Computer Model Emulation with High-Dimensional Functional Output in Large-Scale Observing System Uncertainty Experiments

Pulong Ma*1,2, Anirban Mondal3, Bledar A. Konomi4, Jonathan Hobbs5, Joon Jin Song6, and Emily L. Kang4

1Statistical and Applied Mathematical Sciences Institute
2Duke University
3Case Western Reserve University
4University of Cincinnati
5Jet Propulsion Laboratory, California Institute of Technology
6Baylor University

Abstract

The observing system uncertainty experiments (OSUEs) have been widely used as a cost-effective way to make retrieval quality assessment in NASA’s Orbiting Carbon Observatory-2 (OCO-2) mission. One important component in the OCO-2 retrieval algorithm is a full-physics forward model that describes the relationship between the atmospheric variables such as carbon dioxide and radiances measured by the remote sensing instrument. This forward model is complicated and computationally expensive but a large-scale OSUE requires evaluation of this model numerous times, which makes it infeasible for operational usage. To tackle this issue, we develop a statistical emulator to facilitate efficient large-scale OSUEs in remote sensing. This emulator represents radiances output at irregular wavelengths via a linear combination of basis functions and random coefficients. These random coefficients are then modeled with a nearest-neighbor Gaussian process with built-in input dimension reduction via active subspace. The proposed emulator reduces dimensionality in both input space and output space, so that fast computation is achieved within a fully Bayesian inference framework. Validation experiments demonstrate that this emulator outperforms a reduced order model that approximates the full-physics forward model.

*Correspondence to: Pulong Ma, Postdoctoral Fellow at the Statistical and Applied Mathematical Sciences Institute and Duke University, 79 T.W. Alexander Drive, P.O. Box 110207, Research Triangle Park, NC 27709. Email: pulong.ma@duke.edu.
Keywords: Active subspace; Functional basis representation; Nearest-neighbor Gaussian process; OCO-2 mission; Retrieval quality assessment; Uncertainty quantification

1 Introduction

With space-based observations, remote sensing technology provides a wealth of information for understanding geophysical processes with unprecedented spatial and temporal coverage. Quantitative inference for the global carbon cycle has been bolstered by greenhouse gas observing satellites. NASA’s Orbiting Carbon Observatory-2 (OCO-2) collects tens of thousands of observations of reflected sunlight daily. These observed spectra, or radiances, are used to infer the atmospheric carbon dioxide (CO$_2$) at fine spatial and temporal resolution with substantial coverage across the globe (e.g., Bösch et al., 2017; Crisp et al., 2014). Estimates of atmospheric CO$_2$ are computed from the observed radiances using an inverse method known as a retrieval algorithm. The mathematical and computational framework for the retrieval is problem-specific but often involves an optimization with a physical forward model of moderate computational complexity. OCO-2 has adopted a particular retrieval methodology known as “optimal estimation” (Rodgers, 2000) in remote sensing science. The resulting estimates of geophysical quantities of interest are called retrievals.

Quality assessment and uncertainty quantification of remote sensing retrievals are crucial to the success of using these remote sensing data to reveal valid scientific findings and to answer scientific hypotheses appropriately. However, different from many other disciplines, it is infeasible to perform physical experiments to study the quality of remote sensing retrievals thoroughly because a representative ground truth of atmospheric variables is usually lacking; if there were such a ground truth, there would not have been the need to launch satellites to obtain information for these atmospheric variables from space. Therefore, Turmon and Braverman (2019) suggest using simulation-based experiments for quality assessment and uncertainty quantification of remote sensing retrievals, known as the observing system uncertainty experiments (OSUEs). Such experiments are supposed to be cost-effective, because OSUEs are based on simulation, and we do not need to spend resources obtaining the ground truth. Figure 1 shows a basic diagram to compare the retrieval to the synthetic true atmospheric state in an OSUE for remote sensing retrievals. Hobbs et al. (2017) present an OSUE and use it to study the impact of uncertain inputs on the distribution of the retrieval error in OCO-2.

In order to thoroughly study uncertainty in the retrievals and the impact of parameters
in the retrieval algorithm, it is necessary to perform large-scale OSUEs in which we need to perform simulation and evaluate the forward model $\mathbf{F}(\cdot)$ (Figure 1) many times over a large spatial domain at various specifications of the model inputs. However, this forward model $\mathbf{F}(\cdot)$, designed to represent the complicated physical relationship between atmospheric variables and radiances, usually involves integro-differential equations and thus is computationally expensive, making the computational costs of large-scale OSUEs prohibitive. Hobbs et al. (2017) present a surrogate model that is computationally efficient and substitute this new model for the original forward model in order to make a large-scale OSUE practical. This surrogate model in Hobbs et al. (2017), referred to as a reduced order model (ROM) hereafter, preserves some key physical laws in the original $\mathbf{F}(\cdot)$, referred to as the full-physics (FP) forward model, but uses a low-dimensional representation of the atmospheric state and simplified numerical routines to solve the nonlinear equation of radiative transfer. Readers are referred to Hobbs et al. (2017) for details of this ROM and the retrieval algorithm for OCO-2.

In this work, we focus on developing an efficient statistical emulator for the FP forward model $\mathbf{F}(\cdot)$ in OCO-2. There is a well-established paradigm and vast literature of constructing statistical emulators for many expensive computer models (e.g., Bayarri et al., 2007; Conti et al., 2009; Higdon et al., 2008; Mak et al., 2018; Sacks et al., 1989; Santner et al., 2018; Guillou et al., 2018; Welch et al., 1992), but developing a statistical emulator for the OCO-2 FP forward model presents a few challenges: (1) The outputs of the FP forward model are radiances at irregular wavelengths, varying across spatial locations. Moreover,
these radiance outputs from the FP forward model are at up to a total of 1016 wavelengths for the three near infrared spectral bands, termed the O₂-A band, the weak CO₂ band and the strong CO₂ band, respectively, but there are often values missing at some wavelengths irregularly due to the instrument’s characteristics and its interaction with its environment in low-Earth orbit. (2) The input of the FP forward model in OCO-2 is a 66-dimensional vector. Such high-dimensional inputs can pose problems in emulator construction, especially for Gaussian process (GP) emulation, and thus dimension reduction is necessary (e.g., Constantine et al., 2014; Constantine and Gleich, 2014; Constantine et al., 2016; Liu and Guillas, 2017). (3) We have about n = 10,849 runs from the FP forward model based on which we need to build a statistical emulator. This size of data can cause computational issues for a GP emulator, since its estimation and prediction will require storing and factorizing matrices of size n × n. Note that in this article we use n = 10,849 simulation runs to demonstrate our methodology but in general millions of FP forward runs are available from previous simulations, which can be incorporated in the study.

This article provides a unified framework of overcoming these three challenges in order to construct a computationally efficient statistical emulator for large-scale OSUEs, and the proposed method is applied in building a statistical emulator for the FP forward model in OCO-2. In particular, motivated by Bayarri et al. (2007) and Higdon et al. (2008), we treat the radiances as functional outputs and choose to perform functional principal component analysis (FPCA) (Ramsay and Silverman, 2005) to achieve dimension reduction. Similar to the principal component analysis (PCA) approach in Higdon et al. (2008), FPCA is used as a data-driven approach to obtain a basis representation of the output; the associated weights are then modelled using a Gaussian process. Unlike PCA, which is mainly designed for outputs at a regularly-spaced grid without missing values, FPCA can handle outputs over irregularly spaced and varying wavelengths with missing values via basis functions of functional forms. Bayarri et al. (2007) use fixed wavelets basis functions placed over a dyadic grid for functional representation of the output. In our approach, B-spline basis functions are used for functional representation of the output, then the principal components are obtained as functions of wavelength using the data-driven FPCA. To reduce the dimensionality of inputs X, we choose to use the active subspace method (Constantine et al., 2014, 2017) because the gradient of the FP forward model in OCO-2 is available, and the active subspace method tends to give comparable or better results compared to other input dimension reduction methods including sufficient dimension reduction (Cook, 1994, 2009) and effective dimension reduction (Li, 1991) as demonstrated in literature (e.g., Constantine et al., 2014).
To alleviate the computational difficulty related to GP emulation with large $n$, we adopt the nearest neighbor Gaussian process (NNGP; Datta et al., 2016) developed in the area of spatial statistics and extend it to our context of computer model emulation. We further demonstrate that the resulting GP emulator based on NNGP achieves efficient inference within a fully Bayesian framework at a linear computational cost and gives satisfactory performance compared to a GP emulator based on the local Gaussian process approximation (Gramacy and Apley, 2015) and the ROM in Hobbs et al. (2017).

This article is organized as follows. Section 2 describes the OCO-2 FP forward model and the OCO-2 data. Section 3 introduces the statistical framework of building an emulator for the FP forward model in OCO-2 retrieval. In Section 4, we compare the proposed emulator with other statistical methods and validate the statistical emulator with a reduced order model. Section 5 is concluded with discussion and possible extensions.

2 The OCO-2 Data

The OCO-2 instrument carries three imaging grating spectrometers measuring solar radiation reflected from the Earth’s surface in the infrared portion of the spectrum. The spectrometers measure the radiation intensity in the three relatively small wavelength bands (O$_2$-A, WCO$_2$, and SCO$_2$). Then bands, O$_2$-A, weak CO$_2$, strong CO$_2$, are centered near 0.76 µm, 1.61 µm, and 2.06 µm, respectively. In what follows, we will refer to these bands as O$_2$ band, WCO$_2$ band and SCO$_2$ band, respectively. Each band is a function of different wavelengths with length 1,016. The collection of observed radiances from the three spectral bands at a particular time makes up a sounding. Figure 2 shows three spectral bands from five different OCO-2 soundings.

![Fig. 2. Example of five OCO-2 soundings at three different spectrum bands.](image-url)
The instrument collects eight soundings over its 0.8-degree-wide swath every 0.333 seconds (Bösch et al., 2017). The observed radiances result from the interaction between the radiation and the composition of the atmosphere and the Earth’s surface along the path from the top of the atmosphere to the surface and back to the satellite.

The emulator developed in this work is based on the FP forward model implemented in Version 7 of the OCO-2 data products (Eldering et al., 2017). The retrieved state vectors, observed radiances, and forward model evaluations are part of the OCO-2 Level 2 diagnostic data products, available at the NASA Goddard Earth Science Data and Information Services Center (GES DISC, https://disc.gsfc.nasa.gov/OCO-2). We utilize data that include a state vector $x$ of dimension 62 and augmented with the FP forward model Jacobians, consistent with the configuration used by Connor et al. (2016). The state vector characterizes the CO$_2$ vertical profile, surface pressure, surface albedo, aerosols, temperature and water vapor profile offsets, solar induced fluorescence, and wavelength offsets for the instrument. In developing the emulator, we also include the FP forward model dependence on viewing geometry parameters $b$. These include instrument azimuth angle, instrument zenith angle, solar azimuth angle, and solar zenith angle.

3 Model Formulation

Let $y(x, b; \omega)$ be a spectrum radiance at wavelength $\omega$ generated form the FP forward model at input $(x, b)$, where $x$ is a 62-dimensional vector containing the atmospheric state, and $b$ is a 4-dimensional vector containing viewing geometry. For the O$_2$ band, $\omega$ takes 1,016 different values near 0.76$\mu$m; for the WCO$_2$ band, $\omega$ takes 1,016 different values near 1.61 $\mu$m; for the SCO$_2$ band, $\omega$ takes 1,016 different values near 2.06$\mu$m. Then $y(x, b) := (y(x, b; \omega_1), \ldots, y(x, b; \omega_q))^\top$ denotes a sounding consisting of radiances at O$_2$ band, WCO$_2$ band and SCO$_2$ band, where $q = 3,048$.

The goal here is to construct an emulator that can approximate the FP forward model output at arbitrary input settings over the input space. To deal with multivariate output, separability between input space and output has often been assumed to reduce computational burdens (e.g., Conti and O’Hagan 2010, Bilionis et al. 2013, Gu and Berger 2016, Guillas et al. 2018), that is, the covariance matrix of data can be written as a Kronecker product of the input covariance matrix and the output covariance matrix. The separability assumption has been proven quite successful in many real applications when outputs are generated over regularly-spaced grid for different inputs. However, the output in the OCO-2 data is a
function of variable wavelengths with missing values. These methods cannot be applied readily in this situation.

To model such high-dimensional functional output, we adopt a basis-function representation approach similar to the approaches in [Bayarri et al. 2007] and [Higdon et al. 2008]. Such basis-function representation allows for non-separability between input space and output space (e.g., [Fricker et al. 2013]). Specifically, we assume that the functional output $y(\mathbf{x}, \mathbf{b}; \omega)$ has a truncated basis-function representation:

$$y(\mathbf{x}, \mathbf{b}; \omega) = \mu(\omega) + \sum_{i=1}^{p_y} \eta_i(\omega) z_i(\mathbf{x}, \mathbf{b}), \quad (3.1)$$

where $\mu(\omega)$ is a mean function, $\eta_i(\omega)$’s are the basis functions at wavelength $\omega$, and $z_i(\mathbf{x}, \mathbf{b})$’s are the FPCA weights that are assumed to be independent GPs over the input space. Unlike the characteristics of computer model outputs in [Higdon et al. 2008], the OCO-2 outputs are radiances at irregular wavelengths, varying across spatial locations and containing many missing values. Thus, we cannot construct the basis functions via principal component analysis directly.

To deal with irregularly-spaced output at varying wavelengths and missing values, we use a functional principal component analysis (FPCA) to choose basis functions. In the FPCA approach, the mean function $\mu(\omega)$ and the basis functions $\eta(\omega)$ are estimated as functions of wavelength, which can deal with the irregularly-spaced and/or missing data. Note that [Bayarri et al. 2007] use a functional data analysis approach, where a wavelet decomposition is used for irregularly spaced functional outputs. However, the complex modes of variation in the OCO-2 output can only be explained by a prohibitive number basis functions in such functional representation. Hence we use a FPCA approach, where first a functional representation of the OCO-2 output is obtained using a B-spline basis and then FPCA is applied to the functional data to obtain the main modes of variation through a small number of functional principal components. More specifically, the number of principal components, $p_y$, is chosen such that more than 99% of the variation is captured. Our numerical results show that very few functional principal components (at most 3) can explain more than 99% of the variability for each of the three bands. The detailed procedure of FPCA is described in Section 3.1.1.

After constructing the mean function $\mu(\omega)$ and eigen function $\eta_i(\omega)$, each FPCA weight $z_i(\mathbf{x}, \mathbf{b})$ is then modeled as a mean-zero GP:

$$z_i(\mathbf{x}, \mathbf{b}) \sim \mathcal{GP}(0, \sigma_i^2 R_z((\mathbf{x}, \mathbf{b}), (\mathbf{x}', \mathbf{b}'); \Theta_i)), \quad (3.2)$$
where $\sigma_i^2$ is a variance parameter and $\Theta_i$ contains range parameters $\theta_i$ for input parameter $x$ and $\vartheta_i$ for input parameter $b$ in the correlation function $R_z(\cdot, \cdot)$. As the dimension of $x$ is $m = 62$, $\Theta_i$ will contain $66 = 62 + 4$ parameters in GP emulation. This will cause computational bottlenecks in GP fitting due to the high-dimensionality of the parameter space. To overcome this problem, we project the high-dimensional input space $\mathcal{X} = \{x : x \in \mathbb{R}^m\}$ into a low-dimensional space $\mathcal{S} := \{s : s = \mathbf{P}^\top x\}$ via an $m \times p$ projection matrix $\mathbf{P} := [\mathbf{p}_1, \ldots, \mathbf{p}_p]$ with $p < m$. Let $d = \sqrt{\sum_{j=1}^p (s_j - s'_j)^2 / \theta_{ij}^2}$ be the Mahalanobis distance between $s$ and $s'$ with $\theta_i := (\theta_{i1}, \ldots, \theta_{ip})^\top$. Then it is easily to recognize that

$$d = \sqrt{(x - x')^\top \left( \sum_{j=1}^p \frac{\mathbf{p}_j \mathbf{p}_j^\top}{\theta_{ij}^2} \right) (x - x')}.$$  

Many correlation families can be written as a function of $d$, see Appendix A for various examples. The form of correlation function $\rho_s(s, s'; \Theta) = \rho_s(d; \Theta)$ is known as the range anisotropic correlation in geostatistics (Zimmerman, 1993). The induced correlation function $\rho_x(x, x'; \theta_i) = \rho_s(d; \theta_i)$ is a geometrically anisotropic correlation function (see, e.g., Cressie, 1993). Thus, if $\rho_s(s, s'; \theta_i)$ is a function of $d$ on $\mathcal{S} \times \mathcal{S}$, it induces a correlation function $\rho_x(x, x'; \theta_i)$ on $\mathcal{X} \times \mathcal{X}$. Assuming a product form of correlation functions between input parameters $x$ and $b$, the correlation function $R_z((x, b), (x', b'); \Theta)$ on $(\mathcal{X} \times \mathcal{B}) \times (\mathcal{X} \times \mathcal{B})$ can be written as

$$R_z((x, b), (x', b'; \Theta_i)) = \rho_s(s, s'; \theta_i)\rho_b(b, b'; \theta_i),$$  

where $\rho_b(\cdot, \cdot; \theta_i)$ is a correlation function for input parameter $b \in \mathcal{B}$.

The key to maintaining this dual form is to identify the projection matrix $\mathbf{P}$. In Section 3.1.2, we give a choice of the projection matrix via the active subspace method (e.g., Constantine et al., 2014). With the projection matrix $\mathbf{P}$, we can directly work with the independent GPs for $z_i(s, b)$ on the input space $\mathcal{L} \equiv \{\ell : \ell = (s, b), s \in \mathcal{S}, b \in \mathcal{B}\}$ instead of the space $\{(x, b) : x \in \mathcal{X}, b \in \mathcal{B}\}$. In the OCO-2 application, the input dimension is reduced from 66 to 8 via the active subspace approach. The number of model parameters is reduced significantly, resulting in better mixing and less Metropolis-Hasting steps.

After performing dimension reduction for both output space and input space, GP modeling needs to be performed based on $n = 10,849$ model runs for the data $\{z_i(\ell_j) : j = 1, \ldots, n\}$ with $i = 1, \ldots, p_y$. For each FPCA weight, the Cholesky decomposition of a $10,849 \times 10,849$ covariance matrix is computationally too demanding in the likelihood evaluation. So, we adopt a fast Gaussian process approximation method to reduce the computational challenges.
in model fitting. Various GP approximations have been proposed in GP modeling. Here, we choose the nearest neighbor Gaussian process (NNGP), since it has a linear computational cost in the number of model runs, and it possesses good predictive accuracy (Datta et al., 2016). The NNGP has an advantage of enabling fully Bayesian inference over other local GP approximations (e.g., Gramacy and Apley, 2015). The detailed description of NNGP is given in Section 3.2.

### 3.1 Dimension Reduction

The model formulation in the basis function representation requires dimension reduction in both output space and input space for the OCO-2 application. In what follows, we describe these dimension reduction approaches.

#### 3.1.1 Functional principal component analysis

The FP forward model output consists of radiances at 3048 irregularly-spaced wavelengths. To respect the nature of spectral outputs, a functional data analysis approach is suitable to account for the correlations between radiances across wavelengths and to deal with irregularly sampled data with missing values (Ramsay and Silverman, 2005). The radiance vectors are transformed to a functional form (as a function of wavelength) by fitting cubic B-spline basis to the observed data. Then functional principal components are applied to this functional data for dimension reduction.

The central idea of FPCA is to find the set of orthogonal functions, the so-called functional principal components, while retaining as much of the variation in the data as possible. Let \( y(x, b; \omega) \) be a random function on a compact interval \( I \) and belonging to the \( L^2 \) space. Mercer’s theorem implies a spectral decomposition of the covariance function \( V(\omega, \omega') \),

\[
V(\omega, \omega') = \sum_{l} \lambda_l \eta_l(\omega) \eta_l(\omega'),
\]

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \) are the ordered eigenvalues and the \( \eta_k(\cdot) \)'s are the corresponding eigen functions. The Karhunen-Loeve expansion (Karhunen, 1947) allows the representation of a random curve \( y(x, b; \omega) \) as an infinite linear combination of orthogonal functions,

\[
y(x, b; \omega) = \sum_{l=1}^{\infty} z_l(x, b) \eta_l(\omega),
\]

where the coefficient \( z_l(x, b) = \int y(x, b; \omega) \eta_l(\omega) d\omega \) is called the the functional principal
component score corresponding to the lth FPC \( \eta_l(\omega) \). Note that \( z_l \) are uncorrelated random variables with \( E(z_l) = 0 \) and \( Var(z_l) = \lambda_l \). The eigenvalues \( \lambda_l \) measure the variation in \( y(x, b; \omega) \) in the \( \eta_l \) direction. FPCA can achieve dimension reduction by retaining only the first \( p_y \) eigencomponents, eigen functions \( \eta(\omega) \) and eigenvalues \( \lambda \), in (3.6).

To estimate the eigencomponents, mean and covariance functions for the functional data are first estimated. The estimated mean function is \( \bar{y}(\omega) = \frac{1}{n} \sum_{j=1}^{n} y(x_j, b_j; \omega) \) and the estimator of the covariance function \( V(\omega, \omega') \) is \( \hat{V}(\omega, \omega') = \frac{1}{n} \sum_{j=1}^{n} (y(x_j, b_j; \omega) - \bar{y}(\omega))(y(x_j, b_j; \omega') - \bar{y}(\omega')) \). An empirical version of (3.5) is given by

\[
\hat{V}(\omega, \omega') = \sum_{l=1}^{\infty} \hat{\lambda}_l \hat{\eta}_l(\omega^l) \hat{\eta}_l(\omega'), \tag{3.7}
\]

where \( \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq 0 \) and \( \hat{\eta}_k(\cdot) \)'s are the ordered eigenvalues and the corresponding eigen functions of \( \hat{V}(\omega, \omega') \). These eigenvalues and eigen functions can be computed by solving the eigen equations

\[
\int \hat{V}(\omega, \omega') \hat{\eta}_l(\omega') d\omega' = \lambda_l \hat{\eta}_l(\omega), \ l = 1, 2, \ldots \tag{3.8}
\]

To implement the FPCA, the centered radiances, \( y^c(x_j, b_j; \omega) = y(x_j, b_j; \omega) - \bar{y}(\omega) \), are transformed to a functional form using a basis function representation given by \( y^c(x_j, b_j; \omega) = \sum_{g=1}^{G} a_{jg} \phi_g(\omega), j = 1, 2, \ldots n, \) where \( \phi \)'s are a series of basis functions and the \( a_{jg} \)'s are the corresponding coefficients. This expression can be written in matrix notation,

\[
y^c(\omega) = A \phi(\omega). \tag{3.9}
\]

There are many choices for \( \phi(\omega) \), such as polynomial basis, exponential basis, spline basis, Fourier basis etc. For the radiance data B-splines basis functions are used where the knot points are selected equidistantly over the range of wavelengths in each of the three bands. The coefficients \( a_{jg} \)'s are obtained using the method of penalized least squares. A similar basis representation is also obtained for the mean radiance function, given by \( \bar{y}(\omega) = \frac{1}{n} \sum_{g=1}^{G} \bar{a}_g \phi_g(\omega) \). Thus the mean function defined in (3.1) is estimated by \( \hat{\mu}(\omega) = \frac{1}{n} \sum_{g=1}^{G} \bar{a}_g \phi_g(\omega) \), where \( \bar{a}_g \)'s are obtained using method of penalized least squares. Substituting the basis representation (3.9) in the covariance function representation (3.7) we have

\[
\hat{V}(\omega, \omega') = n^{-1} \phi^T(\omega) A^T A \phi(\omega'). \tag{3.10}
\]

Let the \( l \)th eigen function in (3.7) be expressed as

\[
\hat{\eta}_l(\omega) = d_l^T \phi(\omega). \tag{3.11}
\]
Substituting this in (3.8), the eigen equations become
\[ n^{-1} \phi^T(\omega) A^T A d_l = \lambda_l \phi^T(\omega) d_l, \quad l = 1, 2, \ldots \]  \hspace{1cm} (3.12)
where \( J = \int \phi(\omega') \phi^T(\omega') d\omega' \). Furthermore, Equation (3.12) has to be true for all argument values \( \omega \), and consequently
\[ n^{-1} A^T A d_l = \lambda_l d_l, \quad l = 1, 2, \ldots \]  \hspace{1cm} (3.13)
In practice, Equation (3.13) is solved numerically to obtain the \( d_l \)'s, which are then substituted in (3.11) to obtain the eigen functions \( \hat{\eta}_l(\omega) \). The functional principal component scores are then estimated by numerical integration of \( \hat{\eta}_l(x, b) = \int y^c(x, b; \omega) \hat{\eta}_l(\omega) d\omega \).

3.1.2 Active subspace

The high-dimensional input imposes computational challenges in GP modeling, so we need to perform dimension reduction for the input space. The viewing geometry characterizes the measurement process, which has large impact on outputs. We only reduce the dimensionality of the atmospheric state input \( x \). As the gradient information for input \( x \) is available, we focus on the active subspace approach, although other dimension-reduction approaches such as sufficient dimension reduction and effective dimension reduction can be alternatives. Below is a brief review of the active subspace method.

The relationship between \( y \) and \( x \) can be captured through a nonlinear physical forward model \( F(\cdot, \cdot) \) and a Gaussian noise term \( e \sim \mathcal{N}(0, \Sigma_e) \) as follows:
\[ y(x, b) = F(x, b) + e, \]  \hspace{1cm} (3.14)
where \( b \) denotes a vector of viewing geometry. The state vector \( x \) is assumed to be a Gaussian random vector with known mean \( \mu_x \) and covariance matrix \( \Sigma_x = D_x R_x D_x^\top \).

Let the Jacobian of the forward model w.r.t. \( x \) be \( K(x) \equiv \frac{\partial F(x, b)}{\partial x} \). Let \( \tilde{x} = D_x^{-1}(x - \mu_x) \). Then \( \tilde{x} \sim \mathcal{N}_m(0, R) \) with \( \tilde{x} \in \mathcal{X} \). We define the following data misfit function \( f \):
\[ f(\tilde{x}) \equiv [y - F(x, b)]^\top \Sigma_e^{-1} [y - F(x, b)]. \]  \hspace{1cm} (3.15)
Then the derivative of the function \( f \) w.r.t. \( \tilde{x} \) is
\[ \nabla f \equiv \frac{\partial f(\tilde{x})}{\partial \tilde{x}} = [K(x) D_x]^\top \Sigma_e^{-1} [y - F(x, b)]. \]  \hspace{1cm} (3.16)
Following the procedure in Constantine et al. (2014), we give the details of the active subspace in Algorithm 1. We first generate \( n \) Monte Carlo samples \( \{x_j : j = 1, \ldots, n\} \) from
Algorithm 1 Active subspace

**Input:** Data \{ (x_j, y_j) : j = 1, \ldots, n \} and gradient information \{ K_j : j = 1, \ldots, n \}.

**Output:** Projection matrix \( P \).

1. Draw \( n \) independent samples \{ \tilde{x}_j : j = 1, \ldots, n \} from its prior distribution \( \pi(\tilde{x}) \). This step is accomplished with samples \{ x_j : j = 1, \ldots, N \}, since \( \tilde{x}_j = D^{-1}_x(x_j - \mu_x) \).
2. Compute the gradient \( \nabla f_j \equiv \nabla f_j(\tilde{x}_j) \) for \( j = 1, \ldots, n \).
3. Compute eigenvalue decomposition of the \( M \times M \) matrix \( \Sigma \equiv 1/N \sum_{j=1}^{n} (\nabla f_j)(\nabla f_j)^\top \):
   \[
   \Sigma \equiv W \Lambda W^\top,
   \]
   where \( W \) is the orthogonal matrix and \( \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_m\} \) is a diagonal matrix of ordered eigenvalues with \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0 \). The matrix \( \Sigma \) is a Monte Carlo approximation of the sensitivity matrix \( E_{\pi(\tilde{x})}[(\nabla \tilde{x} f(\tilde{x}))(\nabla \tilde{x} f(\tilde{x}))^\top] \).
4. Let \( W \equiv [W_1, W_2] \), and \( \Lambda \equiv \text{blockdiag}\{\Lambda_1, \Lambda_2\} \), where \( \Lambda_1 \) contains the largest \( p \) eigenvalues and \( W_1 \) contains the corresponding \( p \) eigenvectors. The value of \( p \) can be determined based on the gap of eigenvalues and the summation \( \sum_{j=1}^{p} \lambda_j \).
5. Return the project matrix \( P := W_1 \).

the input distribution \( \pi(x) \). The number of Monte Carlo samples is set to be large enough so that it is larger than the recommended values in Constantine et al. (2014) and meanwhile these Monte Carlo samples are well spread across the input space. Then the FP forward model is used to generate outputs \{ y_j : j = 1, \ldots, n \} at these inputs. As the active subspace requires inputs to be drawn from the same input distribution, we standardize them to obtain a new collection of inputs \{ \tilde{x}_j : j = 1, \ldots, n \}.

Once the dimension of the subspace, \( p \), is selected, the column space of \( P \) is called an active subspace, denoted by \( \mathcal{S} \equiv \{ s : s \equiv P^\top \tilde{x}, \tilde{x} \in \mathcal{X} \} \), where \( s \equiv P^\top \tilde{x} \) is called active variable. The function \( f \) is most sensitive to the active subspace \( \mathcal{S} \), and is almost flat in the space \( \mathcal{S}^\perp = \{ u : u \equiv W_2^\top \tilde{x}, \tilde{x} \in \mathcal{X} \} \) on average.

To select the active subspace, the evaluation of the gradient function \( \nabla f \) in (3.16) is required. The construction of the active subspace needs \( n \) times evaluation of the gradient \( \nabla f \), forward function \( F \), and Jacobian \( K \), which in total requires \( 2n \) times evaluation of the forward function \( F \) and Jacobian \( K \). The constructed active subspace depends on the radiance vector \( y \). For different \( y \), different active subspaces will be selected. The selected subspace \( \mathcal{S} \) will be treated as an input space in Gaussian process modeling in Section 3.2.
3.2 Nearest Neighbor Gaussian Process

This section describes the nearest neighbor Gaussian process (NNGP) that is used to emulate the relationship between the low-dimensional input space and the FPCA weights.

For notational convenience, the index $i$ is suppressed for the $i$th FPCA weights. Suppose that the FPCA weight $z(\ell)$ on $\mathcal{L}$ is modeled with a Gaussian process with three different components:

$$z(\ell) = h^\top(\ell)\beta + w(\ell) + \epsilon(\ell), \quad \ell \in \mathcal{L}, \quad (3.17)$$

where $h(\cdot)$ is a vector of fixed basis function and $\beta$ is the corresponding unknown regression coefficients. $w(\cdot)$ is a Gaussian process with zero mean and covariance function $C(\cdot, \cdot)$ with correlation parameters $\Theta$ and marginal variance parameter $\sigma^2$.

$\epsilon(\cdot)$ is a Gaussian white-noise process accounting for the nugget effect with variance $\tau^2$. This error term can potentially capture the truncation error in the basis function representation (3.1). The covariance function is assumed to be a product form between input $s$ and input $b$: $C(\ell, \ell') = \sigma^2 \rho_s(s, s') \rho_b(b, b')$, where $\rho_s(s, s')$ is chosen to be a range anisotropic Matérn correlation function, and $\rho_b(b, b')$ is chosen to be the same correlation family.

As we have each FPCA weight $z(\cdot)$ over $n$ input locations $\mathcal{L}^O = \{\ell_1, \ldots, \ell_n\} \subset \mathcal{L}$. Let $z = (z(\ell_1), z(\ell_2), \ldots, z(\ell_n))^\top$ be a vector of corresponding outputs. Then the likelihood function of $z$ given model parameters $\beta, \Theta, \tau^2, \sigma^2$ is

$$L(z | \beta, \Theta, \tau^2, \sigma^2) \propto |C + \tau^2 I|^{-1/2} \exp \left\{ - (z - H\beta)^\top(C + \tau^2 I)^{-1}(z - H\beta)/2 \right\}, \quad (3.18)$$

where $H \equiv (h(\ell_1), \ldots, h(\ell_n))^\top$ and $C = [C(\ell_i, \ell_j)]_{i,j=1,\ldots,n}$.

Model fitting and prediction through either a likelihood approach or a Bayesian approach typically require the repeated evaluation of the log-likelihood, which involves calculating the inverse and determinant of large, dense, and unstructured $n$-by-$n$ covariance matrix that requires $O(n^3)$ flops and $O(n^2)$ storage. However, memory limitations and computational complexity grow with $n$ as well as with the dimensionality of the input space, making these approaches impractical for large datasets.

To reduce the computational complexity of the full Gaussian process, many methods have been proposed to tackle this issue. We will employ the nearest-neighbor Gaussian process (NNGP; Datta et al., 2016) in this article. Specifically, the NNGP extends Vecchia’s approximation (Vecchia, 1988) ideas based on localized information. To approximate $w(\cdot)$, the NNGP model is constructed in two steps. First, we specify a multivariate Gaussian distribution over a fixed set of $r$ points in the domain $\mathcal{L}$, say, $\mathcal{L}^* = \{\ell_1^*, \ldots, \ell_r^*\}$, which is
referred to as the reference set. For instance, the reference set can be chosen to coincide with the set of input locations. Then we extend this finite-dimensional multivariate normal distribution to a stochastic process over the domain based on the reference set.

To construct the NNGP, the process $w(\cdot)$ is assumed to depend on a few (say $k$) nearest neighbors in the reference set $L^*$, where the nearness is defined according to some topological order for points in the reference set $L^*$. Now define a history set $H(L_i^*)$ of point $L_i^*$, as follows: $H(L_i^*)$ is the empty set, and $H(L_i^*) \equiv \{L_1^*, \ldots, L_i^*\}$ for $i = 2, \ldots, r$. Then we define $N(L_i^*)$ to be a neighbor set of $L_i^*$, which is a subset of $H(L_i^*)$, and contains only a few points. Its definition is given as follows:

$$
N(L_i^*) = \begin{cases} 
\emptyset & \text{for } i = 1 \\
\mathcal{H}(L_i^*) = \{L_1^*, \ldots, L_{i-1}^*\} & \text{for } i = 2, 3, \ldots, k \\
k \text{ nearest neighbors of } L_i^* \text{ among } \mathcal{H}(L_i^*) & \text{for } i = k + 1, \ldots, n
\end{cases}
$$

(3.19)

Based on the reference set equipped with the ordering mechanism specified above, the conditional distribution of $w(L)$ is assumed to depend only on its neighbor sets $w(N(L))$, where $N(L)$ is a collection of $k$ nearest locations in the reference set $L^*$. Given $N$ input locations $L^O \equiv \{L_1, \ldots, L_n\}$, the joint distribution of $w(\cdot)$ over $L^O$ is

$$
p(w(L^O)) = p(w(L_1)) \prod_{i=2}^{n} p(w(L_i) \mid w(N(L_i)))
$$

$$
= \mathcal{N}(w(L^O) \mid 0, \tilde{C}),
$$

where $\tilde{C} \equiv (I - A)^{-1}D(I - A)^{-\top}$. The matrix $A$ is a strictly lower-triangular matrix with no more than $k$ elements in each row. The non-zero elements in the $i$th row are given by $C^{-1}_{N(L_i),N(L_i)}C_{L_i,N(L_i)}$. The matrix $D$ is a diagonal matrix with $i$-th diagonal element given by $C(L_i, L_i) - C_{L_i,N(L_i)}C^{-1}_{N(L_i),N(L_i)}C_{N(L_i),L_i}$. Replacing the full covariance matrix $C$ by $\tilde{C}$ in the likelihood function (3.18) leads to $O(nk^3)$ computational cost and $O(nk^2)$ memory cost. As $k$ is much smaller than $n$, the computations in both model fitting and prediction are reduced tremendously. For detailed model formulation, see Datta et al. (2016).

Let $L_0$ be a new input location. The predictive distribution of $z(L_0)$, given model parameters $\{\beta, \Theta, \tau^2, \sigma^2\}$ and $h(L_0)$, is a normal distribution with predictive mean and variance given below:

$$
\hat{z}(L_0) = h(L_0)^\top \beta + c_0^\top \Sigma_0^{-1} (z_{N(L_0)} - X_{N(L_0)}\beta),
$$

(3.20)

$$
\hat{\sigma}^2(L_0) = \sigma^2 - c_0^\top \Sigma_0^{-1} c_0 + \tau^2,
$$

(3.21)

where $c_0 \equiv (C(L_0, L_1), \ldots, C(L_0, L_n))^\top$ and $\Sigma_0 \equiv C_{N(L_0),N(L_0)}$. 

14
For a typical GP model, the computational cost of performing Cholesky decomposition of the covariance matrix would be $10849^3 \approx 1.3 \times 10^{12}$ flops for each FPCA weight. In contrast, the NNGP model requires $8.7 \times 10^7$ flops with 20 nearest neighbors. The computational complexity is reduced substantially based on the active subspace and FPCA via the NNGP model compared to the typical Gaussian process model.

4 Application to the FP Forward Model

In this section, we construct the emulator in Section 3 with the NASA OCO-2 data. As the input space is of high dimensionality, a large number of samples from the input space are required, so that the input space is well explored by these inputs. In addition we must ensure that the samples from the input space span realistic conditions that OCO-2 could plausibly encounter. Moreover, the number of samples need to be large enough in order to select the active subspace that not only well represent the input space but also reduce the Monte Carlo approximation error. Therefore we have assembled a sample of $n = 10,894$ state vectors and FP forward model evaluations from the full OCO-2 record of land retrievals for February 2015. This collection has also been used by the OCO-2 team for retrieval error analysis, following Connor et al. (2016), and gives suitably global coverage for constructing the emulator. The viewing geometry $b$ parameters are determined from the location and time of an individual sounding, along with the satellite’s orientation.

To represent the functional output via a basis function expansion, we perform FPCA on the radiances in log scale separately for each of the three bands, viz. $O_2$ band, WCO$_2$ band and SCO$_2$ band, since the wavelengths of these three bands do not overlap. The B-spline basis functions are used to smooth the simulated radiance data and convert them into a functional form of wavelength using the R package fda (Ramsay et al., 2018). The eigenvalues in Figure 3 show that the first three principal components for the $O_2$ band can explain more than 99% of variance in the output. For the WCO$_2$ and SCO$_2$ bands, the first principal component can explain more than 99% of variability. Figure 4 shows the true radiances and the reconstructed radiances via the FPCA approach for three different soundings. We can see that the reconstructed radiances, using the first three functional principal components for $O_2$ band and one component each for WCO$_2$ and SCO$_2$ band, are very close to the true radiances. This indicates that the main modes of variation of the OCO-2 radiance vectors can be explained through a very small number of functional principal components. This will lead to a significant computation gain in building the emulator, as we only need to model a
To reduce the dimensionality of the input space, we first use principal component analysis (PCA), which is a popular and well-established tool for extracting important information from multivariate data (Jolliffe 2011). We perform PCA on the standardized input data $\{(x_i, b_i)\}_{i=1}^N$. Figure 5 shows that the first 13 principal components can explain 94.8% of the total variation. Notice that the PCA approach only uses input information for input dimension reduction. In contrast, the active subspace method is a better alternative dimension reduction approach than the PCA approach in the OCO-2 application, since it incorporates gradient information of the FP forward model to perform dimension reduction. As we only have gradient information of the FP forward model with respect to the state vector $x$, the active subspace approach is only applied to state vectors. For the viewing geometry, we do not perform any dimension reduction on it, since it only has dimension 4. Figure 6 shows that four active variables are enough to capture above 95% of the variability. There is a sharp change in the eigenvalues starting from the fourth eigenvalue. This suggests that four active variables can be used with theoretical justification given in Constantine et al. (2014). It is expected that the active subspace approach results in a smaller number of important variables, since this approach uses the gradient information of the forward model to select active variables that have most impact on the outputs. The predictive performance of the
Fig. 4. The FP forward model simulated radiances and the corresponding reconstructed radiances using the FPCA over the three bands at sounding indices 15, 19, and 50 from top to bottom.

emulator using the PCA approach and the active subspace approach is investigated in the next section.
Fig. 5. Diagnostics in the PCA method. The left panel shows the eigenvalues. The right panel shows the cumulative proportion of variation explained.

Fig. 6. Diagnostics in the active subspace method. The left panel shows the eigenvalues. The right panel shows the percentage of cumulative variations.
4.1 Model comparison with PCA and LaGP

After dimension reduction is performed for both input space and output space, we use the NNGP to model the relationship between the FPCA weights and the low-dimensional inputs. As a comparison, we also use the local approximate Gaussian process approximation (Gramacy and Apley, 2015). This method will be referred to as LaGP hereafter. LaGP can be seen as an update on the local kriging approach in spatial statistics (Vecchia, 1988). For both NNGP and LaGP, the predictive performance is evaluated for $n^* = 1,000$ randomly held-out inputs based on PCA and active subspace.

To measure the predictive performance, we computed three criteria: root-mean-squared-prediction error (RMSPE), coverage probability of 95% credible interval ($P_{CI}(95\%)$), and continuous-rank-probability score (CRPS; Gneiting and Raftery, 2007). The RMSPE measures the discrepancy between the forward model output and predictive mean. The coverage probability of the 95% credible interval measures the percentage of 95% credible intervals covering the forward model output. The CRPS measures the uncertainties in the predictive distribution. They are defined as follows:

$$\text{RMSPE} = \left\{ \frac{1}{n^*} \sum_{i=1}^{n^*} (z_i - \hat{z}_i)^2 \right\}^{1/2},$$

$$P_{CI}(95\%) = \frac{1}{n^*} \sum_{i=1}^{n^*} I(z_i \in [\hat{z}_{i.025}, \hat{z}_{i.975}]),$$

$$\text{CRPS} = \frac{1}{n^*} \sum_{i=1}^{n^*} \hat{\sigma}_i \left\{ -\frac{1}{\sqrt{\pi}} + 2\phi \left( \frac{z_i - \hat{z}_i}{\hat{\sigma}_i} \right) + \frac{z_i - \hat{z}_i}{\hat{\sigma}_i} \left( 2\Phi \left( \frac{z_i - \hat{z}_i}{\hat{\sigma}_i} \right) - 1 \right) \right\},$$

where $z_i$ is the FP forward model output for the $i$th held out input, $\hat{z}_i$ is the (posterior) predictive mean for $z_i$, and $\hat{\sigma}_i$ is the (posterior) predictive standard deviation for $z_i$. In addition, $\hat{z}_{i.025}$ and $\hat{z}_{i.975}$ are the 2.5% and 97.5% percentiles of the (posterior) predictive distribution of $z_i$. Further, $\phi$ and $\Phi$ denote the probability density function and the cumulative distribution function, of a standard Gaussian random variable, respectively.

In the cross-validation study, the NNGP is implemented with 20 nearest neighbors in MATLAB. The reference set is chosen to be the set of training inputs $\mathcal{L}^O$ whose elements are ordered by the first active variable. Other types of ordering strategies give similar prediction results, and increasing the number of nearest neighbors achieves negligible predictive performance gain. This is also consistent with the findings in Datta et al. (2016). The LaGP is implemented with R package laGP (Gramacy, 2016). It involves approximating the predic-
tive equations at a particular generic location, via a subset of the data, where the sub-design is (primarily) comprised of a neighbourhood close to the location. These design points are chosen using a sequential approach using different criteria, viz. active learning Cohn (ALC), minimum mean square prediction error (MSPE) and nearest neighbors (NN).

In terms of the choice between PCA and active subspace, the results in Table 1 and Table 2 show that the active subspace approach gives a slightly better predictive performance than the PCA approach in terms of RMSPE, coverage probability and CRPS. The emulator with the PCA approach requires about five more parameters than the emulator with the active subspace approach. This saves significant computing time when fully Bayesian inference is performed in the NNGP.

To compare NNGP and LaGP, we fit the LaGP emulator using all the three suggested design criteria. As expected the NN method gives the worst performance whereas the ALC and MSPE methods give comparable results. The maximum number of neighbors for each prediction point is kept at 100. Only a negligible reduction is obtained in the mean square prediction error when the maximum number of neighbors for each prediction point is increased to more than 100. The prediction results for the validation data set are reported in Table 1 and 2 for input dimension reduction approaches based on PCA and active subspace, respectively. We can see from these results that NNGP outperforms LaGP in most cases in terms of RMSPE, coverage probability, and CRPS. Note that the coverage probability column is left blank for the LaGP results at the original radiance scale. The R package output for LaGP only provides the mean prediction and the standard error but not the posterior samples. Hence it is difficult to compute the coverage probability in the original scale.

It is worth noting that the NNGP gives better results with just 20 neighbors, much fewer than that of LaGP, which uses a maximum of 100 neighbors. The NNGP emulator is much more expensive to fit when compared to LaGP, since the NNGP uses fully Bayesian inference and the LaGP uses a fast approximation. However, once model fitting is completed, the prediction is almost instantaneous for both methods. Moreover, the LaGP is primarily built for fast prediction at some pre-specified input configurations. In the OCO-2 application, the primary objective to construct an emulator to replace the FP forward model in an entire OSUE. For instance, the prediction from the emulator is needed multiple times in the retrieval algorithm in order to estimate the atmospheric state vectors. Thus, the corresponding input configurations for these predictions are not known beforehand, making the LaGP emulator not applicable in OSUEs. The NNGP is a process based model and hence prediction can be obtained at any arbitrary input configuration from the input space. This flexibility is
another justification for an NNGP emulator instead of a LaGP emulator for OSUEs in the OCO-2 application.

Table 1. Numerical comparison of held-out sets based on NNGP and LaGP using the PCA approach. The predictions are obtained at \( n^* = 1000 \) new input values. For each spectrum band, we compared RMSPE, \( P_{CI}(95\%) \), and CRPS. The results are reported for both functional principal scores and radiance data at original scale.

|                | NNGP         | LaGP         |
|----------------|--------------|--------------|
|                | RMSPE | \( P_{CI}(95\%) \) | CRPS | RMSPE | \( P_{CI}(95\%) \) | CRPS |
| Results for each functional principal score |            |              |      |            |              |
| \( O_2 \) (PC 1) | 0.0231 | 0.9460 | 0.0125 | 0.0339 | 0.896 | 0.0189 |
| \( O_2 \) (PC 2) | 0.0110 | 0.9670 | 0.0061 | 0.0149 | 0.919 | 0.0084 |
| \( O_2 \) (PC 3) | 0.0022 | 0.9990 | 0.0016 | 0.0028 | 0.923 | 0.0015 |
| Weak CO\(_2\) (PC 1) | 0.0356 | 0.9480 | 0.0192 | 0.0488 | 0.927 | 0.0270 |
| Strong CO\(_2\) (PC 1) | 0.0591 | 0.9430 | 0.0316 | 0.0841 | 0.934 | 0.0461 |
| Results for radiances at each spectrum band |            |              |      |            |              |
| \( O_2 \) | 0.2425 | 0.9072 | 0.2308 | 0.3432 | —       | —       |
| Weak CO\(_2\) | 0.2537 | 0.9457 | 0.2398 | 0.3477 | —       | —       |
| Strong CO\(_2\) | 0.3529 | 0.9221 | 0.3225 | 0.4855 | —       | —       |

Table 2. Numerical comparison of held-out sets based on NNGP and LaGP using the active subspace approach. The predictions are obtained at \( n^* = 1000 \) new input values. For each spectrum band, we compared RMSPE, \( P_{CI}(95\%) \), and CRPS. The results are reported for both functional principal scores and radiance data at original scale.

|                | NNGP         | LaGP         |
|----------------|--------------|--------------|
|                | RMSPE | \( P_{CI}(95\%) \) | CRPS | RMSPE | \( P_{CI}(95\%) \) | CRPS |
| Results for each functional principal score |            |              |      |            |              |
| \( O_2 \) (PC 1) | 0.0196 | 0.9560 | 0.0109 | 0.0312 | 0.948 | 0.0194 |
| \( O_2 \) (PC 2) | 0.0096 | 0.9650 | 0.0053 | 0.0150 | 0.927 | 0.0085 |
| \( O_2 \) (PC 3) | 0.0022 | 0.9990 | 0.0016 | 0.0029 | 0.969 | 0.0016 |
| Weak CO\(_2\) (PC 1) | 0.0330 | 0.9500 | 0.0177 | 0.0631 | 0.930 | 0.0354 |
| Strong CO\(_2\) (PC 1) | 0.0588 | 0.9420 | 0.0308 | 0.1120 | 0.931 | 0.0633 |
| Results for radiances at each spectrum band |            |              |      |            |              |
| \( O_2 \) | 0.2114 | 0.8978 | 0.1910 | 0.3464 | —       | —       |
| Weak CO\(_2\) | 0.2351 | 0.9502 | 0.2238 | 0.4486 | —       | —       |
| Strong CO\(_2\) | 0.3486 | 0.9334 | 0.3181 | 0.7415 | —       | —       |

4.2 Model validation with a reduced order model

Surrogate models have been developed to approximate expensive computer models from a statistical perspective and mathematical perspective in the Uncertainty Quantification
community. The emulator developed in this article is based on a statistical model that allows assessment of uncertainties in predicting the real-world physical process in the OCO-2 application, while the ROM introduced by Hobbs et al. (2017) is formulated by simplifying the physical laws in the real-world process. It is of fundamental and practical interest to make comparison between these two types of surrogate models. In this section, we only focus on the comparison of predictive performance between the statistical emulator and the ROM.

Following Section 4.1, we run the ROM over these 1,000 held-out inputs. We compare the RMSPE obtained based on NNGP and ROM for radiance data. Figure 7 shows boxplots of RMSPE over 1000 held-out inputs. We can see that the statistical emulator based on NNGP outperforms the ROM at each band in terms of RMSPE. For the O2 band, the statistical emulator based on NNGP gives smaller RMSPE than the ROM at most of the soundings. For the WCO2 and SCO2 bands, the NNGP based statistical emulator outperforms the ROM, with the exception that the former gives much larger RMSPE at a few soundings. Figure 8 shows that the true radiances from the FP forward model, the predictive mean from NNGP and the radiances generated from the ROM for three different soundings which are held out. It is clearly noticed that the predictive mean from the NNGP based emulator is closer to the true radiances from the FP forward model than those from the ROM. The ROM produces noticeably lower radiances at certain wavelengths in all three bands. This discrepancy is likely due to the computational fidelity of the instrument model used in the ROM versus the FP forward model. The instrument model computations add a substantial computational cost to the FP forward model, and this component was simplified extensively in the ROM (Bösch et al. 2017; Hobbs et al. 2017).

Fig. 7. Boxplot of RMSPE based on NNGP and ROM.

As each input is also associated with a longitude and latitude coordinate in space, we can
Fig. 8. The FP forward model simulated radiances, the corresponding ROM predictions, and the predicted radiances using the FPCA and NNGP emulator over the three bands at sounding indices 15, 19, and 50 from top to bottom.
Fig. 9. Spatial map of RMSPE based on NNGP and its difference between ROM and NNGP. The top panels show the RMSPE over 1000 held-out inputs for the O\textsubscript{2} band (top-left panel), WCO\textsubscript{2} band (top-middle panel) and SCO\textsubscript{2} band (top-right panel) based on the emulator with NNGP. The bottom panels show the difference of RMSPE between the ROM and the emulator with NNGP for the O\textsubscript{2} band (bottom-left panel), WCO\textsubscript{2} band (bottom-middle panel) and SCO\textsubscript{2} band (bottom-right panel).

The top panels of Figure 9 show that the O\textsubscript{2} band has comparatively smaller RMSPE than the WCO\textsubscript{2} and SCO\textsubscript{2} bands. There is not clearly spatial pattern on large value of RMSPE across these three bands. The bottom panels of Figure 9 show that the ROM generally gives larger RMSPE than the emulator based on NNGP with exceptions at a few locations for the WCO\textsubscript{2} band. This agrees with previous results in Figure 7. The ROM seems to give larger RMSPE in the middle latitude bands than the emulator based on NNGP. It is worth noting that the ROM can generate radiances for any given input, but there is no uncertainty associated with these radiances. Thus, this NNGP-based emulator also has an advantage over the ROM in terms of assessing the uncertainties in predictions.
5 Discussion

This article focuses on an important scientific application in remote sensing science for quality assessment and uncertainty quantification of remote sensing retrievals. As current assessment of remote sensing retrievals relies on the full-physics forward model, uncertainty quantification for the retrieval problem via OSUEs can be computationally too demanding even with modern high-performance computing resources. To tackle this problem, a reduced order model has been developed to preserves key physical laws in the physical process of interest in the OCO-2 application; see Hobbs et al. (2017). This article takes a statistical approach to constructing an emulator that can inherently assess the accuracy of the approximation to an expensive physical forward model with probabilistic uncertainties. This emulator not only allows fully Bayesian inference but also provides a fast probabilistic approximation so that the quality assessment and uncertainty quantification of remote sensing retrievals can be facilitated. This methodology can be readily extended to other remote sensing instruments.

The proposed emulator is formulated in three steps, resulting in a computationally efficient non-separable GP model with high-dimensional functional output. First, we introduce the functional principal component analysis to deal with high-dimensional output at irregularly-spaced wavelengths with missing values. Second, we use the gradient information to reduce the input space dimension via active subspace approach that induces a geometrically anisotropic correlation function to model each input dimension. Third, we adapt a computationally efficient NNGP in the computer model framework and compare it to the local GP approach (Gramacy and Apley, 2015). The current approach to approximating the functional output does not take into account the truncation error explicitly. This can be further improved by adding a white-noise error process. The dimension reduction approaches used in the output space and input space are currently performed separately, which allows fast computations. A joint dimension reduction approach for both input space and output space can be appealing, but this can be challenging.

The proposed emulator has been proven successful in the OCO-2 application. Cross-validation studies have been used to compare the statistical emulator with other statistical methods and a reduced order model. Validating the statistical emulator with an existing physics-based surrogate model developed by Hobbs et al. (2017) is particularly interesting in the UQ community. Our results suggest that the statistical emulator has very competitive performance and even gives better prediction results than the reduced order model. This
is especially encouraging in the sense that a statistical emulator that is purely data-driven can be advantageous in remote sensing applications over a reduced order model that is built based on simplified physical laws in the real-world process.

Acknowledgements

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute. Part of this work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and NASA.

Appendix

A Correlation functions

The following correlation functions can be written as functions of a Mahalanobis distance:

- The Matérn family:
  \[ \rho(s, s') = \frac{2^{1-v}}{\Gamma(v)} \left( \sqrt{2vd} \right)^v K_v \left( \sqrt{2vd} \right), \]
  where \( d = \sqrt{\sum_{i=1}^{p} (s_i - s_i')^2 / \theta_i^2} \) is the distance between two inputs. \( K_v(\cdot) \) is the modified Bessel function of the second kind, \( \theta_i \) is a range parameter, and \( v \) is a smoothness parameter controlling the differentiability of the Gaussian process. This correlation function is an anisotropic Matérn correlation function.

- The power exponential family:
  \[ \rho(s, s') = \exp\{-d^\alpha\}, \]
  where \( d = \sqrt{\sum_{i=1}^{p} (s_i - s_i')^2 / \theta_i^2} \) is the distance between two inputs. \( \theta_i \) is a range parameter. \( \alpha \) is the smoothness parameter. When \( \alpha = 2 \), the correlation becomes an anisotropic Gaussian correlation function; when \( \alpha = 1 \), the correlation becomes an anisotropic exponential correlation function.
While implementing FPCA the radiance vectors are transformed to a functional form (as a function of wavelength) by fitting cubic B-spline basis to the observed data, i.e., in (3.9) we consider $\phi_j(\omega) = N_{j,3}(\omega)$, where, $N_{j,3}(\omega)$'s are obtained by the recursive formula

$$N_{j,0}(\omega) = \begin{cases} 1, & \text{if } \omega_j < t < \omega_{j+1} \\ 0, & \text{otherwise} \end{cases}$$

$$N_{j,p}(\omega) = \frac{\omega - \omega_j}{\omega_{j+p} - \omega_j} N_{j,p-1} + \frac{\omega_{j+p+1} - \omega}{\omega_{j+p+1} - \omega_{j+1}} N_{j+1,p-1}, p = 1, 2, 3.$$ 

Here $\omega_1, \omega_2, ..., \omega_{K+1}$ are the called the knot points and are selected equidistantly over the range of wavelengths in each of the three bands. A similar basis representation is also obtained for the mean radiance vector $\bar{y}(\omega)$.

References

Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, 35(5):1874–1906.

Bilionis, I., Zabaras, N., Konomi, B. A., and Lin, G. (2013). Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. *Journal of Computational Physics*, 241:212 – 239.

Bösch, H., Brown, L., Castano, R., Christi, M., Connor, B., Crisp, D., Eldering, A., Fisher, B., Frankenberg, C., Gunson, M., Granat, R., et al. (2017). Orbiting Carbon Observatory (OCO-2) Level 2 Full Physics Algorithm Theoretical Basis Document. URL: [https://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/OCO2_L2_ATBD.V8.pdf](https://docserver.gesdisc.eosdis.nasa.gov/public/project/OCO/OCO2_L2_ATBD.V8.pdf).

Connor, B., Bösch, H., McDuffie, J., Taylor, T., Fu, D., Frankenberg, C., O’Dell, C., Payne, V. H., Gunson, M., Pollock, R., Hobbs, J., Oyafuso, F., and Jiang, Y. (2016). Quantification of uncertainties in OCO-2 measurements of XCO$_2$: simulations and linear error analysis. *Atmospheric Measurement Techniques*, 9.

Constantine, P. and Gleich, D. (2014). Computing active subspaces with Monte Carlo. *arXiv.org*, page arXiv:1408.0545.

Constantine, P. G., Dow, E., and Wang, Q. (2014). Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524.

Constantine, P. G., Eftekhar, A., Hokanson, J., and Ward, R. A. (2017). A near-stationary subspace for ridge approximation. *Computer Methods in Applied Mechanics and Engi-
Constantine, P. G., Kent, C., and Bui-Thanh, T. (2016). Accelerating Markov Chain Monte Carlo with active subspaces. *SIAM Journal on Scientific Computing*.

Conti, S., Gosling, J. P., Oakley, J. E., and O’Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, 96(3):663–676.

Conti, S. and O’Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140(3):640–651.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177.

Cook, R. D. (2009). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, volume 482. John Wiley & Sons, New York, NY.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition.

Crisp, D., Bösch, H., Brown, L., Castano, R., Christi, M., Connor, B., Eldering, A., Fisher, B., Frankenberg, C., Gunson, M., McDuffie, J., Miller, C., Natraj, V., O’Dell, C., O’Brien, D., Polonski, I., Osterman, G. B., Oyafuso, F., Smyth, M., Thompson, D., Toon, G. C., and Spurr, R. (2014). OCO (Orbiting Carbon Observatory)-2 level 2 full physics retrieval algorithm theoretical basis. Technical report, NASA Jet Propultion Laboratory, OCO D-65488, Pasadena. http://co2.jpl.nasa.gov/#mission=OCO-2.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Eldering, A., O’Dell, C. W., Wennberg, P. O., Crisp, D., Gunson, M., Viatte, C., Avis, C., Yoshimizu, J., et al. (2017). The Orbiting Carbon Observatory-2: First 18 months of science data products. *Atmospheric Measurement Techniques*, 10(2):549–563.

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gramacy, R. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *Journal of Statistical Software, Articles*, 72(1):1–46.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Gu, M. and Berger, J. O. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *The Annals of Applied Statistics*, 10(3):1317–1347.

Guillas, S., Sarri, A., Day, S. J., Liu, X., and Dias, F. (2018). Functional emulation of high resolution tsunami modelling over Cascadia. *The Annals of Applied Statistics*, 12(4):2023–2053.
Higdon, D., Gattiker, J., and Williams, B. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583.

Hobbs, J., Braverman, A., Cressie, N., Granat, R., and Gunson, M. (2017). Simulation-based uncertainty quantification for estimating atmospheric CO\(_2\) from satellite data. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):956–985.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, New York, NY.

Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung, volume 37. Sana.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316.

Liu, X. and Guillas, S. (2017). Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):787–812.

Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V., and Wu, C. F. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal of the American Statistical Association*, 113(524):1443–1456.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, NY.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2018). *fda: Functional Data Analysis*. R package version 2.4.8.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding*. World Scientific.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2018). *The design and analysis of computer experiments; 2nd ed*. Springer series in statistics. Springer, New York, NY.

Turmon, M. and Braverman, A. (2019). Uncertainty quantification for JPL retrievals. Technical report, Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2019. http://hdl.handle.net/2014/45978.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B*, 66(2):297–312.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34(1):15.

Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453–470.