Theoretical Basis Document: Detecting common patterns and systematic differences between flux maps obtained under different scenarios

Jonathan Hobbs Matthias Katzfuss Hai Nguyen Vineet Yadav

March 11, 2020

Abstract

...

Keywords: ...

1 Introduction

Flux estimates are subject to several sources of uncertainty, including observational errors, spatio-temporal representation uncertainty, and model transport error (Engelen et al., 2002). Some flux solutions attempt to account for these sources in their representation of the posterior uncertainty, but these are not always available and a coherent probabilistic assessment becomes challenging in the presence of multiple flux estimates with varying assumptions. The statistical methodology in this work provides a framework for this common situation.

Different flux maps can arise from combinations of multiple categorical factors. In this work we are particularly interested in flux estimates derived from different inversion systems, such as those investigated in model intercomparison projects (MIPs) (Thompson et al., 2016; Gaubert et al., 2019; Crowell et al., 2019). A second factor of interest is the makeup of the CO_2 concentration data used in the inversions. Our effort contrasts inversions that use Level 2 satellite retrievals directly versus inversions that use Level 3 products produced through data fusion (Nguyen et al., 2017).

Given a set of flux maps obtained under different scenarios, or combinations of these factors of interest, our goal is to find common features among the scenarios, and to identify systematic ways or regions in which fluxes from different scenarios differ. Analysis of variance (ANOVA) is a statistical modeling framework that facilitates the estimation of the common and factor-specific effects. It further characterizes the magnitude of the differences within factors relative to the inherent variability within a scenario. Model assumptions dictate the estimation of this within-scenario variability and will be an additional focus of our

[0]A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration (80NM0018F0527).

investigation. ANOVA methodology has been extended to functional data, such as time series and spatial fields, where it can provide a coherent depiction of space/time patterns and anomalies due to various factors (Kaufman and Sain, 2010).

2 ANOVA Model for Flux Estimates

Consider a spatial flux field $Y_\ell(s)$ for setting ℓ over a spatial region, which might be the entire globe or some focused region, such as the Transcom regions compared in Crowell et al. (2019).

In a spatial ANOVA approach (Kaufman and Sain, 2010), we assume that the setting is due to a number of "factors." For example, consider inversion system i and data source j. Then, we could assume

$$Y_{ijk}(s) = \mu(s) + \alpha_i(s) + \beta_j(s) + (\alpha\beta)_{ij}(s) + \epsilon_{ijk}(s), \tag{1}$$

where μ is the mean field representing spatial features in the common response, α_i quantifies the variation around μ due to the inversion system i, β_j quantifies the variation around μ due to data source j, $(\alpha\beta)_{ij}$ is an interaction effect, ϵ_{ijk} quantifies the internal variability within each scenario, and k is the replicate index.

We will make a Gaussian process assumption for each spatial field, whose mean and covariance functions can account for nonstationarity, if present. We condition on sum-to-zero constraints to ensure identifiability. For example, if there are are I inversion systems under consideration, the constraint becomes

$$\sum_{i=1}^{I} \alpha_i(s) = 0,$$

for all locations s.

A minimum collection of levels for these two factors is outlined in Table 1. These include two inversion systems. Additional inversion results from intercomparison projects (Crowell et al., 2019) could be incorporated into this framework. The levels for the data source factor include native Level 2 retrievals and fused products.

Table 1: Planned ANOVA comparison factors and corresponding factor levels.

Factor	Level
Inversion System, i	CMS-Flux (Liu et al., 2014) PCTM 4D-VAR (Baker et al., 2010)
Data Source, j	OCO-2 Level 2 Retrievals OCO-2 Fused, Gap-Filled

2.1 Inference

The model is fitted using Markov chain Monte Carlo (MCMC), which produces samples from the posterior distribution. From these samples, it is straightforward to produce maps including uncertainties of each of the spatial fields in (1). It is also possible to produce maps that show how and where different sources of variability contribute to the observed fluxes. This will allow us to identify regions in which there are systematic differences between different scenarios. For example, we could compute maps which at each location indicate the posterior probability that the effect of the inversion system is larger than the internal variability at that location.

The approach can also be extended to more formal testing by including indicator variables and computing their posterior probabilities.

2.2 Data

We will have daily, global flux maps for each scenario for a number of years. We will split the data into 12 subsets by month of the year, and conduct a separate analysis for each month, but we will have some replicates due to the different years. For a given month and year, we can either consider the individual days as replicates as well, or we simply average over all days within that month (i.e., consider monthly averages).

2.3 Computational issues

The considered data will be quite large, which will lead to substantial computational challenges. For example, when considering data on a global 1×1 degree grid, we will have $n \approx 65{,}000$ grid points or spatial locations. As standard GP calculations scale cubically in n, approximations will be necessary.

The Vecchia approximation (Vecchia, 1988; Stein et al., 2004; Katzfuss et al., 2017) has been applied to individual processes and reduces the number of computations to scale linearly in n. It should also be possible to apply the Vecchia approximation to the more complicated spatial ANOVA model considered here, although care must be taken due to the sum-to-zero constraints, which can be met by conditional simulation. If there are only two levels per factor, we can also easily reparameterize the model to avoid the constraints.

2.4 Additional considerations

In addition to the computational challenges mentioned above, the nature of the outputs from flux inversion systems lead to some additional practical considerations. For example, flux estimates from different inversion systems are available at different native spatial resolutions (Liu et al., 2014; Baker et al., 2010). The ANOVA data model (1) will need to incorporate an appropriate change of spatial support for each combination of factors.

3 Simulation study

We will carry out a simulation study to assess the behavior of the methodology in the two-factor, two-level setup outlined in Table 1. This simulation will serve as a test case for the implementation of the MCMC for the functional ANOVA model.

4 Currently unused material

If frequentist p-values are required, we can consider spatial multiple testing and enhanced false discovery rate procedures (Shen et al., 2002; Huang et al., 2019).

Our initial example is representing flux anomalies in time for a single inversion system. Figure 1 shows flux extimates over land for June 2015 and June 2016. A possible model for the flux $Y_i(s)$ at location s and year i is

$$Y_i(s) = \mu(s) + \alpha_i(s) + \epsilon_i(s)$$

Here we may be interested in estimating locations where the overall time-invariant mean flux $\mu(s)$ is nonzero as well as locations where the annual anomalies $\alpha_i(s)$ are nonzero.

Acknowledgments

This research was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. Support was provided by the MEASURES program.

A Appendix/Proofs

References

Baker, D. F., Bösch, Doney, S. C., O'Brien, D., and Schimel, D. S. (2010). Carbon source/sink information provided by column CO_2 measurements from the Orbiting Carbon Observatory. *Atmospheric Chemistry and Physics*, 10:4145–4165.

Crowell, S., Baker, D., Schuh, A., Basu, S., Jacobson, A. R., Chevallier, F., Liu, J., Deng, F., Feng, L., McKain, K., Chatterjee, A., Miller, J. B., Stephens, B. B., Eldering, A., Crisp, D., Schimel, D., Nassar, R., O'Dell, C. W., Oda, T., Sweeney, C., Palmer, P. I., and Jones, D. B. A. (2019). The 2015–2016 carbon cycle as seen from OCO-2 and the global in situ network. *Atmospheric Chemistry and Physics*, 19:9797–9831.

Engelen, R. J., Denning, A. J., Gurney, K. R., and TransCom3 (2002). On error estimation in atmospheric CO_2 invsersions. *Journal of Geophysical Research*, 107(D22).

Gaubert, B., Stephens, B. B., Basu, S., Chevallier, F., Deng, F., Kort, E. A., Patra, P. K., Peters, W., Rodenbeck, C., Saeki, T., Schimel, D., van der Laan-Luijkx, I., Wofsy, S., and Yin, Y. (2019). Global atmospheric CO_2 invsere models converging on neutral tropical land exchange, but disagreeing on fossil fuel and atmospheric growth rate. *Biogeosciences*, 16:117–134.

Huang, H.-C., Cressie, N., Zammit-Mangion, A., and Huang, G. (2019). False discovery rates to detect signals from incomplete spatially aggregated data. NIASRA, University of Wollongong Working Paper 05-19.

Katzfuss, M., Hammerling, D., and Smith, R. L. (2017). A Bayesian hierarchical model for climate-change detection and attribution. *Geophysical Research Letters*, 44(11):5720–5728.

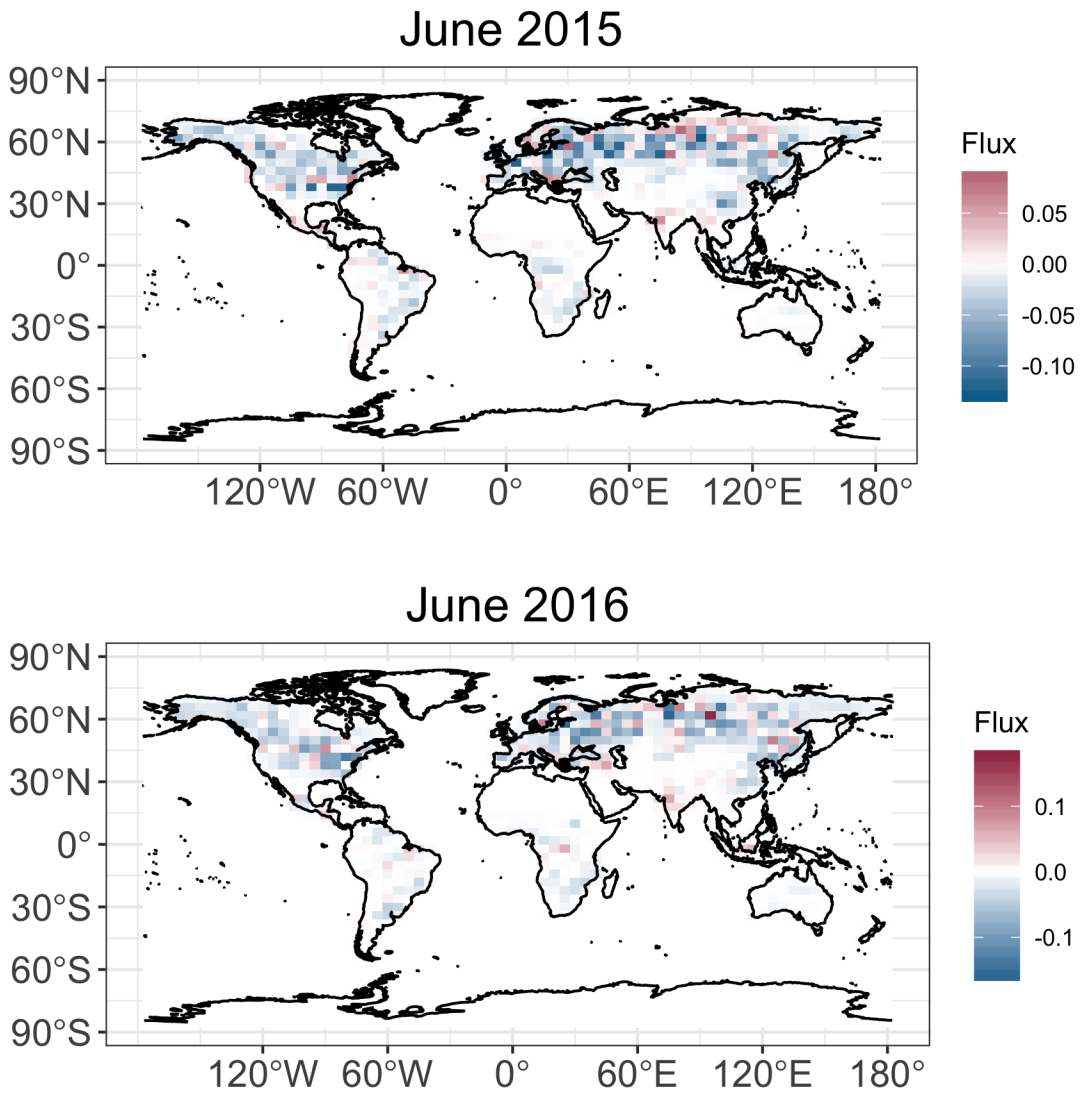

Figure 1: Flux estimates over land for June 2015 (top) and June 2016 (bottom).

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5:123–150.

Liu, J., Bowman, K. W., Lee, M., Henze, D. K., Bousserez, N., Brix, H., Collatz, G. J., Menemenlis, D., Ott, L., Pawson, S., Jones, D., and Nassar, R. (2014). Carbon monitoring system flux estimation and attribution: Impact of ACOS-GOSAT XCO_2 sampling on the inference of terrestrial biospheric sources and sinks. *Tellus B: Chemical and Physical Meteorology*, 66.

Nguyen, H., Cressie, N., and Braverman, A. (2017). Multivariate spatial data fusion for very large remote sensing datasets. *Remote Sensing*, 9(142).

Shen, X., Huang, H.-C., and Cressie, N. (2002). Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, 97:1122–1140.

Stein, M. L., Chi, Z., and Welty, L. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 66(2):275–296.

Thompson, R. L., Patra, P. K., Chevallier, F., Maksyutov, S., Law, R. M., van der Laan-Luijkx, I. T., Peters,

W., Ganshin, A., Zhuravlev, R., Maki, T., Nakamura, T., Shirai, T., Ishizawa, M., Saeki, T., Machida, T., Poulter, B., Canadell, J. G., and Ciais, P. (2016). Top-down assessment of the Asian carbon budget since the mid 1990s. *Nature Communications*, 7.

Vecchia, A. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B*, 50(2):297–312.