Bayesian Statistical Modeling of Spatial Error Correlations in Atmospheric Tracer Inverse Analysis

Prasad Kasibhatla
Chiranjit Mukherjee and Mike West
Duke University

5^{th} International GEOS-Chem Meeting
May 3, 2011

Funding: NASA ACMAP and Carbon Cycle Science Program

Bayesian Atmospheric Tracer Inversion

Bayesian Inverse Analysis

Update prior knowledge of fluxes [p($\mathbf{x}$)] based on measurements [$\mathbf{y}$] to estimate posterior pdf [p($\mathbf{x}|\mathbf{y}$)] – we are usually interested in moments (e.g. mean, variance-covariance) of this posterior distribution.

Linear problem (e.g. CO with prescribed OH)

$\mathbf{y} = \mathbf{K}\mathbf{x} + \boldsymbol{\varepsilon}$

where,

$\mathbf{K}$ is the Jacobian matrix derived from a CTM (K_{ij} is the contribution of a unit source at location/time = j to measurement at location/time = i).

Gaussian assumptions

$\boldsymbol{\varepsilon} \sim N[\mathbf{0}, \mathbf{S}_\varepsilon]$; $\mathbf{x} \sim N[\mathbf{x}_{prior}, \mathbf{S}_{prior}]$, with known $\mathbf{S}_\varepsilon$, $\mathbf{x}_{prior}$, $\mathbf{S}_{prior}$ → the posterior pdf of interest is also Gaussian:

$\mathbf{x}|\mathbf{y} \sim N[\mathbf{x}_{post}, \mathbf{S}_{post}]$,

$\mathbf{x}_{post} = \mathbf{x}_{prior} + (\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}\mathbf{K} + \mathbf{S}_{prior}^{-1})^{-1}\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_{prior})$

$\mathbf{S}_{post} = (\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}\mathbf{K} + \mathbf{S}_{prior}^{-1})^{-1}$

Incorporating Spatial Observation Error Correlation Structure

Covariance Matrix with Unknown Parameters

We can consider $\mathbf{S}_\varepsilon$ to be a function of unknown *structural parameters* $\boldsymbol{\theta}$ as

$\varepsilon|\boldsymbol{\theta} \sim N[\mathbf{0}, \mathbf{S}_\varepsilon(\boldsymbol{\theta})]$

→ The estimation problem then involves estimating $\mathbf{x}$ as well as $\boldsymbol{\theta}$.

Inverse Analysis

$\boldsymbol{\theta}$ can first be estimated using a maximum-likelihood approach, and $p(\mathbf{x}|\mathbf{y})$ can be approximated as $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}_{ML})$. Alternatively, $\boldsymbol{\theta}$ also be treated as a random variable with a prescribed prior distribution $[p(\boldsymbol{\theta})]$ and a fully Bayesian approach can be used to estimate the moments of the joint pdf $p(\mathbf{x},\boldsymbol{\theta}|\mathbf{y})$ based on a large sample generated from the joint pdf using MCMC.

Computational Burden

In either case, an iterative approach involving the calculation of $\mathbf{S}_\varepsilon^{-1}$ at each iteration is needed → this is a computationally intensive step when the dimension of $\mathbf{S}_\varepsilon$ is large.

Conditional Autoregressive (CAR) Spatial Models

CAR modeling

An alternative approach that involves conditional error modeling → $\mathbf{S}_\varepsilon^{-1}$ is sparse and calculations involving $\mathbf{S}_\varepsilon^{-1}$ are not computationally intensive.

Spatial CAR Model

$(\varepsilon_i | \varepsilon_j, j \neq i) \sim N[\,\Sigma\, (\rho\, w_{ij} / w_{i+})\, \varepsilon_j, \tau_c^2 / w_{i+}\,]$,

where,

w_{ij} are elements of a proximity matrix $\mathbf{W}$, with $w_{i+} = \Sigma_j\, w_{ij}$,

and

ρ and τ_c are unknown parameters.

Under this specification, it can be shown that the joint distribution p($\boldsymbol{\varepsilon}$) is

$\boldsymbol{\varepsilon} \sim N\, [\mathbf{0}, \mathbf{U}^{-1}]$,

where,

$\mathbf{U} = \tau_c^2\, (\mathbf{D}_w - \rho\, \mathbf{W})$, with $\mathbf{D}_w = \text{diag}(w_{1+}, w_{2+}, \ldots)$

→**U** is sparse and calculations involving **U** are not computationally intensive

Statistical Computation

Posterior inferences from MCMC-generated large-sample of p($\mathbf{x}$, ρ, τ | $\mathbf{y}$)

Incorporating Non-Normal Priors

Positive fluxes

We are often interested in inverse estimates of emissions, which are positive by definition → e.g. fluxes of CO, CO_2, etc. from vegetation fires, fossil-fuel combustion.

Specification of Prior Distribution

An alternative approach is to use a truncated normal distribution for the prior →$x_i \sim N(x_{a,i}, S_{a,i})\ I(x_i > t_i)$, where $I(.)$ is the indicator function [$I(.)$ is 1 when (.) is true and 0 when (.) is false].

Here, we choose $t_i = x_{prior,i}/4$, and choose $x_{a,i}$ and $S_{a,i}$, so that the mean of x_i is equal to $x_{prior,i}$ and the variance of x_i is equal to $S_{prior,i}$.

Computation

Sampling from the truncated normal distribution is straightforward to implement in the MCMC algorithm.

Synthetic Data Test

Synthetic Data Generation

→ Generate $\mathbf{x}_{true}$ for 15 source categories by sampling from truncated normal prior

→ Generate $\boldsymbol{\varepsilon}$ by sampling from
$\boldsymbol{\varepsilon}_{syn} \sim N(\mathbf{0}, \mathbf{S}_{\varepsilon})$
with $S_{\varepsilon}(i,j) = \sigma^2 \exp(-d_{ij}/L)$,
and σ = 20% of annual, global average prior model CO

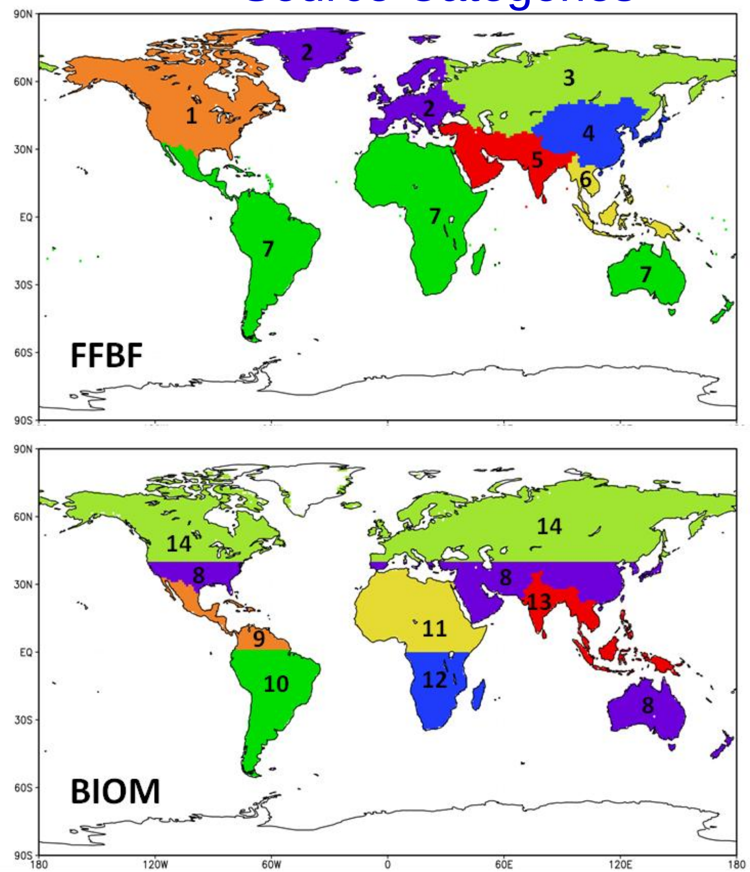
→ Generate $\mathbf{y}_{syn} = \mathbf{K}\mathbf{x}_{true} + \boldsymbol{\varepsilon}_{syn}$

→ Repeat for 6 different values of L (100, 200, 500, 1000, 2000, and 5000 km)

→ Repeat for 1000 synthetic datasets for each value of L

→**K** from Arellano et al. (2004) for April – December 2000

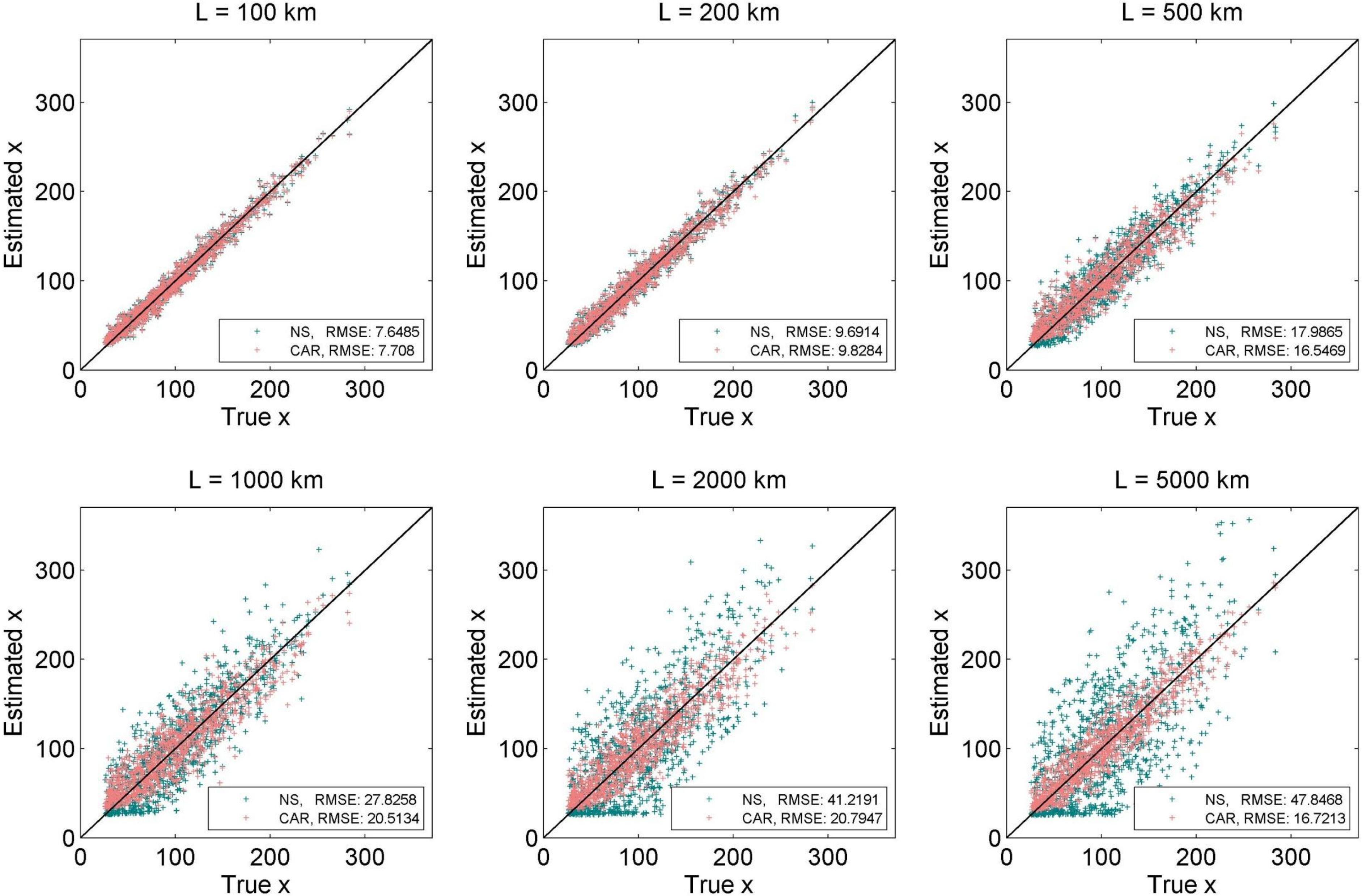
Source Categories



1. FFBF-NAM
2. FFBF-EUR
3. FFBF-RUS
4. FFBF-EAS
5. FFBF-SAS
6. FFBF-SEA
7. FFBF-ROW
8. BIOM-ROW
9. BIOM-NLA
10. BIOM-SLA
11. BIOM-NAF
12. BIOM-SAF
13. BIOM-SSA
14. BIOM-BOR
15. BIOG

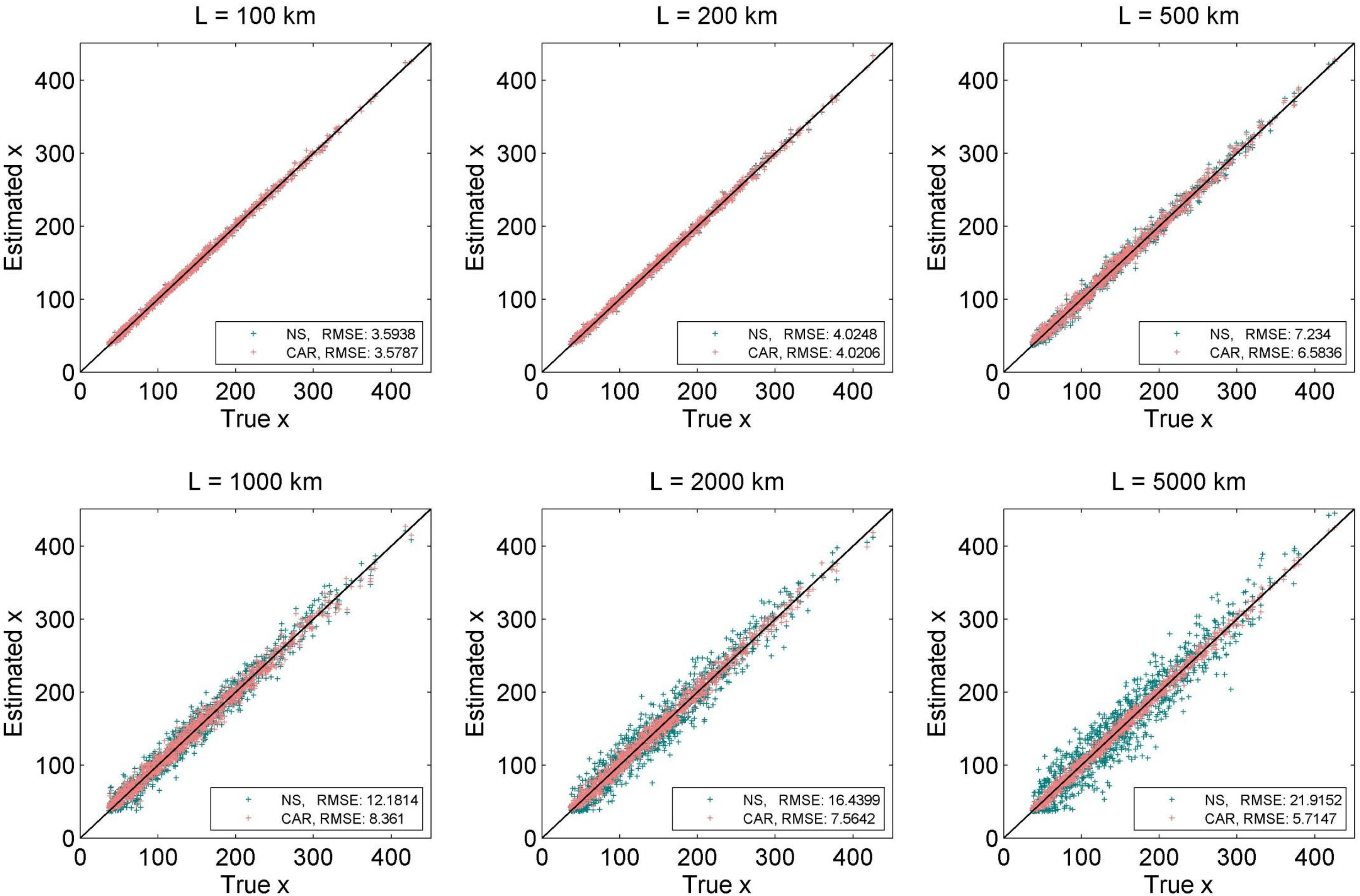
Compare inversion results from non-spatial and CAR models

Synthetic Inversion Results: FFBF-NAM



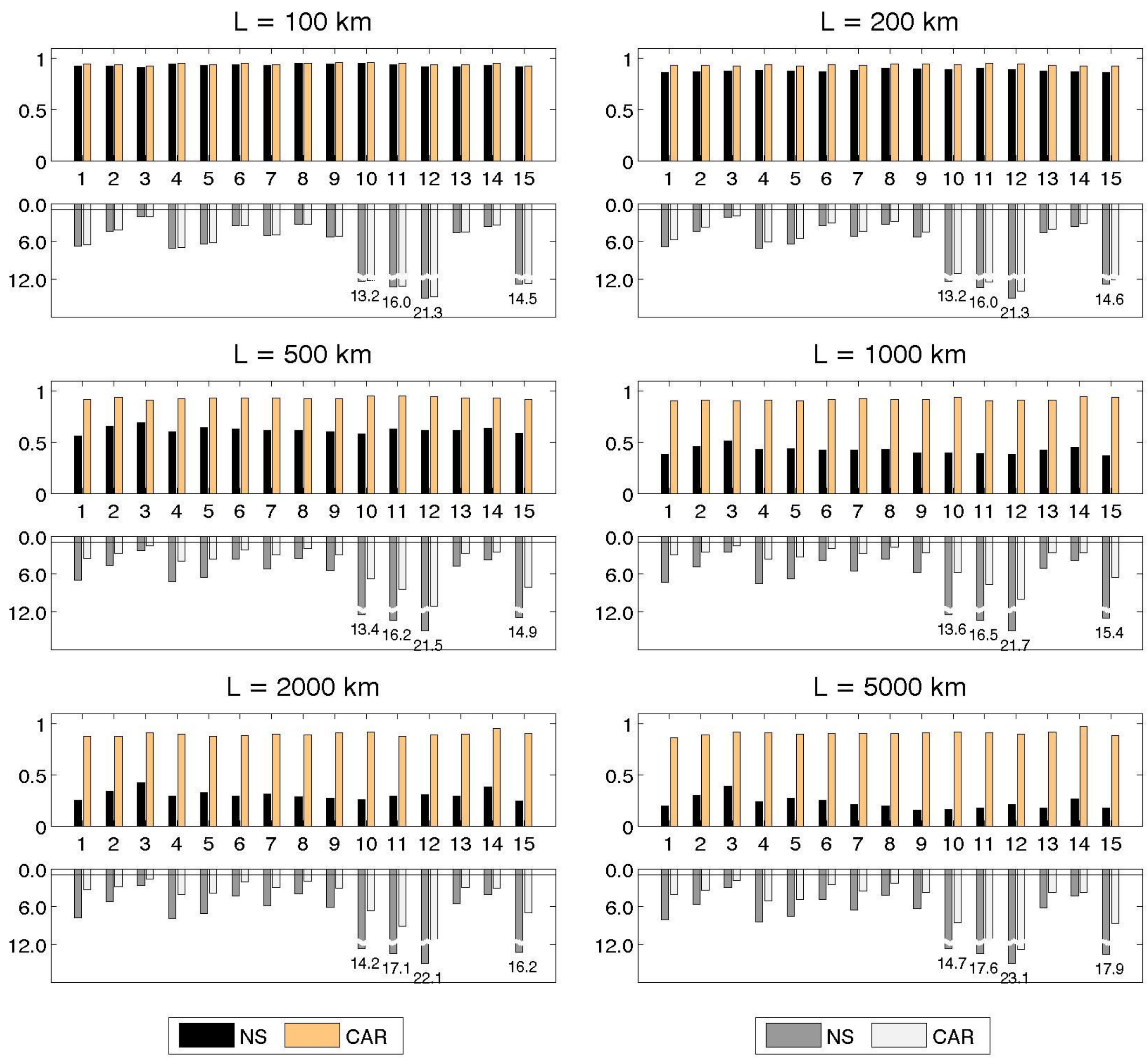
Blue: Non-spatial; Red: CAR

Synthetic Inversion Results: BIOM-SAF



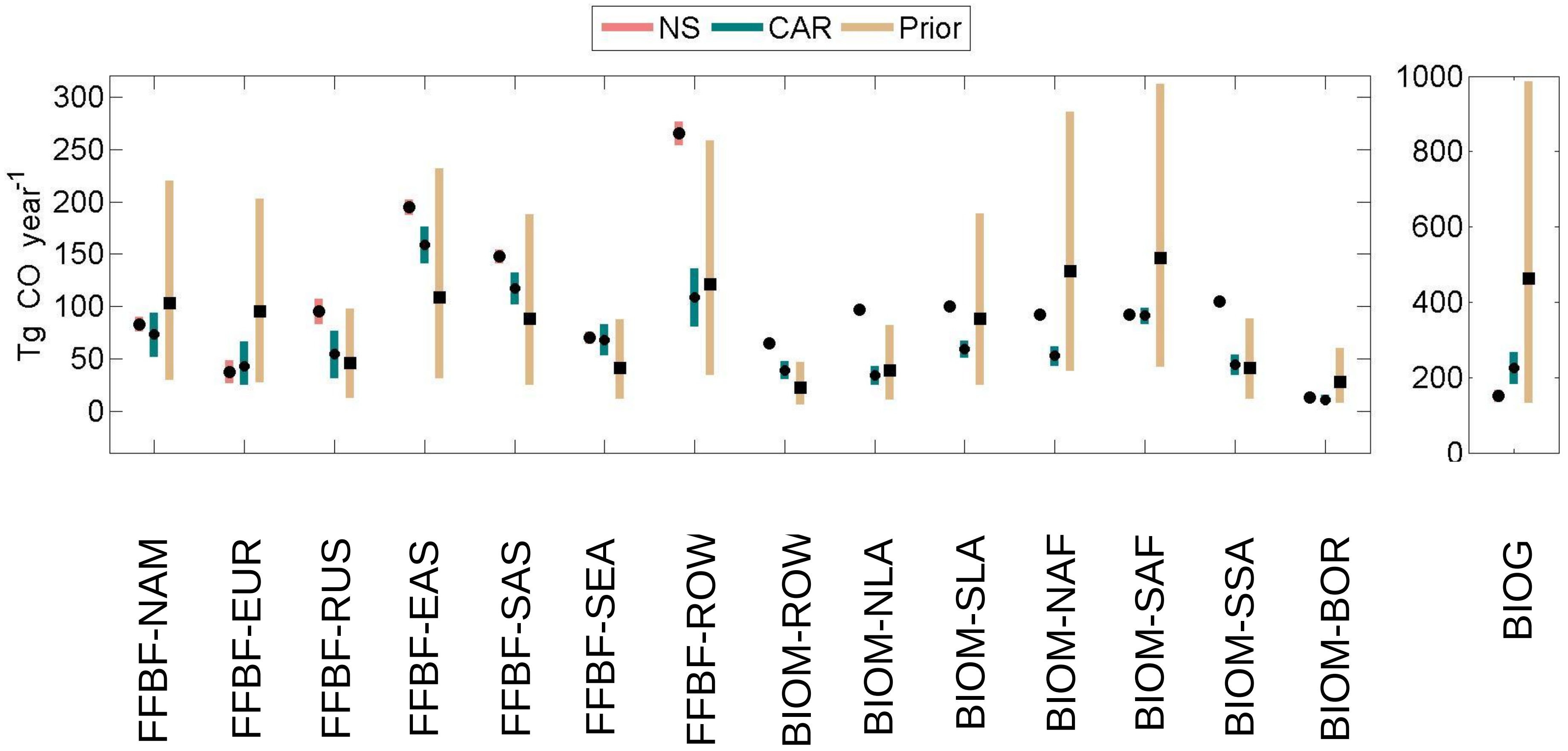
Blue: Non-spatial; Red: CAR

Synthetic Inversion Results
Success Rates (top) and Learning Ratios (bottom)



Real MOPITT Data Inversion Results

Mean (symbols) and 95% credible intervals (lines)



Summary

→ Not accounting for spatial observation error correlations can lead to biased posterior source estimates.

→ CAR spatial modeling offers a flexible and computationally tractable approach to account for spatial error correlation structure in Bayesian atmospheric inverse modeling.

→ Future work

Extend to multiple correlation length scales (e.g lat-dep. L)

Extend to consider time correlations

Integrate into on-going EnKF work for grid-scale inversions

m	FLOPs
10	70404728
100	704064128
1000	7040658128
10000	70406598128