k$, with the (i, j)th term of $\mathbf{R}_l$ being $K_l(\mathbf{x}_i, \mathbf{x}_j)$. For any inputs $\mathbf{x}_a := (x_{a1}, ..., x_{ap})$ and $\mathbf{x}_b := (x_{b1}, ..., x_{bp})$, we assume a product covariance (Bayarri et al., 2007)

$$K_l(\mathbf{x}_a, \mathbf{x}_b) = \prod_{t=1}^{p_x} K_{l,t}(x_{at}, x_{bt}), \tag{10}$$

where each $K_{l,t}(\cdot, \cdot)$ is a one-dimensional kernel function for the correlation between the t^{th} coordinate of any two inputs for the source l, $l = 1, ..., k$. Denote $d_t = |x_{at} - x_{bt}|$. One popular choice is the Matérn correlation, which has a closed form expression with the roughness parameter $\alpha = (2k + 1)/2$ for $k \in \mathbb{N}$. That is, the Matérn correlation function with $\alpha = 5/2$ (Handcock and Stein, 1993) has the expression below,

$$K_{l,t}(d_t) = \left(1 + \frac{\sqrt{5}d_t}{\gamma_{l,t}} + \frac{5d_t^2}{3\gamma_{l,t}^2}\right) \exp\left(-\frac{\sqrt{5}d_t}{\gamma_{l,t}}\right). \tag{11}$$

A desirable feature of the Matérn correlation is that the sample path of the process is $\lfloor \alpha \rfloor$ differentiable. Concerning the present scientific goal, we also note that previous works have argued that Matérn correlation functions are suitable for modeling atmospheric noise in InSAR data (Knospe and Jonsson, 2010). However, we do not limit ourselves to any specific correlation function, and the methods discussed in this work apply to all such functions.

2.4 Prior distributions and posterior sampling

We assume the calibration model follows (1) with data sets from multiple sources, where users can choose either GaSP or S-GaSP to model the discrepancy function. The marginal likelihood and predictive distributions are provided in Section S2 in the supplementary materials. Here, the parameters contain the calibration parameters, mean parameters, range and variance parameters of the discrepancy and measurement bias. For computational purposes, we define the nugget parameter $\eta_l := \sigma_{0l}^2/\sigma_l^2$, the inverse range parameters $\beta_t = 1/\gamma_t$, and

$\beta_{l,t} = 1/\gamma_{l,t}$, for $l = 1, ..., k$ and $t = 1, ..., p_x$.

We assume the prior of the parameters below:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\beta}_{1:k}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \boldsymbol{\beta}, \tau^2) \propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{\beta})}{\tau^2} \prod_{l=1}^{k} \left\{ \frac{\pi(\boldsymbol{\beta}_l, \eta_l)}{\sigma_l^2} \right\}, \tag{12}$$

where the prior of the calibration parameters $\boldsymbol{\theta}$ often depends on experts' knowledge, as the calibration parameters have scientific meanings. For the simulated examples and the real example of calibrating the geophysical model of Kīlauea Volcano, we assume $\pi(\boldsymbol{\theta})$ is a uniform distribution on the domain of the calibration parameters. The mean and scale parameters are assigned a usual location-scale prior in (12), i.e., $\pi(\tau^2) \propto 1/\tau^2$ and $\pi(\mu_l, \sigma_l^2) \propto 1/\sigma_l^2$, for $l = 1, ..., k$. Furthermore, we assume a jointly robust prior for the range and nugget parameters in the measurement bias functions and discrepancy function (Gu, 2018).

We have implemented the posterior sampling for Bayesian model calibration and prediction using single or multiple data sets in the `RobustCalibration` R package, available on CRAN. Users can specify the model with or without the measurement bias. Both the GaSP and discretized S-GaSP discrepancy models are implemented for users to choose as well. Furthermore, the geophysical model used in this study is computationally inexpensive, but that is often not the case (e.g., Anderson and Segall, 2011). In such cases, a statistical emulator can be used to approximate the expensive computer model. The GaSP emulator from the `RobustGaSP` package is implemented in the `RobustCalibration` package for emulating costly computer models with scalar or vectorized outputs. These tools can be used for different applications of model calibration and prediction.

3 Simulated examples

We study simulated examples of model calibration and prediction in this section. An example comparing the GaSP and S-GaSP models of the discrepancy function is provided in Section S5 of the supplementary materials, where the reality is from a deterministic function. Here

we discuss an example where the discrepancy function and measurement bias are sampled from GaSPs, and we compare model calibration based on full and aggregated data.

Example 1. *Assume data is sampled from model (1), where* $f^M(x,\theta) = \sin(\theta x)$*, with* $\theta = \pi/2$*. The model discrepancy and measurement bias are sampled from* $\boldsymbol{\delta} \sim \text{MN}(\mathbf{0}, \tau^2\mathbf{R})$ *and* $\boldsymbol{\delta}_l \sim \text{MN}(\mathbf{0}, \sigma_l^2\mathbf{R}_l)$*, where* $\tau = 0.2$ *and* $\sigma_l = 0.4 + 0.4(l-1)/(k-1)$*.* $\mathbf{R}$ *and* $\mathbf{R}_l$ *are both parameterized by the Matérn kernel in (11) with* $\gamma = 0.1$ *and* $\gamma_l = 0.02$*, respectively. The standard deviation of the noise is* $\sigma_{0l} = 0.05$*, and the observations are equally spaced at* $x_i \in [0,1]$ *for* $i = 1, ..., n$ *and* $l = 1, ..., k$*. We let* $n = 100$ *and implement* $N = 200$ *experiments at three configurations with* $k = 5$*,* $k = 10$*, and* $k = 15$*.*

Example 1 illustrates that modeling the individual data is more accurate than the aggregated data when measurement bias exists. We record the performance of three models. The first and second approaches are the GaSP calibration and S-GaSP calibration based on the full data, where the discrepancy function is modeled by GaSP in (7) and the discretized S-GaSP model with the marginal distribution in (8), respectively. Here the GaSP calibration model is the true sampling model, as the discrepancy is sampled from a GaSP. We include S-GaSP calibration model to illustrate that the S-GaSP model has comparable performance to the GaSP model even if the true discrepancy function is sampled from the GaSP calibration. Also included is the GaSP calibration using aggregated data, i.e., the averages of all sources of data. We draw $20,000$ posterior samples of the parameters for each approach, with the first $4,000$ posterior samples used as the burn-in samples. To reduce storage space, we thin posterior samples by ten times.

The mean square error (MSE) of measurement bias, discrepancy functions, reality, and the squared error (SE) of the calibration parameter of each experiment using different approaches for Example 1 are shown in Figure 2. First, even though the data is sampled from the GaSP calibration model, the MSEs of the S-GaSP calibration and GaSP calibration are similar in estimation. We are not trying to show S-GaSP calibration can outperform GaSP calibration in this example, as the GaSP calibration model is the true sampling model. We

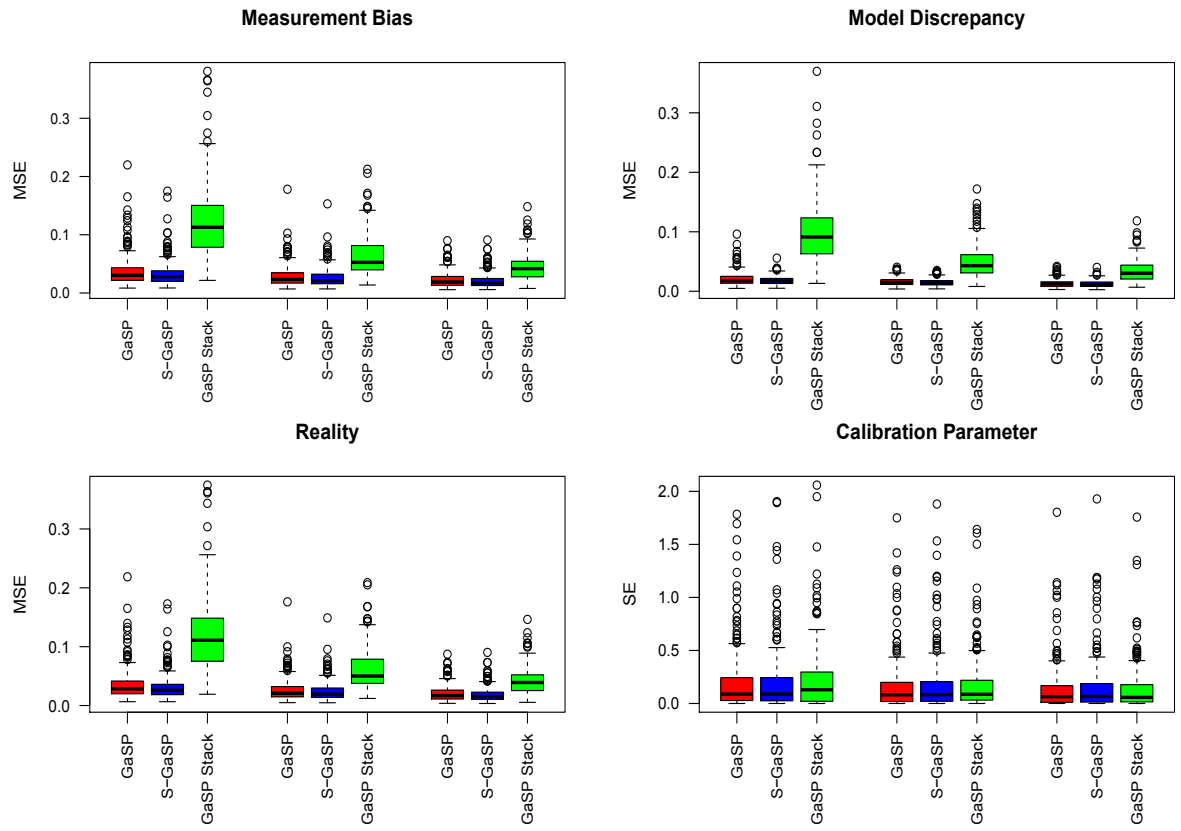

Figure 2: MSE of measurement bias discrepancy function, reality and SE of the calibration parameters for Example 1. In each panel, the first three boxes, the middle three boxes, and the right three boxes are the results when $k = 5$, $k = 10$, and $k = 15$, respectively. The MSE of GaSP calibration, S-GaSP calibration based on the full data, and the GaSP calibration based on the aggregated data are colored as red, blue, and green.

include a simulated study in Example S2 in the supplementary materials, to illustrate the identifiability problem of GaSP prior of the discrepancy function and the better performance of the S-GaSP calibration model. Second, both methods based on the full data are better than the GaSP calibration based on the aggregated data, as averaging different sources of data causes loss of information due to the presence of the measurement bias and the unknown variance of the noise, discussed in Section 2.1. The estimation of the calibration parameter by the three methods is similar. When the number of sources of data increases, all methods become more accurate in estimation. The decrease of SEs of the calibration parameter is small when the number of sources of the observations increases because of the relatively large variance and correlation in the measurement bias.

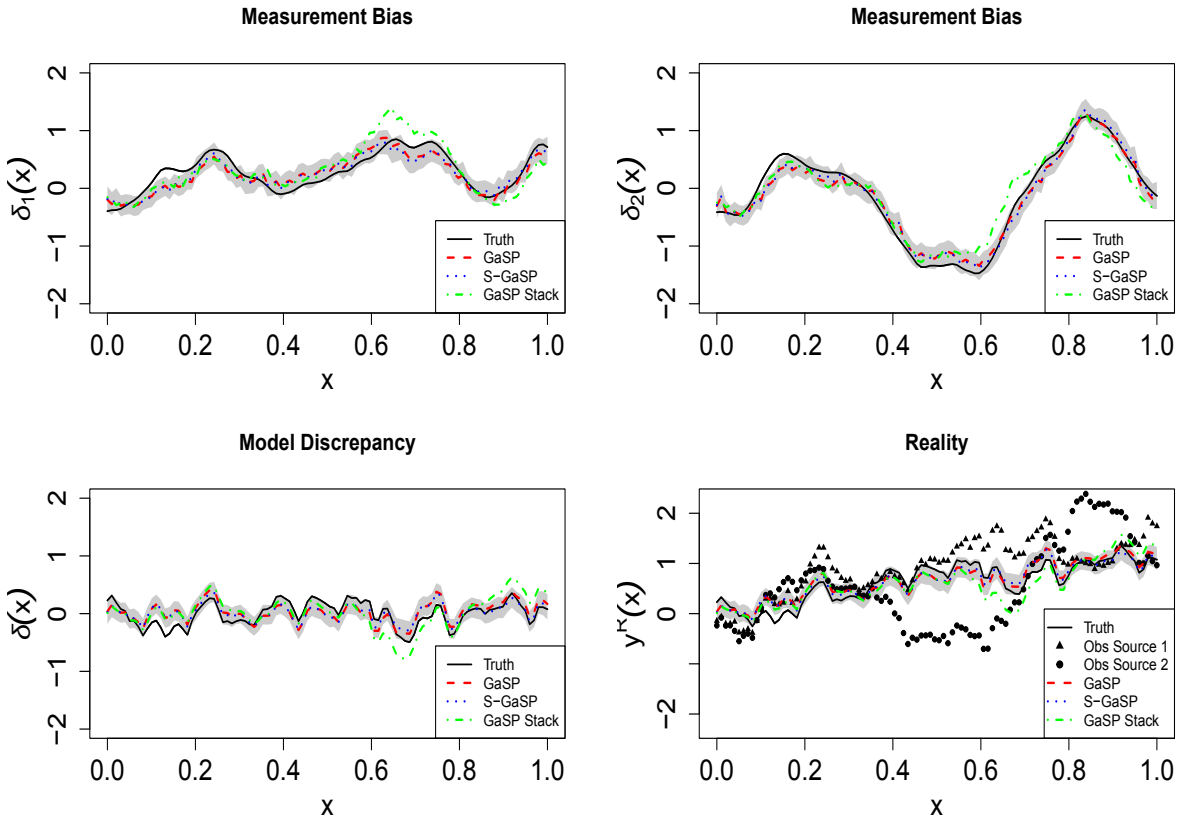

Figure 3: The measurement bias in the first two sources, the model discrepancy function, and the reality in the first experiment of Example 1 when $k = 10$ are graphed in the upper panels, lower left panel, and lower right panel, respectively. The truth and the estimation by the GaSP and S-GaSP calibrations based on the full data are graphed as the black solid lines, the red dashed lines, and blue dotted lines, respectively. The estimation of the GaSP calibration based on aggregated data, denoted as GaSP Stack, is graphed as the green dotted and dashed lines. The 95% posterior credible intervals from the S-GaSP calibration is graphed as the shaded area. The observations from the first two sources are graphed as black triangles and dots, respectively, in the lower right panel. The black, red, and blue lines almost overlap in all panels, indicating higher accuracy based on individual data than the aggregated data, when the observations contain measurement bias.

We graphed the measurement bias, the model discrepancy, reality, and their estimations in the first simulated experiment of Example 1 with $k = 10$ in Figure 3. All methods seem to capture the patterns of the measurement bias, model discrepancy, and reality. The estimation conducted by the GaSP and S-GaSP calibrations based on the full data are more accurate than the GaSP calibration using the aggregated data. This is because the true model contains the measurement bias and unknown variance of the noise. The inference based on

Table 1: Input variables, calibration parameters of the geophysical model, and other model parameters in calibration.

Input variables ($\mathbf{x}$)	Description
x_1	East-west spatial coordinate
x_2	North-south spatial coordinate
Calibration parameters ($\boldsymbol{\theta}$)	**Description**
$\theta_1 \in [-2000, 3000]$	East-west spatial coordinate of chamber centroid (m)
$\theta_2 \in [-2000, 5000]$	North-south spatial coordinate of chamber centroid (m)
$\theta_3 \in [500, 6000]$	Depth of the chamber (m)
$\theta_4 \in [0, 0.15]$	Volume change rate of the reservoir (m^3/s)
$\theta_5 \in [0.25, 0.33]$	Poisson's ratio (host rock property)
Model parameters	**Description**
$\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$	Mean parameters
$\boldsymbol{\beta}_{1:k} = (\beta_{1,1}, \beta_{1,2}, ..., \beta_{k,1}, \beta_{k,2})$	Inverse range parameters of the measurement bias
$\boldsymbol{\eta} = (\eta_1, ..., \eta_k)$	Nugget parameters of the measurement bias
$\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_k^2)$	Scale parameters of the measurement bias
$\boldsymbol{\beta} = (\beta_1, \beta_2)$	Range parameters of the model discrepancy
τ^2	Scale parameter of the model discrepancy

the full data is thus more precise in this scenario. Furthermore, Example 1 indicates that we can estimate the measurement bias and model discrepancy functions based on multiple sources of data.

4 Model calibration by multiple InSAR interferograms

InSAR data have been widely used at Kīlauea and other volcanoes to estimate the locations and volume changes of magma reservoirs and intrusions (e.g., Poland et al., 2014; Anderson et al., 2019). In this section, we study the performance of the aforementioned approaches in calibrating a geophysical model of Kīlauea Volcano using interferograms spanning late-2011 to mid-2012 (Figure 4). During this time, the summit of the volcano inflated due to the storage of magma supplied from the Earth's mantle (Anderson and Poland, 2016). Our goal is to use InSAR observations to obtain an improved characterization of the location of the magma reservoir and its volume change, which is important for hazard assessments and for resolving the rate of magma supply to the volcano. Our work extends previous analysis

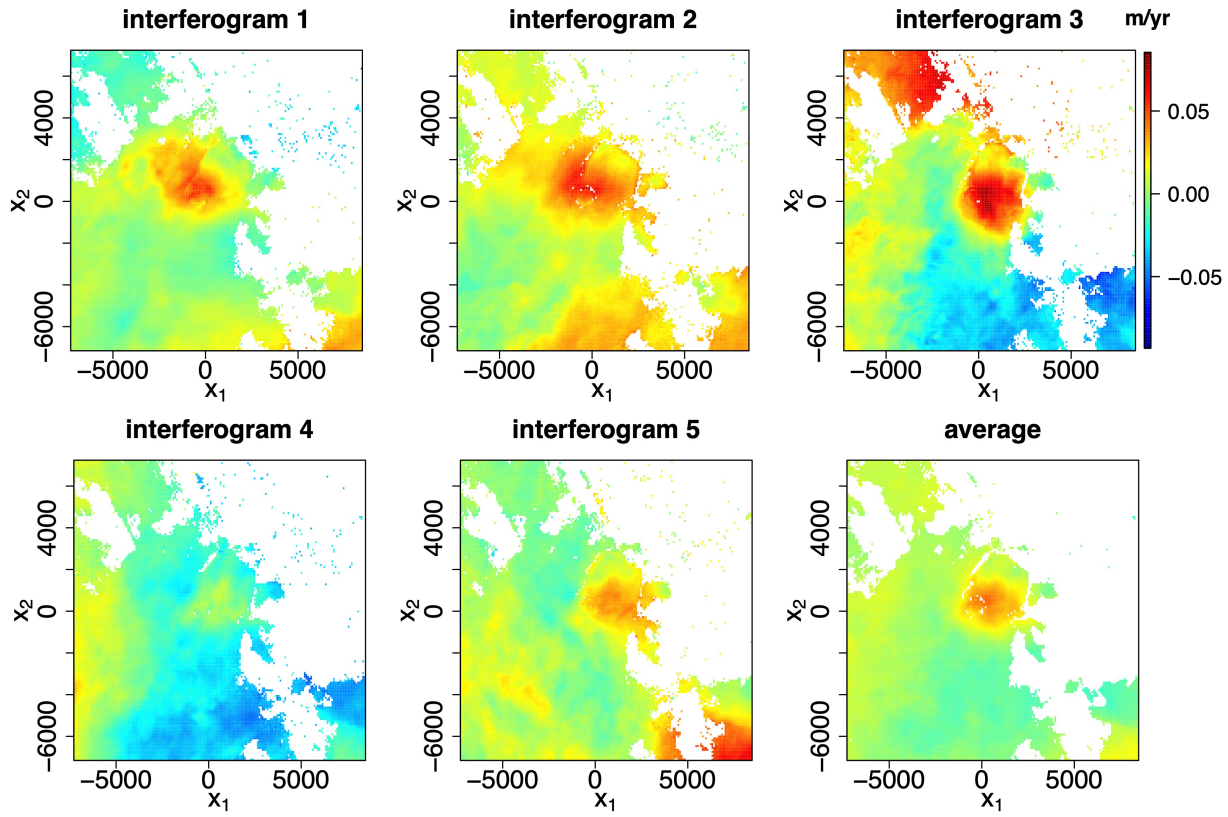

Figure 4: Five COSMO-SkyMed satellite interferograms spanning the following time periods: 1) 17 Oct 2011 - 04 May 2012; 2) 21 Oct 2011 - 16 May 2012; 3); 20 Oct 2011 to 15 May 2012; 4) 28 Oct 2011 to 11 May 2012; 5) 12 Oct 2011 - 07 May 2012. Interferograms 1 and 2 have an ascending-mode look angle, while the rest are descending-mode. Horizontal position is in meters relative to a chosen point in Kīlauea Caldera; vertical scale is m/yr. The last figure shows the stack (average) of 6 images.

(Anderson and Poland, 2016) by utilization of additional interferograms and consideration of spatially correlated data, noise, and discrepancy functions.

Consistent with past work, we model the InSAR data by a geophysical model of volume change of a spherical magma reservoir embedded in an elastic medium (Mogi, 1958; Anderson and Poland, 2016), parameterized by the 3D location of its centroid (east distance, north distance, and depth), its volume change, and the Poisson's ratio of the elastic medium. The input variables, calibration parameters, and other model parameters are listed in Table 1 for calibrating the geophysical model of Kīlauea Volcano using multiple interferograms.

Ground deformation velocities computed from five interferograms captured by the COSMO-

SkyMed satellite spanning late 2011 to mid-2012 are shown in Figure 4 (in our model, the rate of ground deformation is assumed constant over the complete time range, and the small misalignment between the start and end dates across different interferograms should have only negligible effects). We notice a relatively large measurement bias from the atmospheric error in interferograms, which makes using multiple interferograms necessary for model calibration.

We use model (1), which includes the measurement bias term and the discrepancy function, and we compare the difference between using a GaSP and an S-GaSP model for the discrepancy function. The range of the calibration parameters using GaSP and S-GaSP calibrations are given in Table 1. In both models, 50,000 posterior samples are drawn with the first 10,000 posterior samples used as the burn-in samples. The posterior samples in every 10th step are saved to reduce storage space and autocorrelation in Markov chains. We present the results based on 400 uniformly sampled pixels here, and the results using quadtree subsampling and stacked interferograms in Section S6 in the supplementary materials.

Figure 5 graphs the posterior samples of the model calibration parameters. Estimates of the chamber depth (θ_3) and volume change rate (θ_4) are larger when the discrepancy function is modeled using a GaSP than with an S-GaSP. This is because, for a given variance, the GaSP prior places a large probability mass on smooth discrepancy functions. Here the deeper magma chamber and larger volume change rate from the GaSP calibration yield a discrepancy function with relatively smooth ground deformation over a large region. In comparison, the estimated depth and volume change rate by S-GaSP is more consistent with other studies using different sources of data (Poland et al., 2014), as the S-GaSP prior has more probability mass on smaller L_2 loss between reality and the computer model.

The MSE of the predictions on each interferogram based on 400 uniformly sampled pixels are given in Table 2. The mean parameters $\boldsymbol{\mu}$ are treated as a part of the geophysical model for making predictions. As shown in Figure 6 and Figure S3, the predictive mean using both the calibrated computer model, discrepancy, and measurement bias terms is accurate

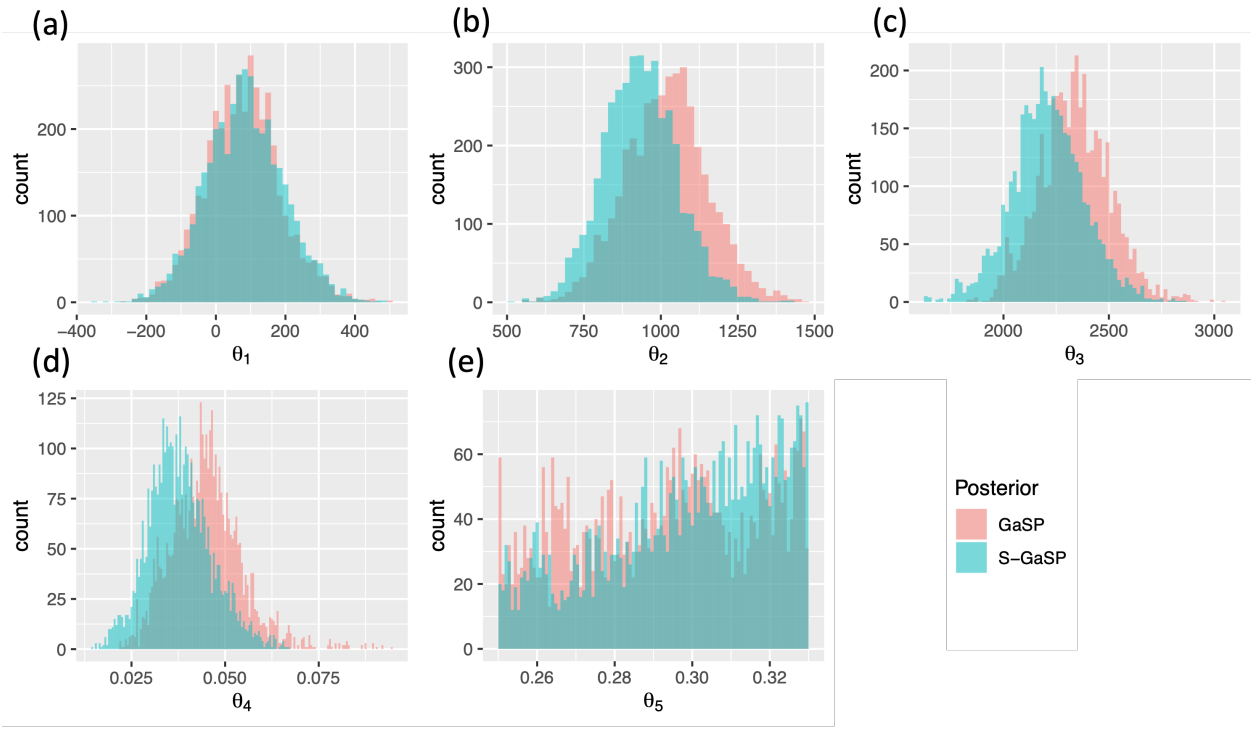

Figure 5: The posterior samples of $\boldsymbol{\theta}$ in the GaSP and S-GaSP calibrations. The range of the parameter θ_5 (Poisson's ratio) is consistent with many rock types, but the geophysical model is relatively insensitive to this parameter (Mogi, 1958).

in both GaSP and S-GaSP calibration. However, the computer model calibrated by S-GaSP is more accurate than the GaSP, as shown in Figure S6 in the supplementary materials. In S-GaSP calibration, the prior of the discrepancy has more probability mass near zero, allowing the calibrated geophysical model to explain more variability in the observations than GaSP calibration. Consequently, the calibrated geophysical model by S-GaSP is more accurate in prediction (at points not used for calibration) than the GaSP-calibrated model, shown in the first two rows of Table 2. Furthermore, we also explored other pioneering methods, such as LS and L_2 calibration (Tuo and Wu, 2015; Wong et al., 2017). However, the convergence of numerical optimization seems to be a challenging issue when there are multiple sources of data. In addition, the measurement bias, which is the focus of this work, was not considered in previous methods.

The predictive mean of each interferogram and stacked (averaged) interferogram from the S-GaSP calibration is shown in Figure 6. Predictions are very close to the real interferogram

Table 2: Predictive mean squared error (MSE) in the prediction of the full interferograms using the GaSP and S-GaSP models based on 400 uniformly sampled pixels in each interferogram. MSE_{f^M} is the MSE using the calibrated geophysical model for prediction, and $\text{MSE}_{f^M+\delta+\delta_l}$ is the MSE using the combined calibrated geophysical model, discrepancy function, and measurement bias for prediction. The number is by 10^{-4}. Bold font indicates a smaller error.

MSE_{f^M}	image 1	image 2	image 3	image 4	image 5
GaSP	1.26	1.63	7.80	4.33	1.97
S-GaSP	**1.21**	**1.45**	**7.66**	**4.05**	**1.76**
$\text{MSE}_{f^M+\delta+\delta_l}$	image 1	image 2	image 3	image 4	image 5
GaSP	0.116	0.115	**0.264**	0.134	**0.120**
S-GaSP	**0.109**	**0.112**	0.267	**0.131**	0.123

(with around 99% of pixels held out), and performance is better than the GaSP calibration method. In the supplementary materials, we provide estimated measurement bias and model discrepancy for GaSP and S-GaSP calibrations, a detailed comparison between GaSP and S-GaSP, and trace plots of all the parameters.

Finally, we compare our results with previous studies of Kīlauea Volcano. The second and third interferograms shown in Figure 4 were also used for calibrating the same geophysical model as part of a broader geophysical study (Anderson and Poland, 2016). However, that work did not consider spatially correlated noise in the data or a discrepancy function. The same two interferograms were used in Gu and Wang (2018) for calibration with a discrepancy function, but the interferogram measurement bias was neglected. Of all the images, the ones used in the previous studies show the largest apparent volcanic ground displacement. As a result, the reservoir volume change rate (θ_4) we estimate here in the S-GaSP calibration using all five images is smaller than in those studies (0.02 m^3/s vs. 0.04–0.05 m^3/s, respectively).

The estimated reservoir position depends on the spatial pattern of displacement but not the rate. We estimate a reservoir location ∼500 m east and ∼800 m north of the reference position (southeast rim of Halema‘uma‘u Crater) at 1.9 km depth. Many previous studies have examined reservoir locations using a variety of data sets over many years. Despite the relatively low signal-to-noise ratio in the data, our estimated depth is consistent with these studies (e.g., Poland et al. (2014)). The horizontal position of our most likely reservoir

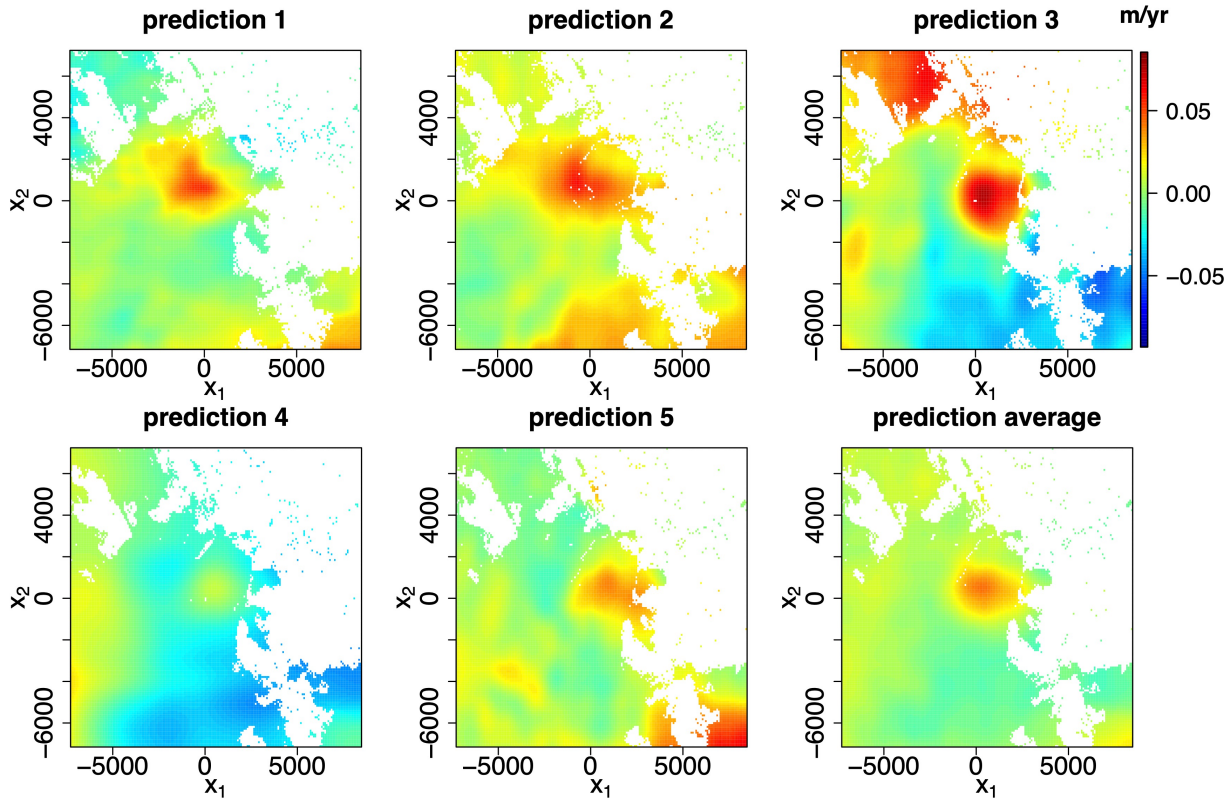

Figure 6: Predictive mean of each interferogram and stack image in S-GaSP calibration.

centroid is several hundred meters north of the most commonly accepted location near the east rim of Halemaʻumaʻu Crater. However, it is closer than the position estimated previously in Anderson and Poland (2016) and Gu and Wang (2018) using two interferograms without modeling the measurement bias; this confirms the importance of addressing the uncertainty in the measurement bias. Future studies may combine not only multiple measurements of a single type but also multiple types of data (for instance, GPS or ground tilt), and may also utilize more sophisticated geophysical models with larger numbers of calibration parameters. Furthermore, advanced sampling algorithms and design techniques (e.g., Mak and Joseph (2018)) may be used to represent the posterior distribution.

5 Concluding remarks

We have introduced a statistical framework to estimate measurement bias, model discrepancy, and calibration parameters using multiple sources of data. In addition, we studied data reduction by aggregating different sources of data and reducing the number of observations in one source of data. We have shown that modeling the full data is more efficient than the aggregated data when either variance of the noise or measurement bias is unknown. We have also shown that certain data reduction approaches, such as the quadtree algorithm, can be very useful for reducing the computational cost of modeling the InSAR interferograms for volcanic hazard quantification. Numerical results based on simulated experiments and real observations validate these findings.

There are several possible future extensions. First, the computation based on the full data scales linearly with the number of data sources when the inputs (i.e., spatial coordinates in InSAR interferograms) are aligned. When the inputs are misaligned, it will be helpful to design an algorithm for scalable computation. Second, quadtree processing is used widely to downsample satellite interferograms. A theoretical study on how to model quadtree images that properly takes into account the size of the boxes and the measurement bias will be beneficial. It will also be interesting to study whether quadtree-processed images improve calibration and prediction accuracy compared to alternate designs (e.g., Fukushima et al., 2005). Lastly, for volcanological applications, more work is required to fuse diverse data types, such as gas emissions, GPS data, and InSAR data, with geophysical models for Bayesian inversion.

Supplementary materials

The supplementary materials contain 6 sections. Section S1 and S2 give the proof of Lemma 1 and derivations of the marginal likelihood and predictive distribution of the calibration model, respectively. An example to illustrate the inconsistent maximum likelihood estima-

tion of the GaSP calibration model is discussed Section S3. We introduce the discretization of S-GaSP calibration model in Section S4, and provide additional numerical results for simulation studies and real data analysis in Section S5 and Section S6, respectively.

Acknowledgements

This research is supported by the National Science Foundation under Award No. DMS-2053423. The authors thank the editor, AE, Chuck Wicks, and two referees for their comments that substantially improved the article. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Anderson, K. and Segall, P. (2011). Physics-based models of ground deformation and extrusion rate at effusively erupting volcanoes. *Journal of Geophysical Research: Solid Earth*, 116(B7):B07204.

Anderson, K. R., Johanson, I. A., Patrick, M. R., Gu, M., Segall, P., Poland, M. P., Montgomery-Brown, E. K., and Miklius, A. (2019). Magma reservoir failure and the onset of caldera collapse at Kīlauea volcano in 2018. *Science*, 366(6470):eaaz1822.

Anderson, K. R. and Poland, M. P. (2016). Bayesian estimation of magma supply, storage, and eruption rates using a multiphysical volcano model: Kīlauea volcano, 2000–2012. *Earth and Planetary Science Letters*, 447:161–171.

Arendt, P. D., Apley, D. W., and Chen, W. (2012a). Quantification of model uncertainty: Calibration, model discrepancy, and identifiability. *Journal of Mechanical Design*, 134(10):100908.

Arendt, P. D., Apley, D. W., Chen, W., Lamb, D., and Gorsich, D. (2012b). Improving

identifiability in model calibration using multiple responses. *Journal of Mechanical Design*, 134(10):100909.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49(2):138–154.

Bürgmann, R., Rosen, P. A., and Fielding, E. J. (2000). Synthetic Aperture Radar Interferometry to Measure Earth's Surface Topography and Its Deformation. *Annual Review of Earth and Planetary Sciences*, 28(1):169–209.

Casella, G. and Berger, R. L. (2002). *Statistical Inference, 2nd Edition*. Cengage Learning.

Chen, C. W. and Zebker, H. A. (2001). Two-dimensional phase unwrapping with use of statistical models for cost functions in nonlinear optimization. *Journal of the Optical Society of America A*, 18(2):338.

Fukushima, Y., Cayol, V., and Durand, P. (2005). Finding realistic dike models from interferometric synthetic aperture radar data: The February 2000 eruption at Piton de la Fournaise. *Journal of Geophysical Research: Solid Earth*, 110(B3):B03206.

Gong, W., Lu, Z., and Meyer, F. (2016). Uncertainties in estimating magma source parameters from InSAR observation. In Riley, K., Webley, P., and Thompson, M., editors, *Natural Hazard Uncertainty Assessment: Modeling and Decision Support, Geophysical Monograph 223*, chapter 7, pages 89–104. John Wiley & Sons, Inc.

Gu, M. (2018). Jointly robust prior for Gaussian stochastic process in emulation, calibration and variable selection. *Bayesian Analysis*, 14(1).

Gu, M. (2022). Robustcalibration: Robust calibration of computer models in R. *arXiv preprint arXiv:2201.01476*.

Gu, M. and Wang, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1555–1583.

Gu, M., Xie, F., and Wang, L. (2022). A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1435–1460.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of Kriging. *Technometrics*, 35(4):403–410.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Hooper, A., Segall, P., and Zebker, H. (2007). Persistent Scatterer InSAR for Crustal Deformation Analysis, with Application to Volcan Alcedo, Galapagos. *Journal of Geophysical Research*, 112:1–19.

Jonsson, S., Zebker, H., Segall, P., and Amelung, F. (2002). Fault Slip Distribution of the 1999 Mw 7.1 Hector Mine, California, Earthquake, Estimated from Satellite Radar and GPS Measurements. *Bulletin of the Seismological Society of America*, 92(4):1377–1389.

Joseph, V. R. and Mak, S. (2021). Supervised compression of big data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(3):217–229.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Knospe, S. and Jonsson, S. (2010). Covariance estimation for dinsar surface deformation

measurements in the presence of anisotropic atmospheric noise. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4):2057–2065.

Lohman, R. B. and Simons, M. (2005). Some thoughts on the use of insar data to constrain models of surface deformation: Noise structure and data downsampling. *Geochemistry, Geophysics, Geosystems*, 6(1):Q01007.

Mak, S. and Joseph, V. R. (2018). Support points. *Annals of Statistics*, 46(6A):2562–2592.

Massonnet, D., Briole, P., and Arnaud, A. (1995). Deflation of Mount Etna monitored by spaceborne radar interferometry. *Nature*, 375(6532):567–570.

Mogi, K. (1958). Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. *Bulletin of the Earthquake Research Institute, University of Tokyo*, 36:99–134.

Montgomery-Brown, E. K., Wicks, C. W., Cervelli, P. F., Langbein, J. O., Svarc, J. L., Shelly, D. R., Hill, D. P., and Lisowski, M. (2015). Renewed inflation of Long Valley Caldera, California (2011 to 2014). *Geophysical Research Letters*, 42(13):5250–5257.

Neal, C. A., Brantley, S., Antolik, L., Babb, J., Burgess, M., Calles, K., Cappos, M., Chang, J., Conway, S., Desmither, L., et al. (2019). The 2018 rift eruption and summit collapse of Kīlauea volcano. *Science*, 363(6425):367–374.

Pinel, V., Poland, M., and Hooper, A. (2014). Volcanology: Lessons learned from Synthetic Aperture Radar imagery. *Journal of Volcanology and Geothermal Research*, 289:81–113.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285.

Poland, M. P., Miklius, A., and Montgomery-Brown, E. K. (2014). Magma Supply, Storage, and Transport at Shield-Stage Hawaiian Volcanoes. In *Characteristics of Hawaiian Volcanoes*, chapter 5, pages 179–234. U.S. Geological Survey Professional Paper 1801.

Pritchard, M., Simons, M., Rosen, P., Hensley, S., and Webb, F. (2002). Co-seismic slip from the 1995 July 30 Mw = 8.1 Antofagasta, Chile, earthquake as constrained by InSAR and GPS observations. *Geophysical Journal International*, 150(2):362–376.

Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT Press.

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206.

Simons, M., Fialko, Y., and Rivera, L. (2002). Coseismic deformation from the 1999 Mw 7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. *Bulletin of the Seismological Society of America*, 92(4):1390–1402.

Simons, M. and Rosen, P. A. (2007). Interferometric synthetic aperture radar geodesy. In *Treatise on Geophysics*, volume 3, pages 391–446. Elsevier Press.

Tuo, R. and Wu, C. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

Tuo, R. and Wu, C. J. (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795.

Wang, C., Ding, X., Li, Q., and Jiang, M. (2014). Equation-based InSAR data quadtree downsampling for earthquake slip distribution inversion. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2060–2064.

Wang, W., Tuo, R., and Jeff Wu, C. (2020). On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930.

Wong, R. K., Storlie, C. B., and Lee, T. (2017). A frequentist approach to computer model

calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:635–648.

Zebker, H. A., Rosen, P. A., and Hensley, S. (1997). Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. *Journal of Geophysical Research: Solid Earth*, 102(B4):7547–7563.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Supplement of "Calibration of imperfect geophysical models by multiple satellite interferograms with measurement bias"

All the formulas in the supplementary materials are cross-referenced in the main body of the article.

S1 Proof of Lemma 1

Proof of Lemma 1. Note that the likelihood of the model for the full data in (2) follows

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\delta}, \boldsymbol{\theta}, \mu, \sigma_0^2) &= (2\pi\sigma_0^2)^{-nk/2} \exp\left(-\frac{\sum_{l=1}^{k} (\mathbf{y}_l - \boldsymbol{\delta} - \mathbf{f}_\theta^M - \mu \mathbf{1}_n)^T (\mathbf{y}_l - \boldsymbol{\delta} - \mathbf{f}_\theta^M - \mu \mathbf{1}_n)}{2\sigma_0^2} \right) \\ &= (2\pi\sigma_0^2)^{-n(k-1)/2} k^{-n/2} \exp\left(-\frac{\sum_{l=1}^{k} (\mathbf{y}_l - \bar{\mathbf{y}})^T (\mathbf{y}_l - \bar{\mathbf{y}})}{2\sigma_0^2} \right) \\ &\quad \times (2\pi\sigma_0^2/k)^{-n/2} \exp\left(-\frac{k(\bar{\mathbf{y}} - \boldsymbol{\delta} - \mathbf{f}_\theta^M - \mu \mathbf{1}_n)^T (\bar{\mathbf{y}} - \boldsymbol{\delta} - \mathbf{f}_\theta^M - \mu \mathbf{1}_n)}{2\sigma_0^2} \right) \end{aligned}$$

Marginalizing out $\boldsymbol{\delta}$ based on prior $\boldsymbol{\delta} \sim \text{MN}(\mathbf{0}, \tau^2\mathbf{R})$, we have

$$
\begin{aligned}
& p(\mathbf{y} \mid \boldsymbol{\theta}, \mu, \sigma_0^2, \tau^2, \mathbf{R}) \\
=& (2\pi\sigma_0^2)^{-n(k-1)/2} k^{-n/2} \exp\left(-\frac{\sum_{l=1}^{k} (\mathbf{y}_l - \bar{\mathbf{y}})^T (\mathbf{y}_l - \bar{\mathbf{y}})}{2\sigma_0^2} \right) \\
& \times (2\pi)^{-n/2} \left| \tau^2\mathbf{R} + \frac{\sigma_0^2}{k}\mathbf{I}_n \right|^{-1/2} \exp\left(-\frac{(\bar{\mathbf{y}} - \mathbf{f}_\theta^M - \mu\mathbf{1}_n)^T (\tau^2\mathbf{R} + \frac{\sigma_0^2}{k}\mathbf{I}_n)^{-1} (\bar{\mathbf{y}} - \mathbf{f}_\theta^M - \mu\mathbf{1}_n)}{2} \right) \\
=& (2\pi\sigma_0^2)^{-n(k-1)/2} k^{-n/2} \exp\left(-\frac{\sum_{l=1}^{k} (\mathbf{y}_l - \bar{\mathbf{y}})^T (\mathbf{y}_l - \bar{\mathbf{y}})}{2\sigma_0^2} \right) \bar{p}(\bar{\mathbf{y}} \mid \boldsymbol{\theta}, \mu, \sigma_0^2, \tau^2, \mathbf{R})
\end{aligned}
$$

where $\bar{p}(\bar{\mathbf{y}} \mid \boldsymbol{\theta}, \mu, \sigma_0^2, \tau^2, \mathbf{R})$ is the marginal density of aggregated data after integrating out $\boldsymbol{\delta}$, from which we have concluded the proof. □

S2 Marginal distribution and predictive distribution of the multi-calibration model

The observations from the lth source of data are denoted as $\mathbf{y}_l^F = \left(y_l^F(\mathbf{x}_1), ..., y_l^F(\mathbf{x}_n)\right)^T$ at $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, for $l = 1, ..., k$. For the InSAR interferograms, each entry of $\mathbf{y}_l^F$ represents a line-of-sight displacement at a point on the Earth's surface. The following lemma gives the marginal distribution of model (1) after integrating out the random measurement bias functions.

Lemma S1. *After integrating out $\boldsymbol{\delta}_l$ in model (1), $l = 1, ..., k$, one has the following distributions:*

1. *For each source $l = 1, ..., k$, the marginal distribution of the field data follows a multivariate normal distribution*

$$
(\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2) \sim \text{MN}(\mathbf{f}_\theta^M + \mu_l\mathbf{1}_n + \boldsymbol{\delta}, \sigma_l^2\mathbf{R}_l + \sigma_{0l}^2\mathbf{I}_n), \tag{S1}
$$

where $\mathbf{f}_{\boldsymbol{\theta}}^M = (f^M(\mathbf{x}_1, \boldsymbol{\theta}), ..., f^M(\mathbf{x}_n, \boldsymbol{\theta}))^T$ *and* $\boldsymbol{\delta} = (\delta(\mathbf{x}_1), ..., \delta(\mathbf{x}_n))^T$*, with* $\mathbf{I}_n$ *being an* $n \times n$ *identity matrix.*

2. *The marginal posterior distribution of the discrepancy function follows a multivariate normal distribution*

$$\left(\boldsymbol{\delta} \mid \{y_l^F, \sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \sim \mathrm{MN}(\hat{\boldsymbol{\delta}},\, \hat{\boldsymbol{\Sigma}}), \tag{S2}$$

where

$$\hat{\boldsymbol{\Sigma}} = \left(\sum_{l=1}^k \tilde{\boldsymbol{\Sigma}}_l^{-1} + (\tau^2\mathbf{R})^{-1}\right)^{-1} \quad \textit{and} \quad \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\Sigma}} \sum_{l=1}^k \tilde{\boldsymbol{\Sigma}}_l^{-1} \tilde{\mathbf{y}}_l,$$

with

$$\tilde{\mathbf{y}}_l = \mathbf{y}_l - \mathbf{f}_{\boldsymbol{\theta}}^M - \mu_l \mathbf{1}_n \quad \textit{and} \quad \tilde{\boldsymbol{\Sigma}}_l = \sigma_l^2 \mathbf{R}_l + \sigma_{0l}^2 \mathbf{I}_n.$$

Proof of Lemma S1. Marginalizing out δ_l, by the laws of the total expectation and total variance, the marginal distribution of $\mathbf{y}_l^F$ is a multivariate normal distribution with the mean

$$\begin{aligned}
\mathbb{E}[\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2] &= \mathbb{E}[\mathbb{E}[\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2,\, \boldsymbol{\delta}_l]] \\
&= \mathbb{E}[\mathbf{f}_\theta^M + \mu_l \mathbf{1}_n + \boldsymbol{\delta} + \boldsymbol{\delta}_l \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2] \\
&= \mathbf{f}_\theta^M + \mu_l \mathbf{1}_n + \boldsymbol{\delta},
\end{aligned}$$

and the covariance matrix

$$
\begin{aligned}
&\mathbb{V}[\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2] \\
=&\,\mathbb{V}[\mathbb{E}[\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2,\, \boldsymbol{\delta}_l]] + \mathbb{E}[\mathbb{V}[\mathbf{y}_l^F \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2,\, \boldsymbol{\delta}_l]] \\
=&\,\mathbb{V}[\mathbf{f}_\theta^M + \mu_l \mathbf{1}_n + \boldsymbol{\delta} + \boldsymbol{\delta}_l \mid \boldsymbol{\delta},\, \boldsymbol{\theta},\, \sigma_l^2, \mathbf{R}_l,\, \mu_l,\, \sigma_{0l}^2] + \sigma_{0l}^2 \mathbf{I}_n \\
=&\,\sigma_l^2 \mathbf{R}_l + \sigma_{0l}^2 \mathbf{I}_n,
\end{aligned}
$$

from which (S1) follows.

After marginalizing out $\{\boldsymbol{\delta}_l\}_{l=1}^k$, the posterior distribution of $\boldsymbol{\delta}$ follows a multivariate normal distribution with the mean and covariance matrix given in (S2), from which the proof is complete.

□

Further marginalizing out $\boldsymbol{\delta}$, the marginal distribution of the field data $\mathbf{Y}_v^F := ((\mathbf{y}_1^F)^T, ..., (\mathbf{y}_k^F)^T)^T$ is given in the following lemma.

Lemma S2. *After integrating out both $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_l$, $l = 1, ..., k$, the marginal distribution of the field data follows a multivariate normal distribution*

$$
\left(\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \sim \mathrm{MN}(\mathbf{1}_k \otimes \mathbf{f}_{\boldsymbol{\theta}}^M + \boldsymbol{\mu} \otimes \mathbf{1}_n^T,\, \tau^2 \mathbf{1}_k \mathbf{1}_k^T \otimes \mathbf{R} + \boldsymbol{\Lambda}), \tag{S3}
$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$, and $\boldsymbol{\Lambda}$ is a $kn \times kn$ block diagonal matrix, with the lth diagonal block being $\tilde{\boldsymbol{\Sigma}}_l$ defined in Lemma S1, $l = 1, ..., k$. The density

of (S3) can be expressed as

$$
\begin{aligned}
& p\left(\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \\
&= (2\pi)^{-\frac{nk}{2}} \tau^{-n} |\mathbf{R}|^{-\frac{1}{2}} |\hat{\boldsymbol{\Sigma}}|^{\frac{1}{2}} \prod_{l=1}^{k} |\tilde{\boldsymbol{\Sigma}}_l|^{-\frac{1}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2} \left(\sum_{l=1}^{k} \tilde{\mathbf{y}}_l^T \tilde{\boldsymbol{\Sigma}}_l^{-1} \tilde{\mathbf{y}}_l - (\sum_{l=1}^{k} \tilde{\boldsymbol{\Sigma}}_l^{-1} \tilde{\mathbf{y}}_l)^T \hat{\boldsymbol{\Sigma}}^{-1} (\sum_{l=1}^{k} \tilde{\boldsymbol{\Sigma}}_l^{-1} \tilde{\mathbf{y}}_l) \right) \right\},
\end{aligned} \tag{S4}
$$

where $\hat{\boldsymbol{\Sigma}}$, $\tilde{\mathbf{y}}_l$, *and* $\tilde{\boldsymbol{\Sigma}}_l$ *are defined in Lemma S1, for* $l = 1, ..., k$.

Proof of Lemma S2. After marginalizing out δ_l, $l = 1, ..., k$, one has

$$
\left(\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \sim \mathrm{MN}(\mathbf{1}_k \otimes \mathbf{f}_\theta^M + \boldsymbol{\mu} \otimes \mathbf{1}_n + \mathbf{1}_k \otimes \boldsymbol{\delta},\, \boldsymbol{\Lambda}), \tag{S5}
$$

where $\boldsymbol{\Lambda}$ is a block diagonal matrix with the lth diagonal block being $\tilde{\boldsymbol{\Sigma}}_l$, $l = 1, ..., k$.

Note $(\boldsymbol{\delta} \mid \tau^2, \mathbf{R}) \sim \mathrm{MN}(\mathbf{0}, \tau^2 \mathbf{R})$. Further marginalizing out $\boldsymbol{\delta}$, the marginal distribution of $(\mathbf{Y}^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R})$ follows a multivariate normal distribution, with the mean

$$
\begin{aligned}
\mathbb{E}[\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}] &= \mathbb{E}[\mathbb{E}[\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau^2, \mathbf{R}]] \\
&= \mathbf{1}_k \otimes \mathbf{f}_\theta^M,
\end{aligned}
$$

and the posterior variance

$$
\begin{aligned}
& \mathbb{V}[\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}] \\
=& \mathbb{V}[\mathbb{E}[\mathbf{Y}_v^F \mid \{y_l^F, \tilde{\boldsymbol{\Sigma}}_l, \mu_l\}_{l=1}^k, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau^2, \mathbf{R}]] + \mathbb{E}[\mathbb{V}[\mathbf{Y}_v^F \mid \{y_l^F, \tilde{\boldsymbol{\Sigma}}_l, \mu_l\}_{l=1}^k, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau^2, \mathbf{R}]] \\
=& \mathbb{V}[\mathbf{1}_k \otimes \mathbf{f}_\theta^M + \boldsymbol{\mu} \otimes \mathbf{1}_n + \mathbf{1}_k \otimes \boldsymbol{\delta} \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}] + \boldsymbol{\Lambda} \\
=& \mathbf{1}_k \mathbf{1}_k^T \otimes (\tau^2 \mathbf{R}) + \boldsymbol{\Lambda},
\end{aligned}
$$

from which (S3) follows. Note that the density of $(\mathbf{Y}_v^F, \boldsymbol{\delta})$ follows

$$
\begin{aligned}
&p\left(\mathbf{Y}_v^F, \boldsymbol{\delta} \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \\
&\quad = \prod_{l=1}^{k} \left\{ (2\pi)^{-\frac{n}{2}} |\tilde{\boldsymbol{\Sigma}}_l|^{-\frac{1}{2}} \exp\left(-\frac{(\tilde{\mathbf{y}}_l - \boldsymbol{\delta})^T \tilde{\boldsymbol{\Sigma}}_l^{-1} (\tilde{\mathbf{y}}_l - \boldsymbol{\delta})}{2} \right) \right\} \\
&\qquad \times (2\pi\tau^2)^{-\frac{n}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp\left(-\frac{\boldsymbol{\delta}^T \mathbf{R}^{-1} \boldsymbol{\delta}}{2\tau^2} \right).
\end{aligned}
$$

Marginalizing out $\boldsymbol{\delta}$ from the above equation yields the density of $(\mathbf{Y}_v^F \mid \{\sigma_l^2, \mathbf{R}_l, \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, \mathbf{R})$.

□

Both Lemma S1 and Lemma S2 can be used for computing the likelihood given the other parameters in the full Bayesian analysis. Lemma S2 may be used to develop the maximum likelihood estimator, as both the random model discrepancy and measurement bias functions are marginalized out explicitly. Note that the computational complexity of the marginal density of the field data is $O((k+1)n^3)$ in both lemmas, from inverting $k+1$ covariance matrices each with the size $n \times n$, rather than $O((kn)^3)$; this is the case even if the covariance matrix in (S3) is $nk \times nk$ in Lemma S2. Such simplification is the key to proceeding without approximations to compute the likelihood if n is not very large. Note that the simplifications of computation rely on the aligned measurements of each source of field data. When the measurements are misaligned, approximations might be needed when the number of sources is large.

Since the discrepancy function between the mathematical model and reality is often scientifically important, one can draw and record δ using Lemma S1 in the posterior sampling. The following theorem gives the predictive distribution at any input $\mathbf{x}$, given the parameters and posterior samples of $\boldsymbol{\delta}$.

Theorem S1. *For any $\mathbf{x}^* \in \mathcal{X}$, one has the following predictive distributions for model (1):*

1. *The predictive distribution of the model discrepancy at any input $\mathbf{x}^*$ follows a normal*

distribution

$$(\delta(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)) \sim \mathrm{N}(\hat{\delta}(\mathbf{x}^*),\, \tau^2 \hat{K}^{**}),$$

where $\hat{\delta}(\mathbf{x}^*) = \mathbf{r}(\mathbf{x}^*)^T \mathbf{R}^{-1} \boldsymbol{\delta}$, $\mathbf{r}(\mathbf{x}^*) = (K(\mathbf{x}^*, \mathbf{x}_1), ..., K(\mathbf{x}^*, \mathbf{x}_n))^T$, *and* $\hat{K}^{**} = K(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{r}(\mathbf{x}^*)^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}^*)$.

2. *For each source* l, $l = 1, ..., k$, *the predictive distribution of the measurement bias at any input* $\mathbf{x}^*$ *follows a normal distribution*

$$(\delta_l(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)) \sim \mathrm{N}(\hat{\delta}_l(\mathbf{x}^*),\, \sigma_l^2 \hat{K}_l^{**}),$$

where $\hat{\delta}_l(\mathbf{x}^*) = \sigma_l^2 r_l(\mathbf{x}^*)^T \tilde{\boldsymbol{\Sigma}}_l^{-1} (\tilde{\mathbf{y}}_l^F - \boldsymbol{\delta})$ *and* $\hat{K}_l^{**} = K_l(\mathbf{x}^*, \mathbf{x}^*) - \sigma_l^2 \mathbf{r}_l(\mathbf{x}^*)^T \tilde{\boldsymbol{\Sigma}}_l^{-1} \mathbf{r}_l(\mathbf{x}^*)$, *with* $\mathbf{r}_l(\mathbf{x}^*) = (K_l(\mathbf{x}^*, \mathbf{x}_1), ..., K_l(\mathbf{x}^*, \mathbf{x}_n))^T$, $\tilde{\mathbf{y}}_l^F$ *and* $\tilde{\boldsymbol{\Sigma}}_l^{-1}$ *being defined in Lemma S1, for* $l = 1, ..., k$.

3. *For each source* l, $l = 1, ..., k$, *the predictive distribution of the field data at any input* $\mathbf{x}^*$ *follows a normal distribution*

$$\begin{aligned}(y_l^F(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)) \\ \sim \mathrm{N}(\hat{y}_l^F(\mathbf{x}^*),\, \tau^2 \hat{K}^{**} + \sigma_l^2 \hat{K}_l^{**} + \sigma_{0l}^2),\end{aligned}$$

where $\hat{y}_l^F(\mathbf{x}^*) = \hat{\delta}(\mathbf{x}^*) + \hat{\delta}_l(\mathbf{x}^*) + f^M(\mathbf{x}^*, \boldsymbol{\theta}) + \mu_l$.

Proof of Theorem S1. We only verify the third claim, and the previous two claims can be

verified similarly. For the third claim, the mean follows

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{y}_l^F(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)\right] \\
=&\mathbb{E}\left[\delta(\mathbf{x}^*) + \delta_l(\mathbf{x}^*) + f^M(\mathbf{x}^*.\boldsymbol{\theta}) + \mu_l \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)\right] \\
=&\mathbf{r}(\mathbf{x}^*)^T\mathbf{R}^{-1}\boldsymbol{\delta} + \sigma_l^2\mathbf{r}_l(\mathbf{x}^*)\tilde{\boldsymbol{\Sigma}}_l^{-1}(\mathbf{y}_l^F - \mathbf{f}_{\boldsymbol{\theta}}^M - \mu_l\mathbf{1}_n - \boldsymbol{\delta}) + f^M(\mathbf{x}^*, \boldsymbol{\theta}) + \mu_l \\
=&\hat{\delta}(\mathbf{x}^*) + \hat{\delta}_l(\mathbf{x}^*) + f^M(\mathbf{x}^*, \boldsymbol{\theta}) + \mu_l,
\end{aligned}
$$

and the variance is

$$
\begin{aligned}
&\mathbb{V}\left[\mathbf{y}_l^F(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)\right] \\
=&\mathbb{V}\left[\mathbb{E}\left[\mathbf{y}_l^F(\mathbf{x}^*) \mid \boldsymbol{\delta}, \{\mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l\}_{l=1}^k, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot), \delta(\mathbf{x}^*), \delta_l(\mathbf{x}^*)\right]\right] + \sigma_{l0}^2 \\
=&\mathbb{V}\left[\delta(\mathbf{x}^*) + \delta_l(\mathbf{x}^*) + f^M(\mathbf{x}^*, \boldsymbol{\theta}) + \mu_l \mid \boldsymbol{\delta}, \mathbf{y}_l^F, \sigma_l^2, K_l(\cdot,\cdot), \sigma_{0l}^2, \mu_l, \boldsymbol{\theta}, \tau^2, K(\cdot,\cdot)\right] + \sigma_{l0}^2 \\
=&\tau^2\hat{K}^{**} + \sigma_l^2\hat{K}_l^{**} + \sigma_{l0}^2.
\end{aligned}
$$

The claim soon follows by noticing the predictive distribution is a multivariate normal distribution. □

S3 Inconsistent estimation of MLE when discrepancy function is modeled as a GaSP

Here we provide a closed-form example of inconsistent estimation when the discrepancy function is modeled as a GaSP.

Example S1. *Assume field data of the source l at input x_i follows*

$$y_l^F(x_i) = f^M(x_i, \theta) + \delta(x_i) + \epsilon_l(\mathbf{x}_i),$$

where $f^M(x_i, \theta) = \theta$, $\delta(\cdot) \sim GaSP(0, \tau^2 K(\cdot,\cdot))$, with $K(x_i, x_j) = \exp(-|x_i - x_j|/\gamma)$, and

$\epsilon_l(\mathbf{x}_i) \sim \mathrm{N}(0, \sigma_0^2)$ *is an independent Gaussian noise for each* x_i, $i = 1, ..., n$, *and for* $l = 1, ..., k$. *Assume the observations* $y_l^F(x_i)$ *are equally spaced from* $[0, 1]$, *i.e.* $x_i = (i-1)/(n-1)$, *for* $i = 1, ..., n$ *and* $l = 1, ..., k$. *Further assume* γ, τ^2 *and* σ_0^2 *are known and finite.*

The following results show that MLE is an inconsistent estimator of θ in this scenario.

Lemma S3. *Assume* $\tau^2 > 0$ *and* $\gamma > 0$ *are both finite. When* $n \to \infty$ *and* $k \to \infty$, *after integrating out* $\boldsymbol{\delta}$, *the MLE of* θ *in Example S1 has the limiting distribution:*

$$\hat{\theta}_{MLE} \xrightarrow{d} \mathrm{N}(\theta, \frac{2\tau^2\gamma}{2\gamma + 1}). \tag{S6}$$

Proof of Lemma S3. First, after marginalizing out δ, the marginal likelihood is given by equation (4), with $\mu = 0$ and $f^M(x, \theta) = \theta$. Thus, the MLE of the calibration parameter is the same as using the full data or the aggregated data $\bar{\mathbf{y}}^F$.

When $k \to \infty$, the limiting distribution of $\bar{\mathbf{y}}^F$ in model (1) of the aggregate data follows

$$\left(\bar{\mathbf{y}}^F \mid \mu, \boldsymbol{\theta}, \tau^2, \mathbf{R}\right) \sim \mathrm{MN}\left(\mathbf{f}_{\boldsymbol{\theta}}^M + \mu \mathbf{1}_n,\, \tau^2 \mathbf{R}\right). \tag{S7}$$

Thus the MLE of the calibration parameter is $\hat{\theta}_{MLE} = (\mathbf{1}_n^T \mathbf{R}^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n^T \mathbf{R}^{-1} \bar{\mathbf{y}}^F$ with the sampling distribution when $k \to \infty$ follows

$$\hat{\theta}_{MLE} \sim \mathrm{N}(\theta, \tau^2 (\mathbf{1}_n^T \mathbf{R}^{-1} \mathbf{1}_n)^{-1}).$$

Denote $\rho_n = \exp(-\frac{1}{n\gamma})$. The inverse correlation matrix can be computed explicitly (Gu

et al., 2018)

$$\mathbf{R}^{-1} = \frac{1}{1-\rho_n^2}\begin{pmatrix} 1 & -\rho_n & 0 & 0 & ... & 0 \\ -\rho_n & 1+\rho_n^2 & -\rho_n & 0 & ... & 0 \\ 0 & -\rho_n & 1+\rho_n^2 & -\rho_n & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & -\rho_n & 1+\rho_n^2 & -\rho_n \\ 0 & 0 & ... & 0 & -\rho_n & 1 \end{pmatrix}.$$

When $n \to \infty$, the variance of the sampling distribution of $\hat{\theta}_{MLE}$ is

$$\begin{aligned} \mathbb{V}[\theta_{MLE}] &= \lim_{n\to\infty} \tau^2 (\mathbf{1}_n^T \mathbf{R}^{-1} \mathbf{1}_n)^{-1} \\ &= \lim_{n\to\infty} \tau^2 \frac{1-\exp(-\frac{2}{n\gamma})}{(1-\exp(-\frac{1}{n\gamma}))(n-(n-2)\exp(-\frac{1}{n\gamma}))} \\ &= \frac{2\tau^2\gamma}{2\gamma+1}, \end{aligned}$$

where the last line follows from the Taylor expansion. This completes the proof. □

The result in Lemma S3 is surprising as both k and n go to infinity, yet reasonable, because the discrepancy function is shared across all experiments. Thus, even if the estimation of the summation of the model output and discrepancy (i.e., the reality) is consistent, the estimation of the calibration parameters or discrepancy function is often inconsistent.

The MSE between the MLE of θ and truth in Example S1 is graphed as the red triangles at different number of observations in Figure S1 when $k \to \infty$. In the left panel, when $\gamma = 0.1$, the MSE quickly converges when the sample size increases. In the right panel, the MSE converges to a smaller value when $\gamma = 0.02$, because the correlation is smaller. Both MSEs converge to the limiting value, $\frac{2\tau^2\gamma}{2\gamma+1}$ (black horizontal line), found in Lemma S3.

Though the closed-form expression of the limiting distribution of the MLE of the parameter in Lemma S3 relies on the exponential kernel function, the MLE is inconsistent for the

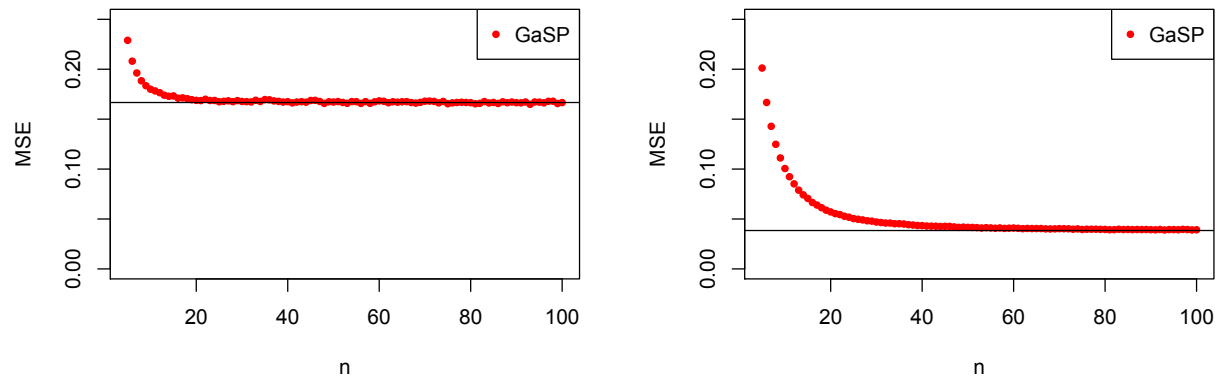

Figure S1: Mean squared error (MSE) of the MLE of θ in Example S1 for different numbers of observations when k is large. The limiting variance of the MLE for GaSP data in Lemma S3 is graphed as the black horizontal lines. 10^5 simulations are implemented to calculate each value. The range parameter is assumed to be $\gamma = 0.1$ and $\gamma = 0.02$ in the left and right panels, respectively, both assuming $\tau_0^2 = 1$.

mean parameter of the GaSP model with many other kernel functions, such as the Matérn kernel. We refer the reader to Chapter 4.2 in Stein (2012) for the detailed discussion on this topic. Example S1 points out that even an infinite number of repeated samples cannot solve the identifiability problem between calibration parameters and discrepancy function.

S4 Discretized S-GaSP

The discretized S-GaSP of the discrepancy function can be modeled below Gu and Wang (2018):

$$\begin{aligned} \delta_z(\mathbf{x}) &= \left\{ \delta(\mathbf{x}) \mid \frac{1}{n}\sum_{i=1}^{n} \delta^2(\mathbf{x}_i) = Z \right\}, \\ \delta &\sim \text{GaSP}(0, \tau^2 K(\cdot,\cdot)),\ Z \sim p_{\delta_z}(\cdot), \end{aligned} \tag{S8}$$

where $p_{\delta_z}(\cdot)$ is the density of the random squared error between the reality and mathematical model. Given $Z = z$, S-GaSP is a GaSP constrained at the space $\sum_{i=1}^{n} \frac{\delta^2(\mathbf{x}_i)}{n} = z$. The idea is to assign more probability mass on the small squared error by $p_{\delta_z}(\cdot)$. Denote $p_\delta(Z = z \mid \boldsymbol{\gamma}, \tau^2)$

the density of $Z = z$ induced by the GaSP model in (7), where τ^2 and $\boldsymbol{\gamma}$ are the variance and range parameters in the covariance function, respectively. We let $p_{\delta_z}(\cdot)$ proportional to $p_\delta(\cdot)$, but scaled by a exponential function:

$$p_{\delta_z}(Z = z|\boldsymbol{\gamma}, \tau^2, \lambda_z) = \frac{\exp\left(-\frac{\lambda_z z}{2\tau^2 \text{Vol}(\mathcal{X})}\right) p_\delta\left(Z = z \mid \boldsymbol{\gamma}, \tau^2\right)}{\int_0^\infty \exp\left(-\frac{\lambda_z z}{2\tau^2 \text{Vol}(\mathcal{X})}\right) p_\delta\left(Z = t \mid \boldsymbol{\gamma}, \tau^2\right) dt}, \tag{S9}$$

where λ_z is a positive scaling parameter and $\text{Vol}(\mathcal{X})$ is the volume of the input domain $\mathcal{X}$. The GaSP with any covariance function is a special case of S-GaSP when $f_Z(\cdot)$ is a constant function, or equivalently $\lambda_z = 0$.

S5 A simulated example to compare different models of discrepancy

Here we discuss an example to compare the identifiability issue between GaSP and S-GaSP models of discrepancy function.

Example S2. *Assume* $y^F(\mathbf{x}) = y^R(\mathbf{x}) + \epsilon(\mathbf{x})$ *where* $\mathbf{x} = (x_1, x_2) \in [0,1]^2$, $\epsilon(\mathbf{x}) \sim N(0, 0.05^2)$ *is an independent Gaussian noise for each* $\mathbf{x}$ *and reality is assumed to be (Lim et al., 2002):*

$$y^R(\mathbf{x}) = \frac{1}{6}\left\{(30 + 5x_1 \sin(5x_1))(4 + \exp(-5x_2)) - 100\right\}.$$

Let $f^M(\mathbf{x}, \boldsymbol{\theta}) = \theta_1 + \theta_2 \sin(5x_1)$*, where* $\boldsymbol{\theta} = (\theta_1, \theta_2)$ *are unknown calibration parameters. Field data* $y^F(\mathbf{x}_i)$ *at* $\mathbf{x}_i$*,* $i = 1, ..., 30$*, is drawn from the maximin Latin hypercube design (Santner et al., 2003). The goal is to estimate* $\boldsymbol{\theta}$ *and predict the reality at all* $\mathbf{x} \in [0,1]^2$*.*

For Example S2, the true values of the calibration parameters are not well-defined because the discrepancy function is a deterministic function. We thus compare GaSP and S-GaSP calibrations based on two criteria of predictions on $y^R(\mathbf{x}_i^*)$ at the held-out $\mathbf{x}_i^*$, uniformly

Table S1: Predictive mean squared errors and the MLE of the parameters in GaSP and S-GaSP calibration models in Example S2.

	MSE_{f^M}	$\text{MSE}_{f^M+\delta}$	$\hat{\boldsymbol{\theta}}$	$\hat{\tau}^2$	$\hat{\boldsymbol{\gamma}}$	$\hat{\sigma}_0^2$
GaSP	152	0.0100	$\{16.0, 2.06\}$	81.8	$\{1.52, 2.02\}$	0.00297
S-GaSP	1.58	0.0102	$\{3.76, 1.89\}$	349	$\{1.52, 2.13\}$	0.00815

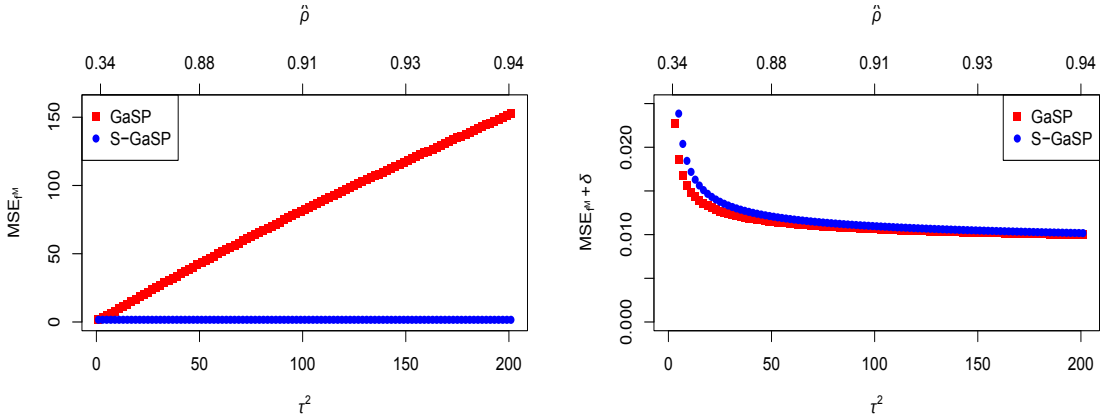

Figure S2: Predictive mean squared errors of Example S2 when τ^2 is fixed (lower x coordinate) in the GaSP and S-GaSP calibrations. The upper x coordinate is the estimated median value of the correlation matrix $\mathbf{R}$ for each τ^2 in the GaSP calibration model.

sampled from $[0,1]^2$ for $i = 1, ..., 1000$ below:

$$\text{MSE}_{f^M} = \frac{1}{n^*}\sum_{i=1}^{n^*}(f^M(\mathbf{x}_i^*, \hat{\boldsymbol{\theta}}) - y^R(\mathbf{x}_i^*))^2 \quad \text{and} \quad \text{MSE}_{f^M+\delta} = \frac{1}{n^*}\sum_{i=1}^{n^*}(\hat{y}^R(\mathbf{x}_i^*) - y^R(\mathbf{x}_i^*))^2,$$

where $\hat{\boldsymbol{\theta}}$ are estimated calibration parameters and $\hat{y}^R(\mathbf{x}_i^*)$ is the prediction of the reality combining mathematical model and discrepancy function. Only the calibrated mathematical model is used to predict the reality in calculating MSE_{f^M}, whereas both the calibrated mathematical model and discrepancy function for predictions can used to calculate $\text{MSE}_{f^M+\delta}$.

The predictive error of Example S2 using two models is given in Table S1, where parameters are estimated by the MLE via the low-storage quasi-Newton optimization method (Nocedal, 1980) with 10 different initializations. The $\text{MSE}_{f^M+\delta}$ is similar using both GaSP and S-GaSP calibration, while MSE_{f^M} by S-GaSP is much smaller than the one by GaSP.

Denote ρ as the median value in the correlation matrix $\mathbf{R}$, and let $\hat{\rho}$ be the estimated value in the GaSP calibration model. After plugging in $\hat{\boldsymbol{\gamma}}$ in Table S1, we found $\hat{\rho} \approx 0.92$ indicating

relatively large estimated correlation. To further explore the cause of the large MSE_{f^M} in the GaSP calibration, we fix the scale parameter τ^2 at different values and estimate the rest of the parameters by the MLE in the GaSP calibration. We use the same range parameters in the covariance matrix in the S-GaSP calibration, and compute the MLE for the calibration parameters. The MSE_{f^M} and $\text{MSE}_{f^M+\delta}$ due to the change of τ^2 are shown in Figure S2. When τ^2 is fixed at a small value, the median estimated correlation in the GaSP calibration, shown in the upper x coordinate in Figure S2, is small. When the correlation is small, the MSE_{f^M} is small (left panel), whereas the $\text{MSE}_{f^M+\delta}$ is comparatively large (right panel). When τ^2 is fixed at a large value, the $\text{MSE}_{f^M+\delta}$ becomes small by both models, whereas the MSE_{f^M} becomes incredibly large in the GaSP calibration due to the large estimated correlation. The MSE_{f^M} is always small in the S-GaSP model, shown in Figure S2.

Example S2 shows that the calibrated mathematical model can be far from the reality in the GaSP model when the estimated correlation is large. When the correlation is small, the predictive distribution of the mathematical model and discrepancy function may sometimes be less precise to predict the reality in this scenario. In comparison, the calibrated mathematical model by the S-GaSP calibration is still close to reality when the correlation is large. Thus the small MSE_{f^M} and $\text{MSE}_{f^M+\delta}$ may not be simultaneously obtained in the GaSP calibration with some frequently used kernel functions, but they can be achieved at the same time in the S-GaSP calibration.

S6 Additional results for real data analysis

We give more results of the real data analysis in this Section. The predictive mean of each interferogram and stack image in GaSP calibration is shown in Figure S3. They are also close to the truth, but they have larger predictive errors than those from S-GaSP.

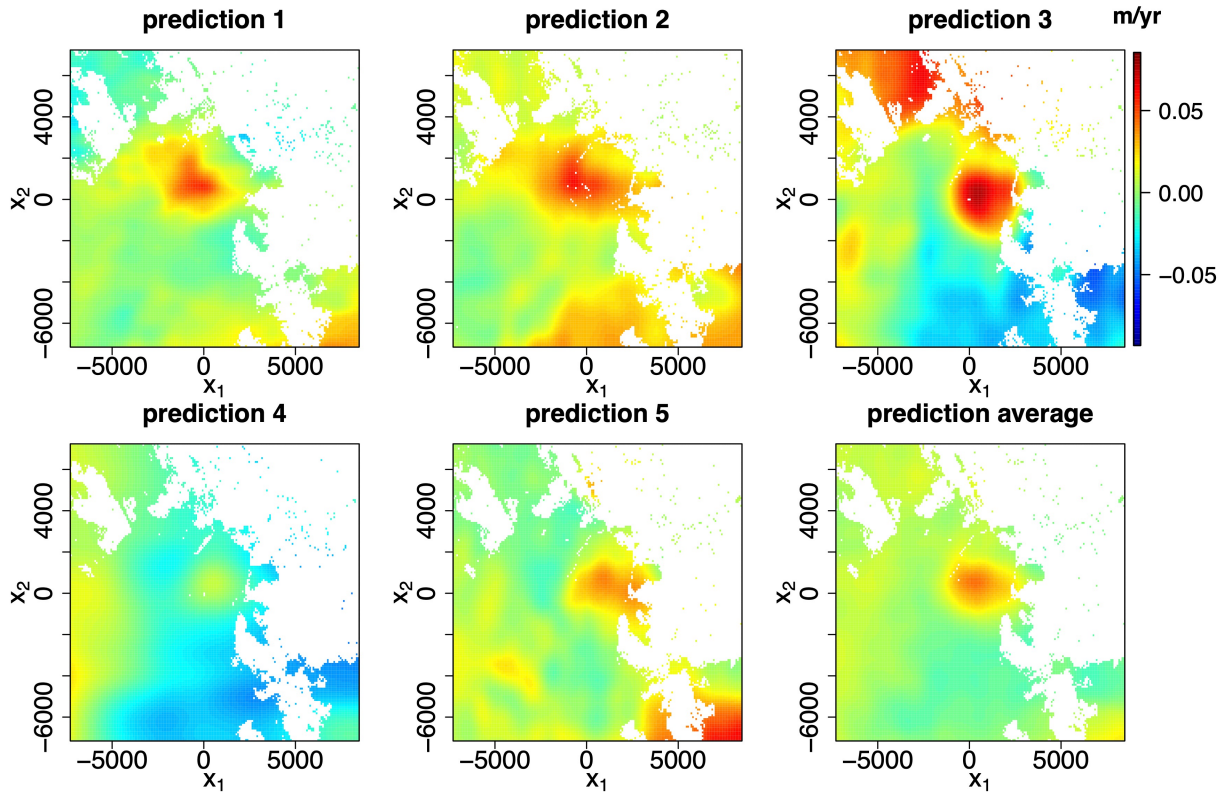

Figure S3: Predictive mean of each interferogram and stack image in GaSP calibration.

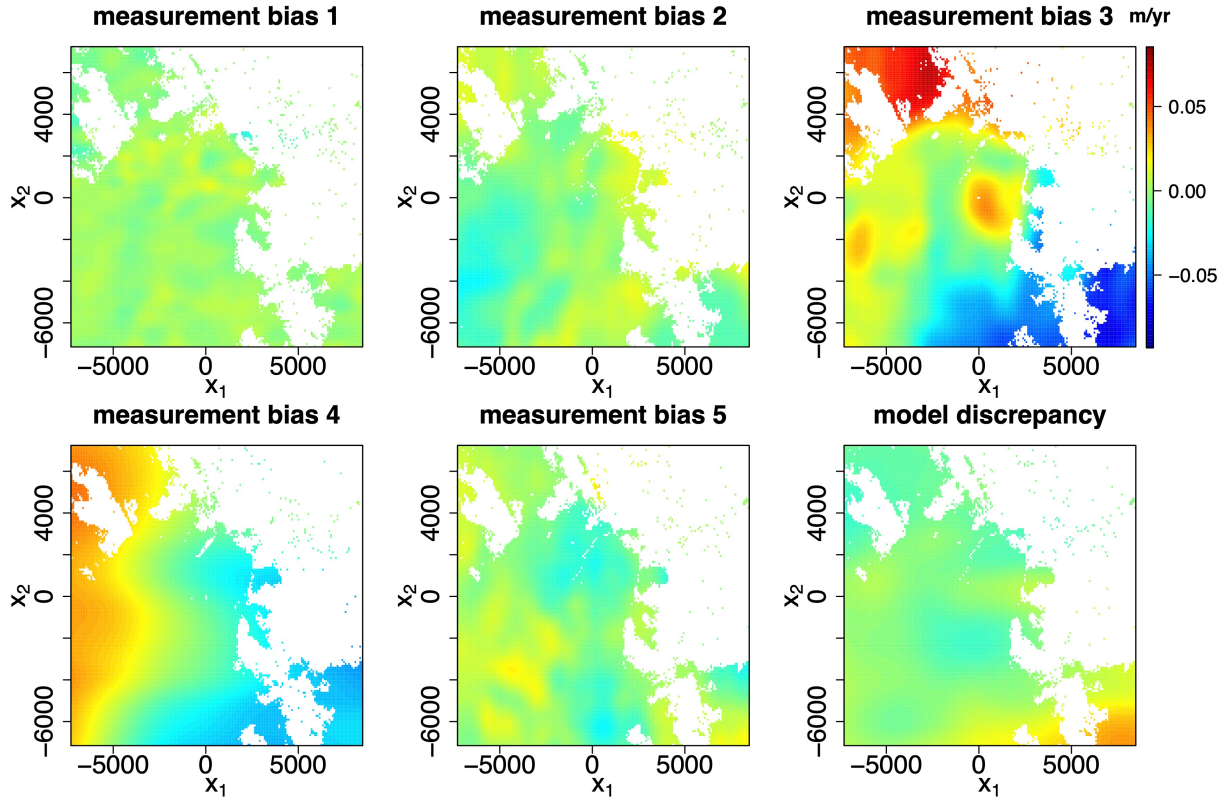

Figure S4: Estimated measurement bias and model discrepancy in the GaSP calibration.

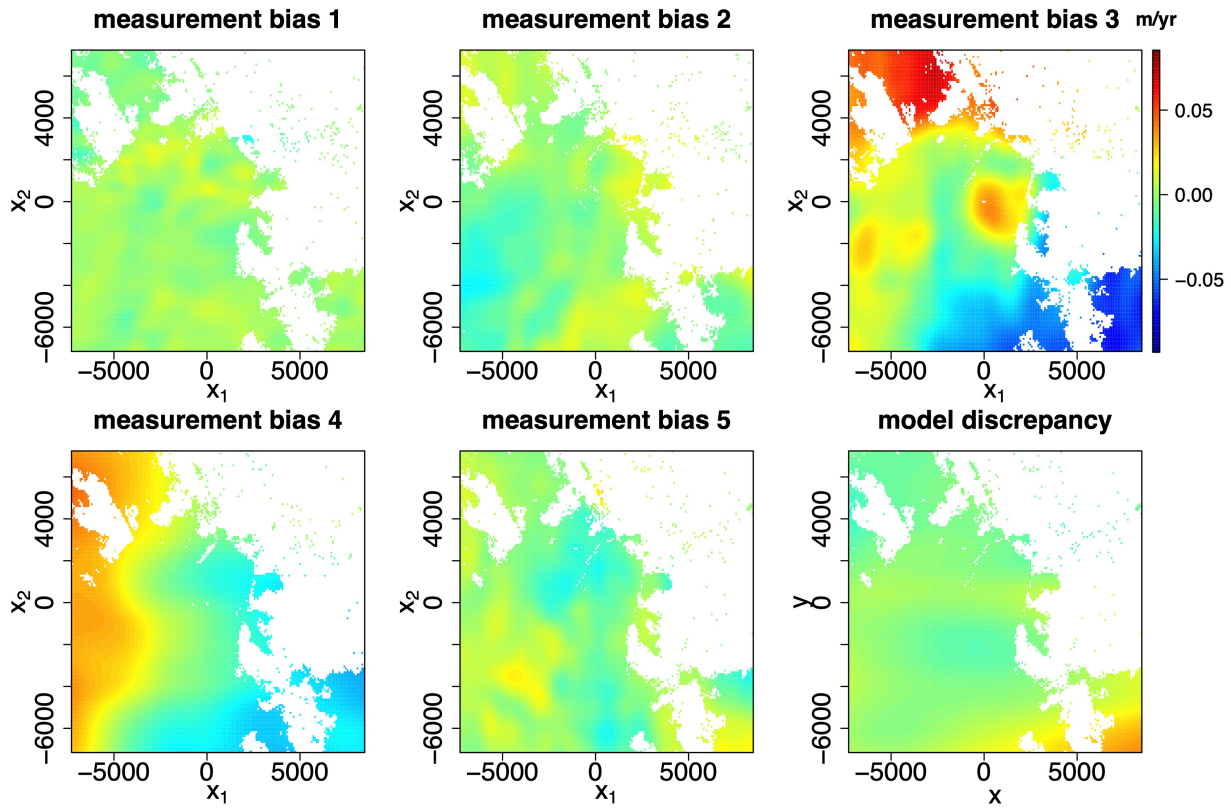

Figure S5: Estimated measurement bias and model discrepancy in the S-GaSP calibration.

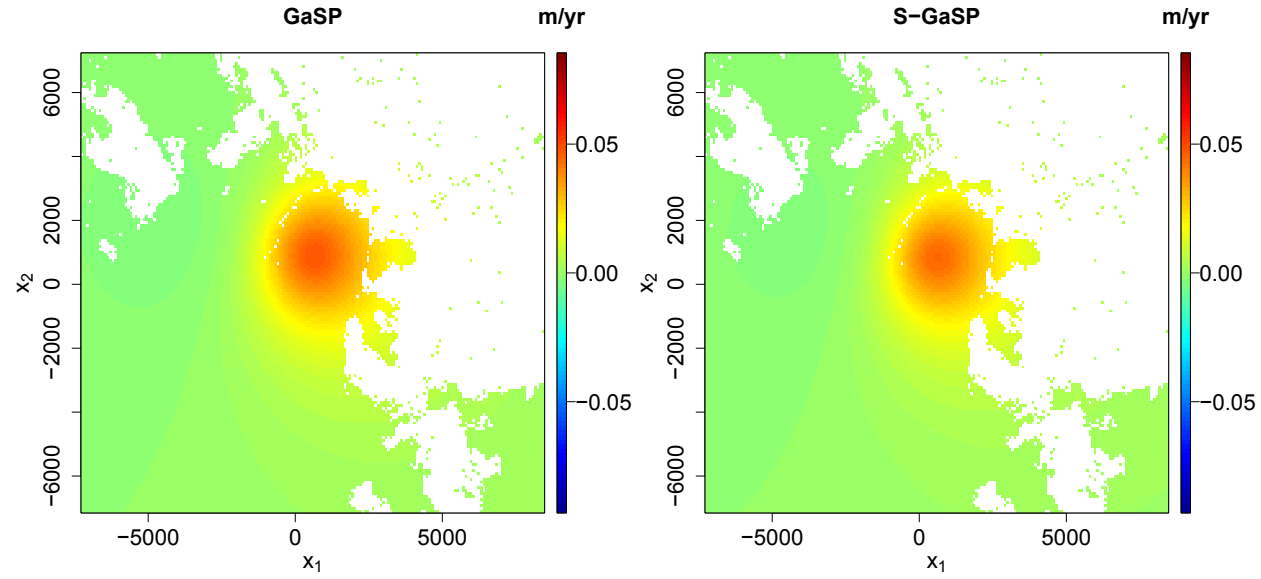

Figure S6: The calibrated geophysical model by the GaSP and S-GaSP are given in the left and right panels.

Estimated measurement bias and model discrepancy for GaSP and S-GaSP calibrations are shown in Figures S4 and S5, respectively. The calibrated computer models by the GaSP and S-GaSP calibration models of the interferograms with measurement bias are shown in Figure S6. The estimated model discrepancy in the GaSP calibration suggests that the calibrated geophysical model may underestimate the ground displacement in the southeast region. However, this is likely caused by the atmospheric artifact appearing in the first, second and fifth panels in Figure 5. In comparison, the atmospheric artifact seems to be properly explained as measurement bias in the S-GaSP calibration shown in Figure S5.

Although stacked images can reduce the measurement bias and noise in the observations, one usually loses some information in estimating the measurement bias and discrepancy function using aggregated data. Among all approaches, the S-GaSP calibration based on the full data seems to be both robust in estimating the calibration parameters, and accurate in separating measurement bias and model discrepancy from the observations.

For Figure S7 and Figure S8, the first row and second row are the trace plots of the calibration parameters and mean parameters, respectively. The trace plots of the inverse range parameters of the measurement bias are shown in the third and fourth rows. The fifth and sixth rows show the trace plots of the nugget and scale parameters of the measurement bias, respectively. The last three panels in the last row give the trace plots of the inverse range parameters and the scale parameters in the discrepancy function. Finally, the trace plots of all the parameters in the GaSP calibration and S-GaSP calibration are given in Figure S7 and Figure S8, respectively. Most of the posterior samples seem to mix reasonably well.

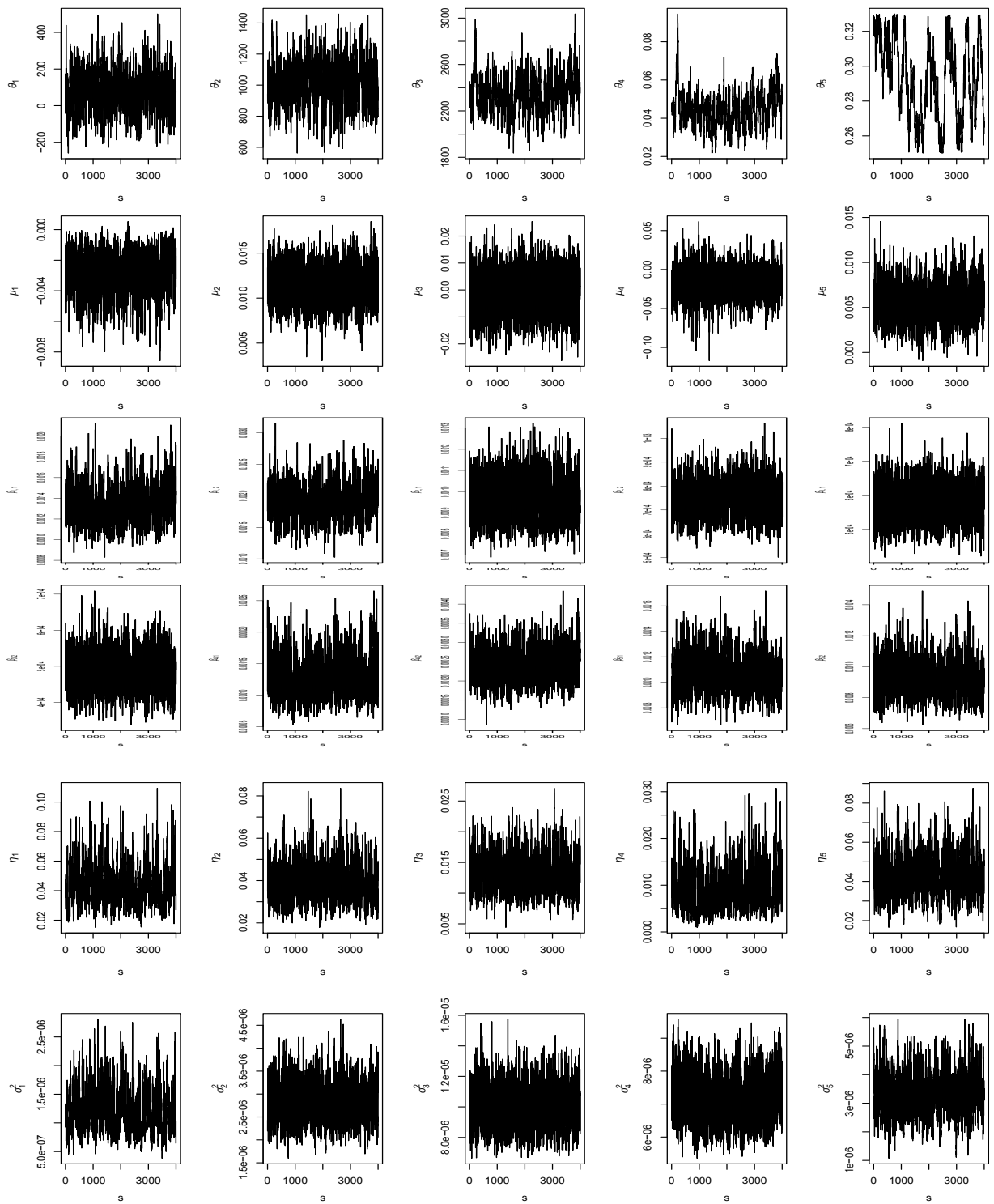

Figure S7: The trace plots of the parameters in the GaSP calibration.

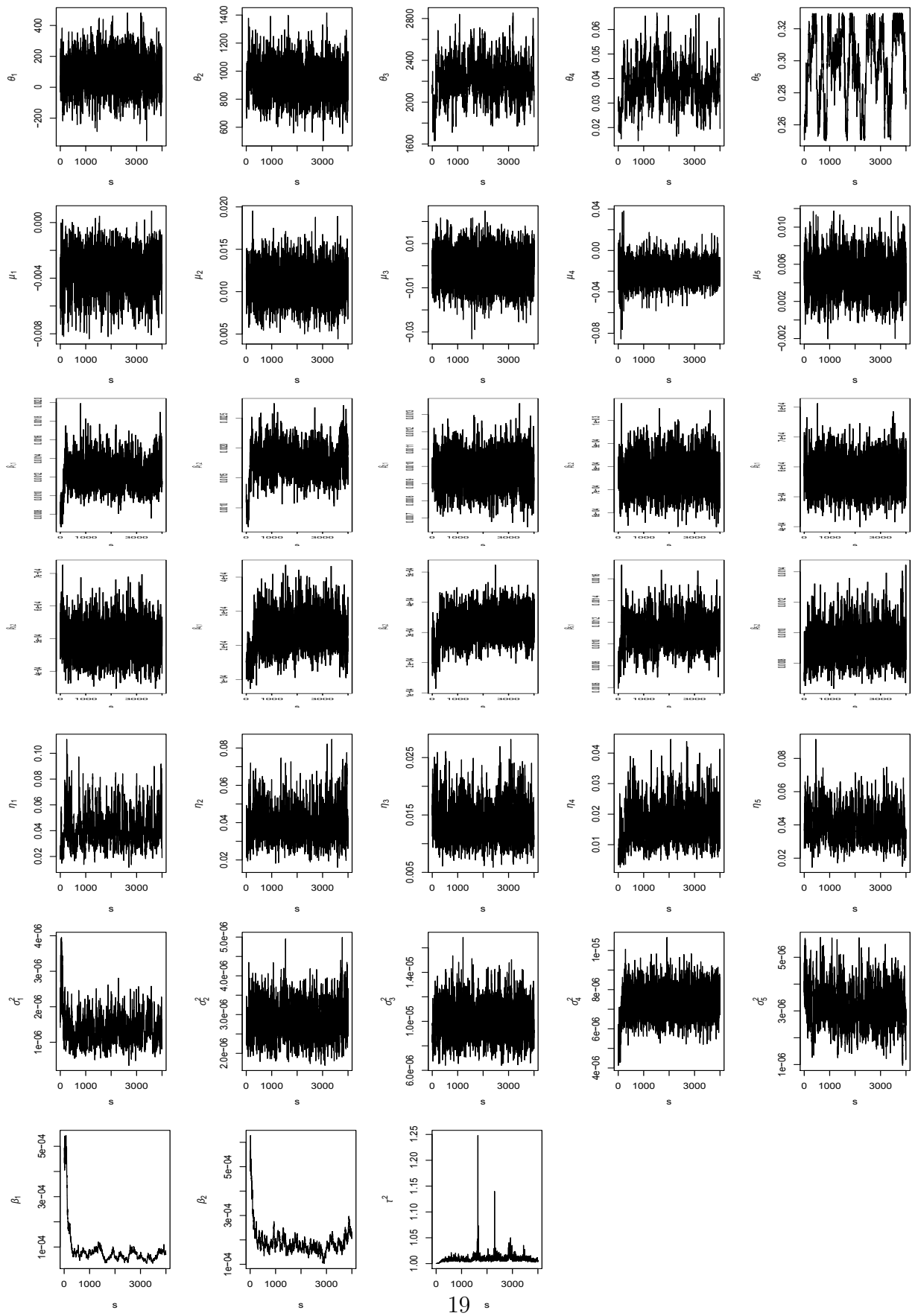

Figure S8: The trace plots of the parameters in the S-GaSP calibration.

References

Gu, M. and Wang, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1555–1583.

Gu, M., Wang, X., and Berger, J. O. (2018). Robust Gaussian stochastic process emulation. *Annals of Statistics*, 46(6A):3038–3066.

Lim, Y. B., Sacks, J., Studden, W., and Welch, W. J. (2002). Design and analysis of computer experiments when the output is highly correlated over the input space. *Canadian Journal of Statistics*, 30(1):109–126.

Nocedal, J. (1980). Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The design and analysis of computer experiments*. Springer Science & Business Media.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.