https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

Efficient multi-scale Gaussian process regression for massive remote sensing data with satGP v0.1

Jouni Susiluoto[1,2,3], Alessio Spantini[1], Heikki Haario[2,3], and Youssef Marzouk[1]

[1]Massachusetts Institute of Technology, Department of Aeronautics and Astronautics, 77 Massachusetts Avenue, 33-207, Cambridge MA 02139 USA
[2]Lappeenranta University of Technology, School of Engineering Science, P.O. Box 20, FI-53851 Lappeenranta, Finland
[3]Finnish Meteorological Institute, Erik Palménin aukio 1, FI-00560 Helsinki, Finland

Correspondence: Jouni Susiluoto (jsusiluo@mit.edu)

Abstract. Satellite remote sensing provides a global view to processes on Earth that has unique benefits compared to measurements made on the ground. The global coverage and the enormous amounts of data produced come, however, with the price of spatial and temporal gaps and less than perfect data quality. Meaningful statistical inference from such data requires overcoming these problems and that calls for developing efficient computational tools.

We design and implement a computationally efficient multi-scale Gaussian process (GP) software package, satGP, geared towards remote sensing applications. The software is designed to be able to handle problems of enormous sizes and is able to compute marginals and sample from a random process with at least over hundred million observations.

The mean function of the Gaussian process is described by approximating marginals of a Markov random field (MRF). For covariance functions, Matern, exponential, and periodic kernels are utilized in a multi-scale kernel setting to describe the spatial heterogeneity present in data. We further demonstrate how winds can be used to inform the covariance kernel formulation. The covariance kernel parameters are learned by calculating an approximate marginal maximum likelihood estimate and this is utilized to verify the validity of the multi-scale approach in synthetic experiments.

For demonstrating the techniques above, data from the Orbiting Carbon Observatory 2 (OCO-2) satellite is used. The satGP program is released as open source software.

1 Introduction

Climate change is one of the most important current global environmental challenges, to the point where it is drawing constant widespread attention even in mainstream media. The underlying reason is the anthropogenic carbon emissions: among the well-mixed greenhouse gases, carbon dioxide (CO2) has currently the strongest effect on warming the planet, with the radiative forcing of ca. 1.68 W m^{-}2 according to the latest IPCC report (IPCC, 2013).

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

The resulting global interest in atmospheric carbon along with technological advances has resulted in several CO2-measuring satellites continuously monitoring the Earth and producing enormous quantities of data, which are processed to local estimates of CO2 by solving a complicated inverse problem (Crisp et al., 2012). These include the Greenhouse gases Observing SATellite (GOSAT) from Japan (Yokota et al., 2009), which has been operational since January 2009, the OCO-2 from NASA, launched in July 2014, and the Chinese TanSat (Yi et al., 2018), which was launched in December 2016. October 2018 saw the launch of GOSAT-2, and in May 2019 the OCO-3 instrument was taken to the International Space Station. In addition to the CO2-measuring instruments, also other types of data are produced by remote sensing. For instance the European TROPOspheric Monitoring Instrument (TROPOMI) produces measurements of nitrogen dioxide, formaldehyde, carbon monoxide, aerosols, methane, and ozone.

Common denominators among most non-gridded remote sensing data sets include: a large number of observations, global coverage but small area observed at any given time, sensitivity to prevailing weather conditions and cloud cover, unknown and/or unreported error covariances, and predetermined positioning that rules out freely observing at a given time and location. These shortcomings can be partly remedied with computational statistics. The many steps of producing carbon flux estimates from readings produced by satellites are summarized by e.g. Cressie (2018). In this work a tool to solve one of those steps, the production of gridded level 3 data sets with uncertainties from pointwise level 2 column integrated dry air CO2 mole fraction (XCO2) data, is introduced. Even though we demonstrate the capabilities of the software with OCO-2 data, the methods are not constrained by the quantity of interest observed.

The purpose of this manuscript is four-fold. First, to introduce satGP, a fast computer program that estimates Gaussian process covariance and mean function parameters from data, computes posterior marginal distributions, and samples from GP priors and posteriors conditioning on over hundred million observations in situations where several hundred million marginals need to be computed. While lots of advances have recently been made in the field, we are not aware of any literature or software solving problems of quite this scale so far. Second, computational methods that allow the solution of problems of such scales are introduced. Third, covariance function and mean function formulations, some of which we have not seen used in the remote sensing community, are presented. In particular, the multi-scale formulation avoids excessive smoothing, allowing one to see local effects where observations become available. Fourth, these methods are demonstrated with the XCO2 data from the OCO-2 satellite.

Several interesting kriging studies have been published before in the context of satellite measurements of CO2. Zeng et al. (2013) analyzed the variability of CO2 in both space and time over China producing monthly maps from GOSAT data with slightly over 10000 observations. Nguyen et al. (2014) used a four times larger set of observations with Kalman Smoothing in a reduced dimension with GOSAT and the Atmospheric InfraRed Sounder (AIRS) data from NASA. A map of atmospheric carbon dioxide derived from GOSAT data was presented at the higher resolution of 1×1.25 degrees in space and 6 days in time by Hammerling et al. (2012). In another publication by the same authors, synthetic OCO-2 observations were considered with the same spatial resolution.

A global dataset derived from GOSAT was presented by Zeng et al. (2017), with the spatiotemporal resolution of three days and one degree. The results were validated against both Total Carbon Column Observing Network (TCCON) and modeling

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

results from CarbonTracker and the Goddard Earth Observing System with atmospheric chemistry (GEOS-Chem). This study evaluated also the temporal trend of the XCO2. Similarly Tadić et al. (2017) describe a moving window block kriging algorithm to introduce time dependence into GOSAT-based XCO2 map construction process using a quasi-probabilistic screening method for subsampling observations, thinning the data for computational reasons. Other recent studies have also contained analyses of OCO-2 data. For example, Zammit-Mangion et al. (2018) present fixed rank kriging (FRK) results based on OCO-2 data using a 16-day moving window. The results again appear very smooth.

An interesting approach is presented by Ma and Kang (2017), who describe a fused Gaussian process, combining a graphical model with a Gaussian process and applying that to sea surface temperature data. Another interesting approach for atmospheric trace gas inversion is presented by Zammit-Mangion et al. (2015), who simultaneously model both flux fields and concentrations using a bivariate spatiotemporal model, utilizing Hamiltonian Monte Carlo (Neal, 2011) for sampling the posterior. However, due to computational challenges the footprint area is very small.

For overcoming the difficulties posed by large numbers of data, various methods have been proposed. Lindgren et al. (2011) provide an explicit link between some random fields arising as solutions to certain stochastic partial differential equations and Markov random fields. A recent review of Vecchia-type approximations (Vecchia, 1988) is given by (Katzfuss et al., 2018) and a comparison of the performance of several recently developed methods is given by Heaton et al. (2018), with applications to MODIS data. The difficulty of ordering the observations for effective inference with Gaussian processes, especially as the dimension of the inputs grows, is underlined by Ambikasaran et al. (2016).

In this work we describe an approach to solve spatial statistics problems with hundreds of millions of data points. We do this by combining various ideas and techniques that come close to those applied in Vecchia-type and nearest neighbor Gaussian processes while utilizing random sampling and aggressive pre-filtering of uninformative data when possible. The presentation of the general Gaussian process problem is based on the one given by Santner et al. (2003) and Rasmussen and Williams (2006).

A generic space and time dependent mean function of the Gaussian process is found by solving marginals of a Markov random field (MRF). For covariance modeling, a multi-scale covariance kernel formulation is given. The validity of the multi-scale approach is established via a synthetic study. Approximate methods to learn the parameters of both the covariance kernel and the mean function as implemented in satGP are outlined. Additionally, a non-stationary covariance kernel formulation for utilizing wind data for computation, partly inspired by (Nassar et al., 2017), is proposed.

The capabilities of this early version satGP are demonstrated in practice by computing global XCO2 concentrations for a duration of 1526 days at 0.5° spatial and daily temporal resolution with XCO2 data from OCO-2 utilizing over 116 million observations. The number of computed marginals is over 350 million. An example of how these results look like is given by Fig. 7.

The key advances of this work are the capability to compute Gaussian process predictions with enormous remote sensing data sets, a practical way of learning the multi-scale kernel parameters and mean function parameters from data, and introduction of the flexible open source software, of which this is a first released version. Describing these developments is approached from the perspective of how the various parts of computation are implemented in the current version of satGP.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019


Geoscientific Model Development Discussions

The rest of the manuscript is organized in the following manner: Section 2 describes the methods both generally and as implemented in satGP. An overview of computation in satGP is given in Sect. 3, and Sect. 4 presents and discusses simulation results, including a multi-scale synthetic parameter identifiability study and two applications to the OCO-2 v9 data set. In the concluding Sect. 5 some possible future directions are briefly mentioned.

2 Methods

In geosciences, kriging (Cressie and Wikle, 2001; Chiles and Delfiner, 2012) is often used for performing spatial statistics tasks such as gap-filling or representing data in a grid. The semivariogram models used in kriging are closely related to the covariance models used in the Gaussian process formalism (Santner et al., 2003; Rasmussen and Williams, 2006; Gelman et al., 2013), where instead of learning the variogram model from the data, a form of a covariance function is prescribed and its parameters learned.

Intuitively, one would like to learn properties of a spatio-temporal surface from some observational data of some quantity of interest. To each point in space and time corresponds a Gaussian distribution of that quantity, whose mean and variance can be calculated by solving a local regression problem at each desired point. This can also be crudely thought about as optimally solving a spatio-temporal interpolation problem when the observations have Gaussian errors.

The underlying theory related to Bayesian statistics, Gaussian processes, and Markov random fields is well known and therefore the novel aspects in this section have to do with the computational methods and modifications that are presented, such as observation selection schemes in Sect. 2.6 or approximate marginal maximum likelihood computation in Sect. 2.7. These modifications trade precision for tractability, but in a way that the results still remain valid. Due to the size of the problem, some sacrifices need to be made in order to be able to obtain any solution.

This section goes through the Gaussian process formalism, and both generic and the satGP-specific forms of mean and covariance functions are described. This is followed by discussion of how observation selection is carried out and how model parameters are learned.

2.1 Gaussian process regression

A Gaussian process is a stochastic process, which can be thought of as an infinite-dimensional Gaussian distribution in that the joint distributions at any finite set A of space-time points are multivariate normal. We denote the vector of these points by $x \in \mathbb{R}^q$ and underline that they contain both space and time components. In this work $q = 3$, even though this restriction can be overcome if needed, and satGP does have limited support for space-only problems.

The Gaussian process is denoted by

$$\Psi(x) \sim \mathrm{GP}(m(x;\beta), k(x,x';\theta)), \tag{1}$$

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

where $m : \mathbb{R}^q \to \mathbb{R}$ and $k : \mathbb{R}^{q2} \to \mathbb{R}$ are respectively the mean and covariance functions of the process parameterized by hyperparameter vectors $\beta \in \mathbb{R}^{n_\beta}$ and $\theta \in \mathbb{R}^{n_\theta}$. Note, that with these functions x and x' refer to coordinates of a single location in the spatio-temporal domain, while below it may also refer to multiple locations, depending on context.

The function m above is called the drift in kriging literature, and the expected value of the process in areas with no data will tend to the value of the mean function in that area. It is chosen to reflect the deterministic patterns in the data, and these choices also affect how the function k and parameters θ in Eq. (1) need to be chosen. With inadequate modeling of the mean function, the obtained uncertainty estimates may end up being unnecessarily large. For instance linear trends, constant factors, seasonal and other periodic fluctuations should be included if they are known. An example of what is used with the OCO-2 data is shown later in Eq. (11).

The covariance function $k(x, x'; \theta)$ controls the smoothness of the draws ψ from Ψ. The parameter vector θ typically contains at least one scale parameter ℓ and a parameter controlling the maximum covariance τ^2. The ℓ parameters correspond to the length scales of the random fluctuations of the realizations around the mean function, and the τ parameters describe the amplitude of that fluctuation. The functions m and k are fully described in Sect. 2.3 and 2.5, respectively. Additional practical guidelines are given in Appendix A.

In what follows the domain $\mathbb{R}^q \ni x$ is divided into two disjoint parts, one of which, $\mathcal{X}^{\text{train}} \subset \mathbb{R}^q$, contains the part where observation data (training data) was measured, and another one, $\mathcal{X}^{\text{test}} = \mathbb{R}^q \backslash \mathcal{X}^{\text{train}}$, where observations were not made. Any $x \in \mathcal{X}^{\text{test}}$ is below called *test input* as is often done in the GP literature, and these points are generally denoted by x^*.

In practice marginals of the random function Ψ in Eq. (1) or samples ψ from it are evaluated (computed) only at a finite set of points. Let $\psi^{\text{obs}} \in \mathbb{R}^n$ denote a vector of observations — synthetic or real — generated by the Gaussian process at locations $x^{\text{obs}} \in \mathbb{R}^{n \times q}$. Given a set of functions f_i for constructing the mean function, the matrix with elements $f_i(x_j; \delta(x_j^{\text{s}}))$ corresponding to locations x_j with regression coefficients $\beta(x_j^{\text{s}})$ is denoted by $F(x)$. For a single input, instead of $F(x)$ the notation $f : \mathbb{R}^q \to \mathbb{R}^{n_\beta}$ is used, and with that, $f(x^*) = [f_1(x^*), \ldots, f_{n_\beta}(x^*)]^T$. The joint distribution of the field at observed locations is then given by

$$\psi^{\text{obs}} \sim \mathcal{N}\left(F(x^{\text{obs}})\beta, K)\right), \tag{2}$$

where the covariance matrix K is defined by its elements $K_{i,j} = k(x_i^{\text{obs}}, x_j^{\text{obs}}; \theta)$. For the mean function, in this work a specific form

$$m(x; \beta, \delta) = f(x; \delta)^T \beta(x) \equiv \tilde{f}(x^{\text{t}}; \delta(x^{\text{s}}))^T \beta(x^{\text{s}}) \tag{3}$$

is used, where the superindexes s and t refer to the spatial and temporal parts of the generic coordinate x, respectively, and $\delta(x^{\text{s}})$ are auxiliary parameters which are potentially space-dependent. The purpose of the function $\tilde{f}$ is purely illustrative, showing that given the parameters δ, the function f does not depend on the spatial part of x, and similarly that the β parameters do not depend on x^{t}. This definition of m is very general and can describe in practice a large number of realistic scenarios. However, the form of Eq. (3) imposes the strong assumption of separation of space and time in that the β and δ parameters do not depend on time. The explicit form of functions f_i used to model the OCO-2 data are given below in Sect. 2.3.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

Bayesian statistics is a standard paradigm for analyzing data and uncertainties, and it is also widely used in geosciences (Rodgers, 2000; Gelman et al., 2013). From the vantage point it provides, given the observed data $\Psi^{\text{obs}} = \psi^{\text{obs}}$ at some finite set of points x^{obs}, the object of interest of the inference problem in this work is the joint posterior distribution of the Gaussian process and the parameters,

$$p(\psi, \beta, \delta, \theta | \psi^{\text{obs}}) = \frac{p(\psi^{\text{obs}} | \psi, \beta, \delta, \theta) p(\psi | \beta, \delta, \theta) p(\beta, \delta, \theta)}{p(\psi^{\text{obs}})}, \tag{4}$$

where $p(\psi|\beta,\delta,\theta)$ is the Gaussian process prior and $p(\beta,\delta,\theta)$ is a prior on the Gaussian process hyperparameters. This calculation is not generally tractable for a huge number of inputs x, but posterior estimates of the GP, $p(\psi|\psi^{\text{obs}}, \hat{\beta}, \hat{\delta}, \hat{\theta})$, can be calculated by conditioning on parameter point estimates $\hat{\theta}$, $\hat{\beta}$, and $\hat{\delta}$. The first of these may be found by minimizing some loss function $\mathcal{L}$, described below in Sect. 2.7,

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta), \tag{5}$$

and for the second a closed-form expression, given a point estimate of the parameters θ and δ, is given by

$$\mathbb{E}[\beta | \Psi^{\text{obs}} = \psi^{\text{obs}}, \theta, \delta] = (F^T K^{-1} F)^{-1} F^T K^{-1} \psi^{\text{obs}} \tag{6}$$

$$\mathbb{V}[\beta | \Psi^{\text{obs}} = \psi^{\text{obs}}, \theta, \delta] = (F^T K^{-1} F)^{-1}. \tag{7}$$

The δ parameters can be found approximately by finding a point estimate of parameters β and δ before computing Eq. (6), and by re-calibrating δ alone after. In practice this produces stable results with the OCO-2 data, and for pathological data sets, repeated alternating optimization of the parameters may be performed.

Even though a full posterior distribution of the parameters is not obtained this way, the solution of the Gaussian process itself is Bayesian in that the posterior marginals at each x are found by conditioning on the observations. In the satGP software, the space-dependent β and δ parameters are fitted first, and any learning of the covariance parameters is done only after that.

For prediction in the context of Gaussian random functions, the properties of multivariate normal distributions are exploited for calculating marginals of the random field Ψ at any set of points x.

The posterior distribution $p(\psi^* | \psi^{\text{obs}}, \hat{\theta}, \hat{\beta})$ of the Gaussian process at a finite set of test inputs x^* can, given point estimates $\hat{\beta}$ and $\hat{\theta}$, be modeled according to Eq. (2) with

$$\begin{pmatrix} \Psi^* \\ \Psi^{\text{obs}} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} f(x^*)^T \\ F(x^{\text{obs}}) \end{bmatrix} \hat{\beta}, \begin{bmatrix} K(x^*, x^*) & K(x^*, x^{\text{obs}}) \\ K(x^{\text{obs}}, x^*) & K(x^{\text{obs}}, x^{\text{obs}}) \end{bmatrix} \right) \tag{8}$$

where Ψ and x have been divided into two parts - one for the test inputs x^*, and the other one for the observations x^{obs}. The predictive distribution at x^* can then be written as $\Psi^* | \hat{\beta}, \hat{\theta} \sim \mathcal{N}(\mu^*, \Sigma^*)$, where its moments are given by

$$\mu^* = f(x^*)^T \hat{\beta} + K(x^*, x^{\text{obs}}) K(x^{\text{obs}}, x^{\text{obs}})^{-1} (\psi^{\text{obs}} - F\hat{\beta}) \tag{9}$$

and

$$\Sigma^* = K(x^*, x^*) - K(x^*, x^{\text{obs}}) K(x^{\text{obs}}, x^{\text{obs}})^{-1} K(x^{\text{obs}}, x^*), \tag{10}$$

and where the covariance Σ^* is the Schur complement of $K(x^*, x^*)$.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

2.2 Overview and objectives of satGP

The satGP program is meant to be a general purpose Gaussian process toolbox with emphasis on applicability to large remote sensing datasets. It features a selection of covariance kernels and routines for learning space-dependent mean function parameters and covariance parameters from data. With a given set of parameters, it computes posterior marginals and uncertainties at the spatial resolution desired by the user, or generates samples from the process. Drawing samples from the prior is also supported, and this can be utilized for devising synthetic data experiments to study the identifiability of the GP covariance kernel parameters. This section goes through these capabilities and relevant computational details. Since the softwere is applied in Sect. 4 to OCO-2 data, details pertaining to that particular case are included for illustration.

2.3 Mean functions in satGP

The most general mean function form available in satGP is given by Eq. (3). The functions f_i above are user-defined and, for ease of use, functionality for using a zero mean function, a spatially independent mean function, and an arbitrary gridded array of values are available. The specific forms of f_i used for the OCO-2 experiments in Sect. 4 are given by

$$\left.\begin{aligned} f_1(x) &= \sin\left(2\pi x^{\mathrm{t}}\Delta_{\mathrm{year}}^{-1} + \delta_{x^{\mathrm{s}}}\right) \\ f_2(x) &= \cos\left(4\pi x^{\mathrm{t}}\Delta_{\mathtt{year}}^{-1} + \delta_{x^{\mathrm{s}}}\right) \\ f_3(x) &= 1 \\ f_4(x) &= x^{\mathrm{t}} \end{aligned}\right\} \tag{11}$$

where Δ_{year} is the duration of one year, and $\delta_{x^{\mathrm{s}}}$ is a space-dependent phase shift. The function f_1 fits the summer-winter cycle, and f_2 fits the semiannual cycle. It is assumed that these can be modeled with the same $\delta_{x^{\mathrm{s}}}$ parameters. The constant term is given by f_3, and f_4 gives the slow global trend. The fit to the global mean values of XCO2 from OCO-2 can be seen in Fig. 1.

2.4 Learning $\beta(x^{\mathrm{s}})$ as a Markov random field

When not learning GP covariance parameters or generating synthetic training sets, the finite set of test inputs x^* for GP calculation is taken in satGP to be a grid with predefined geographical and temporal extents and resolution. Solving the GP marginalization and sampling problems then amounts to solving Eq. (9) and (10) at each corresponding space-time point. Since e.g. sources, sinks and timing of seasons are local, the mean function should be different from one spatial grid point to another. This is achieved by modeling the $\beta(x^{\mathrm{s}})$ parameters as a Markov random field, which are often used in geophysics as a computational tool to solve large spatial statistics or inference problems. In practice what follows explains how the spatial dependence can be resolved using computational statistics. The MRF imposes the condition that neighboring grid cells should not be too different from each other. How different they are allowed to be is a modeling choice, see Appendix A.

This MRF is an undirected graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Lauritzen, 1996) with the set of vertices $\mathcal{V} = \{\nu^{ij} | i = 1 \dots n_{\mathrm{lat}}, j = 1 \dots n_{\mathrm{lon}}\}$ and edges $\mathcal{E} = \{(\nu^{i,j}, \nu^{i+1,j}) | i = 1 \dots n_{\mathrm{lat}} - 1, j = 1 \dots n_{\mathrm{lon}}\} \cup \{(\nu^{k,l}, \nu^{k,l+1}) | k = 1 \dots n_{\mathrm{lat}}, l = 1 \dots n_{\mathrm{lon}} - 1\}$. The vertices ν^{ij} correspond to the mean function parameters β^{ij} at grid point (i, j). This Markov property implies that the prob-

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions
Open Access
EGU

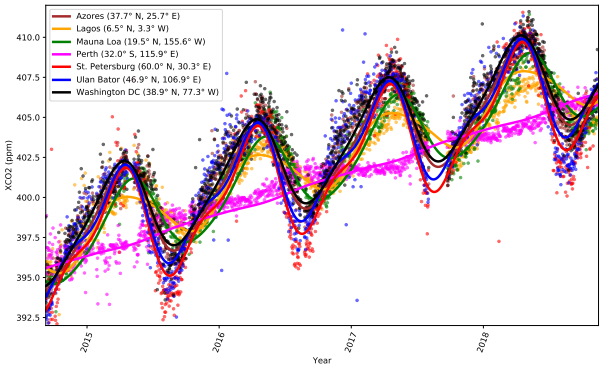

Figure 1. Mean function m with components of f given by Eq. (11). The solid lines give the mean function value, fitted to local data, and the corresponding daily means are shown as dots of the same color. The fit is not perfect at all times due to e.g. smoothness constraints of the field, but it works well as the Gaussian process mean function.

ability of the β-parameters of latitude i and longitude j is given by $p(\nu^{ij}) = \int_{\partial\nu^{ij}} p(\nu^{ij}|\partial_{\nu^{ij}})p(\partial_{\nu^{ij}})$, where $\partial\nu^{ij} = \{\nu \in \mathcal{V}|(\nu,\nu^{ij}) \in \mathcal{E}\}$.

Since the maximal cliques of this graph are the connected pairs of vertices, according to Hammersley and Clifford (1971) the full joint distribution of the graph $p(\mathcal{V})$ factors as $\prod_{(\nu,\nu')\in\mathcal{E}} \frac{1}{Z}\phi(\nu,\nu')$, where Z is called a partition function and ϕ are compatibility functions. One reasonably efficient way to solve marginals for each vertex in such a graph is to use the variable elimination algorithm, which is an exact standard algorithm suitable for undirected graphs of moderate size. To make the computation faster, satGP currently uses a modified version to compute each diagonal in the graph in parallel from $\nu^{0,0}$ to $\nu^{n_{\mathrm{lat}},n_{\mathrm{lon}}}$ and back, conditioning each ν^{ij} on the previously evaluated vertices in $\partial\nu^{ij}$ without introducing the diagonal edges of the reconstituted graph, as would be normally done. The program also inversely weights the edges exponentially according to the distances between the (geographical) coordinates corresponding to the connected nodes. This rate of exponential decay is user-configurable. The structure of the MRF and the approximate elimination order are shown in Fig. 2.

In the particular form used for OCO-2 data in Eq. (11), the phase-shift parameter δ cannot be estimated with regression like β in Eq. (9) and (10). For this reason, the nonlinear space-dependent δ-parameters are found with an optimization algorithm from the NLOpt package, by default the BFGS algorithm, before finding $\hat{\beta}$ with Eq. (9) and (10), and after obtaining $\hat{\beta}$ the δ parameter is re-optimized given the $\hat{\beta}$. For calibrating the δ parameters for vertex ν, the quantity $\sum_{j=1}^{n}(m(x_\nu;\beta_\nu,\delta_\nu) - \psi_j)^2 + \sum_{j'\in\partial\nu}(\delta_\nu - \delta_{j'})^2$ is minimized. Here the first sum runs over the training data selected by the observation selection method described in Sect. 2.6. This optimization problem is very simple since there are few β or δ parameters for the individual vertices. The complexity introduced by the interactions described by the edges is taken care of by the approximate elimination algorithm described above.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions
Open Access
EGU

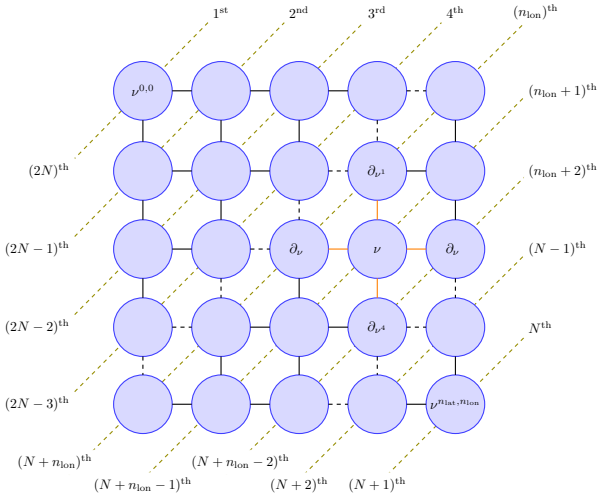

Figure 2. The marginal distribution of vertex ν, $p(\nu)$, is conditional only on the neighbors $\partial_{\nu^1} \dots \partial_{\nu^4}$ (red edges) due to the Markov structure in the pictured lattice graph. Each connected pair is a maximal clique in this particular case. For effective solving, the vertices on the diagonal dashed lines are computed simultaneously making the algorithm non-exact. The order numbers labeling the diagonal lines represent an ordering in which the diagonals can be computed in parallel to get all the marginals in $\mathcal{O}(N)$ wall time, where $N = n_{\text{lat}} + n_{\text{lon}} - 1$. The $(N+1)^{\text{th}}$ computation in the corner is not conditioned on already-computed neighbors to avoid double counting data.

2.5 Covariance functions in satGP

The smoothness, amplitude, and length scale of the Gaussian process are determined by the covariance kernel used, and this choice much determines how the result of the computation looks like. The satGP program supports several different types of covariance function components for forming the full covariance function k in Eq. (1). The options available reflect the properties that can be expected in remote sensing data – varying smoothness and meridional and zonal length scales, potential periodicity, and changing the orientation of the data-informed and uninformed axes according to wind speed and direction. This section lists the available covariance function formulations. For further intuition regarding the parameters, also see Appendix A.

For convenience, let

$$\xi^{\gamma}_{\ell_I}(x, x') = \sum_{c \in I} \left| \frac{x^c - x'^c}{\ell_c} \right|^{\gamma} = \|P^I(x) - P^I(x')\|^{\gamma}_{\Gamma}, \tag{12}$$

where $\gamma > 0$ is the exponent, $I \subseteq \{x^{\text{s}}, x^{\text{t}}\}$ is a set of dimensions of the input, with x^{s} referring to latitude and longitude and x^{t} to time. The P^I matrix projects x onto indices I, and Γ is a diagonal covariance matrix with elements ℓ^{γ}_c, and the notation $\|r\|_{\Gamma}$ stands for $\sqrt{r^T \Gamma^{-1} r}$. The space-only variables are denoted I_S and spatial and temporal variables together are denoted I_{ST}.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

The exponential family of covariance functions with parameters $\theta = (\gamma, l, \tau)$ is defined by the covariance function

$$k_{\text{exp}}(x, x'; \theta, I) = \tau^2 \exp\left(-\xi^{\gamma}_{\ell_I}(x, x')\right). \tag{13}$$

The exponent γ controls the smoothness of the samples from the Gaussian process, with $\gamma = 2$ yielding infinitely differentiable realizations.

The Matérn family of covariance functions, with $\theta = (\nu, \ell_I, \tau)$ is given by the covariance

$$k_{\text{M}}(x, x'; \theta) = \frac{\tau^2 s^{\nu}}{\Gamma(\nu) 2^{\nu-1}} K_{\nu}(s), \tag{14}$$

where $s = 2\sqrt{\nu}\xi^{1}_{\ell_I}(x, x')$ and ν controls the smoothness parameter usually denoted by α via $\alpha = \nu + \frac{q}{2}$. The function K_{ν} is the modified Bessel function of the second kind of order ν. With $q = 1$, the value $\nu = \infty$ corresponds to the squared exponential kernel and $\nu = 0.5$ to the exponential kernel with $\gamma = 1$. Despite this similarity between the Matérn and exponential kernels, the realizations of the random function from the processes with values $\frac{1}{2} < \nu < \infty$ do not correspond to those with the kernel k_{exp} with any value of γ.

A periodic kernel with $\theta = (\tau, \ell_{\text{per}}, \theta_{\text{exp}})$ is defined in satGP by

$$k_{\text{per}}(x, x'; \theta, I) = \\ = \tau^2 \exp\left(-\frac{2\sin^2\left(\pi \left[\frac{x^t - x^{t'}}{\Delta_{\text{period}}}\right]\right)}{\ell^2_{\text{per}}} - \xi^{\gamma}_{\ell_S}(x, x') \right), \tag{15}$$

and the term θ_{exp} defines the parameters for the exponential functions ξ, while ℓ_{per} controls the periodic (inter-period) covariance length. While the periodic kernel is not utilized with the OCO-2 case studies below, it can be a useful tool in many other situations, such as with OCO-3, which due to not being on a Sun-synchronous orbit will make observations at varying local times.

An additional covariance function formulation available in satGP is one based on local wind information. The underlying rationale is that winds affect how quantities of interest such as gases in the atmosphere or algae blooms in the surface water spread. Therefore, if wind data is available, it is natural to use it in the Gaussian process.

The wind-informed covariance has parameters $\theta = (\tau, \ell_I, \rho, w^*)$ and is defined by

$$k_W(x, x'; \theta, I) = k_{\text{exp}}(x_W, x'_W; \theta^W, ST), \tag{16}$$

where the difference between x_W and x'_W is represented using transformed axes parallel and perpendicular to the wind direction at the test input x^*. The spatial scaling parameters in Eq. (13) for k_W, corresponding to the parallel to wind and perpendicular to wind directions, are given by

$$\ell^{\parallel} = \ell\sqrt{1 + w^*\rho} \qquad \ell^{\perp} = \ell, \tag{17}$$

where w^* is the wind velocity at the test input x^* and ρ scales the effect of the wind. The parameter vector for the exponential kernel $\theta^W = (\tau, \gamma, \ell^{\parallel}, \ell^{\perp}, \ell_t, 2)$, where the last element denotes the exponent γ used by the exponential kernel. The resulting covariance ellipses are shown in Fig. 3 for several wind vectors and values of ρ.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions
Open Access
EGU

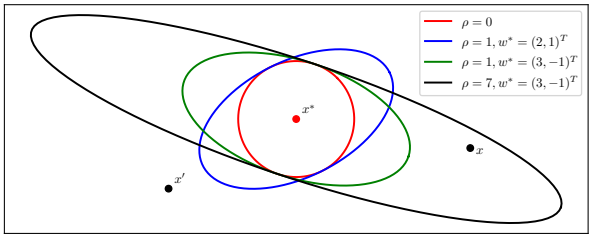

Figure 3. Equicovariance ellipses from the wind-informed kernel with various wind vectors w^* and values of ρ. The wind values are taken at the test input x^*, but the covariance function k is evaluated also for each pair of observations x and x'.

The covariance functions used in this work to model Ψ are sums of several kernels - sums of valid Gaussian process kernels remain valid kernels. The general form of this *multi-scale kernel* is given by

$$k(x,x';\theta) = \delta_{x,x'}\sigma_x^2 + k_{\text{per}}(x,x';\theta,I_S) + k_{\text{M}}(x,x';\theta) + \\ + k_{\text{exp}}(x,x';\theta,I_{ST}) + k_W(x,x';\theta,I), \tag{18}$$

where the first term, which in kriging is called the nugget, contains the observation error variances, and the parameter θ is understood to be different for each component. Not all kernels are included in all experiments - rather, the simulations in Sect. 4 utilize kernels with one to three components. The kernel components of a multi-scale kernel are below called *subkernels*.

2.6 Covariance localization and observation allocation for the multi-scale kernel

Using a large number of observations makes solving the Gaussian process Eq. (9) and (10) untractable as the cost of inverting the covariance matrix scales as $\mathcal{O}(n_{\text{obs}}^3)$. This creates a need for finding approximate solutions while introducing as little error as possible. In satGP covariance localization is used to utilize only a subset of observations for computing Eq. (9) and (10). To do this, a maximum subkernel covariance matrix size κ and minimum covariance parameter σ_{min}^2 are defined by the user.

Assume that the multi-scale kernel defined by the user contains n_{ker} subkernels. For each test input x^* and for each subkernel k_l let the set of observations feasible for inclusion in K in Eq. (8) be

$$A_{*,l}^{\text{obs}} = \{\psi_i \in \psi^{\text{obs}} | k_l(x_i,x^*) < \sigma_{\text{min}}^2,\ \psi_i \notin A_{*,j}^{\text{obs}}\ \forall j < l\}, \tag{19}$$

where the last condition prevents observations from being added by several subkernels. From these candidate observations, $\min(|A_{*,l}^{\text{obs}}|,\kappa)$ are selected, either greedily selecting the κ observations with highest $k(x_i,x^*)$, or choosing the observations uniformly randomly sampling from those training data for which the minimum covariance threshold is exceeded, see Appendix A for additional details. When $|A_{*,l}^{\text{obs}}| < \kappa$ and $l < n_{\text{ker}}$, the parameter κ will be grown for the next kernel to compensate for the deficit by setting $\kappa \leftarrow \kappa + (\kappa - |A_{*,l}^{\text{obs}}|)$. This is done to allow the full kernel size to grow to $n_{\text{ker}}\kappa$ when possible.

Since the kernels are handled sequentially, the order of the different kernels may slightly affect which observations are selected due to the exclusion in Eq. (19), and to grow the full kernel to size $n_{\text{ker}}\kappa$ as often as possible, it is recommended to

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

specify the subkernel with the largest ℓ parameters as the last one. After selecting all observations for all kernels, the covariance matrix K is constructed by evaluating the full covariance function k according to Eq. (18) for all pairs of selected observations.

For learning the locally varying parameters in the mean function with Eq. (6) – (7), the observation selection is performed by disregarding the time component, i.e. setting $x_i^{\mathrm{t}} \leftarrow x^{*\mathrm{t}}$ for all x_i.

Observation allocation could be done also by selecting observations based on values of k instead of each k_l individually, or by other approaches, such as the one presented by Schäfer et al. (2017). However, while the method of observation selection does have an effect on the inferred posterior marginals, the screening property of Gaussian processes ensures that this effect is not major as long as observational noise is small and the nearest observations are included in all directions.

Out of the two methods available in satGP, random selection avoids observation sorting and is therefore faster, especially if a huge number of data are near the test input x^*. This comes at the cost of producing slightly noisier fields of marginal posterior means. For covariance parameter estimation random selection works well. The current nearest-neighbor-in-covariance approach is only one possibility, but is justified by the parameter identifiability results in Sect. 4.1.

2.7 Learning the covariance parameters $\boldsymbol{\theta}$

From Eq. (4), the log marginal likelihood of observations ψ^{obs} given a set of parameters θ, β and δ is given by

$$2\log p(\psi^{\mathrm{obs}}|\beta,\delta,\theta) = -\|(\psi^{\mathrm{obs}} - F\beta)\|_K - \log|K| - n_{\mathrm{obs}}\log(2\pi), \tag{20}$$

where the covariance function parameters θ are implicitly in K and the non-linear mean function parameters in F, for which the shorthand notation $F = F(x^{\mathrm{obs}})$ is used in this section. The maximum (marginal) likelihood estimate (MLE) $\hat{\theta}$ of θ can be found via minimizing

$$\mathcal{L}(\theta) = \left\{\|(\psi^{\mathrm{obs}} - F\hat{\beta})\|_K + \log|K| + n\log(2\pi)\right\} \tag{21}$$

as stated in context of Eq. (5).

In the presence of a huge number of observations, calculating the determinant of the full covariance $|K|$ is not feasible, and the log likelihood is approximated with the block diagonal form, resulting in

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta} \sum_{x_i \in E_{\mathrm{ref}}} \left\{\|(\psi_i^{\mathrm{obs}} - F_i\hat{\beta})\|_{\tilde{K}_i} + \log|\tilde{K}_i|\right\}, \tag{22}$$

where E_{ref} is a set of randomly sampled points from the specified spatio-temporal domain. While the selection of inputs included in E_{ref} has an effect on the obtained parameter estimate, that effect has proven in simulations to be small. The vector $\psi_i^{\mathrm{obs}} \in \mathbb{R}^{d_i}$ contains observations closest in covariance to x_i, chosen according to the observation allocation rules outlined in Sect. 2.6. The last term in Eq. (21) is dropped, since while varying θ in Eq. (22) changes d_i, the size of ψ^{obs} stays the same.

While this method is suitable for finding point estimates for the parameters θ, the log-likelihood has an unknown scaling factor resulting in an unknown multiplicative factor for the variance term in the exponent of the Gaussian distribution, and hence information about the true size of the posterior of the covariance parameters $p(\theta|\psi^{\mathrm{obs}},\beta,\delta)$ is lost.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

The covariance parameter optimization can be performed by using optimization algorithms such as COBYLA or SBPLEX available in NLOpt (Johnson, 2014). An alternative is to explore the scaled posterior by using the Adaptive Metropolis (AM) Markov chain Monte Carlo (MCMC) algorithm (Haario et al., 2001), an implementation of which is included in the satGP source code. Using MCMC is feasible since the forward model is just sampling from a multivariate normal distribution which is very fast, and also due to that the parameter dimension is moderate, even with multiple subkernels.

3 Overview of Computation

The satGP code is written in C with visualization scripts written in Python and parallellization implemented with OpenMP directives. The program reads data from netCDF files and the configuration from a C header file. For linear algebra, the C interfaces of LAPACK and BLAS, LAPACKE and CBLAS, are utilized and for optimization tasks, algorithms in the NLOpt library are used. The computations are carried out in single precision both in order to save memory resources with the largest data sets and also in anticipation of implementing the covariance function routines in a way that allows computation on graphics processing units.

The most important configuration variables are listed in Table 1. The user needs to define whether parameters are learned or prescribed and whether marginals or samples from the GP are to be computed. The mean and covariance kernel need to be defined by initializing corresponding structs with parameters and their limits if calibration is to be performed. For computing GP marginals or drawing samples from the random process, the geographic and temporal extents need to be specified and the mean function and the covariance kernel used must be given. For more details than is described below, see Appendix A.

For computational efficiency, several parameters can be tweaked, including all of those in the second and last sections of Table 1. The first main bottleneck for computing a marginal at x^* is sorting the observations for selecting the most informative ones to be used in the covariance matrices, see Sect. 2.6. This requires roughly $\mathcal{O}(r_l \log r_l + \kappa \log \kappa)$ operations, where $r_l \propto \prod_{i=1}^{q} \ell_i^l$ is the number of grid locations (test inputs) x_i^* in the spatial grid such that for the l^{th} subkernel, $k_l(x_i^*, x^*) < \sigma^2_{\min}$. Here the parameters ℓ_i^l are the corresponding length scale parameters over all the dimensions of the inputs x – this controls the size of the hypersphere inside which observations are considered for each x^*. The second bottleneck is calculating the Cholesky decompositions of the covariance matrices K with cost $\mathcal{O}((n_{\text{ker}}\kappa)^3)$. The cost of calculating the means and variances of the GP in a grid for a set of n_{times} points on the time axis is therefore given by

$$\text{cost} = \mathcal{O}\left(\frac{A n_{\text{times}}}{\omega^2} \left[(n_{\text{ker}}\kappa)^3 + \sum_{l=1}^{n_{\text{ker}}} (r_l \log r_l + \kappa \log \kappa) \right] \right), \tag{23}$$

where A is the grid area in degrees squared and ω is the grid resolution. When the random observation selection method mentioned in Sect. 2.6 is used, the $r_l \log r$ in Eq. (23) becomes just r_l.

The execution of the program is presented in Fig. 4. The names of the subprograms here deviate from those in the code to improve readability.

The function `AddToState()` reads observations (asynchronously) into a state object that tracks the proximity of each observation to each grid point. Only part of data is added, and what part, is controlled on l. 6 by the parameter η^i_{train}, which

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

Table 1. Most important satGP control variables and high level C structs: first section contains parameters for program logic, second for domain specification, third for covariance and mean function definition, and last for observation handling. This list is by no means exhaustive – the configuration file contains lots of variables that can control the program. Some additional tweaking is possible by changing hard-coded values directly in the source code, such as those listed in Appendix A.

Variable	Type	Low	High	Notes
learn_k	`int`	0	2	(0) Don't train θ, (1) generate observations and learn θ, (2) learn θ from non-synthetic data.
learn_m	`int`	0	1	(0) Don't train local β and δ, (1) find local β and δ as in Sect. 2.4.
sampling	`int`	0	2	(0) Skip sampling, (1) calculate GP marginals at each grid point, (2) sample from GP.
area	`char*`	-	-	Area definition setting longitude and latitude minimum and maximum values
n_{days}	`int`	1	∞	Number of days to be simulated
ω	`float`	> 0	180	1-d grid resolution in degrees – small values degrade esp. posterior sampling performance.
n_{ker}	`int`	1	10	Number of subkernels k_l in k
cfc	`struct*`	-	-	Recursive struct pointer defining $k_1 \dots k_{n_{\text{ker}}}$ and corresponding θ, see Sect. 2.5.
mf	`struct*`	-	-	Struct pointer for defining type of $m(\cdot,\cdot)$ and associated (initial) β and δ, see Sect. 2.3.
ζ_{train}	`float`	0	∞	Determines what fraction of observations are randomly included in ψ^{obs} when learning θ.
ζ_{sample}	`float`	0	∞	Determines what fraction of observations are randomly included in ψ^{obs} when sampling$\neq 0$.
σ^2_{min}	`float`	0	∞	Discard observation at x_i for x^* if $k(x_i, x^*) < \sigma^2_{\text{min}}$, see Sect. 2.6.
n_{ref}	`int`	0	∞	Number of reference points in E_{ref} in Eq. (22) for training θ
$n_{\text{synthetic}}$	`int`	0	∞	Number of random locations where synthetic data is generated for training θ
$\sigma^2_{\text{synthetic}}$	`float`	0	∞	Variance of Gaussian noise added to synthetic observations
κ	`int`	1	∞	Maximum subkernel size, values $\kappa > n_{\text{ker}}^{-1}1000$ will be slow due to $\mathcal{O}(\kappa^3)$ scaling.

corresponds to the inclusion probability of each observation. This probability depends on ζ_{train} in Table 1 via

$$\eta^i_{\text{train}} = \frac{d(x_i, x_{i_{\text{prev}}})}{\omega \zeta_{\text{train}}} \wedge 1, \tag{24}$$

where $d(x_i, x_{i_{\text{prev}}})$ is the Euclidean distance to the previous added point and $\wedge$ is the standard notation for minimum. Hence with $\zeta = 0$, all observations will be added.

For computing the marginals, the spatial domain can be decomposed with `Decompose()`, line 23, into several spatial subdomains (sd) so that arbitrary-size grids can be computed. This makes solving large problems with limited amount of memory possible, but only works with sampling $= 2$.

The state object is emptied by `ReInitializeState()` which also potentially sets new subdomain extents. Function `SampleFromPrior()` actually performs the computations on lines 30-37, but with the set of points x^*_i in a random pattern instead of in a grid as is the case in l. 27-38.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019


Geoscientific Model Development Discussions

The `AddSubdomainData()` method on l. 29 adds data as on lines 3-9, but only to the current subdomain. After that, the `SelectObservations()` method (l. 31) carries out selecting the best observations as described in Sect. 2.6. For constructing the set of potential observations, the grid is searched for locations that may have informative observations for the current test input stored in the state object. These locations are first ordered into categories with decreasing potential covariance and for the best locations, that together hold at least 2κ observations, the covariance function with the test input is evaluated. Out of these, the κ best are chosen. The factor 2 can be increased for the wind-informed kernel and the value 8 is used in the demonstration of the wind-informed kernel in Sect. 4.7.

The function `ComputeMarginal()` constructs the covariance matrix K, inverts via the Cholesky decomposition, and solves Eq. (9) and (10) to find the marginal distribution at any test input x^*. That function returns the negative log likelihood and is therefore directly used in learning the covariance parameters θ in `FindCovfunCoeffs()` on line 18.

The Gaussian process algorithm is an interpolation algorithm when observation noise is zero, and interpolation algorithms may misbehave when used for extrapolation. In a spatio-temporal large grid, when sampling $= 2$, i.e. when draws of the Gaussian process are generated in a regular spatio-temporal grid, computing conditionals based on the previous predictions would amount to extrapolation if done in order. For this reason, a deterministic sparse ordering is used, which ensures that test inputs corresponding to simultaneous predictions are far from each other so that their mutual covariance is negligible. For this reason conditioning on already computed values is for the vast majority of GP evaluations interpolation instead of extrapolation.

4 Results and discussion

In this section, several simulation studies are presented. In the first experiment, parameter identifiability with the multi-scale kernel is examined with satGP-generated data. After that, the MRF of mean function β coefficients is trained with OCO-2 data and those fields are then briefly analyzed.

Based on a locally varying mean function of the form in Eq. (3), the covariance parameters of the OCO-2 XCO2 spatio-temporal field are learned. Knowing both the mean and the covariance functions, the Gaussian process is then solved globally in a grid and snapshots of the mean and uncertainty fields are presented. The section is concluded by a demonstrating how the wind-informed kernel works. The covariance function parameters are learned from data.

4.1 Parameter identifiability with the multi-scale kernel

A synthetic study was performed to confirm the identifiability of the multi-scale covariance function parameters. For this, sampling with a random spatial pattern from the prior was carried out, adding 1% noise, and then estimating the parameters by computing the posterior mean estimates using Adaptive Metropolis.

The identifiability experiment was performed with various kernels, and the more complex the kernel, the more difficult recovering the true parameters was. With a single Matern, exponential, or periodic kernel, the parameters could be recovered very easily. This was also true for a combination of exponential and Matern kernels with a relatively small κ parameter.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

Data: filelist containing files with observation data $y_i = (\mu_{\psi_i}, \sigma^2_{\psi_i})$ indexed by location x_i, input variables from Table 1.

Result: Optimized β parameters for mean function and θ parameters for covariance kernel, gridded Gaussian process marginal means and variances or a sample from the Gaussian process evaluated in a grid.

Initialization: Create grid according to area *and* ω, *define* $k(x, x')$ *and* $m(x, t)$, *initialize* state;
if learn_m $= 1$ **or** learn_k $= 2$ **then**
  for file **in** filelist **do**
    D $\leftarrow$ ReadData (file);
    for $(x_i, y_i) \in$ D **do**
      if Bernoulli(η^i_{train}) **then**
        AddToState(state, x_i, y_i);
      end
    end
  end
end
if learn_m **then** FindLocalMeanfunCoeffs (state);
if learn_k $= 1$ **then**
  ReInitializeState (state, fulldomain);
  for $i \leftarrow 1$ **to** $n_{\text{synthetic}}$ **do**
    $(x_i, y_i) \leftarrow$ SampleFromPrior ( );
    AddToState(state, x_i, y_i);
  end
end
if learn_k $\neq 0$ **then**
  FindCovfunCoeffs (n_{ref})
end
if not sampling **then**
  $(\mathsf{n}_{\text{sd}}, (\mathsf{sd}_i)_{i=1}^{\mathsf{n}_{\text{sd}}}) \leftarrow$ Decompose($\mathsf{n}^{\max}_{\text{dom}}$, area, ω);
else
  assert ($\mathsf{n}_{\text{gp}} < \mathsf{n}^{\max}_{\text{dom}}$);
end
if sampling **then** **for** $i \leftarrow 1$ **to** n_{sd} **do**
  ReInitializeState (state, sd_i);
  AddSubdomainData (state, filelist, sd_i, η_{sample});
  for $x^* \in \mathsf{sd}_i$ **do**
    $A^{\text{obs}}_* \leftarrow$SelectObservations(state, x^*);
    $\mu^*, \sigma^2_* \leftarrow$ ComputeMarginal(x^*, A^{obs}_*);
    if sampling $= 2$ **then**
      $\widehat{\psi^*} \leftarrow$ Normal(μ^*, σ^2_*);
      AddToState(state, x^*,$(\widehat{\psi^*}, \sigma^2_{\text{synthetic}})$)
    end
  end
end

Figure 4. Overview of satGP. After initialization data is read for training m and k, after which possible MRF computation is carried out. This is followed by sampling the prior if a synthetic study is performed, and learning the θ parameters controlling k. Gaussian process marginals are then computed in a grid, potentially by decomposing the domain for large grids. Finally, samples from the GP may be drawn.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

The covariance kernel parameters were still recoverable with a combination of three kernels – Matern with $\nu = \frac{5}{2}$, exponential, and periodic, but for this, a larger κ was needed – the simulation shown used $\kappa = 256$. With small κ, some of the parameters had a tendency to end up at the lower boundary, possibly due to effects of the covariance cutoff on the determinant of the covariance matrix in Eq. (20). Optimization using minimization algorithms such as Nelder-Mead, COBYLA, or BOBYQA tended to often end up in local minima, and for this reason MCMC was used instead. The number of random reference points in E_{ref} in Eq. (22) was set to 12, which was enough to reliably recover parameters close to the true value.

The parameter limits, true values, and posterior means of the synthetic experiment with three kernels are given in Table 2. In total 200,000 observations were created in the region between -10 and 10 latitude and -10 and 10 longitude over a period of four years according to the values in Table 2. A total of 10 million Metropolis-Hastings iterations were carried out to make sure that the posterior covariance stabilized. The posterior, with first 50% of the chain discarded as burn-in, is shown in Fig. 5

Table 2. Lower and upper limits, with true and estimated parameter values. The three-kernel synthetic covariance function parameter estimation problem is already very difficult, here resulting in slight overestimation of the parameters of the smallest kernel.

	low	high	true	est	$\frac{\text{est}-\text{true}}{\text{true}}$
τ^{mat}	0.05	1	0.5	0.652	0.304
ℓ_{lat}^{mat}	0.003	0.02	0.007	0.00989	0.413
ℓ_{lon}^{mat}	0.003	0.02	0.01	0.0135	0.350
ℓ_{t}^{mat}	1d	14d	7d	8.06d	0.15
τ^{per}	0.01	2	1	1.073	0.073
ℓ_{lat}^{per}	0.001	0.04	0.02	0.0207	0.035
ℓ_{lon}^{per}	0.001	0.04	0.02	0.0220	0.1
ℓ_{per}	0.01	0.3	0.1	0.1075	0.075
τ^{exp}	0.5	3	1	0.927	-0.077
ℓ_{lat}^{exp}	0.005	0.1	0.025	0.0352	0.408
ℓ_{lon}^{exp}	0.005	0.1	0.04	0.0405	0.0125
ℓ_{t}^{exp}	7d	30d	21d	24.83d	0.182

4.2 The OCO-2 v9 data

The simulations with real remote sensing data utilize the v9 data from the OCO-2 satellite. The OCO-2 satellite was launched in 2014, and it orbits the Earth on a Sun-synchronous orbit (Crisp et al., 2012; O'Dell et al., 2012). The footprint area of each measurement is roughly 1.29 by 2.25 kilometers, but the data is very sparse in time and in space. The satellite completes 14.57 revolutions around Earth overpasses in one day. In the presence of clouds, the satellite is not able to produce measurements, and this poses a challenge for areas with persistent cloud covers, such as Northern Europe in the winter.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

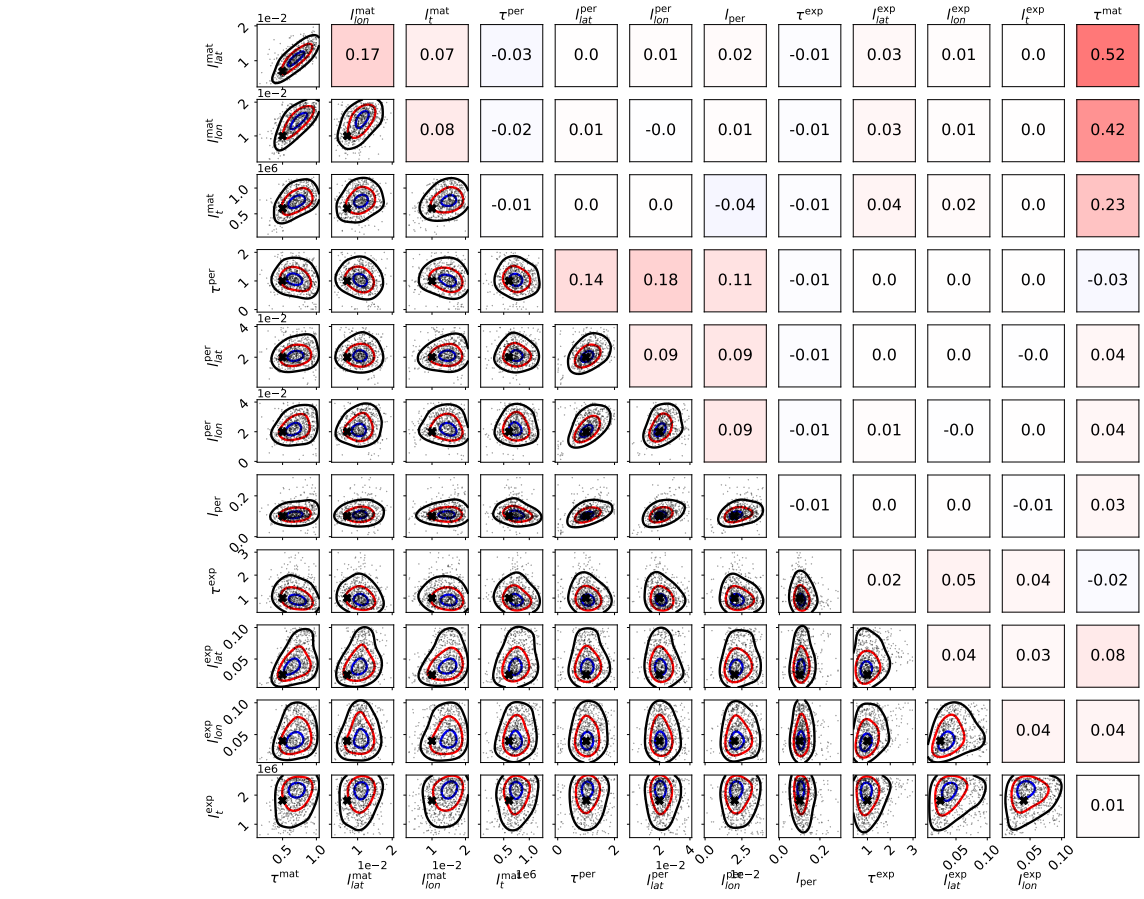

Figure 5. Scaled MCMC posteriors from a synthetic study showing identifiability of multi-scale Gaussian process kernel parameters. On lower left, the pairwise marginal distributions are shown, with the black crosses denoting true values. The axis labels are on the left and below the figure. On upper right, sample correlations are shown, with axis labels on the left and on the top. Small within-kernel component positive covariances are present. The contours shown include 85% (black), 50% (red) and 15% (blue) of the posterior mass.

The present work utilizes the XCO2 data, its reported uncertainties, associated coordinate information, and zonal and meridional wind speeds that are contained in the data files. Only observations flagged good are used, and there are in total 116489342 such observations for the time period considered.

4.3 Solving the mean function for OCO-2 v9

Solving the mean function from OCO-2 v9 XCO2 data, as described in Sect. 2.4, produces best estimates for the coefficients of Eq. (11) shown in Fig. 6. The upper left quadrant shows the semiannual seasonality of the XCO2 concentration, which explains the color shift along the equator. The lower left quadrant shows the amplitude of the twice faster oscillations, and like β_1, also β_2 shows the highest amplitude oscillations in the boreal region.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

The constant term β_3 in the upper right quadrant shows the background concentration. Some of the reddest areas such as East China, both coasts of the United States, Central Europe, and the Persian Gulf stand out and are areas where major emission sources are known to exist. The observation of a local elevated concentration compared to the surrounding areas approaches the work of Hakkarainen et al. (2016), where empirically defined time-integrated local XCO2 anomalies are interpreted as possible emission sources.

The phase shift is modeled separately, and the field in the lower right quadrant is obtained by optimization, conditioning on the β factors. This partly explains the slightly different spatial pattern. The figure shows how the phases of the XCO2 annual cycles differ in some regions, such as the Amazon or the Central African rain forests and the Sahel. The trend component β_4 was here set to be constant, as CO2 over time mixes in the atmosphere.

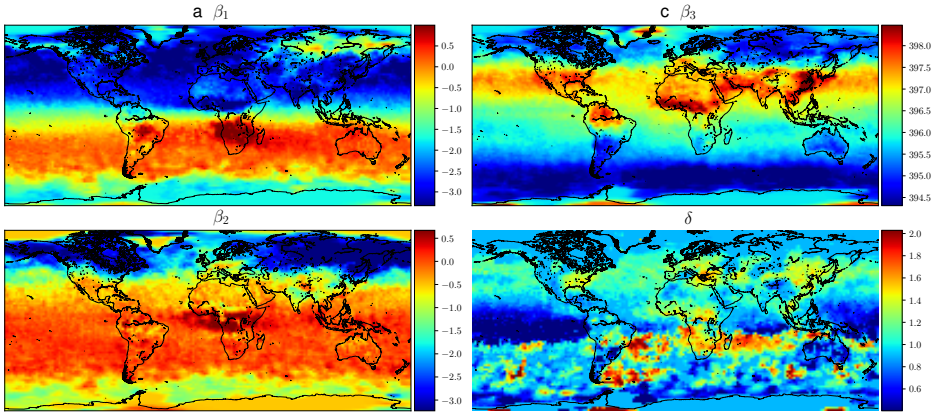

Figure 6. Mean values of mean function coefficients that were described as a Markov random field, calculated in a $2° \times 2°$ grid. The β_i coefficients multiply the f_i functions in Eq. (11). Panel (d) shows how the phase parameter δ can vary more in the southern hemisphere where β_1 and β_2 are small. The mean function and fitted daily means for several locations with the corresponding mean function parameters are shown in Fig. 1.

4.4 Covariance parameters of the OCO-2 v9 data

The OCO-2 data has several natural length scales, both spatially and temporally. The distance between adjacent observations is only one to two kilometers in space and some hundredths of a second in time, but the distance between consecutive orbits is thousands of kilometers in space and several hours in time. On consecutive days the satellite passes close to the trajectory of the previous day at a distance of tens to three hundred kilometers depending on the latitude. The Earth has natural temporal diurnal and annual cycles, but since OCO-2 is Sun-synchronous, only the latter matters. Since the annual cycle is already fitted in the particular form of the mean function used, Eq. (11), a periodic kernel component is not included, and the data is modeled with a kernel consisting of a larger-scale exponential and smaller scale Matern component.

The covariance parameters for the two-component kernel, which are the median values from sampling the posterior with MCMC, are given in Table 3. With learning the parameters from a data set with natural length scales, the posterior may appear

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

multi-modal, with some of the modes only having relatively little mass. In such a case, the median provides a more robust estimate for the parameters than the mean. The $\ell^{(\cdot)}_{\mathrm{lon}}$ and $\ell^{(\cdot)}_{\mathrm{t}}$ parameters of the posterior mean were slightly larger, which would result in slower computation. Selecting the median is further justified by the slight overestimation of some parameters in the synthetic study in Sect. 4.1.

Learning the covariance parameters from OCO-2 v9 data used the following configuration parameters for satGP: $\zeta_{\mathrm{train}} = 0$, $\kappa = 256$, and $n_{\mathrm{ref}} = 12$. A total of 1.1184 million MCMC iterations were completed, with the first 50% discarded as burn-in to produce statistics. The reference points were randomly picked from a rectangle with corners at (0°S, 65°E) and (60°N, 145°E). While using the whole globe would have been a principled choice, MCMC requires lots of iterations, and for any claim of global coverage n_{ref} would have needed to be much larger.

Table 3. Covariance function parameter values learned from OCO-2 data. First column shows the Matern kernel parameters, and the second column the exponential kernel parameters. The length scale along the parallels, $\ell^{(\cdot)}_{\mathrm{lon}}$ is much larger than that along the meridians, $\ell^{(\cdot)}_{\mathrm{lat}}$.

	$(\cdot) = \mathrm{mat}$	$(\cdot) = \exp$
$\tau^{(\cdot)}$	0.899	2.72
$\ell^{(\cdot)}_{lat}$	0.00513	0.0418
$\ell^{(\cdot)}_{lon}$	0.0363	0.397
$\ell^{(\cdot)}_{t}$	20h 22min	16d 20h 12min

4.5 Posterior predictive distributions of XCO2 from the OCO-2 v9 data

The marginal posterior predictive distribution at test points x^*, given by Eq. (9) and (10), were calculated globally in a half-degree grid between 80°S and 80°N at a daily resolution. The first day of simulation was September 6 2014, and the last day was November 10 2018, spanning in total 1526 days. For each day, 230400 marginals were computed, resulting in a collective 351 million inverted covariance matrices. The satGP parameters used were $\zeta_{\mathrm{sample}} = 0$ and $\kappa = 256$, and the covariance kernel used was the one learned in Sect. 4.4, with parameters given in Table 3. The simulation time was 25 days on a moderately fast Intel i7-8700K CPU utilizing the available 12 CPU threads and 32 GiB memory.

Global fields of the mean values and marginal uncertainties are presented in Fig. 7 and 8, with a subset (to avoid excessive over-drawing) of observations shown as a scatter plot. For this simulation, a maximum distance of 1100 km (10° on the equator) was specified for speeding up searching for closest observations in the direction along parallels. This constraint can be seen as discontinuities in uncertainty when no observations are nearby, especially close to the poles. The (b) parts of the figures show how uncertainty is reduced with the overpass of OCO-2. This uncertainty reduction diminishes fast due to the Matern component of the multi-scale kernel having a very short length scale parameter in the time dimension. In the upper figures, the background color (posterior mean) usually matches the observations. Due to observational noise, the GP mean is not strictly interpolation, however.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

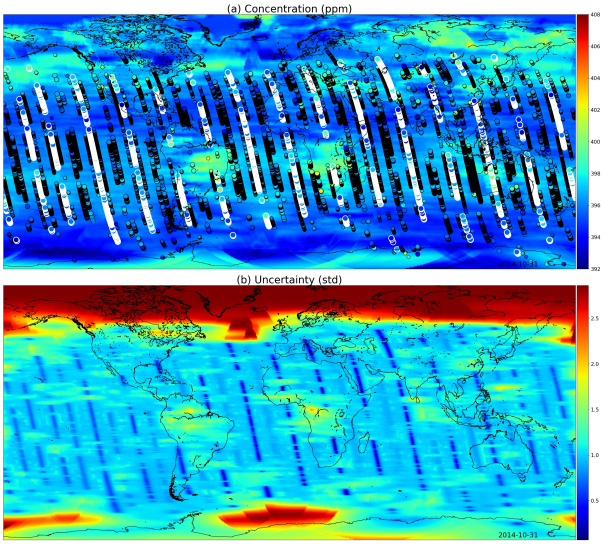

Figure 7. Global XCO2-distribution posterior mean values (a) and their uncertainties (b) on last day of October 2014. The most informative observations are shown with the concentrations, with the large white circles being from October 31st 2014, medium circles from one day before or after, and small circles from two days before or after. The OCO-2 utilizes sunlight for retrieval, and that is why there are very few observations above 60°N.

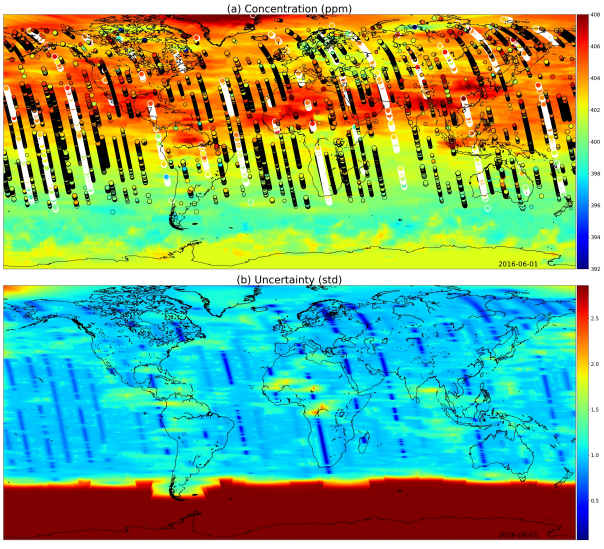

Figure 8. Global XCO2-distribution posterior mean values (a) and their uncertainties (b) on June 1st 2016. While photosynthesis in the Northern Hemisphere is already reducing the carbon dioxide concentrations globally, the observations condition the Gaussian process to higher mean values than in Fig. 7. In the summer months the uncertainty stays high close to the South Pole.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019


Geoscientific Model Development Discussions

4.6 Comparison of single- and multi-scale kernels with OCO-2 data

How the multi-scale kernel formulation affects the predictive posterior distributions can be demonstrated with OCO-2 data. In Fig. 9 posterior marginals from September 15 2014 are shown. The first row (a-b) contains results from the multi-scale kernel described in Sect. 4.4, and the second row (c-d) shows fields from only the exponential part of the multi-scale kernel. The parameters of the multi-scale kernel are shown in Table 3. The bottom row (e-f) contains the difference fields between the first and the second rows. The single-kernel uncertainty is very low in Fig. 9 (d) since lots of observations fall into regions of high covariance with almost any test input, with the exception of the Northern side of Ireland, which does not have any observations nearby. Since the covariance kernel parameters were trained for the multi-scale kernel, the parameters used for the single kernel are not the ones describing the XCO2 field best.

Figure (a) shows that as intended, the multi-scale approach leads to local enhancements of the XCO2 mean field. Far from the measurements, the smaller Matern kernel no longer reduces the predicted marginal uncertainties, and this leads to an increase in uncertainty in these areas. Figure (e) shows additional enhancements of the XCO2 mean fields, which are in this case due to the different maximum covariances between the multi-scale and single-scale kernels.

The total kernel size was kept at 1024 ($\kappa = 512$ for (a-b) and $\kappa = 1024$ for (c-d)) in both experiments. Additionally $\zeta_{\text{sample}} = 5$, and $\omega = 0.5°$ in this case. The very same observations were used in both cases.

4.7 Wind-informed kernel with OCO-2 data

The wind-informed kernel, Eq. (16), lets local wind data at test input x^* rotate and scale the coordinate axes. Modeled winds are included with OCO-2 data, and they can be used to produce gridded winds that can then be used locally with the computation of each marginal posterior predictive distribution.

The covariance parameters for a single wind kernel were learned by taking the median of an MCMC posterior, similarly as was done in Sect. 4.4. The resulting parameters were $\tau = 2.07$, $\ell = 0.038$, and $\rho = 56.7$. The variance of ρ was high, possibly due to the square root in the current formulation in Eq. (17). For this simulation, $\zeta = 1$, $\kappa = 1024$, and $\omega = 0.7$, and the simulation time for the area from $(27°N, 115°E)$ to $(40°N, 145°E)$ for the single day was 2.652s (walltime) on the i7-8750H laptop CPU.

The simulation results are shown in Fig. 10. Low uncertainties shown in blue color on the right spread with the winds, as do the concentration estimates on the left both due to the high reading in South Korea and the low reading close to Shanghai.

Optimally the wind-informed kernel should utilize winds that are not recomputed from the observations as was done for convenience, but directly from a weather or climate model. The satGP program contains configuration options for doing this. The optimal covariance function parameter values are conditional on the wind data, so the values should be learned separately for each new application and wind data set.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions
Open Access
EGU

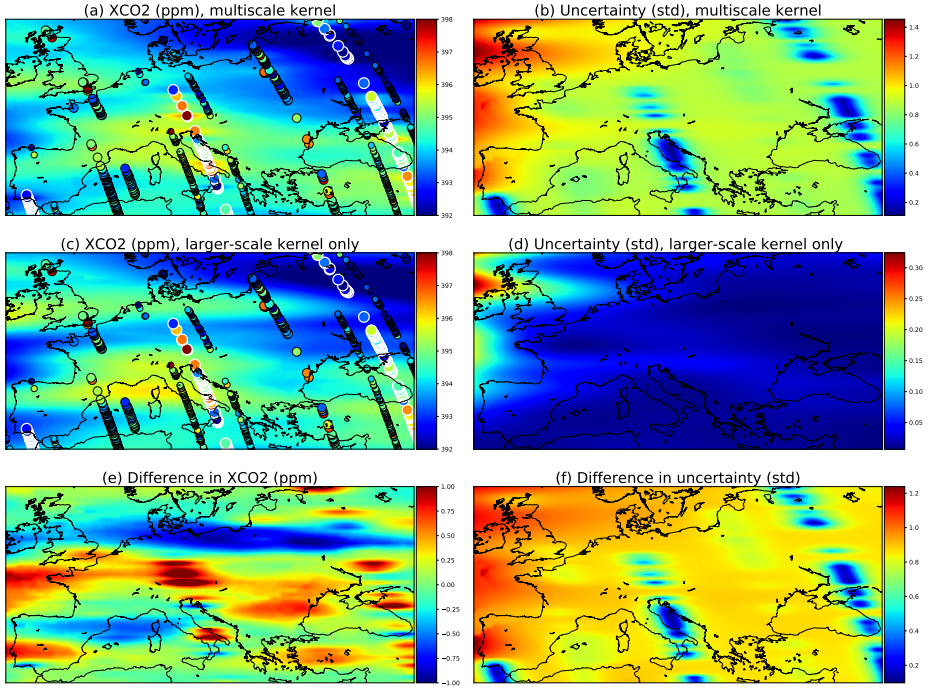

Figure 9. Comparison of a multi-scale kernel with the two components described in Sect. 4.4 and a single component kernel defined by the parameters of the exponential kernel. These parameters were given in Table 3. The observations used are the same and are shown in panels (a) and (c) as circles. The large ones with white borders are observations from the present day, September 15^{th} 2014, medium circles are observations from 14^{th} and 16^{th}, and small circles from 13^{th} and 17^{th}.

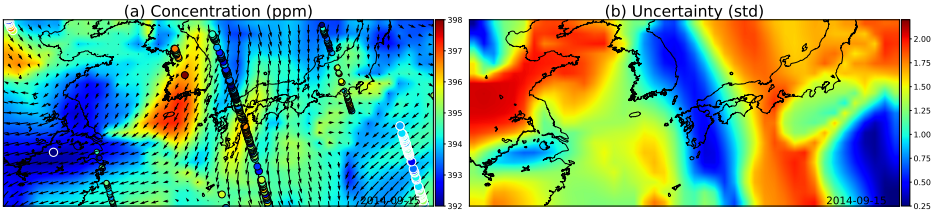

Figure 10. (a) GP posterior mean of XCO2 and (b) its uncertainties with the wind-informed kernel. The area shown contains the Korean peninsula in the center, China on the left, and Japan on the center-right. The large circles with the white edges are present-day observations, medium circles are observations from adjacent days, and the smallest ones are observations from two days away. Wind direction and magnitude are given by the black arrows, and uncertainty is clearly reduced where wind is blowing directly towards or away from the observations.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

5 Conclusions and future work

In this work we have introduced the first version of a new fast Gaussian process software, satGP v.0.1. It aims at being a general purpose Gaussian process toolbox, especially meant to be used with remote sensing data. The software solves the problems of spatially varying mean function, learning its parameters via computation of marginals of an MRF, and also allows learning the parameters of the multi-scale covariance function using either optimization algorithms or adaptive Markov chain Monte Carlo. On top of these, satGP allows to conduct synthetic parameter identification studies via sampling Gaussian process prior and posterior distributions, and this can be done with any kernel prescribed, including a non-stationary wind-informed kernel. We are not aware of open source remote sensing-oriented software that would provide this combination of features. The satGP program was demonstrated with the enormous data set produced by the NASA Orbiting Carbon Observatory 2.

There are various aspects of satGP that could be improved in future versions. These include addition of routines for doing model selection to select the components of the multi-scale kernel, improving the observation selection/thinning scheme for statistical optimality, and finding joint posterior predictive distributions. For the last one, a multi-grid version can be developed, and this could be potentially useful for flux inversion studies.

The satGP software utilizes various approximations for computational tractability, and the connection between parameters such as length scales ℓ, thinning parameter ζ, maximum kernel size κ, and prediction accuracy could be studied further, as well as changing the grid resolution according to density of observations.

The methodology and code presented can be also used with other data sources. For instance, combining data from the various satellites that measure CO2, such as GOSAT, GOSAT-2, OCO-2, TANSAT, and the OCO-3, would be particularly interesting. That more and more instruments are about to provide data from the orbit in the near future will lead to a need to understand the properties of even larger data sets.

Code and data availability. The satGP code is available from the contact author upon a reasonable request. The OCO-2 v9 data used is freely available directly from JPL/NASA.

Appendix A: Input parameters and variables in satGP

The satGP software by design allows for lots of flexibility for defining how to model the quantity of interest as a Gaussian random field. In this section the possibilities are discussed along with some recommendations. The parameters in Table 1 are described in more detail than earlier, along with some other configuration variables in the configuration file `config.h`. Practical aspects of defining mean functions and covariance kernels are also included. Some of the details in this section may change for future versions of satGP.

Of the four sections in Table 1, the first is obvious, as those parameters control the main logic of satGP. It is recommended to first learn the mean function, then with that mean function learn the covariance function, and only after that calculate the means

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

and variances of the Gaussian process with sampling $= 1$. The setting sampling $= 2$ can be used for illustration purposes, for understanding how the different realizations of the random function would look like.

The area parameter defines the longitude-latitude extents of the domain where satGP is wished to be used. The strings and the corresponding areas are defined in the beginning of the file `gaussian_proc.h`, and can be changed there as needed. Current available areas contain e.g. `NorthAmerica`, `Europe`, `EastAsia`, `World`, and `TESTAREA`.

The parameter n_{days} defines how many days are to be simulated after the starting day. Currently the starting day is hardcoded in the code base to the first day of OCO-2 data. However, if `use_daylist`$\neq 0$ in the configuration file, a list of days can be used. This list can quite easily generated by modifying a python script `create_daylist.py`, which is included with satGP.

The ω parameter determines how much spatial detail is resolved when sampling or computing marginals of the random field. A small value like 0.1 will make computing very expensive, and using such values might be unnecessary when the smallest covariance subkernel length scale parameters are large. These ℓ parameters are in the scale of distances on the unit ball, and therefore on the equator an ℓ parameter of 0.05 corresponds to a length scale of around $2.9°$, so the ω parameter should rarely be much less than half of that. On the other hand, if the observations are spatially very close to each other and describing local variation is aimed for, then the ℓ parameters need also to be small. Given computational constraints, larger values or different area parameters may need to be used.

In the third section of Table 1, the first parameter n_{ker} denotes the number of subkernels. Even though the hard limit is set at 10, in practice this should be between one and three since the parameters of more than three subkernels are not necessarily identifiable. More kernels means also more computational cost, due to the κ parameter, which is the last one in the table and discussed later.

The parameters cfc and mf are not strictly input variables, but C struct pointers that are created based on input variables. These variables are described in the configuration file, and they amount to choosing the covariance kernels from prescribed types (e.g. Matern, exponential, and periodic), and then defining the parameters for those kernels. The best parameters are those that are learned with learn_k $= 2$ when non-synthetic data is used.

The learning is best performed with MCMC, and the posterior mean and median have proven to be a useful values. For unimodal posterior distributions these values are very close. The number of MCMC iterations is controlled by the variable `mcmc_iters`, for which 10^6 is a large enough value, and for computing the log-likelihood in Eq. (22), the number of reference points n_{ref} in the set E_{ref} can be set to a low value of e.g. number of CPU threads, if at least 12 are available. If with MCMC the chain gets stuck in local minima, the value of the `mcmcs->scalefactor` in the `mcmc()` function in `mcmc.h` may be shrunk, and equally well, if the posterior ends up being flat with respect to many parameters, it may be increased.

For learning the covariance parameters, parameter limits need to be given. These should correspond to the expected length scales in the data – e.g. long-range fluctuations with low amplitude, and short-scale variations due to local effects. It is in practice best if the parameter ranges do not overlap.

If the exponent of the exponential kernel needs to be changed, that needs to be done by changing the `exponent` variable in the `covfun_dyn()` function in the file `covariance_functions.h`. Similarly, if the order of the Matern kernel needs to

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

be changed, that can be done by changing the variable `n` in functions `covfun_matern52()` and `initialize_covfunconfig()` in that same file.

For constructing the mean function, the configuration file contains the parameter `mftype`. The possible values are: 0) a zero mean function is used, 1) a mean function that changes only in time is used, 2) a (time-dependent) field is read in and used - this can be e.g. the mean value from a previous Gaussian process simulation, and 3) a space and time dependent mean function is used. The function itself is given as a function pointer to variable `mean_function` in the configuration file, and this function needs to be defined somewhere – e.g. in the file `mean_functions.h`. For the mean function, another variable, `mfcoeff`, needs to be set. This is the total number of parameters (β and δ in Eq. (3)) if `mftype` $\in \{1,3\}$. If the mean function parameters are learned, the parameter `nnonbetas`, the number of mean function non-linear δ parameters, needs to be set to the appropriate value in the function `fit_beta_parameters_with_unc()` in `mean_functions.h`. For global mean function coefficients, the values of those coefficients are given in the configuration file. Additionally, parameter limits for learning the space-dependent mean function parameters are set in the configuration file. Finally, when learning the space-dependent mean function parameters, the smoothness of the field may be controlled by changing the `dscale` parameter in the configuration file, and to a lesser extent by modifying the `dfmin` and `dfmax` parameters in function `fit_beta_parameters_with_unc()` in file `mean_functions.h`.

In the last section, the ζ_{train} parameter controls data thinning when learning covariance kernel parameters and the ζ_{sample} parameter has the same effect for when sampling $\neq 0$. How the thinning takes place was explained in the context of Eq. (24). While with few observations no thinning needs to be done at all, i.e. $\zeta_{\cdot}$ may be set to zero, with large data sets the representability of data may be improved when a coarse grid is used for computation, and also memory bottlenecks may be avoided. These parameters may be also increased if faster execution is required, e.g. for debugging purposes.

The σ^2_{min} parameter controls which observations are not considered at all when computing at a location x^*, as described by Eq. (19). The higher this is, the more data is discarded. Setting σ^2_{min} to a very low value makes searching for candidate observations slow, while picking too high a value may make posterior fields look edgy. In practice values between 10^{-7} and 10^{-3} seem to work well. This parameter is not actually meant to be changed, and it is for that reason set in `create_config()` in the file `gaussian_proc.h`.

The variable $n_{\text{synthetic}}$ defines how many synthetic observations are generated when learn_k $= 1$. Very large values are once again expensive, and instead a smaller area should rather be used with more moderate values of $n_{\text{synthetic}}$. Those values can be in practice up to 10^5 or more. With very low values, it may be that spatial patterns specified by the prescribed covariance kernel are not represented appropriately, and therefore values less than 10^4 should be avoided, except for maybe in settings with only a single subkernel. If $\sigma^2_{\text{synthetic}}$ is high, parameter identifiability suffers. Varying this parameter could be used for understanding how complex a multi-scale kernel can be useful with particular data sets. The values also depend on the maximum covariance parameters of the Gaussian process, given by the τ^2 parameters in the formulas of Sect. 2.5.

The last parameter in Table 1, κ, defines the maximum subkernel size. The larger this parameter is, the more data is included for constructing the covariance matrix K, whose Cholesky decomposition needs to be computed to solve the local regression problem inherent to Gaussian processes. In practice the full kernel size should be kept under 1000, and in order to compute GP

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

calculations fast, a full kernel size of less than 500 is recommended. However, with a very small number of marginals, values up to 10^4 may be experimented with. When $n_{\mathrm{ker}}\kappa < 64$, the speed-up due to solving the GP formulas faster decreases, since at that point computing Cholesky decompositions no longer takes up majority of computing time. This lower bound depends on the CPU architecture and the sizes of the various CPU caches.

Whether the observations for computing the local values are chosen at random or greedily is determined by the variable `select_closest` in function `pick_observations()` in file `covariance_functions.h`. The value used should normally be non-zero, since with random selection adjacent grid points often do not utilize the best available observations closest by, leading to noisiness or graininess in the computed mean field.

In addition to the parameters and variables listed here, there are also other parameters in the configuration file and in the code, even though those should not need to be changed. Any variables that the user might want to tweak are generally accompanied by at least some comments describing their effects.

In the current version, the satGP program is run with the script `gproc.sh`, whose comments describe the various options. Compiling and running require a modern GCC version (such as version 8) and the meson build system, and additionally all the needed libraries listed in Sect. 3. The current low version number reflects the fact that as of now, installing and using the software will require a degree of technical knowledge, including some Python, C, and BASH programming skills.

Author contributions. JS, AS, HH, and YM designed the study. JS prepared this manuscript, wrote the satGP code, chose, tested and implemented the computational methods, and performed the simulations.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Geoscientific Model Development Discussions

References

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., and O'Neil, M.: Fast Direct Methods for Gaussian Processes, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, 252–265, https://doi.org/10.1109/TPAMI.2015.2448083, 2016.

Chiles, J.-P. and Delfiner, P.: Geostatistics, Wiley, 2012.

Cressie, N.: Mission CO2ntrol: A Statistical Scientist's Role in Remote Sensing of Atmospheric Carbon Dioxide, Journal of the American Statistical Association, 113, 152–168, https://doi.org/10.1080/01621459.2017.1419136, 2018.

Cressie, N. and Wikle, C.: Statistics for Spatio-Temporal Data, Wiley, 2001.

Crisp, D., Fisher, B. M., O'Dell, C., Frankenberg, C., Basilio, R., Bösch, H., Brown, L. R., Castano, R., Connor, B., Deutscher, N. M., Elder-ing, A., Griffith, D., Gunson, M., Kuze, A., Mandrake, L., McDuffie, J., Messerschmidt, J., Miller, C. E., Morino, I., Natraj, V., Notholt, J., O'Brien, D. M., Oyafuso, F., Polonsky, I., Robinson, J., Salawitch, R., Sherlock, V., Smyth, M., Suto, H., Taylor, T. E., Thompson, D. R., Wennberg, P. O., Wunch, D., and Yung, Y. L.: The ACOS CO_2 retrieval algorithm; Part II: Global XCO2 data characterization, Atmospheric Measurement Techniques, 5, 687–707, https://doi.org/10.5194/amt-5-687-2012, https://www.atmos-meas-tech.net/5/687/2012/, 2012.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.: Bayesian Data Analysis, Chapman and Hall/CRC, 3rd edn., 2013.

Haario, H., Saksman, E., and Tamminen, J.: An Adaptive Metropolis Algorithm, Bernoulli, 7, 223–242, http://www.jstor.org/stable/3318737, 2001.

Hakkarainen, J., Ialongo, I., and Tamminen, J.: Direct space-based observations of anthropogenic CO2 emission areas from OCO-2, Geophysical Research Letters, 43, 11,400–11,406, https://doi.org/10.1002/2016GL070885, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GL070885, 2016.

Hammerling, D. M., Michalak, A. M., O'Dell, C., and Kawa, S. R.: Global CO2 distributions over land from the Greenhouse Gases Observing Satellite (GOSAT), Geophysical Research Letters, 39, https://doi.org/10.1029/2012GL051203, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL051203, 2012.

Hammersley, J. and Clifford, P.: Markov random fields on finite graphs and lattices, unpublished manuscript, 1971.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A.: A Case Study Competition Among Methods for Analyzing Large Spatial Data, Journal of Agricultural, Biological and Environmental Statistics, https://doi.org/10.1007/s13253-018-00348-w, 2018.

IPCC: Summary for Policymakers, book section SPM, pp. 1–30, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.004, www.climatechange2013.org, 2013.

Johnson, S. G.: The NLopt nonlinear-optimization package, http://github.com/stevengj/nlopt, 2014.

Katzfuss, M., Guinness, J., and Gong, W.: Vecchia approximations of Gaussian-process predictions, arXiv e-prints, arXiv:1805.03309, 2018.

Lauritzen, S.: Graphical Models, Oxford Statistical Science Series, Clarendon Press, 1996.

Lindgren, F., Rue, H., and Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73, 423–498, https://doi.org/10.1111/j.1467-9868.2011.00777.x, https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x, 2011.

Ma, P. and Kang, E. L.: Fused Gaussian Process for Very Large Spatial Data, ArXiv e-prints, 2017.

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

Nassar, R., Hill, T. G., McLinden, C. A., Wunch, D., Jones, D. B. A., and Crisp, D.: Quantifying CO2 Emissions From Individual Power Plants From Space, Geophysical Research Letters, 44, 10,045–10,053, https://doi.org/10.1002/2017GL074702, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL074702, 2017.

Neal, R. M.: MCMC using Hamiltonian dynamics, in: Handbook of Markov Chain Monte Carlo, edited by Brooks, S., Gelman, A., Jones, G., and Meng, X., Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, 2011.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A.: Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets, Technometrics, 56, 174–185, https://doi.org/10.1080/00401706.2013.831774, 2014.

O'Dell, C. W., Connor, B., Bösch, H., O'Brien, D., Frankenberg, C., Castano, R., Christi, M., Crisp, D., Eldering, A., Fisher, B., Gunson, M., McDuffie, J., Miller, C. E., Natraj, V., Oyafuso, F., Polonsky, I., Smyth, M., Taylor, T., Toon, G. C., Wennberg, P. O., and Wunch, D.: Corrigendum to "The ACOS CO_2 retrieval algorithm - Part 1: Description and validation against synthetic observations" published in Atmos. Meas. Tech., 5, 99-121, 2012, Atmospheric Measurement Techniques, 5, 193–193, https://doi.org/10.5194/amt-5-193-2012, https://www.atmos-meas-tech.net/5/193/2012/, 2012.

Rasmussen, C. and Williams, C.: Gaussian Processes for Machine Learning, MIT Press, http://www.gaussianprocess.org/gpml/chapters/, 2006.

Rodgers, C.: Inverse Methods for Atmospheric Sounding: Theory and Practice, Series on atmospheric, oceanic and planetary physics, World Scientific, 2000.

Santner, T., Williams, B., and Notz, W.: The Design and Analysis of Computer Experiments, Springer Verlag New York, first edn., 2003.

Schäfer, F., Sullivan, T. J., and Owhadi, H.: Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity, arXiv e-prints, arXiv:1706.02205, 2017.

Tadić, J. M., Qiu, X., Miller, S., and Michalak, A. M.: Spatio-temporal approach to moving window block kriging of satellite data v1.0, Geoscientific Model Development, 10, 709–720, https://doi.org/10.5194/gmd-10-709-2017, https://www.geosci-model-dev.net/10/709/2017/, 2017.

Vecchia, A. V.: Estimation and Model Identification for Continuous Spatial Processes, Journal of the Royal Statistical Society. Series B (Methodological), 50, 297–312, http://www.jstor.org/stable/2345768, 1988.

Yi, L., Jing, W., Lu, Y., Xi, C., Zhaonan, C., Dongxu, Y., Zengshan, Y., Songyan, G., Longfei, T., Naimeng, L., and Daren, L.: TanSat Mission Achievements: from Scientific Driving to Preliminary Observations, Chinese Journal of Space Science, 38, 627, https://doi.org/10.11728/cjss2018.05.627, http://www.cjss.ac.cn/EN/abstract/article_2600.shtml, 2018.

Yokota, T., Yoshida, Y., Eguchi, N., Ota, Y., Tanaka, T., Watanabe, H., and Maksyutov, S.: Global Concentrations of CO2 and CH4 Retrieved from GOSAT: First Preliminary Results, SOLA, 5, 160–163, https://doi.org/10.2151/sola.2009-041, 2009.

Zammit-Mangion, A., Cressie, N., Ganesan, A. L., O'Doherty, S., and Manning, A. J.: Spatio-temporal bivariate statistical models for atmospheric trace-gas inversion, Chemometrics and Intelligent Laboratory Systems, 149, 227 – 241, https://doi.org/https://doi.org/10.1016/j.chemolab.2015.09.006, 2015.

Zammit-Mangion, A., Cressie, N., and Shumack, C.: On Statistical Approaches to Generate Level 3 Products from Satellite Remote Sensing Retrievals, Remote Sensing, 10, https://doi.org/10.3390/rs10010155, http://www.mdpi.com/2072-4292/10/1/155, 2018.

Zeng, Z., Lei, L., Guo, L., Zhang, L., and Zhang, B.: Incorporating temporal variability to improve geostatistical analysis of satellite-observed CO2 in China, Chinese Science Bulletin, 58, 1948–1954, https://doi.org/10.1007/s11434-012-5652-7, 2013.

Zeng, Z.-C., Lei, L., Strong, K., Jones, D. B. A., Guo, L., Liu, M., Deng, F., Deutscher, N. M., Dubey, M. K., Griffith, D. W. T., Hase, F., Henderson, B., Kivi, R., Lindenmaier, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Sussmann, R., Velazco, V. A., Wennberg, P. O.,

https://doi.org/10.5194/gmd-2019-156
Preprint. Discussion started: 30 August 2019

and Lin, H.: Global land mapping of satellite-observed CO2 total columns using spatio-temporal geostatistics, International Journal of Digital Earth, 10, 426–456, https://doi.org/10.1080/17538947.2016.1156777, 2017.