Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-156-AC1, 2020

Geoscientific
Model Development
Discussions

Interactive comment on "Efficient multi-scale Gaussian process regression for massive remote sensing data with satGP v0.1" *by* Jouni Susiluoto et al.

Jouni Susiluoto et al.

jouni.i.susiluoto@jpl.nasa.gov

Received and published: 12 February 2020

We thank the anonymous reviewers and the editor for carefully reading the manuscript and for providing the very valuable comments. We address the comments one by one below. The reviewer comments are pasted verbatim below in italics, and the author responses to these comments can be found immediately under the comments, starting "A:".

Anonymous Referee #1

This manuscript describes a model to analyze large spatio-temporal data. Although analyzing remote sensing data of enormous sizes is no double important and challenge, the manuscript fails to describe the model and its computation details and properties sufficiently or clearly. Please see below my comments that are not necessarily ordered chronologically or by importance:

1. *This manuscript suggest using the mean function of a particular form when analyzing OCO-2 data: $m(x;\beta,\delta) = f(x^t;\delta(x^s))\beta(x^s)$ This mean function is not a linear form of unknown parameters $\{\delta(x^s),\beta(x^s)\}$, noting that they are both dependent (i.e., varying) across locations. I find the description on how to estimate $\delta(x^s)$ and $\beta(x^s)$ extremely confusing.*

- *In Lines 10-20 of Page 6, it states that β will be estimated using the formula of generalized least squared as given in Equation (6), and δ will be calibrated, but no explanation is given on how δ will be calibrated. In addition, the authors did not explain the dimension of the matrices F and K in Equation (6). Are they large so that K^{-1} or $(F^TK^{-1}F)^{-1}$ difficult to compute?*

A1: First, we mention that we find a point estimate for the δ parameters before calibrating β with generalized least squares, and that we then still one more time calibrate the δ parameters. We agree that the wording could be better, and we will clarify the alternating optimization in the sentence under (7) for the revised manuscript, adding that we use optimization algorithms for the task. We also give a reference to a later section for the full description of the procedure. Second, the reviewer is absolutely correct about that the matrices describing the joint probability density of all the β coefficients are too large for direct inversion. In the OCO-2 simulations the size of K is up to the order of $10^8 \times 10^8$. We will clarify the sizes of these matrices in the text.

- *How is $\beta(x^s)$ estimated for a location x^s? For a location x^s,test without data/observation, can we estimate $\beta(x^s, test)$ and how?*

A2: The β^s is estimated via the Markov Random Field, by fitting the parameters to match the mean function to local observations, and by conditioning on the parameter values at neighboring spatial locations. When there is no data nearby, the values of the parameters will be determined by prior values (if any – we use a flat prior) and the parameters at neighboring nodes in the MRF. We agree that the description in Section 2.4 is at the moment not very clear, and we will describe the calibration procedure more clearly in the revised version.

- *Although the authors have included Section 2.4 on learning $\beta(x^s)$ as a Markov random field, this section is not connected to other parts of the manuscript but only adds confusion. It is unclear what the authors meant by modeling $\beta(x^s)$ as a Marko random field. Does this mean that the authors no longer use Equation (6) to estimate $\beta(x^s)$? What are the assumptions of this Markov random field (MRF)? What are the parameters in this MRK and how is this MRK fitted?*

A3: The β parameters are still computed with equations (6) and (7), but in addition to just computing a mean field approximation, we condition each vertex by the neighbors. This also imposes some smoothness on the posterior field of the β parameters and and regularizes the problem. The fitting procedure is actually described on p. 8 l. 6-11 and in the caption of Fig. 2. Additionally, the conditioning on the neighbors is briefly explained in the text around p. 7 l. 27 - p. 8 l. 2. However, we agree that this description could be made clearer, and for this reason we will rewrite section 2.4 as needed. Regarding the parameters of the MRF, the MRF is over the β parameters, and for the δ parameters we only obtain point estimates by fitting the parameters before and after obtaining the local β values (amounting to a very short alternating optimization of β and δ). The smoothness of the fitting is controlled by the `dscale` parameter mentioned on p. 26 l. 12-15. The

MRF is fitted according to the procedure described in the caption of Fig. 2. We realize that even though how the fitting is exactly done is not so critical for how the *a posteriori* Gaussian process fields look like, this procedure should be more carefully explained, and not in a figure caption. We will integrate the description in the rewritten section 2.4.

- *It is also confusing how the parameters $\delta(x^s)$ are estimated.*

A4: The fitting of the δ parameters is carried out by optimizing them when computing the MRF as explained on p. 8 l. 12-18. While we think that the procedure is currently described in the text, it could be worded better, and we will do our best to also clarify this part of the text.

- *Line 14 of Page 8: " . . . finding $\hat{\beta}$ with Eq. (9) and (10), ..." Is this a typo? Should it be Eq. (6) and (7)?*

A5: Yes, this is a typo, which has now been fixed.

- *Page 8 Line 15: The objective function $\sum_{j=1}^{n}(m(x_\nu;\beta_\nu,\delta_\nu) - \psi_j)^2 + \sum_{j' \in \partial\nu}(\delta_\nu - \delta_{j'})^2$ and the optimization procedure are poorly explained. It should be noted that the mean function $m(\cdot;\cdot,\cdot)$ involved δ and β. It is very confusing how or why this function is used to estimate δ or β individually or both of them jointly, and why it should be used this way.*

A6: This part of the text describes fitting the phase-shift parameters δ, also mentioned above. For the "why" question, it is mentioned in the text that the nonlinear parameters cannot be calibrated the same way the β parameters are dealt with. The first term blindly fits the mean function to data, while the second term imposes smoothness on the δ-field. For simplicity and speed we don't use a dense error covariance matrix for the first term (as in ordinary least squares as opposed to generalized least squares), since for the δ parameters we are not interested in uncertainties. This is a modeling choice with which we aim to satisfy two objectives:

first, to get reasonable estimates of the δ field (for total column CO2 we expect that the spatial variation of the phase parameter should be be smooth) so that we do not end up fitting noise, and second, to perform this without computational complications. While computing $(m(x_\nu;\beta_\nu,\delta_\nu)-\psi_j)^T K^{-1}(m(x_\nu;\beta_\nu,\delta_\nu)-\psi_j)$ instead of the first term might be slightly more principled, this would come at the cost of needing to compute the Cholesky decomposition of the (often large) matrix K at each optimization iteration. For large grids, the cost of finding the parameters would then be much higher. Regarding the loss in precision due to this compromise: with e.g. OCO-2 data, if there are observations that inform the calibration of the parameters at a certain location, there is a large number of those observations, and even if the full covariance matrix K is used for the weighting, the parameters will end up being constrained enough that the most likely values from the generalized least squares would not differ by much from what is currently used. For this reason we are not concerned about the precision of our mean function.

The “how” part of the question was addressed in the comments above in A1, A3, and A4. We will still add a note about how the graph structure could be solved with algorithms such as generalized belief propagation, implementation of which is not yet included in satGP. This is future work that we hope to tackle in the near future.

As a final note we’d like to point out that the form of the mean function is generally data set specific, and it is the task of the modeler to understand the mean behaviour of the field before learning the GP parameters. While other data sets may require different, perhaps more complicated, mean function formulations, it is also possible to supply the mean function to satGP directly as an array.

2. *The notations in this manuscript are very confusing overall. For example, the authors sometimes use $\beta(^s)$ and later use β_ν. The covariance parameters are even more confusing. There are l, l_c, and l_I . Even the definition of I is not consistent: It is*

originally stated $I \subset \{x^s, x^t\}$, but later used as $I = ST or I = S$, and $I = ST$. Also, the authors used Δ_{year} in Equation (11) and stated Δ_{year} is the duration of one year, does this mean $\Delta_{\text{year}} = 365$? Similarly, in Equation (15), the authors used Δ_{period} ; is it 365 as well?

A7: First, we agree that using both $\beta(x^s)$ and β_ν may be confusing. We use ν to refer to a generic vertex on a graph, whereas we use $\beta(x^s)$ on p. 7 l. 22 to underline that the β parameters are space-dependent. We will remove this latter and explain the connection of the β_ν to the spatiality of the problem better.

Second, regarding the different ℓ variables, we’ll do our best to make the notation more consistent. Due to the many length scales and dimensions in the problem, there is, however, a need to use some kind of subindex notation to differentiate between them. Despite this, the reviewer is correct to point out that more clarity is needed.

Third, regarding the index set notation with the letter I, we agree that this is not optimal, and that the notation is not consistent (there is e.g. both ST and I_{ST} etc.). We will do our best simplify the notation and improve the readability by minimizing redundancy and explaining what the terms stand for in all cases.

Fourth, the Δ_{year} vs. Δ_{period} is an intentional discrepancy: we use a period length of one year for the OCO-2 data (this is a modeling choice) but for other data products other period lengths could be used. For instance the OCO-3 is on the International Space Station and therefore its orbit is not Sun-synchronous and the local time varies: for this reason a period of one day could be used to model the diurnal variation. Since satGP is intended to be a general purpose software, we describe the covariance functions in generic terms and for this reason would like to keep the Δ_{period} notation. However, we will mention that the period can be changed and that Δ_{year} is a modeling choice for OCO-2 data.

3. *The authors suggest the multi-scale covariance function given in Equation (18): $k(x,x';\theta) = \delta(x,x')\sigma_x^2 + k_{\text{per}}(x,x';\theta,I_S) + k_M(x,x';\theta) + k_{\text{exp}}(x,x';\theta,I+ST) +$*

$k_{\mathsf{W}}(x, x'; \theta, I)$.

- *First, I am not sure multi-scale is an accurate way to describe this covariance function. I feel this function is to add different types of covariance functions together, but these components not necessarily differ in terms of scales.*

A8: It is true that the combined covariance function works by adding different covariance functions together. However, how we decided to call the combined covariance function had a lot to do with the intended use of satGP: in the OCO-2 case we are particularly interested in finding the different length scales in the data induced by both spatial sparsity and underlying processes. Furthermore, remote sensing data often describes data from processes that involve different various characteristic length scales, as presented in e.g. figure 9. We could of course call the full kernel “multi-component”, but we would like to emphasize that we are in particular interested in the different length scales. Note, that even if the kernel components are of different types, they still may describe processes at different length scales. A non-multiscale kernel would arise in a situation, where a kernel utilizes, say, an exponential and a periodic kernel component with the same length scale parameters. Such usage, while possible, would likely be slightly unusual. For this reason we’d like to keep the terminology that we currently have. We will, however, add a note that the kernel could also be called “multi-component”, and briefly explain the reasoning behind the multi-scale name.

- *The authors did not explain clearly the component $kW(x, x0; \theta, I)$. Although Equation (16) states it is equal to $k\exp(x_W, x'_W; \theta^W, ST)$, the authors fail to explain x_W or the quantifies in Equation (17) especially, l, l^t, $l_{||}$, and $l_\perp$, and how these parameters are chosen/estimated.*

A9: We agree that this explanation is not adequate. The subindexes W are spurious in x_W, and those will be dropped, e.g. in (16). Also, we will clarify how the rotated

kernels function and rephrase this part of the text to improve clarity. As with other covariance kernels, also these parameters may be found by maximum likelihood. This procedure is outlined in Section 2.7, but we will add a note that it applies also to the wind-informed kernel parameters.

- *What will happen if there are missing data in wind velocity?*

A10: In case of OCO-2 (and with many other products), the wind data is included with the data files. The satGP code also includes running a Gaussian process for the wind data (and the output can then be utilized with k_W). Wind data may also be read from an external file. We will add a note about these capabilities in the text.

- *Why isn’t there an I involved in the Matérn component $k_M(\cdot, \cdot; \cdot)$?*

A11: Yes, there should of course be. This will also make the Matern description consistent with how the other kernels are described.

- *For the exponential component, the definition given in Equations (12) and (13) are not clear. At least there are two ways to define this component:*

$$k\exp(x, x_0; \theta, I_{ST}) = \tau^2 \exp\left(-\left|\frac{x - x'}{l_{ST}}\right|^{\gamma}\right)$$

or

$$k\exp(x, x_0; \theta, I_{ST}) = \tau^2 \exp\left(-\left|\frac{x^s - x^{s'}}{l_s}\right|^{\gamma_s}\right) \exp\left(-\left|\frac{x^t - x^{t'}}{l_t}\right|^{\gamma_t}\right)$$

dependent on whether the spatial and temporal components share the scale or exponent parameters. I don’t know what the authors have used, and there is no justification of their choice.

A12: Each temporal dimension has its own scale length parameter. This is what the subindex c in the sum and also in the term ℓ_c in (12) refers to. The sum is over the dimensions in the set I, and while we think this is quite clearly presented, we will still try to clarify. This means that the second version listed above is what is being used, with the caveat that the exponents γ are the same. If needed, this restriction can of course be quite easily lifted. For the OCO-2 experiments the exponent 2 was used.

- *The authors need to provide a better description of these components in the covariance function and explain why they are identifiable based on their formulations and definitions. Also, it is necessary to clarify whether some parameters are the same or vary across these components, such as τ^2 , γ, and l.*

A13: We will clarify that parameters such as τ are different for each kernel component. They can be found from the data, as was shown in the OCO-2 case. Of course the reviewer is correct that parameters of an arbitrary set of kernels would not necessarily be identifiable. However, what set of kernel components are chosen, is up to the modeler and depends on the data used. In the synthetic experiments we show that length scales of even three kernel components are recoverable, even though some parameters were slightly overestimated. We did perform additional tests, according to which parameters of two-component kernels are recoverable without such overestimation. We will add a comment on the modeler's role in picking the set of kernel components, underline that the synthetic studies verify the identifiability of the parameters, and furthermore do our best to improve the description of the kernels in general.

4. *I find Sections 2.6 and 2.7 quite difficult to understand. It seems that the authors use local kriging, that is, using a subset of data close to a prediction location x^* to estimate the covariance parameters and to make prediction.*

A14: This is correct. We use a set of hyperspheres in the space of the inputs x, within which we fit the kernel parameters.

Furthermore, it appears that the authors use different subsets of data to estimate the components in the covariance function. Why not using a single subset data to estimate the entire covariance function? Or, were the authors trying to avoid identifiability issue by using different data sets to estimate different covariance components? If a subset of data are used, I assume the size of this chosen subset is not too large, but why is there a need to use a block diagonal matrix $\tilde{K}$ as in Equation (22)? This approximation is not clearly explained, neither is E_{ref} in Equation (22). 2 ?

A15: We use the same subset of data to fit all the components at once, otherwise we could hardly claim that the parameters we choose are somehow optimal or correct. The sequentiality of the observation selection is due to something different: when we choose the (one and only) set of observations for fitting covariance parameters, we need to pick them so that all the (expected) length scales are represented in the data set. For instance, if the length scales are 10 kilometers and 1000 kilometers, we need to include both local dense data, and data from further away: if for instance we only include the closest observations, we don't really have leverage to say much about the longer-lengthscale behavior. We would like to point out more generally, that parameter identifiability is conditional on the data, so with some data (for instance with only one or zero observations) there will always be identifiability issues. While we think that we actually do explain what E_{ref} is on p. 12 l. 24, we agree that the description is short, and that the block-diagonality is explained only implicitly (or not at all). We will clarify these points and include a better description of the $\tilde{K}$ matrices in the revised manuscript.

Moreover, in Equation (19), should it be $> \sigma_{min}$ rather than $< \sigma_{min}$?

A16: This is definitely true and has now been fixed.

5. *The authors mentioned the nearest neighbor Gaussian process, but did not cite the reference correspondingly.*

A17: Thank you for pointing this out, we will of course add a proper reference.

6. *It is unclear where or why MCMC is needed and how it is implemented (prior specification etc.). The authors described optimization in Section 2 and also in the first paragraph of Page13. However, later in Page 17, the authors stated that MCMC is used instead. Section 2 does not describe MCMC.*

A18: The likelihood for learning the covariance parameters is noisy due to the observations selected changing with changing parameter values. For this reason optimization algorithms tend to get stuck in local minima. This is actually mentioned on p. 17 l. 4-5. We do mention that an Adaptive Metropolis implementation is included in the code, and that that can be used for finding the parameters (p. 13 l. 1-5). It is true that the priors are not described. We use flat priors, and will add information about them in the text in sections 4.1 and 4.4. We will also add a short description of MCMC to section 2.7.

7. *It should be Matérn covariance function, instead of Matern.*

A19: This has been fixed.

Anonymous Referee #2

A20: We thank Anonymous Referee #2 for appreciating our work. (No corrections or clarifications were requested.)

Executive Editor Comment

. . . *Therefore please provide the satGP v0.1 code or provide the reasons why the code can not be made publicly available in your revised submission to GMD.*

A21: We thank the executive editor for pointing out the code availability policy. We will make sure that the final revision conforms to the journal policies as requested.