Atmos. Chem. Phys. Discuss., 11, C3192–C3206, 2011
www.atmos-chem-phys-discuss.net/11/C3192/2011/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Bayesian statistical modeling of spatially correlated error structure in atmospheric tracer inverse analysis" *by* C. Mukherjee et al.

C. Mukherjee et al.

chiranjit@stat.duke.edu

Received and published: 12 May 2011

Response to comments of Referee #1

General comments

1. *"... inversion set-up for the MOPITT illustration, that yields overly-optimistic posterior uncertainties."*

   The credible posterior intervals based on the MCMC approach are consistent with the particular set-up considered here; based on the model and data assumptions,

they are accurate uncertainty assessments. The reviewer is quite correct in suggesting that the posterior uncertainties will likely be larger for higher resolution (in space and time) inversions. This will also be true in future extensions of spatial models to represent additional complexity in spatial structure that higher resolution models can capture. For the current paper, however, the specific model and data drawn from prior studies are used to present a first, proof-of-concept and example of our approach for accounting for spatial error correlation structures; the uncertainty results presented are accurate for the given model and data setup. We add statements making this clear in Section 3 and Section 4 of the revised manuscript.

2. *"...computational performance of the MCMC approach for real-size application: one may doubt that Monte Carlo methods are appropriate for very large (10^5) state vectors."*

   Computations increase roughly linearly with n, the size of the source vector, given our MCMC strategy. Counting floating point operations (FLOPs) necessary for sampling $x_1, \ldots, x_n$ in each iteration of the MCMC (using Minka's *Lightspeed* Matlab toolbox) gives results in the table below. Additional MCMC steps are based on $\mathbf{Kx}$ previously computed, so they are independent of n.

   The table reports FLOP counts for different n values where the posterior computation is based on a single epoch data (one month's data in our paper) on a 26×72 lattice. We also report stopwatch time (*tic toc*) and elapsed CPU time (*cputime*) for running the Matlab code on an Intel® Core™2 Duo CPU E8600 computer equipped with 8 GB memory.

   We observe a linear increase in FLOP counts, stopwatch and CPU times with n, the best one can expect. Moreover, matrix multiplications constitute most of the increase; these can be trivially parallelized on multi-core machines or clusters, or via exploitation of GPU parallelization, for very substantial gains with larger n. We have added a comment on this issue in Section 5 of the revised manuscript.

ACPD
11, C3192–C3206, 2011

n	FLOPs	Stopwatch time	Elapsed CPU time
10	70404728	0.0449	0.0839
100	704064128	0.0927	0.1700
1000	7040658128	0.4914	0.5784
10000	70406598128	5.0850	5.2070

Detailed comments

1. *p. 1672, l.11: 'know' should be 'known'.*

   This typo is corrected in the revised manuscript.

2. *p. 1674, l.23: The time scale for this numbers should be given. If they refer to a year worth of data for instance, the four powers of ten should be at least multiplied by 10 for real applications ($m \approx 10^5 - 10^6,\ n \approx 10^2 - 10^5$.)*

   This is corrected in the revised manuscript.

3. *p. 1675, l.24: disproportionate with respect to what?*

   We clarify in the revised manuscript that we mean disproportionate relative to surface and lower tropospheric mixing ratios.

4. *p. 1676, l.26-28: is the computational situation really better with MCMC?*

   MCMC based posterior computation for Gaussian process models (using a distance-based correlation function) involves inversion of a $m \times m$ correlation matrix in each iteration for computation of $\mathbf{U}$. Computational complexity of the inversion step is $\mathcal{O}(m^3)$, a prohibitively large load for a reasonable sized grid. CAR models, on the other hand, bypass the matrix inversion by modelling the inverse correlation matrix directly. Further, the CAR model has an inherently

ACPD
11, C3192–C3206, 2011

sparse $\mathbf{U}$ matrix — this leads to additional and considerable computational gains in an analysis with large m.

5. *Eq. (4): τ_c has not been defined.*

   A definition is added in the revised manuscript: *"The spatial dependence parameter ρ defines spatial association of each element of $\boldsymbol{\epsilon}^{CAR}$ with its neighbors, based on a specified neighborhood structure, and the scaling parameter τ_c^2 defines ranges of variation in these individual elements. Both parameters are to be estimated."*

6. *p. 1681: the sentence is not clear. Why are the missing rows interpolated and does this procedure introduce any artificial information to the inversion system?*

   The matrix $\mathbf{K}$ is calculated based on the model-generated Jacobian matrix and the instrument averaging kernels which characterize the vertical smoothing of each retrieval (see page 1674, lines 18-20 of the original manuscript). Averaging kernels are not available for missing measurements, and hence the corresponding elements of the matrix $\mathbf{K}$ are interpolated.

   The reason is that we need the full $\mathbf{K}$ matrix for our analysis as it is based on spatial modelling of the precision matrix. In contrast to properties of a covariance matrix, the precision matrix of $\mathbf{y}_H$ is *not* the corresponding submatrix of the full precision matrix. Computation of the reduced dimensional precision matrix for $\mathbf{y}_H$ requires high-dimensional matrix inversions; we have bypassed this through imputation $\mathbf{y}_M$ and subsequently working with the full precision matrix. The imputation step is computationally inexpensive as shown in Appendix A1.2; further, Appendix A1.2 shows why we need a one-time imputation of $\mathbf{K}(M,:)$ values.

   We do not perform any prediction at the missing sites; rather the imputation is used as a computational device. It is true that the specific values interpolated play a role in the analysis, and in particular on the implicit predictions of the missing outcomes $\mathbf{y}_M$ were we to be interested in those values (we are not,

ACPD

11, C3192–C3206, 2011

inherently). Other forms of interpolation could certainly be investigated, and the same question would arise with any alternative approach.

7. *p. 1682, l.16: the expression 'non-spatial' here relates to the absence of spatial correlations but may be not clear enough for some readers. The distinction with the CAR vs. GP comparison of Section 2.2 should be made more obvious.*

   We have clarified this in the revision; see the first paragraph of Section 3.

8. *Section 3.1: Only 15 degrees of freedom for the CO emission fields are adjusted by the 9-month inversion. Doing so, the system combines temporal and spatial aggregation errors, even though they do not seem to be taken into account. It should be stated at the beginning or at the end of the section that the set-up defined is very crude.*

   See response to general comment #1 above.

9. *p. 1684, l.22: two-sigma errors bars are shown, but one-sigma figures are more common in this field. This should be mentioned.*

   In this paper we have reported 95% posterior credible intervals calculated from the MCMC posterior sample; these intervals are not one- or two-sigma error bars, but provide the MCMC based approximations of intervals containing 95% posterior probability.

10. *p. 1685, l.12: 'accuracy' should be 'precision'.*

    Corrected.

11. *p. 1686, l.26: it would be interesting to report the equivalent e-folding lengths of the set-up.*

    Remember that the CAR model is a parametric model on the precision matrix; as such, its goodness-of-fit is best reflected through comparison of the *conditional distributions* of the data generating model and the CAR model. Fig. 1 in

ACPD

11, C3192–C3206, 2011

our manuscript illustrates model fit in the synthetic data example through comparison of conditional regression coefficients — we see a good fit in each case. We understand the question as it is based on the traditional idea of correlation functions; however, a part of our aim here is to broaden thinking about relevant statistical concepts: pairwise correlations are induced from conditional dependencies and the latter are the natural way to investigate CAR and other related spatial models.

That said, we present here some additional figures in response to this comment. Fig. 1 and Fig. 2, presented in the last two pages of this document, plot the data-generating exponential correlation functions and the corresponding *fitted* CAR correlation functions in the synthetic data study. The CAR correlation function is different from the exponential correlation function — it initially decays fast but more slowly later on. If we consider the entire range, the CAR correlation provides a good approximation of the exponential correlation function for each L-value while adding attractive features of fast sparse computation — but the e-folding length is misleadingly high for higher L-values. It is difficult to provide a single point estimate that describes the characteristic of the entire CAR correlation function; the e-folding length is certainly not a good measure. In our opinion the spatial correlation coefficient ρ is the appropriate indicator of strength of spatial association in the data as reflected in the estimated model.

12. *p. 1687, l.12: it may be appropriate to warn the reader that some of the correlated structures may be caused by the very coarse space-time resolution of this inversion (that may induce severe aggregation errors).*

    It is certainly true that the coarse space-time resolution may induce significant aggregation errors in the inversion. We have added a comment to that effect in the revision. However, this is really not an issue given the particular focus of our paper as the same coarse resolution model is used for both the NS and CAR inversions; the poorer posterior fits with the NS model result from not taking into

ACPD
11, C3192–C3206, 2011

account any spatial error correlations in the high resolution measurement field.

13. *p. 1688, l.2: the authors seem to suggest that the previous studies missed the obvious. The computational cost of MCMC methods has actually prevented some scientists from using it.*

    We agree with the second comment here, but stress that should change. As we demonstrate in the paper, using CAR spatial models (and perhaps more general models in future) the MCMC analysis is accessible. See also the response to the general comment on computational performance.

14. *p. 1688, l. 15: the effective dimension of the CO_2 application is even larger than for CO because the fluxes are much more diffuse and can be negative. It is not clear from the paper how MCMC methods can be efficiently applied in this case.*

    As we have indicated in our response to the general comment on computational performance, the analysis scales well with increasing dimension of the application. Negative fluxes are not an issue in that they can be taken into account by appropriately specifying the prior.

15. *p. 1690, l.16: this was debated in the statistical community decades ago, but a prior is always informative, to some extent.*

    We agree and have dropped the word “uninformative" in the revised manuscript.

16. *p.1692, l.14: extra ‘.*

    We have corrected this typo in the revised manuscript.

17. *References: Parasite numbers appear at the end of each citation.*

    We have corrected this in the revised manuscript.

18. *Figure 1: it would be good to give the mesh size (in km) in the legend, together with the found values of ρ and τ_c.*

ACPD

11, C3192–C3206, 2011

This is a nice suggestion. We have added the estimated values of ρ and τ_c in the z-axis of the right panel bar diagrams. Mesh size is not constant over the entire lattice, assuming mean radius r = 6371.0 km the average size of 4° latitude $\times$ 5° longitude grid boxes have approximate size 444.78 km $\times$ 555.97 km. We have added this information in the caption.

19. *Figure 7: it is confusing to see the truth plotted on a bar that represents the prior credible interval. At least this is not consistent with the way the posterior results are represented.*

    The “true" fluxes in the synthetic data analysis were simulated from the prior distribution; this was our reason for plotting the “truth” together with the prior credible interval. This is different to the way posterior point estimate and credible intervals are presented, but the side-by-side comparison of the posterior credible intervals with the “truth" provides a most useful and informative visual cue about whether the “truth" had been “captured" by the posterior inference in the synthetic data study.

20. *Figure 7: the posterior error bars are way too small for this observation type, which indicates that the inversion configuration is not well defined.*

    The posterior error bars are small because we are making inference about 15 source categories using a 9 months of data on a large grid. This data is informative about the “true" emissions, therefore the posterior credible intervals are concentrated around the “truth". This finding actually supports the validity of our inference procedure with MCMC. If one includes more sources the confidence intervals will naturally be wider since there will be fewer of observations per x_i.

21. *Figure 8: Same. As a result, NS posterior results for source category 7 are not statistically consistent with the truth.*

    Please see our reply to comment #20 above.

ACPD

11, C3192–C3206, 2011

22. *Figure 12: the squares are undefined here.*

    The squares represent prior means, now defined in the revised manuscript.

Response to comments of Referee #2

1. *How feasible is the method in cases with many degrees of freedom?*

   See response to general comment 1 and 2 of Referee #1.

2. *In addition, I'm not sure whether the "simply overwhelming statistical evidence" (p 1687 line 6) is actually a property of this example case with few degrees of freedom, and would look very different in a more relevant case with more degrees of freedom.*

   We certainly expect analyses with relevant spatial model components to dominate non-spatial models as a general rule, whatever the dimension of the source vector. We do also agree with what, perhaps, underlies the comment: increasingly realistic models including models with increased source dimension and better representation of flux processes can (and should) be expected to reduce residual spatial structure, as some of the spatial structure observed is due to model misfit. Ultimately, it is as this referee suggests a matter of case-by-case assessment, of course. Our analysis shows that this approach provides the opportunity to make such assessments with new data and new models, including models with more degrees of freedom — see comments on computational scale-up in the reply to general comments of the first referee.

3. *The covariances of the model-data error implied by the presented method correspond to a simple decorrelation with growing distance. How realistically does this describe the actual errors of satellite retrievals and/or model simulations? A*

ACPD

11, C3192–C3206, 2011

discussion of this point seems to be essential, as it touches the applicability of the method.

We agree that the single-dependence parameter CAR model is likely overly simplistic as a representation of spatial errors that combine model misfit and natural local dependencies in satellite retrieval data. To represent additional complexity in spatial structure, especially with regard to the potential opportunities to fit higher resolution models that can capture more refined structure, extensions or alternatives to CAR models will be needed. Indeed, one of our current research projects is to extend the general strategy of the paper to models that permit changes in the local dependency parameter across the spatial region; importantly, such extensions will also need to come through precision matrix models, rather than covariance models, for both the feasibility of computations and, we believe, for practical realism. That said, this is not so relevant from the viewpoint of our main goals in this paper: those of demonstrating an ability to fit a first class of — acceptably limited — spatial CAR models, and to demonstrate their already marked ability to substantially improve over the standard non-spatial models in terms of statistical fit, ability to recover sources in synthetic examples, and predictive match with the real data. We have included comments to this effect in the revision of the concluding section of the paper.

4. *In many places, the paper is written in rather complicated sentences (e.g. p. 1676 line 29, p. 1677 line 9-11, and many more), that are difficult to follow. There are also many filling words, and repetitive argumentation, which should be cut.*

   We have carefully edited the manuscript to address such issues, revising and clarifying text throughout.

5. *The motivation in abstract and introduction is very general and vague. It would be much more useful (and also make a more appealing entrance into the paper) if it would describe specifically what the paper will be about.*

ACPD

11, C3192–C3206, 2011

We have made changes to the abstract and introduction as requested, although we could not see any obvious ways to clarify further. Changes have been made to the wording at the start of the abstract, and in terms of an additional paragraph (now the second paragraph) in the introduction more directly summarizing the main contributions of the paper. We feel the revised abstract and introduction are focused and precise in representing both the overall scope and specific research presented in this paper.

Minor comments

1. *Just for interest: In the synthetic tests with known L, how does the a-posteriori estimated correlation length (i.e., that implied by the estimated ρ and τ_c^2) compare to the known L?*

   Please see the reply to comment #11 of Referee #1 on e-folding length.

2. *Eq (4): I suggest to write the condition in curly brackets, $\epsilon_i|\{\epsilon_j, j \neq i\}$, to indicate that it is the set of all other j.*

   This is a good suggestion. We have revised Eq (4) in the manuscript.

3. *p 1681 line 2: avoid "common" because there are many other choices in the literature.*

   We have modified this in the revised manuscript.

4. *p 1681 line 6: Explain "indicator function"'.*

   The definition of the indicator function is added in the revised manuscript.

5. *p 1681 line 21: Why didn't you just calculate the missing lines of $\mathbf{K}$ from the CTM?*

   Please see the reply to comment #6 of Referee #1.

ACPD
11, C3192–C3206, 2011

6. *p 1682 line 2: "de facto standard" — this may be the case in certain fields but certainly not in biogeochemistry.*

   We have added the phrase "in the statistical community" to clarify.

7. *p 1684 line 1: What is the 20.*

   This was chosen as a typical value based on the previous study by Arellano et al. (2004).

8. *p 1685 line 8: Shouldn't "LearningRatio $>$ 1" always be the case by theory?*

   This is not necessarily true. If the data is indicative of more uncertainty in x_i than the prior represents, then the 95% posterior credible interval will be longer than the 95% prior credible interval resulting in a Learning Ratio less than 1.

   The referee might be a little misled in thinking about Gaussian learning as a baseline here; posteriors are certainly always more precise than priors in Gaussian linear models with Gaussian priors and with no other complications or parameters involved. However, in models with more elaborate structure, such as with uncertainty about variances, spatial parameters and the truncated priors here, this is no longer a surety.

9. *p 1686 line 3: Probably can drop "logged".*

   We have modified his sentence in the revised manuscript.

10. *p 1687 line 12: Where does the noise come from here (prior fluxes, transport/chemistry)?*

    Please see the reply to detailed comment #12 of Referee #1.

11. *p 1690 line 2: Has τ_n been explained anywhere?*

    Please refer to our reply to detailed comment #7 of Referee #1.

ACPD

11, C3192–C3206, 2011

12. *Figs 7 and 8: Maybe 2 values of L are enough.*

We have provided results for all values of L to illustrate as fully as possible the performance of the CAR approach relative to the NS approach.

Other minor correction

The Michalak et al. (2005) references in the text have been corrected to the following Michalak et al. (2004) reference:

Michalak, A. M., L. Bruhwiler, and P. P. Tans (2004), A geostatistical approach to surface flux estimation of atmospheric trace gases, J. Geophys. Res., 109, D14109, doi:10.1029/2003JD004422.

Interactive comment on Atmos. Chem. Phys. Discuss., 11, 1671, 2011.

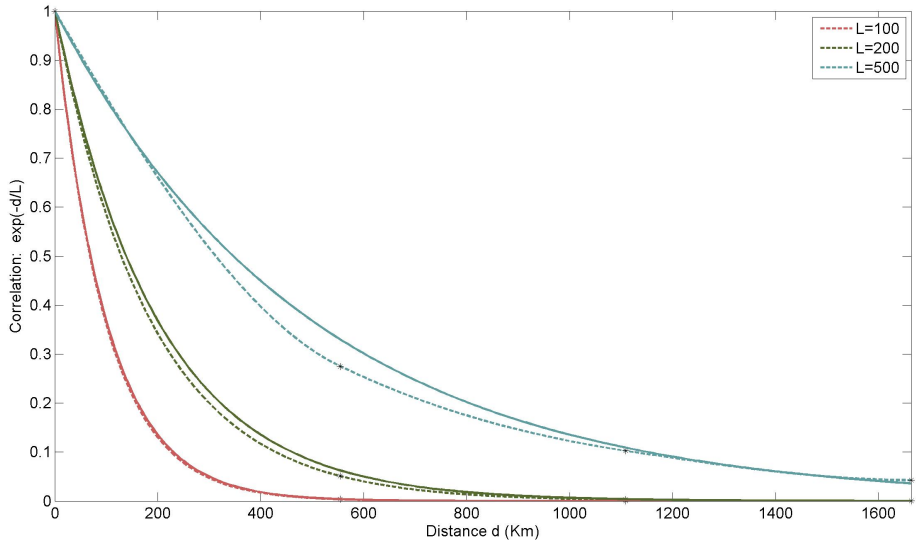

Fig. 1. Plots of the exponential decay correlation function exp(-d/L) for given values of L (in solid line) and corresponding fitted CAR correlation functions (in dashed line).

ACPD
11, C3192–C3206, 2011

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

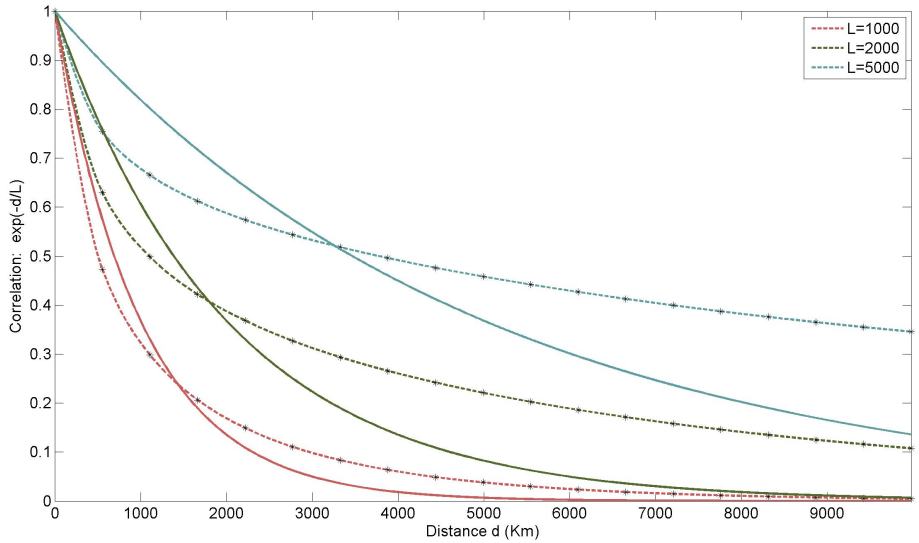

Fig. 2. Plots of the exponential decay correlation function exp(-d/L) for given values of L (in solid line) and corresponding fitted CAR correlation functions (in dashed line).

ACPD
11, C3192–C3206, 2011