SANDIA REPORT

SAND2014-17932
Unlimited Release
Printed September 2014

Greenhouse Gas Source Attribution: Measurements, Modeling, and Uncertainty Quantification

Z. Liu, C. Safta, K. Sargsyan, H. N. Najm, B. G. van Bloemen Waanders, B. W. LaFranchi, M. D. Ivey, P. E. Schrader, H. A. Michelsen, R. P. Bambha

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2014-17932
Unlimited Release
Printed September 2014

Greenhouse Gas Source Attribution: Measurements, Modeling, and Uncertainty Quantification

Z. Liu, C. Safta, K. Sargsyan, H. N. Najm, B. G. van Bloemen Waanders
B. W. LaFranchi, M. D. Ivey, P. E. Schrader, H. A. Michelsen, R. P. Bambha
Sandia National Laboratories
Livermore, CA 94550

Abstract

In this project we have developed atmospheric measurement capabilities and a suite of atmospheric modeling and analysis tools that are well suited for verifying emissions of greenhouse gases (GHGs) on an urban-through-regional scale. We have for the first time applied the Community Multiscale Air Quality (CMAQ) model to simulate atmospheric CO_2. This will allow for the examination of regional-scale transport and distribution of CO_2 along with air pollutants traditionally studied using CMAQ at relatively high spatial and temporal resolution with the goal of leveraging emissions verification efforts for both air quality and climate. We have developed a bias-enhanced Bayesian inference approach that can remedy the well-known problem of transport model errors in atmospheric CO_2 inversions. We have tested the approach using data and model outputs from the TransCom3 global CO_2 inversion comparison project. We have also performed two prototyping studies on inversion approaches in the generalized convection-diffusion context. One of these studies employed Polynomial Chaos Expansion to accelerate the evaluation of a regional transport model and enable efficient Markov Chain Monte Carlo sampling of the posterior for Bayesian inference. The other approach uses deterministic inversion of a convection-diffusion-reaction system in the presence of uncertainty. These approaches should, in principle, be applicable to realistic atmospheric problems with moderate adaptation.

We outline a regional greenhouse gas source inference system that integrates (1) two approaches of atmospheric dispersion simulation and (2) a class of Bayesian inference and uncertainty quantification algorithms. We use two different and complementary approaches to simulate atmospheric dispersion. Specifically, we use a Eulerian chemical transport model CMAQ and a Lagrangian Particle Dispersion Model - FLEXPART-WRF. These two models share the same WRF assimilated meteorology fields, making it possible to perform a hybrid simulation, in which the Eulerian model (CMAQ) can be used to compute the initial condition needed by the Lagrangian model, while the source-receptor relationships for a large state

vector can be efficiently computed using the Lagrangian model in its backward mode. In addition, CMAQ has a complete treatment of atmospheric chemistry of a suite of traditional air pollutants, many of which could help attribute GHGs from different sources. The inference of emissions sources using atmospheric observations is cast as a Bayesian model calibration problem, which is solved using a variety of Bayesian techniques, such as the bias-enhanced Bayesian inference algorithm, which accounts for the intrinsic model deficiency, Polynomial Chaos Expansion to accelerate model evaluation and Markov Chain Monte Carlo sampling, and Karhunen-Loève (KL) Expansion to reduce the dimensionality of the state space.

We have established an atmospheric measurement site in Livermore, CA and are collecting continuous measurements of CO_2, CH_4 and other species that are typically co-emitted with these GHGs. Measurements of co-emitted species can assist in attributing the GHGs to different emissions sectors. Automatic calibrations using traceable standards are performed routinely for the gas-phase measurements. We are also collecting standard meteorological data at the Livermore site as well as planetary boundary height measurements using a ceilometer. The location of the measurement site is well suited to sample air transported between the San Francisco Bay area and the California Central Valley.

Acknowledgment

We are very grateful to our colleagues and collaborators, Drs. Bert J. Debusschere, Fred Helsel and Todd Barnett at SNL, Dr. Joseph Pinto at US EPA, Drs. Tao Zeng and Jim Boylan at Georgia Department of Natural Resources, Drs. Maoyi Huang and Chun Zhao at PNNL, Drs. Huimin Lei, Shishi Liu, Jiafu Mao, Xiaoying Shi, and Yaxing Wei at ORNL, Dr. Christopher R. Schwalm at Northern Arizona University, Drs. Daven Henze and Nicolas Bousserez at University of Colorado, Dr. Jerome Brioude at NOAA, and Alexander J. Turner at Harvard University.

We are also very grateful to Drs. Andrew Jacobson, Anna Michalak, Kevin Schaefer, and Ling Jin for their extensive comments and suggestions on the CMAQ CO_2 simulations. The WRF output and non-CO_2 emission data for the CMAQ CO_2 modeling are products of the Southeastern Modeling, Analysis, and Planning (SEMAP) project (`http://www.metro4-sesarm.org`; accessed January 14, 2013). The CLM4VIC output is a product of the MsTMIP project. The MsTMIP project has been funded has been funded by National Aeronautics and Space Administration (NASA) under grant No. NNX11AO08A and NNH10AN68I, and is a contribution of the North American Carbon Program. PNNL is operated for the US DOE by Battelle Memorial Institute under Contract DE-AC06-76RLO1830. The information in the CMAQ CO_2 chapter has been subjected to review by the National Center for Environmental Assessment, U.S. Environmental Protection Agency, and approved for publication. We thank Arlyn Andrews and the NOAA GMD for making the tower CO_2 data public, and thank NOAA ESRL for making the CarbonTracker-2011 results public (`http://carbontracker.noaa.gov`). We thank Dr. Kevin Gurney and the Vulcan project for generating and sharing the Vulcan emission inventory (`http://vulcan.project.asu.edu`). We also thank Dr. Kevin Gurney and the TransCom project (`http://transcom.project.asu.edu`) for making the TransCom3 data public.

This project is funded by Sandia Laboratory Directed Research and Development (LDRD) Project (No.158809) entitled *Designing Greenhouse Gas Monitoring Systems and Reducing Their Uncertainties*. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the National Nuclear Security Administration under contract DE-AC04-94-AL85000.

This page intentionally left blank.

Contents

Appendix

Figures

Tables

Nomenclature

4D-Var Four dimensional variational data assimilation

a.g.l. above ground level

BAO Boulder Atmospheric Observatory

CAP criteria air pollutant

CASA Carnegie-Ames Stanford Approach

CCDAS Carbon Cycle Data Assimilation System

CDIAC Carbon Dioxide Information Analysis Center

CMAQ Community Multiscale Air Quality

CMS Carbon Monitoring System

DA data assimilation

EnKF ensemble Kalman Filter

HAP hazardous air pollutant

LPDM Lagrangian particle dispersion model

MAP maximum a posteriori

MRV&V Measurement, Reporting, Verification & Validation

MsTMIP Multi-Scale Synthesis and Terrestrial Model Intercomparison

NACP North American Carbon Program

NDVI Normalized Difference Vegetation Index

NEE Net Ecosystem Exchange

NEI National Emission Inventory

NMIM National Mobile Inventory Model

NWP numerical weather prediction

ppm parts per million

PTR-MS proton transfer reaction mass spectrometer

RHS right-hand-side

sccm standard cubic centimeters per minute

SMOKE Sparse Matrix Operator Kernel Emissions

TBM terrestrial biosphere model

UNFCCC United Nations Framework Convention on Climate Change

UQ Uncertainty Quantification

VPRM Vegetation Photosynthesis and Respiration Model

WRF Weather Research and Forecasting model

1 Introduction

1.1 Fossil-fuel CO_2 Emissions Verification: Challenges and Opportunities

The increase of atmospheric CO_2 concentrations, as the largest human-induced climate forcer, is continuing and accelerating [1]. Reducing anthropogenic emissions is the most effective way to mitigate the resulting climate change risks. The success of an international collaborative effort in emissions reduction relies upon accurate information of current emissions in each country and their change over time [2]. Under the United Nations Framework Convention on Climate Change (UNFCCC), all countries are required to report annual anthropogenic emissions and removal of greenhouse gases (GHGs), although developing countries are allowed to report less frequently and in less detail than Annex I (developed) countries. These self-reported national emissions inventories, however, are known to have uncertainties attributable to the incomplete knowledge of the numerous emission sources or inaccurate national and state statistical data (e.g., [3, 4]). Emissions verification, which aims at (1) reducing uncertainties in current emissions inventories, and (2) monitoring and verifying changes in emissions over time [2], has emerged as an urgent need for decision-making by policy makers and business leaders [5, 6].

From the perspective of the global carbon cycle, our understanding of fossil-fuel emissions is generally believed to be better than that of natural carbon sources and sinks [1]; there are currently large gaps of knowledge in the biogeochemistry and physics of the natural carbon cycle [7], whereas fossil-fuel emissions are largely constrained by relatively well-documented global fuel consumption data [8]. Nevertheless, our understanding of some important characteristics of fossil-fuel emissions, such as their spatiotemporal variability, remains elusive [3, 9]. One unique feature of fossil-fuel emissions is their extremely uneven spatial distribution. A striking example pointed out by Marland [5] shows that even a tiny amount of uncertainty, i.e., 0.9% as estimated by two independent sources, in one of the leading emitters like the U.S. is equivalent to total emissions from a very large group of smaller emitters in the world, i.e., 147 countries out of 195 countries analyzed. A refined characterization of fossil-fuel emissions in space and time is necessary for emissions verification, and is also important for better constraining the carbon cycle [10, 11].

A formal approach to the emissions verification problem is inverse modeling, which seeks to improve existing emissions estimates through assimilating information from atmospheric observations [2]. In inverse modeling, a source-receptor relationship, i.e., a relationship between concentrations at a receptor site and emissions strength from sources, is first established by modeling CO_2 fluxes and atmospheric transport. With such a relationship and observations of CO_2 concentrations in the atmosphere, an updated emissions estimate with reduced uncertainties (due to the addition of observational information) can be obtained using various estimation techniques [12].

Despite the rigor of the theory behind inverse assimilation, inverse modeling of atmospheric CO_2 has long been challenged by (1) the sparseness of observational data (e.g., [13]) and (2) inaccurate atmospheric transport modeling (e.g., [14]). The majority of CO_2 flux-inversion studies has focused on natural (i.e., terrestrial biosphere and oceans) carbon fluxes, which are believed to be much more uncertain than fossil-fuel emissions. In turn, inverting for fossil-fuel emissions to

achieve the objective of emissions verification would be challenged by the well-known signal-to-noise problem, namely the strong, highly variable, but quite uncertain interference by the biosphere [2]. As such, emissions verification turns out to be a challenging scientific problem of seeking an optimized combination of state-of-the-art observational techniques and modeling capabilities to pinpoint and quantify fossil-fuel emissions signals in the atmosphere. In order to find an effective and practical observational strategy, a growing number of observations from ground sites (e.g., [15, 16]), aircraft (e.g., [17]), and satellites (e.g., [18]) have been examined for their ability to constrain fossil-fuel emissions. Proxy techniques using isotopologues (e.g., [19, **?**, 2]) and trace-gas species to isolate fossil-fuel CO_2 (e.g., [20, 21, 19, 22]) have also been proposed and investigated. Nevertheless, these previous observational studies were limited to a small number of locales, and thus the larger scale representativeness of their findings is not clear. There is apparently an urgent need for better interpreting these emerging observations and optimizing existing observation networks.

1.2 Inverse Modeling of Atmospheric CO_2 Sources and Sinks: A Short Overview

CO_2 concentrations measured in the atmosphere yield information about CO_2 emissions trends and can provide constraints on magnitudes and strengths of CO_2 sources and sinks in space and time. Earlier pioneering studies in the 1990s (e.g., [23]) used atmospheric CO_2 concentrations and sea-surface partial pressure of CO_2 from global networks to demonstrate a significant discrepancy between the observed and model simulated hemispheric CO_2 gradients. Such a discrepancy revealed a large missing carbon sink over land in the northern hemisphere and motivated extensive subsequent efforts to further utilize in situ atmospheric observations of CO_2 to bound the global and regional budgets of CO_2 e.g., [13]). Whereas in situ networks have been demonstrated to be too sparse to provide adequate constraints, satellite retrievals of column average mixing ratio of CO_2 (xCO_2) that were available in the last decade or so have generated a large pool of novel data awaiting exploration (e.g., [24, 25, 26, 27])]. Meanwhile, as a result of continuing development of inverse modeling algorithms and improvement of computational infrastructure, advanced carbon cycle data assimilation (DA) systems, capable of ingesting huge-volume data from satellites, have been developed to produce CO_2 flux reanalysis products with high spatial and temporal resolution. In addition, some DA systems also have the capability of constraining control variables in terrestrial biosphere models (TBMs) using atmospheric observations [28], making it possible to directly assess the value of atmospheric observations in informing carbon cycle and climate modeling.

Inferring CO_2 fluxes at the Earths surface using observed atmospheric concentrations is a classic ill-conditioned inverse problem. The flux-concentration relation is well-known to be linear, as can be readily shown in a Lagrangian framework, where atmospheric concentrations (observables) are Lagrangian line integrals and the fluxes at sources and sinks (unknowns) are integrands (e.g., [29, 30]). Bayesian methodology provides the most widely adopted framework to formulate and solve the atmospheric CO_2-flux-inversion-problem, in which an optimal estimate of the fluxes based on both observed concentrations and prior knowledge of fluxes is sought. Furthermore, most existing CO_2-flux-inversion algorithms use a least-squares approach [12], which seeks to minimize a cost function in the L2-norm consisting of the penalties for prior and observations

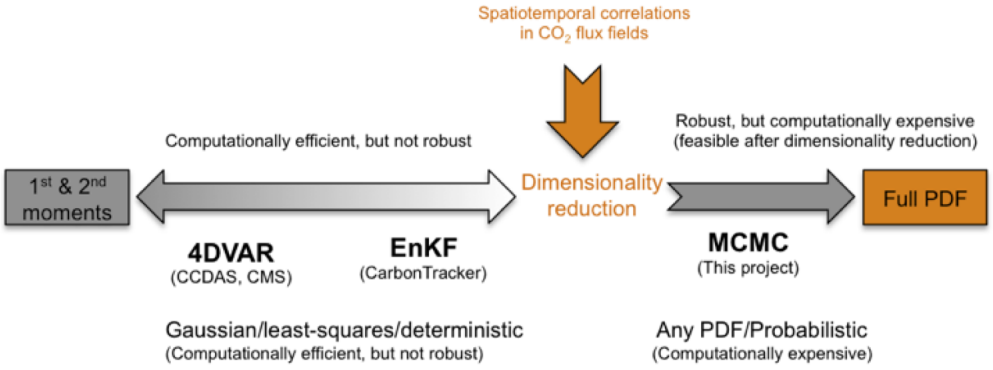


Figure 1. State-of-the-art CO_2 surface flux inversion algorithms and the new Bayesian Probabilistic Inference approach.

weighted by their uncertainties. The least-squares approach is associated with Gaussian assumptions in the Bayesian context. The most attractive consequence of the Gaussian assumption and the least-squares approach to inverse problem is that the solution, consisting of a maximum a posteriori (MAP) estimate of the unknown state vector and a corresponding error covariance matrix, can be obtained analytically.

Various least-squares-based inversion and DA algorithms, which were previously employed for numerical weather prediction (NWP), have been adapted for application to atmospheric CO_2 data assimilation. Three popular algorithms that have been used in production mode include Bayesian synthesis inversion, ensemble Kalman Filter (EnKF), and variational methods rooted in control theory (e.g., 4D-Var), see Fig. 1.

Bayesian synthesis inversion has been adopted most extensively since the earliest efforts of CO_2 flux inversion. The mathematical formulation and experimental protocol of global-scale Bayesian synthesis inversion formally account for key properties of CO_2 (e.g., long lifetime), atmospheric transport characteristics, and spatial distributions of observational data and ecosystems [31, 32, 33, 13], and have been well documented, e.g., during the TransCOM project (`http://transcom.project.asu.edu/`). A notable issue regarding Bayesian synthesis inversion is the effect of aggregation error, which stems from the definition of state vectors representing aggregated fluxes over space and time [34]. Bayesian synthesis inversion is still being employed for many studies because of its theoretical completeness and convenience for implementation.

Compared to Bayesian synthesis inversion, the other two algorithms, EnKF and 4D-Var, have not been studied as well in the context of CO_2 flux inference. Both EnKF and 4D-Var were previously used for NWP and were adapted for CO_2 flux inference only recently [35, 28]. Some fundamental differences in the dynamical models driving meteorology and CO_2 transport [36, 35] have led to current efforts to investigate and optimize various technical aspects of these two approaches to improve their performance and robustness in CO_2 flux inference [37]. The EnKF method assimilates observational data only prior to the time step being analyzed and has a moderate level of statistical completeness attained by sampling from prior space in a Monte Carlo

setting. The 4D-Var is attractive mostly for the high computational efficiency that makes it applicable for producing CO_2 flux analyses operationally at native model resolution. The 4D-Var method, however, requires the adjoint of the transport model. 4D-Var has been adopted by the Carbon Cycle Data Assimilation System (CCDAS) [28], and it was also adopted to produce NASA CMS flux products (`http://cmsflux.jpl.nasa.gov/AS-SystemArchitecture.aspx`). Both 4D-Var and EnKF, while computationally efficient, are statistically incomplete and either rely on ad hoc measures to evaluate the confidence in results without a formal probabilistic basis or do not assess the associated uncertainties (Fig. 1).

Uncertainty quantification and propagation is a critical component of useful inverse modeling of CO_2 fluxes, given the typical sparseness of observational constraints and diffusive nature of atmospheric transport that render the problem under-determined. Although critically important, it has been a long-standing challenge to fully understand and quantify uncertainties in CO_2 flux inversion. Theories have been established regarding the composition of the error budget and physical meanings of various types of errors therein [31, 38, 12]. Ideally in the Gaussian least-squares framework, once the quantitative information of errors in the prior and the model-data mismatch are obtained, the two error covariance matrices can be filled, and these errors will propagate into the a posteriori flux covariance matrices. However, this has turned out to be very challenging. First, it is difficult to quantify these various types of errors. Numerous efforts have been made in this regard, e.g., to characterize the probability distributions of prior flux errors [39, 40], to mitigate the impact of aggregation errors [34], and to quantify model transport errors [41, 42, 43] and representation errors [44]. Secondly, the size of the covariance matrices is prohibitively large for high-resolution flux analysis, such as the Carbon Monitoring System (CMS) flux product, so that it is not feasible to propagate uncertainties in the flux fields at there native resolution, and certain types of truncation have to be implemented [45]. Finally and probably the most challengingly, recent evidence suggests that the probability distributions of the errors in the prior and the likelihood function are not likely to be Gaussian [39, 46, 47], which is in contradiction to the premises of the least-squares approach and the associated methodology for accounting for errors. Chevallier *et.al.* [39, 40] estimated the errors in model predicted fluxes through comparisons with eddy covariance flux observations, and a distribution with long tail is found which contradicts the common assumption of Gaussian error. As Ricciuto *et.al.* [46] argued and confirmed with their model results, propagation of errors through a highly non-linear terrestrial biosphere model (TBM) generates non-Gaussian errors in the output. The uncertainties of anthropogenic fluxes of CO_2 are also inherently non-Gaussian given that they are always positive. A common exercise during inference of biospheric fluxes is to subtract the fossil-fuel CO_2 components from the observations [13]. Previous studies have noted marked sensitivity of biospheric fluxes to seasonal and inter-annual variability [48, 11] and uncertainties [10] in fossil fuel emissions, which highlights the importance of properly accounting for fossil-fuel emissions uncertainties while inverting for natural fluxes.

To address these issues and challenges, it is desirable to develop an alternative flux-inversion algorithm that can (1) reduce the high-dimensionality of the state vector and error covariance matrices, and (2) accommodate and propagate uncertainties with non-Gaussian distributions. In the following, we outline a method to formulate and solve the CO_2 flux inversion problem in a fully Bayesian, probabilistic framework without regularizing assumptions that are not backed by scientific evidence. As detailed in the following sections, the new inversion framework involves two

components/steps. The first step is to substantially reduce the dimensionality of the prior flux fields by some means, e.g., Karhunen-Love expansions, through exploiting the correlations in the prior flux fields in both space and time. Such dimensionality reduction will make it computationally feasible to employ a probabilistic Bayesian inference method. The parameterization of the flux fields will form the new state vector that is to be inferred. The solution of the inverse problem, in the form of full posterior distributions, is obtained via Markov Chain Monte Carlo (MCMC) sampling.

1.3 Bayesian Inference Theories: Beyond the Gaussian Assumption

A key scientific innovation of this project lies in the employment of a class of Bayesian probabilistic methods. It should be noted that the Gaussian least-squares approach commonly adopted in inverse problems is usually regarded as a Bayesian approach in the atmospheric science literature. The Gaussian least-squares approach is, however, a very special case of Bayesian inference among many others, as will be shown in the following. Before getting into that, we first go over the general concept of Bayesian inference.

Generally speaking, Bayesian methods are known to be well-suited for dealing with uncertainties arising from different sources: errors due to model assumptions, parametric uncertainties and experimental errors. They provide convenient means for capturing the state-of-knowledge about quantities of interest before and after the information provided by data. Furthermore, Bayesian techniques are very convenient for dealing with nuisance, secondary parameters, i.e., parameters that are generally unknown and are not of interest. Moreover, Bayesian methods can be generalized to hierarchical dependencies and allow an elegant approach to solving the problems model comparison and model selection. Bayesian methods allow for a formal comparison between various modeling choices. The key relationship for Bayesian inference is Bayes' formula

$$p(\mathbf{f}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{f})\, p(\mathbf{f}) \tag{1}$$

In Eq. (1) the prior probability $p(\mathbf{f})$ and the posterior probability $p(\mathbf{f}|\mathbf{d})$ represent degrees of belief about $\mathbf{f}$ before and after the data $\mathbf{d}$ are available, respectively. The key concept in Eq. (1) is the likelihood $L_{\mathbf{d}}(\mathbf{f}) = p(\mathbf{d}|\mathbf{f})$ that relates the data to the object of inference. The construction of a justifiable likelihood is the most critical step for obtaining the posterior probability distribution. The likelihood essentially measures the match between model-produced results and observed data.

In the current context, the object of inference ($\mathbf{f}$ in Eq. (1)) is the flux of CO_2 from earths surface ($\mathbf{f}$) and the data are in situ observations of CO_2 mixing ratios near the earth's surface and/or column average mixing ratios of CO_2 retrieved from satellites ($\mathbf{d}$). The observed data vector $\mathbf{d} = [d_1, \ldots, d_n]$ relates to the flux products via a linear function $\mathbf{d} = M\mathbf{f}$. Because of the imperfect knowledge of the linear transformation matrix, i.e., sensitivity coefficients, M, as well as measurement errors, there is an error term ε, such that

$$\mathbf{d} = M\mathbf{f} + \varepsilon \tag{2}$$

The likelihood construction casts the error term ε as a random vector, or drawn from a random field, and entails an assumption of its distribution. Assuming ε to be an uncorrelated Gaussian random vector is equivalent to the commonly adopted least-squares algorithm assumptions. Together

with the linearity of the relation in Eq. (1), this leads to analytical representations for either a deterministic least-squares solution that is the best estimate of **f**, or to a Gaussian posterior distribution for the values of **f**. Most commonly, the first, deterministic answer is satisfactory, and the Gaussian least-squares maximum a posteriori (MAP) solution is obtained analytically, if the following conditions are fulfilled, (1) the relationship between data d and flux f is linear, and (2) both the prior and the likelihood are Gaussian distributed [12]. While (1) is guaranteed for CO_2 flux inversion using atmospheric observations, there has been mounting evidence suggesting that CO_2 fluxes are non-Gaussian [39, 46, 47], which means that a more general and flexible approach is needed for the problem. To this end, there have been numerous efforts exploring data assimilations without explicit Gaussian assumptions on both likelihood/fit and prior/regularization terms [49, 50, 51], and their applications to atmospheric tracer inversion [52]. These approaches, however, seek only certain approximations to the true distributions, whereas the approach pursued here is much more robust and flexible, as it allows the use of the true distributions no matter what their shapes are.

The few applications of Bayesian probabilistic inference in atmospheric chemistry include quantification of the uncertainties in global mass balance of CH_4 [53], and more recently, inversion of cloud-aerosol interactions [54], and inverse modeling of atmospheric transport of bacteria emissions [55].

1.4 Markov Chain Monte Carlo

Without explicit Gaussian assumptions on priors and likelihoods, the posterior in Eq.(1) is not analytically tractable. Moreover, the numerical estimation of the posterior can be computationally prohibitive due to the high dimensionality of the object of inference, **f**. To this end, we will employ adaptive Markov Chain Monte Carlo (MCMC) sampling approach to efficiently sample from the posterior distribution of **f** arriving at a probabilistic representation of the flux field [56, 57]. The MCMC algorithms essentially march in the space of possible values of f and produce a sample set that is representative of the posterior distribution. These algorithms are flexible and enable posterior samples regardless of the degree of non-linearity in the model function. It has very wide applications for model parameter estimation, and also has been successfully applied for inferring parameters in the terrestrial biosphere models (TBM) [46], which are highly non-linear and complex. In addition, it also has been used to assess the uncertainties of carbon cycle climate modeling and their impact on climate projection [58].

1.5 Dimensionality Reduction

The general workflow to achieve a robust flux inversion is the following: we will use existing observations and expert opinion on fluxes to form a statistical representation of all possible flux values as a stochastic process, then the reduced-dimensional form of the latter will form the prior in the Bayesian MCMC context where the parameters of reduced form are being inferred. The most general description of the flux field (**f**) is given by its values at a grid of locations, i.e., having a complete representation of the discretized flux field. However, inferring such a large vector

(for instance, inferring fluxes at 1 degree x 1 degree resolution for 12 months gives about 105 unknowns) with MCMC can become computationally prohibitive, even if the model **M** is linear with respect to the object of inference f. One approach to dimensionality reduction is coarsening by region. Regardless, whether or not full or coarsened field descriptions, however, this approach does not properly take into account the information contained in the spatial and temporal correlations between fluxes. Utilizing such information, both in space and time dimensions, can drastically reduce the dimensionality and lead to a more flexible description of the flux field. Indeed, given a single realization, or time series, of the flux field $\mathbf{f}(x,t)$ over several periods (i.e. years), one can construct an effective set of realizations of the field for a duration of a single year. These realizations will provide a necessary sample set for time-correlation estimation. In space, on the other hand, one can either utilize flux field realizations available or build a synthetic correlation measures based on expert information or prior analysis. Next, having the covariance function $C(x,t),(x',t')$ available, one can solve the discretized version of the eigenvalue problem:

$$\int C((x,t),(x',t'))\phi(x',t')dx'dt' = \lambda\phi(x,t) \tag{3}$$

in order to obtain a set of eigenvalue-eigenfunction pairs $(\lambda_j, \phi_j(x,t))$ that serve as an optimal basis for the flux-field expansion [59, 60]

$$f(x,t) = f_0(x,t) + \sum_j \sqrt{\lambda_j}\xi_j\phi_j(x,t) \tag{4}$$

where $f_0(x,t)$ is the mean estimate of the field, and ξ_j are the uncertain representation coefficients with samples available that correspond to prior realizations. The expansion in Eq.(4) is called a Karhunen-Loève (KL) expansion [59, 60], and for sufficiently fast-decaying eigenvalue structure, it can provide a considerably lower dimensional representation. This is very similar to Empirical Orthogonal Function (EOF) method commonly used in the climatology community, which relies on Singular Value Decomposition (SVD) to obtain an expansion in terms of a finite set of eigenfunctions [61]. But the KL expansion will retain the uncertain representation of the coefficients ε_i and will use it as a prior distribution in the full Bayesian setting where the object of inference is the coefficients ξ_j of the KL expansion in Eq.(4). The KL expansion essentially provides a linear transformation of the basis and the likelihood in the Bayesian formulation with a reduced-order representation will capture the goodness-of-fit of model values $\sum_{j=1}^{K}\sqrt{\lambda_j}\xi_j\phi_j$(x,t) from observations $\mathbf{d}_i$, where K is the dimensionality truncation dictated by the eigenvalue structure, since the total variance in the flux field is the sum of the eigenvalues. For example, Zhuravlev *et al.* [62] successfully represent a spatiotemporal CO_2 flux field (2.5 x 2.5 degree spatial resolution, 144 (longitudes) x 72 (latitudes) x 96 (months)) with 30-40 eigenvalues within a 5 percent representation error. *To reiterate, in the new Bayesian probablistic inference method, we build a spatio-temporal covariance structure of the flux field based on expert opinion and prior realizations, followed by dimensionality reduction through a KL expansion. The uncertainties present in the KL expansion coefficients ξ_j will serve as prior distributions for the inference procedure that seeks to obtain posterior distributions on these coefficients in light of observational data.*

1.6 Attributing Fossil-fuel CO_2 using Tracers

The observed CO_2 concentration in the atmosphere, expressed either in terms of average mixing ratio in a column of dry air or in terms of dry air CO_2 mixing ratio at a sampling site near the ground, can be described as follows:

$$[CO_2]_{OBS} = [CO_2]_{BIO} + [CO_2]_{FF} + [CO_2]_{BB} + [CO_2]_{BG} \tag{5}$$

in which the brackets, [], imply dry mixing ratio, $[CO_2]_{OBS}$ is the observed CO_2 concentration, $[CO_2]_{BIO}$ is the biospheric component not including fire emissions, $[CO_2]_{BB}$ is the biomass burning contribution from natural and anthropogenic sources, and $[CO_2]_{BG}$ is CO_2 with relatively long history in the atmosphere sampled by the measurement. Eq. (3) demonstrates that each of the CO_2 components on the right-hand-side (RHS) contributes to the observed values, and this equation implies that inferring any of these components requires information about the contributions from each of the other components. For example, using atmospheric CO_2 measurements to infer fluxes of biosphere-atmosphere exchange has been studied extensively by the carbon-cycle community, and the first step in almost all these studies is to obtain the $[CO_2]_{OBS}$ component by estimating and subtracting $[CO_2]_{FF}$, $[CO_2]_{BG}$ and $[CO_2]_{BB}$ from $[CO_2]_{OBS}$, based on observations (for example those of clean background CO_2 concentration upwind) and/or model simulations. In practice,$[CO_2]_{FF}$, $[CO_2]_{BG}$ and $[CO_2]_{BB}$ are often regarded as perfectly known quantities to alleviate the technical difficulty of the problem (e.g., [33]). However, it has been increasingly recognized that the errors of those companion components can be large and introduce substantial uncertainties to the solution of the inverted biosphere fluxes [48, 11].

The estimated $[CO_2]_{BIO}$ from Eq. (5) will be linked to the biosphere-atmosphere CO_2 fluxes using the source-receptor relationship established by an atmospheric transport model, i.e. Eq.(2) In practice, $\mathbf{M}$ is likely to have large systematic biases [63], because of our imperfect knowledge and model descriptions of atmospheric transport [14], which limit our confidence in the inverted fluxes using atmospheric observations based on Eqs. (5) and (2). Likewise, the two challenges described above, i.e., (a) the lack of constraint for each individual component of CO_2 on the RHS of Eq.(5), and (b) the uncertainty due to transport model errors (errors in k) exist when using atmospheric CO_2 observations to infer fossil-fuel CO_2 emissions. In particular, estimating $[CO_2]_{FF}$ using Eq. (3) is challenged by the large and uncertain biosphere component $[CO_2]_{BIO}$ (e.g., [64]). Addressing these two challenges is the key to a successful MRV&V system. Using proxy tracers for combustion emissions provides a powerful approach.

Multiple authors have investigated the use of tracers, such as $^{14}CO_2$ (e.g., [65]), CO (e.g., [21, 66, 15]), SF_6 (e.g., [19]), C_2Cl_4 [22], and a combination of NOy, SO_2 and CO [20], to isolate and constrain concentrations and fluxes of fossil-fuel CO_2. The theoretical basis for using these tracers has been described in the literature [20, 31, 22, 15].

The premise of defining a species X_t as the tracer of the species of interest X_i is the assumption that the two species have collocated sources and both source strengths (denoted as X_i and X_t for convenience) are proportional to each other with a ratio of β, i.e.,

$$X_i = \beta X_t \tag{6}$$

One can prove that the measured sequences of dry mixing ratios, $[X_i]$ and $[X_t]$, at a given location will also be proportional to each other with a regression slope of the same β after the linear transport processes gone through by the two species in the atmosphere, i.e.,

$$[X_i] = \beta [X_t] \tag{7}$$

In practice, Eqs. (6) and (7) can be used in different ways, depending on what information is available and how reliable it is. As the key parameter in Eqs. (6) and Eq.(7), β can be derived from concurrent measurements of $[X_i]$ and $[X_t]$ by a linear regression based on Eq. (7) , or obtained from bottom-up emission inventories of X_i and X_t. In the case of inferring fossil-fuel CO_2 fluxes, (CO_2FF) using a tracer (X_{FF}), the tracer concentrations, $[X_{FF}]$, are often the measured quantity. In some studies, is obtained using Eq. (4) based on measured $[X_{FF}]$ and $[CO_2]_{FF}$ obtained from Eq. (1) using $[CO_2]_{OBS}$, $[CO_2]_{BG}$ and $[CO_2]_{BIO}$, $[CO_2]_{BB}$, which are measured directly or inferred based on additional assumptions [20, 19]. In the cases where $[CO_2]_{FF}$ can not be obtained independently, must be derived from bottom-up emission inventories of X_{FF} and fossil-fuel CO_2. The object to infer using the tracer can be either concentrations or emission fluxes of fossil-fuel CO_2 [22].

Although it is straightforward to link the tracer species and the species of interest using Eqs. (6-7), this approach introduces uncertainties related to tracer measurements and the various assumptions invoked in the process, especially the assumptions of collocated sources and the value of β. The key measure of efficiency of a tracer method is the magnitude of the uncertainty that needs to be smaller than the uncertainty associated with other comparable methods, with or without explicit usage of other tracers [22, 19, 66]. Therefore, one key objective when using tracers is to quantify the uncertainty in the inferred fossil-fuel CO_2 emissions using our the tracer approach compared to other tracer approaches and non-tracer approaches (e.g., [67, 18, 68]).

2 Toward Verifying Fossil Fuel CO_2 Emissions with the Community Multi-scale Air Quality (CMAQ) Model: Motivation, Model Description and Initial Simulation

2.1 Introduction

Atmospheric Dispersion Theories

Atmospheric dispersion refers to the mathematical description of the behaviors of chemical species released from various sources into the atmosphere. Although t is very common to use atmospheric 'diffusion' instead of atmospheric dispersion, it should be noted that dispersion in turbulence is fundamentally different from ordinary molecular diffusion. Depending on the coordinates used to study atmospheric flows, the equation for the mass balance of a chemical species i in the atmosphere can be written in two different ways as follows.

Eulerian approach

Let ρ_i be the concentration of chemical species i, which is expressed as number density (molecules cm^{-3}), or mass density (kg m^{-3}), it in general holds that

$$\frac{\partial \rho_i}{\partial t} = -\nabla \cdot (\rho_i \mathbf{v}) + D\nabla^2 \rho_i + s_i \tag{8}$$

where s_i is a local rate (molecules cm^{-3} s^{-1} or kg m^3 s^{-1}) of local processes such as chemistry, emissions, dry/wet deposition; $\mathbf{F_{adv}} = \rho_i \mathbf{v}$ is the advective flux driven by wind velocity $\mathbf{v}$; $\mathbf{F_{diff}} = -D\nabla\rho_i$ is the molecular diffusion flux; the divergence of mass fluxes $\nabla\mathbf{F} = \nabla(\mathbf{F_{adv}} + \mathbf{F_{diff}})$ measures what flows out of minus flows into the elemental volume of air. Substituting into Eq. (8) gives

$$\frac{\partial \rho_i}{\partial t} = -\nabla \cdot \mathbf{F} + s_i \tag{9}$$

An scale analysis reveals that $\mathbf{F_{adv}} \gg \mathbf{F_{diff}}$, i.e. molecular diffusion is negligible compared to advection for transport scales larger than $\sim$1 cm in the lower atmosphere including the troposphere and the stratosphere (lower than $\sim$100 km from the earth's surface). Therefore, the molecular diffusive flux $\mathbf{F_{diff}} = -D\nabla^2\rho_i$ is safely neglected in these circumstances, thus Eq.(9) becomes

$$\frac{\partial \rho_i}{\partial t} = -\nabla \cdot (\rho_i \mathbf{v}) + s_i \tag{10}$$

Using number density as the concentration unit, Eq.(8), Eq.(9) and Eq.(10) are *Eulerian flux form* continuity equations. Another unit that is often used in atmospheric chemistry to express chemical concentration is *mole fraction* or *mixing ratio* (μ_i, mol mol^{-1}), defined as the number of moles of the chemical per mole of air. ρ_i and μ_i is related by $\rho_i = \mu_i \rho_a$, $\rho_a = \rho_1 + \rho_1 + \rho_3 + ... + \rho_n$. By replacing ρ_i with μ_i and neglecting the local term s for the air itself in Eq.(10), one can obtain

its equivalent *advective form*:

$$\frac{\partial \mu_i}{\partial t} = -\mathbf{v} \cdot \nabla \mu_i + \frac{s_i}{\rho_a} \tag{11}$$

In order to solve the above continuity equations, information on the wind velocity $\mathbf{v}$ comes from an atmospheric weather model that solves the *Navier-Stokes equation*. It is still not practical to apply these models at a sufficiently small scale to deterministically resolve *turbulence*, the fine-scale ($\sim$1 mm spatially and $\sim$10 Hz temporally) variability of wind velocity. Instead, for meso-scale ($5-10^2$ km) atmospheric problems, $\mathbf{v}$ is commonly decomposed into a deterministic mean component $\langle \mathbf{v} \rangle$ and a stochastic turbulent component $\mathbf{v}'$:

$$\mathbf{v} = \langle \mathbf{v} \rangle + \mathbf{v}'; \langle \mathbf{v}' \rangle = \mathbf{0} \tag{12}$$

ρ_i can be decomposed in a similar way:

$$\rho_i = \langle \rho_i \rangle + \rho_i'; \langle \rho_i' \rangle = 0 \tag{13}$$

It turns out that introducing the random variables into the continuity equation will result in the well-known *closure problem*. Compromises and approximations have been made to make the problem solvable. For instance, the mean advective flux can be decomposed as follows:

$$\langle \mathbf{F_{adv}} \rangle = \langle \rho_i \mathbf{v} \rangle = (\langle \rho_i \rangle + \rho_i')(\langle \mathbf{v} \rangle + \mathbf{v}') = \langle \rho \rangle \langle \mathbf{v} \rangle + \langle \rho_i' \mathbf{v}' \rangle \tag{14}$$

where $\mathbf{F_M} = \langle n \rangle \langle \mathbf{v} \rangle$ is the *mean advective flux* and $\mathbf{F_T} = \langle \rho_i' \mathbf{v}' \rangle$ is the *turbulent flux*. Many current chemical transport models solve the continuity equation with parameterized *turbulent flux* based on the *mixing-length theory*, which states that the turbulent flux and the mean concentration can be related in a way that is analogous to molecular diffusion, by introducing the *eddy diffusivity K*:

$$\langle v_j' \rho_i' \rangle = -K_{jj} \frac{\partial \langle \rho_i \rangle}{\partial x_j}, j = 1,2,3 \tag{15}$$

Lagrangian approach

Transforming the *Eulerian advective form* continuity equation Eq.(11) by introducing the total derivative

$$\frac{d\mu_i}{dt} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \tag{16}$$

gives the the Lagrangian form of continuity equation:

$$\frac{d\mu_i}{dt} = \frac{s_i}{\rho_a} \tag{17}$$

It should be noted that in Eq.(17), s_i describes behaviors of a fluid element that moves in space and time, which is in contrast to the Eulerian framework where the element is fixed in space. Eq.(17) states that μ_i remains unchanged if the fluid element only goes through transport ($s_i(\mathbf{x}, \mathbf{t}) = 0$). Eq.(17) can be integrated along the trajectory of the fluid element:

$$\mu_i(\mathbf{r}_B, t_B) = \mu_i(\mathbf{r}_A, t_A) + \int_A^B \frac{s_i}{\rho_a} dt \tag{18}$$

which states that the mixing ratio of i at the ending location B is the sum of mixing ratio of i at the starting location A and the change of it along the trajectory from A to B due to local processes, such as chemistry. However, this is a over-simplified description of transport in a turbulent atmosphere that may not be very useful, because the trajectory is not likely to be known perfectly because of the highly variable wind velocity fields. As noted earlier, the turbulent component of wind velocity $\mathbf{v}'$ is a random variable, therefore it is natural to use a random process to describe the trajectory of a fluid element as well. The formal Lagrangian continuity equation turns out to serve this need very well. Here the object of interest is representative fluid particles, of which the location $\mathbf{x}$ is the dependent variable of concern and is a function of time t. We introduce Ψ, the probability density function (pdf) for a particle's location at time t:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Psi(\mathbf{x},t)d\mathbf{x}=1 \tag{19}$$

Two pdfs are introduced first:

1. The *transition probability density* $Q(\mathbf{x},t|\mathbf{x},t)$ that describes the likelihood of the particle getting to $\mathbf{x}$ at time t from $\mathbf{x}'$ at t';
2. The *initial probability density* $\Psi(\mathbf{x},t)$ that the starting point of the particle was indeed at $\mathbf{x}'$ at t'.

The probability of having the particle at $\mathbf{x}$ at t can be expressed by the product of these two pdfs:

$$\Psi(\mathbf{x},t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Q(\mathbf{x},t|\mathbf{x}',t')\Psi(\mathbf{x}',t')d\mathbf{x}' \tag{20}$$

So far the probability densities have been defined with respect to a single particle. In this Lagrangian approach, the mean concentration (in unit of mixing ratio, as seen in Eq.(17)) of a chemical species $\langle\mu_i(\mathbf{x},t)\rangle$ at a location $\mathbf{x}$ and time t is naturally quantified by counting the number of particles:

$$\langle\mu_i(\mathbf{x},t)\rangle=\sum_{i=1}^{m}\Psi_i(\mathbf{x},t) \tag{21}$$

The concentration at $(\mathbf{x},t)$ should consist of an initial concentration at $(\mathbf{x}_0,t_0)$ and the concentration changes during $t_0\rightarrow t$ (due to chemical decay, emissions input, and deposition to the Earth's surface). By expressing the pdf $\Psi(\mathbf{x},t)$ in terms of the initial distribution of $\mu_i(\mathbf{x}_0,t_0)$, and spatiotemporal distribution of sources $S(\mathbf{x}_0,t_0)$ (with the units of mixing ratio per time), and then substituting these expressions into Eq.(20), one obtains the *Lagrangian form* of the continuity equation:

$$\langle\mu_i(\mathbf{x},t)\rangle=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}Q_i(\mathbf{x},t|\mathbf{x_0},t)\mu_i(\mathbf{x}_0,t_0)d\mathbf{x_0}+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{t_0}^{t}Q_i(\mathbf{x},t|\mathbf{x}',t')S_i(\mathbf{x}',t')d\mathbf{x}'dt' \tag{22}$$

The key of evaluating Eq.(22) is to obtain the transition probability $Q(\mathbf{x},t|\mathbf{x},t)$. Although the knowledge of turbulence properties needed for defining $Q(\mathbf{x},t|\mathbf{x},t)$ is in general unavailable, it can

be approximated by invoking some simplified assumptions, such as the Gaussian distribution of the turbulent wind component $\mathbf{v}'$.

Eulerian VS. Lagrangian

The two approaches for modeling turbulent dispersion reviewed above have their own advantages and disadvantages. The *Eulerian* approach will face the *closure problem* and will introduce numerical diffusion, whereas the *Lagrangian* approach cannot deal with non-linear chemistry. Concerning the source attribution problem for greenhouse gases, chemistry is unimportant for those long-lived greenhouse gases such as CO_2 and CH_4. The key requirement for a chemical transport model in a source attribution system is to establish the source-receptor relationship efficiently and accurately.

The Role of CMAQ: A State-of-the-Art Eulerian Regional Chemical Transport Model

Global 3-D Eulerian CTMs have been playing a pivotal role in global carbon cycle research. Forward modeling analysis and diagnosis have been applied to understand global CO_2 distributions (e.g., [69]) and transport mechanisms (e.g., [70]). As a first step in CO_2 flux inversion, global CTMs have been routinely applied to calculate the concentration-to-flux response functions at numerous sampling sites around the globe (e.g., [13]). In recent years, there has been an emerging need for resolving finer-scale CO_2 transport and variability due to fossil fuel emissions from urban and point sources in order to refine regional carbon budgets and verify emission estimates. To this end, a regional CTM is more suitable with much higher spatial and temporal resolutions than most of current global models, which are often coarser than 1 degree.

Compared to global modeling, high-resolution CO_2 modeling on regional scales started relatively recently. Previous efforts have demonstrated the feasibility of regional CO_2 modeling and shown some promising achievements with a high-resolution regional CTM. For instance, [71] coupled WRF with a diagnostic biospheric model, i.e., the Vegetation Photosynthesis and Respiration Model (VPRM) and demonstrated the ability of the coupled model WRF-VPRM to capture the observed CO_2 features at a coastal site, especially sea breeze transport of CO_2 respired from vegetation during the previous night. Other regional modeling studies have identified various factors that can affect CO_2 spatiotemporal distributions, such as topography [72], diurnal variations [73] and spatial heterogeneity [74] of biospheric fluxes, covariance of transport and fluxes [75], etc. Some key requirements for regional CO_2 modeling have also been noted, such as using realistic initial and lateral boundary conditions [73], considering the long lifetime of CO_2 in the atmosphere. While these regional modeling studies mostly targeted natural areas where biospheric sources and sinks of CO_2 dominate, issues that are of interest for the emission verification problem, such as the magnitude and spatial extent of fossil fuel signals in atmospheric CO_2, were not addressed by these studies. Jacobson [76, 77] has performed high-resolution CO_2 modeling on global-through-urban nested domains to investigate the impact of local CO_2 domes on O3 and particulate matter (PM) pollution.

Lagrangian particle dispersion models (LPDM) have been widely used in CO_2 flux inference

(e.g., [16] and references therein) on regional to urban scales, because they can conveniently establish the source-receptor relationship needed for flux inversion. In principle, the two modeling approaches, i.e., Eulerian and Lagrangin modeling, could be used simultaneously and complement each other [78].

The Community Multiscale Air Quality (CMAQ) model is a widely-used regional CTM that was originally developed for atmospheric chemistry and air quality research [79]. Source attribution of air pollutants has been one of the main applications of CMAQ (e.g., [80, 81]). Capabilities for forward (i.e., Decoupled Direct Method, or DDM) [82, 83] and adjoint [84] sensitivity analysis have also been developed with CMAQ. Although there had been no previous effort to simulate CO_2 with CMAQ, the highly modulated model structure facilitates the addition of new chemical species and modification of their processes in CMAQ. Adding CO_2 into CMAQ while retaining other model species enables simultaneous simulations and examinations of CO_2 and a full suite of traditionally regulated air pollutants. Because there is abundant observational information and emissions-reduction experience for those air pollutants, it is of interest to explore their utility for facilitating CO_2 source attribution [20].

Goal of This Work

Atmospheric CO_2 has unique characteristics (e.g., long atmospheric lifetime, large background concentration, and strong bidirectional biospheric fluxes) that are distinctly different from other traditionally modeled chemical pollutants. Therefore, it is important to (1) characterize and understand the variability of CO_2 on fine spatial and temporal scales; (2) identify and quantify various model uncertainties associated with CO_2 fluxes, model transport, initial and boundary conditions.

This paper serves as a proof-of-concept for using CMAQ to achieve these goals. Because we only present model results for a single month, sweeping conclusions drawn from these results are not possible, considering the significant seasonal variations of CO_2. Instead, since this is the first time using CMAQ to simulate CO_2, we focus on introducing the methodology, including input data and model experiment design, and will try to interpret our results in the context of conventional understanding and findings from previous studies. The modeling framework here will form the foundation for more comprehensive investigations of CO_2 spatiotemporal variability and modeling uncertainties in future, which will be achieved by analyzing model stimulations for a longer time span using more observational data.

The following sections are organized as follows. In the method section, we first explicitly describe the input data used, highlighting the characteristics (e.g., magnitudes, spatial distributions, etc.) of different types of CO_2 fluxes. Then we describe the design of model sensitivity experiments and the observational data to which the modeling results are compared. In the results and discussion section, we present modeling results from an initial implementation over the contiguous U.S. domain in October 2007. We show the characteristics of spatial patterns of CO_2 near the surface simulated by the model, and perform model sensitivity experiments to understand the roles of meteorology, biosphere-atmosphere exchange, and fossil-fuel emissions in shaping the CO_2 spatial distribution. A comparison of CO_2 concentrations simulated by the model and observed at six tall-

tower sites within the NOAA Earth System Research Laboratory network follows, with focus on one site that is influenced by urban fossil-fuel emissions. Finally, the correlations between model-simulated CO_2 and traditionally regulated air pollutants (i.e., CO, NO_x, and SO_2) are examined and their implications for inverse modeling are discussed.

2.2 Methods

CMAQ Configuration and Input Data for CO_2 Simulation

For the present study, we used CMAQ Version 5.0 with meteorological inputs from WRF model (Version 3.1.1). The CMAQ model domain covers the contiguous U.S. and surrounding regions, and has 36-km spatial resolution and 22 vertical layers from the surface to 50 hPa. The base configurations of WRF and CMAQ are listed in Table A.1 in the supplemental information. CO_2 is added into CMAQ as an inert chemical species, of which the concentrations are determined by four types of CO_2 fluxes, including (1) bidirectional biosphere-atmosphere exchange, (2) bidirectional ocean-atmosphere exchange, (3) fossil fuel emissions, and (4) fire emissions, and atmospheric transport (horizontal and vertical advection and diffusion). The fossil fuel and fire emission fluxes are taken from existing emission inventories. The atmosphere-biosphere and atmosphere-ocean bidirectional fluxes are from terrestrial biosphere model outputs. These four types of fluxes are prescribed in the model in the same manner as the emission fluxes for existing chemical species in CMAQ. It should be noted that, in principle, the biosphere-atmosphere and ocean-atmosphere bidirectional fluxes can be simulated in an inline mode by two processes, i.e., emissions and dry deposition of CO_2, as done for ammonia (NH_3) and mercury (Hg) in CMAQ [85]. Such modeling capability with CMAQ for CO_2 will be developed in future. In the following, various input data that are used for CMAQ CO_2 simulations in this work are described.

Biosphere-Atmosphere Exchange

The bidirectional biosphere-atmosphere exchange of CO_2, Net Ecosystem Exchange (NEE), is the net flux between the biosphere and atmosphere due to CO_2 uptake during vegetation photosynthesis and CO_2 release during respirations, as well as CO_2 releases due to natural and anthropogenic disturbances, such as emissions contributed by fire or conversions in land use if any, which is specific to the terrestrial biosphere model (TBM) used (e.g., see detailed comparison of structural differences among TBMs in [86]). NEE simulated by current TBMs is still highly uncertain, which could be attributed to variations in inputs to the models (e.g., climate forcing and model parameters) and model structure (e.g., the models capability of capturing important processes such as CO_2 fertilization, nitrogen limitation, and disturbances). A major endeavor of carbon-cycle research has involved inter-comparison and evaluations of terrestrial-biosphere models (e.g., [87, 86, 88, 89]), and no model is obviously superior to others in all aspects. Efforts using atmospheric CO_2 observations to constrain NEE fluxes have increased in recent years. NOAAs CarbonTracker model [90] is an example of a data-assimilation system that provides optimized biosphere and ocean CO_2 fluxes using in situ CO_2 observations from a global observation network.

We used three different sets of NEE fluxes as input to CMAQ, including (1) NEE from the

Carnegie-Ames Stanford Approach (CASA), which are used by CarbonTracker-2011 (CT2011, http://carbontracker.noaa.gov) as priors ($1^\circ \times 1^\circ$, 3-hourly), and hereafter denoted by CASA NEE); (2) CT2011 optimized NEE ($1^\circ \times 1^\circ$, 3-hourly, hereafter denoted by CT2011 NEE), which represent an improved estimate based upon CASA NEE, as a result of imposing observational constraints; (3) NEE from the Community Land Model (Version 4) with surface and subsurface runoff parameterizations from the Variable Infiltration Capacity model (CLM4VIC, the baseline global simulation, i.e., BG1 simulation, [91]) ($0.5^\circ \times 0.5^\circ$, 3-hourly, hereafter denoted by CLM4VIC NEE) following protocols for the North American Carbon Program (NACP) Multi-Scale Synthesis and Terrestrial Model Intercomparison (MsTMIP) project (`http://nacp.ornl.gov/MsTMIP.shtml`; [86, 92]). More details of the CLM4VIC BG1 simulation are provided in the supplemental information. NEE from CT2011 and CASA do not include fire emissions. CLM4VIC NEE includes fire emissions but the contribution to NEE in October 2007 is small.

Figure 2 shows the monthly mean NEE fluxes from CT2011, CASA, and CLM4VIC over the model domain in October 2007. During this non-growing (Fall) season, the biosphere acts as a net source of CO_2 in terms of total fluxes over the model domain and the contiguous U.S., as consistently shown by the three sets of NEE (Table 1). A transition from a net source in the north to a net sink in the south can be seen in both CT2011 (Fig. 2a) and CASA (Fig. 2b). The spatial distribution of CLM4VIC NEE is distinctly different. Figure 2c shows a dipole structure in the central and southeastern U.S., and the total flux is only less than a half of the former two CASA-derived NEE fluxes. We note that such inter-model discrepancies of NEE observed here are not surprising, in comparison with previous biosphere-model inter-comparison studies (e.g., [87, 88]). Differences between the CLM4VIC and CASA models include, but are not limited to, land cover and land-use history, meteorological input data, and resolution of model biogeophysical and biogeochemical parameterizations and representations. For example, plant phenology is constrained by satellite observed Normalized Difference Vegetation Index (NDVI) in CASA, but is simulated prognostically in CLM4VIC. All these differences could contribute to the differences in spatial patterns and domain total fluxes. Fully understanding such model discrepancies would require a detailed comparison of the algorithms and input data used in CASA and CLM4VIC, which is outside the scope of this work. Instead, such inter-model differences of NEE from these three representative datasets can be employed as a rough estimate of uncertainties of NEE predicted by current TBMs. The differences of model-simulated atmospheric CO_2 concentrations, as a result of using these three different NEE inputs, will provide a rough estimate of the CO_2 uncertainty caused by uncertainty in NEE [93, 94]. We note that although only spatial distributions are shown in Fig. 2, uncertainty of temporal variability of NEE as an important factor has also been taken into account by using these three NEE inputs.

Fossil fuel, Fire and Ocean Fluxes

Two fossil-fuel emission inventories are used in this work. In the standard model configuration (Table 2), for model grids within the U.S., we use the Vulcan fossil-fuel emission inventory [9]. The Vulcan inventory is a well-documented high-resolution (10-km grid spacing, hourly temporal resolution), process driven, and fuel-specific fossil-fuel CO_2 emission inventory compiled for the U.S. In the Vulcan inventory, eight emission sectors are taken into account, i.e., aircraft, cement, commercial, industrial, nonroad, onroad, residential, and electricity production. Non-road, non-

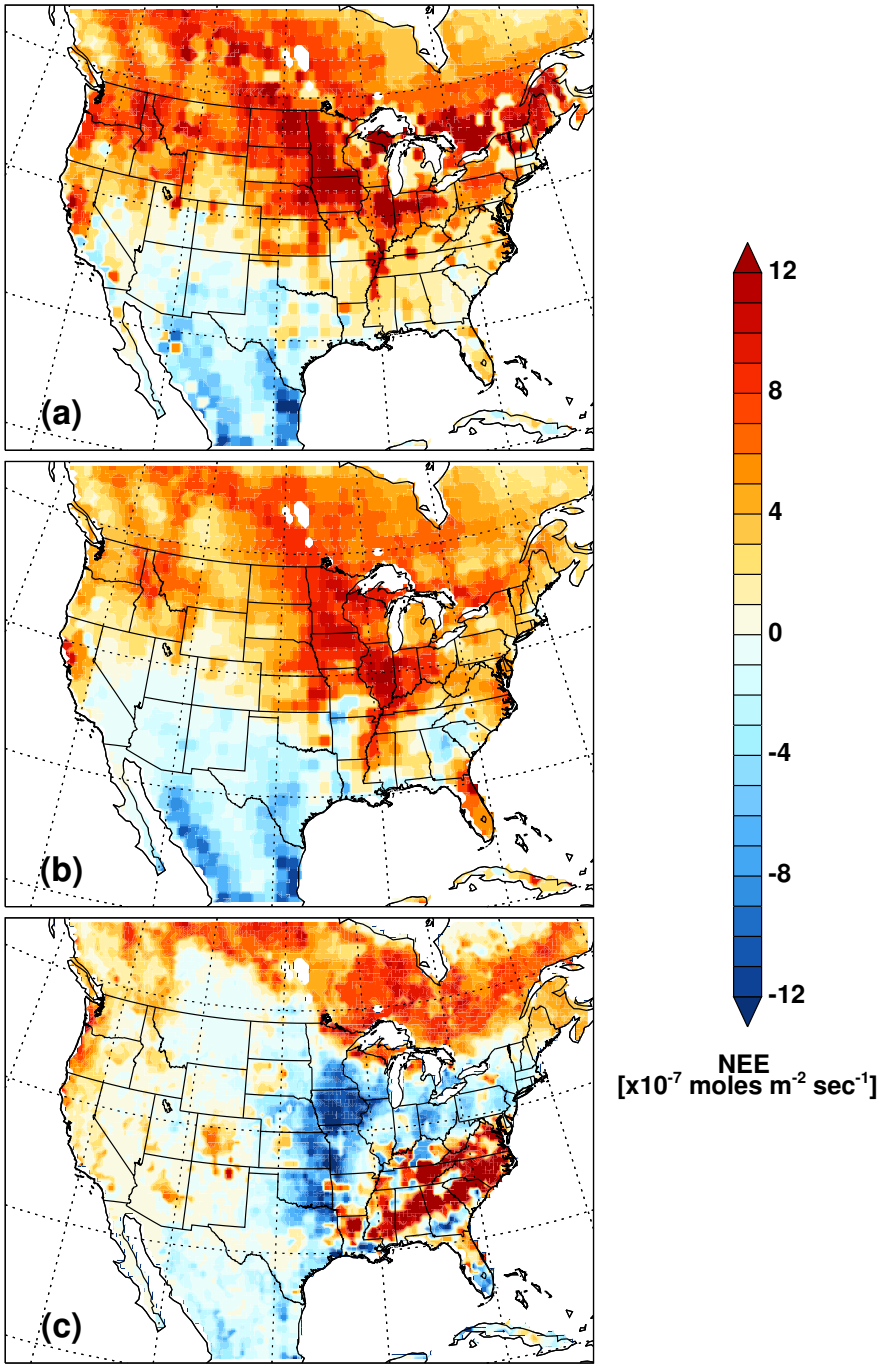


Figure 2. Monthly mean NEE for October 2007 from (a) CT2011, (b) CASA (CT2011 prior), and (c) CLM4VIC-BG1 in the model domain.

Table 1. CO_2 fluxes used in CMAQ simulations

Type of fluxes	Input data source	Total fluxes[a] (Tg Carbon/month)
Biosphere fluxes	(1) CT2011 optim. fluxes based on CASA ($1^\circ \times 1^\circ$; 3 hourly)	163.37 (96.79)
	(2) CASA ($1^\circ \times 1^\circ$; 3 hourly)	125.46 (72.22)
	(3) CLM4VIC ($0.5^\circ \times 0.5^\circ$; 3 hourly)	60.58 (6.74)
Fossil fuel fluxes	(1) VULCAN (2002; 10km; hourly, US only)	115.42 (115.42)
	(2) CDIAC (2007; ($1^\circ \times 1^\circ$; monthly)	132.42 (113.51)
Fire fluxes	GFED ($0.5^\circ \times 0.5^\circ$; monthly)	1.02 (0.79)
Ocean fluxes	CT2011 optimized fluxes	2.84 (0.00)

[a]Shown here are total fluxes over the whole CMAQ domain and within the contiguous U.S. (in parenthesis), respectively

point, point, and airport emission activity data are taken from the EPA National Emission Inventory (NEI) (for the year of 2002), which is a comprehensive inventory of all criteria air pollutants (CAPs) and hazardous air pollutants (HAPs) across the United States. Data from the EPA Emission tracking system/continuous emission monitoring systems (ETS/CEMs) are used for electricity production emissions. National Mobile Inventory Model (NMIM) County Database (NCD) data are used for deriving on-road CO_2 emissions. AERO2K data are used for aircraft emissions [95]. Data for Portland Cement are used for deriving CO_2 emissions from cement-production. More detailed information about the Vulcan inventory can be found by referring to Gurney *et.al.* [9] and the website of Vulcan project (`http://vulcan.project.asu.edu`;accessed January 14, 2013). Since the Vulcan inventory is compiled for the year of 2002 and there is a notable weekday/weekend effect in the data [16], we shifted the days in Vulcan such that the weekday/weekend patterns match the dates in 2007. For model grids outside the U.S., where Vulcan does not have values, we used an emissions inventory for 2007 from the Carbon Dioxide Information Analysis Center (CDIAC), a widely used global emission inventory ($1^\circ 1^\circ$, monthly) in global CO_2 modeling [96]. The Vulcan inventory is based on NEI-2002 and is used here to investigate the advantages of using an inventory with high spatial and temporal resolution. We note that, in principle, CO_2 emissions can be processed together with other pollutants using an emissions processor, e.g., the Sparse Matrix Operator Kernel Emissions (SMOKE), based on EPAs NEI, in future CO_2 regional modeling studies.

Figure 3 shows the monthly mean fossil fuel emissions from Vulcan and CDIAC inventories in the model domain. The differences between Vulcan and CDIAC reflect different spatial resolutions, spatial allocation methods, and reference years of the two inventories (2002 for Vulcan and 2007 for CDIAC). The two inventories, although compiled for different years, have similar (within 3%) total emissions in the contiguous U.S. (Table 1), reflecting the small inter-annual variability of national CO_2 emissions in the U.S. in the latest decade [8]. Another key difference between the two inventories is that Vulcan takes into account temporal variations while CDIAC does not. A comparison of Fig. 2 and Fig. 3 reveals that NEE shares characteristics with area sources, i.e., relatively smooth spatial variability and gradients, whereas fossil-fuel emissions are dominated by point sources and show large spatial heterogeneity and gradients. By comparing CO_2 simulations using these two emission inventories against observations, we examined the benefit of using a high-resolution emission inventory like Vulcan.

For fire emissions, GFED ($0.5^\circ \times 0.5^\circ$; monthly) inventory (as used in CT2011) is used. Fire emissions are highly variable in space and time, but are of minor importance compared to fossil-fuel emissions and to biosphere-atmosphere exchange for the simulation month (Table 1). However, they are likely to be more important during active fire months and regions. Fire emission is turned off when CLM4VIC NEE is used, as it already includes fire emissions. The effects of fire will be examined in future by using different fire emission algorithms such as GFEDv3.1 ($0.5^\circ \times 0.5^\circ$; 3 hourly; `http://www.globalfiredata.org/Data/index.html`, accessed January 14, 2013) and SmartFire.

We used CT2011 optimized estimates for ocean fluxes ($1^\circ \times 1^\circ$; 3 hourly), which are of minor importance compared to the biospheric and fossil fuel fluxes (Table 1).

CO_2 Net Fluxes

Table 2. List of model sensitivity experiments

No.	Notation	Meaning	Biosphere fluxes	Fossil fuel fluxes
1	CMVCT	Standard configuration	CT2011	Vulcan and CDIAC1[a]
2	CMCCT	Replace Vulcan with CDIAC	CT2011	CDIAC
3	CMVCS	Replacing CT2011 NEE with CASA	CASA	Vulcan and CDIAC1[a]
4	CMVLM	Replace CT2011 with CLM4VIC	CLM4VIC	Vulcan and CDIAC1[a]
5	CMBG	With no NEE or fossil-fuel fluxes within the U.S.	None	CDIAC2[b]
6	CMBIO	Without fossil-fuel emission within the U.S.	CT2011	CDIAC2[b]
7	CMFF	Without NEE within the U.S.	None	Vulcan and CDIAC1[a]

[a]Vulcan emissions are used for model grids within the U.S., and CDIAC emissions are used for model grids outside U.S.
[b]CDIAC emissions are used for model grids outside U.S. only

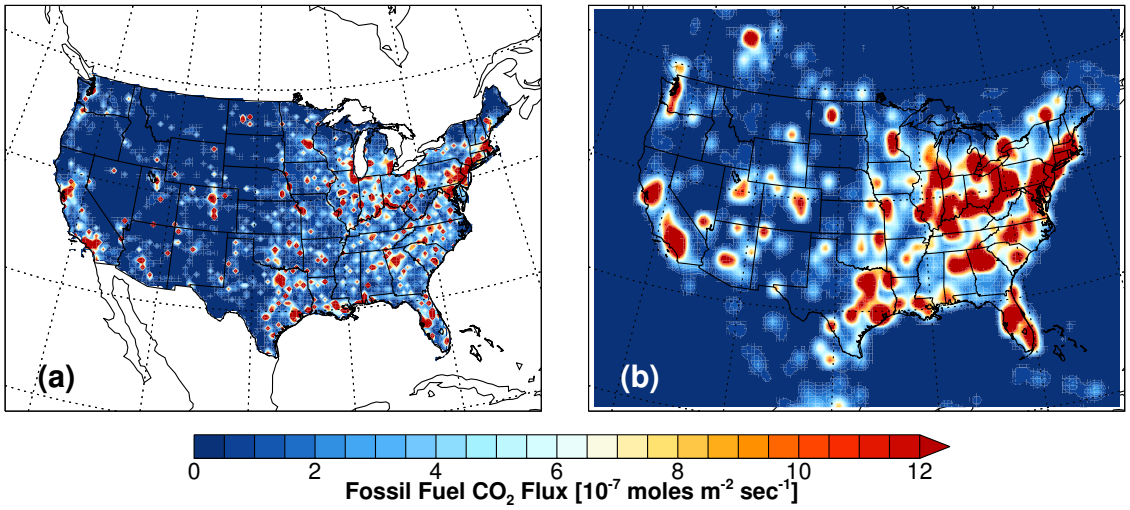


Figure 3. Monthly mean fossil-fuel CO_2 emissions for October 2007 from (a) Vulcan and (b) CDIAC in the model domain.

It is instructive to examine the net CO_2 fluxes as a result of the sum of all four types of fluxes. The spatial pattern of the CO_2 net flux shown in Fig. 4 retains features from NEE (regional pattern) and fossil-fuel emissions (scattered hotspots), due to their comparable flux magnitudes in this month (Figs. 2 and 3). A number of CO_2 emission hotspots shown in Vulcan (Fig. 3a) are still clearly seen in Fig. 4, such as the Los Angeles (LA) Basin and San Francisco Bay Area along the West Coast, the Houston-Galveston-Brazoria area along the Gulf Coast, and Chicago on the shore of Lake Michigan. Smaller and weaker emission sources in Fig. 3a, especially those in the eastern part of the country, do not appear as clearly with the presence of NEE. Neutral (zero) and negative net fluxes are seen in the southwestern U.S. and the Mexican mainland, with all three sets of NEE, as a result of negative NEE and small emission fluxes in these areas. When CLM4VIC NEE is used, negative net fluxes are also observed in central U.S.

The NEE inter-model differences (Fig. 2) lead to different spatial patterns of net fluxes in Fig. 4. Much smaller difference is found along the West Coast than in the eastern U.S. Figure 4c shows a large net CO_2 sink in the central U.S. and a strong net CO_2 source in the southeastern U.S. with CLM4VIC NEE, which are not seen with NEE from CASA or CT2011. In turn, the fossil-fuel emission signals in the southeastern U.S. reflected in the net fluxes with CT2011 (Fig. 4a) and CASA (Fig. 4b) NEE are not seen with CLM4VIC (Fig. 4c). As shown in later sections, such different net fluxes due to the differences in these three NEE inputs indeed lead to different spatial patterns of CO_2 near the surface simulated by CMAQ. The differences of net fluxes in Fig. 4 can facilitate understanding the interference by the uncertainty of NEE with the interpretation of CO_2 simulations and observations.

Initial and Boundary Conditions

Given its long atmospheric lifetime, CO_2 concentrations simulated by a regional model like

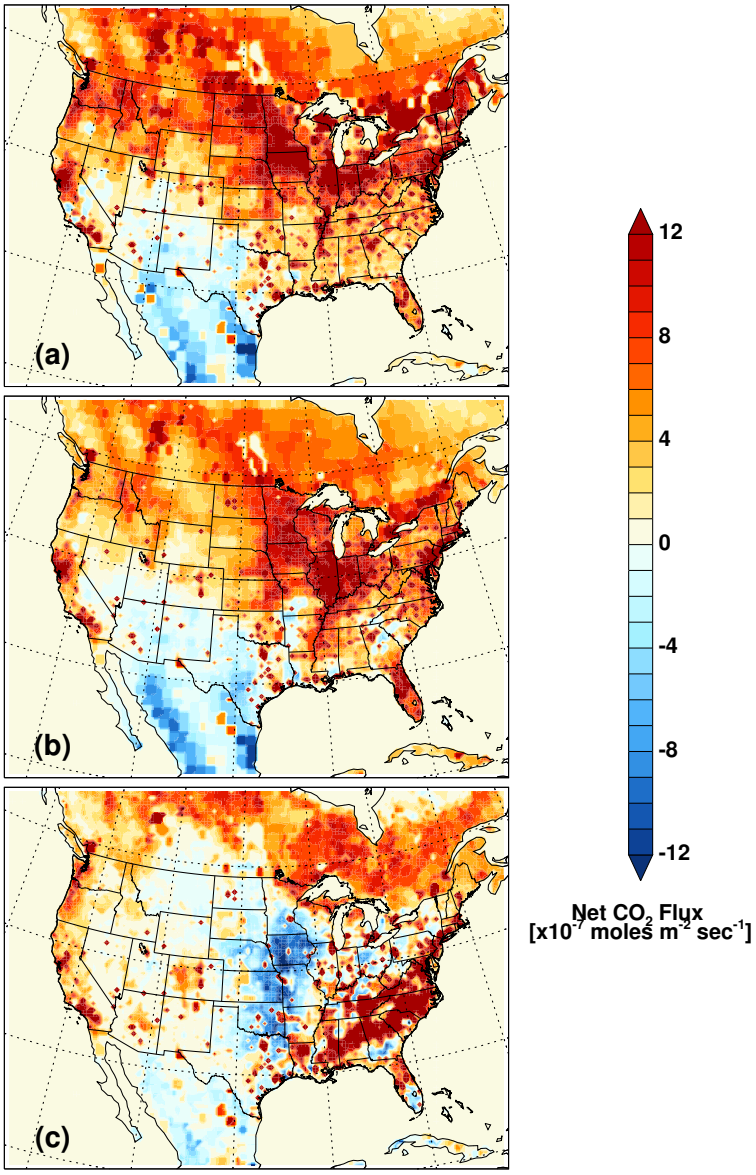


Figure 4. Monthly mean net CO_2 fluxes for October 2007 by adding all four types of fluxes used. Fossil-fuel emissions (Vulcan inside the U.S. and CDIAC outside), fire emissions (GFED), and ocean fluxes (CT2011) are the same for the three model configurations, and NEEs are from (a) CT2011 for CMVCT, (b) CASA for CMVCS, and (c) CLM4VIC-BG1 for CMVLM, respectively.

CMAQ are expected to be sensitive to model initial conditions (IC) and boundary conditions (BC), as has been shown by previous studies [73]. Four-dimensional concentration output from CT2011 using optimized NEE and ocean fluxes at $3^\circ \times 2^\circ$ and 3-hourly resolution were used as lateral and top BC in CMAQ. To minimize the impact of IC uncertainty, we spun up the model for 10 days. An experiment replacing 3-hourly BC with constant BC profiles shows large impact over the whole domain. The sensitivity of the simulation to IC/BC can be assessed in future work using the DDM-3D technique with CMAQ [83].

Model Experiments

We performed simultaneous simulations of CO_2 and a full suite of default chemical species in CMAQ for October 2007. Two model experiments were conducted. The first experiment was designed for comparing the roles of different sources/sinks of CO_2 in regulating CO_2 spatial distributions by decomposing CO_2 into three components, i.e., the background, biosphere, and fossil-fuel components. Specifically, we defined the region of interest to be the contiguous U.S. The background component was defined as CO_2 concentrations as a result of transport, wild fires, ocean fluxes over the whole domain, and fossil-fuel emissions outside the U.S. The biosphere (or fossil-fuel) component was defined to be CO_2 due to NEE (or fossil-fuel emissions) within the U.S. domain in this month. A second model experiment was performed to assess the impact of NEE uncertainty on simulated CO_2 concentrations by comparing model results using three different NEE inputs.

Configurations of the seven model runs for these experiments are tabulated in Table 2, including (1) a standard run (CMVCT) using CT2011 optimized NEE and GFED fire emissions over the entire domain, CDIAC fossil-fuel emissions outside the U.S. and Vulcan fossil-fuel emissions within the U.S.; (2) a CDIAC run (CMCCT) that differs from CMVCT by replacing Vulcan emissions with CDIAC emissions within the U.S.; (3) a CASA run (CMVCS) that differs from CMVCT by replacing CT2011 NEE with CASA NEE; (4) a CLM4VIC run that differs from CMVCT by replacing CT2011 NEE with CLM4VIC NEE; (5) a background run (CMBG) that differs from CMVCT by turning off NEE and fossil-fuel fluxes within the U.S.; (6) a biosphere run (CMBIO) that differs from CMVCT by turning off fossil-fuel emissions within the U.S.; and (7) a fossil-fuel run (CMFF) that differs from CMVCT by turning off NEE in the U.S. The biosphere component was obtained by subtracting concentrations in CMBG from CMBIO, and the fossil-fuel component was obtained by subtracting CMBG from CMFF.

CO_2 Observations from NOAA ESRL Tall Towers

As a component of the NOAA ESRL global sampling network, CO_2 has been continuously measured at a network of tall-tower sites across the contiguous U.S. [97]. These in situ near-surface CO_2 data have been extensively used for carbon cycle research, and are assimilated by CarbonTracker (Peters et al., 2007). The majority of the tall-tower sites are located in remote areas with insignificant influences from local fossil fuel emissions. However, such local fossil-fuel emissions

of CO_2 are of interest to the emission-verification problem. During October 2007, CO_2 data are available from six tall-tower sites in the model domain. Information (location, elevation and sampling altitude) of these sites is given in Table A.2. Five of the six sites are far from fossil emission sources, and thus the observed CO_2 variability is mostly driven by transport and biospheric fluxes. The Boulder Atmospheric Observatory (BAO) (40.05°N, 105.00°W, 300m above ground, 1584 m elevation) is a unique site that frequently receives local emissions from Denver, CO. For this reason, the data from BAO in this month were not assimilated by CT2011, to avoid an artificial scaling factor applied to a larger region due to the misrepresentation of local emission impact by the global model TM5. In this work, we compared model-simulated CO_2 with observations from the six tall tower sites to evaluate the general model performance. In particular, we elaborate on the comparison for BAO to understand the underlying factors driving the observed variability of CO_2 at such a site that is influenced by fossil fuel emissions from a city.

2.3 Results and Discussion

Spatial Distribution of CO_2 Near the Surface Over the Contiguous U.S.

From CarbonTracker to CMAQ

Figure 5 shows the monthly mean spatial distributions of CO_2 from CT2011 (Fig. 5a) and two CMAQ simulations, i.e., CMCCT in Fig. 5b and CMVCT in Fig. 5c. While CMCCT and CT2011 show similar large-scale patterns in general, large difference between the two (as large as 15 - 20 ppmv) exists along the western coast of the Mexican mainland. CMCCT uses CT2011 CO_2 outputs as initial and boundary conditions, and is driven by the same set of CO_2 fluxes as used by CT2011 (Table 2). Therefore, differences between the results from CMCCT (Fig. 5a) and CT2011 (Fig. 5b) can be attributed to the differences of model transport in CMAQ and CT2011, in the following key aspects: (1) assimilated meteorological fields (WRF for CMAQ/CMCCT versus ECMWF forecast for TM5/CT2011), (2) model resolution (36 km 36 km versus $1^\circ \times 1^\circ$) and (3) transport representations (CMAQ versus TM5). As expected, higher spatial resolution of CMCCT allows for resolving fine-scale features that are not seen in CT2011. Further, by replacing the CDIAC inventory in CMCCT with the Vulcan inventory, which has much higher spatial and temporal resolutions, CMVCT simulates numerous hotspots and stronger spatial heterogeneity of CO_2, while retaining the synoptic-scale spatial pattern in CMCCT. Overall, Fig. 5 demonstrates that compared to CT2011, the much-refined descriptions of transport and emissions in CMAQ allows for more detailed characterization of the spatial distribution of CO_2. A spatial map of CO_2 as shown in Fig. 5c can facilitate interpretation of sparse observational data in a regional context. In the next section, the roles of meteorology, biosphere, and fossil-fuel emissions in shaping the spatial pattern of CO_2 simulated by CMAQ are understood through a decomposition of these components, focusing on regions within the U.S.

Decomposition of background, biosphere and fossil-fuel components of CO_2

Figure 6 shows the background, biosphere, and fossil-fuel components of CO_2 simulated by CMAQ, using the methods described in Table 2. First, without biospheric and fossil fuel fluxes in-

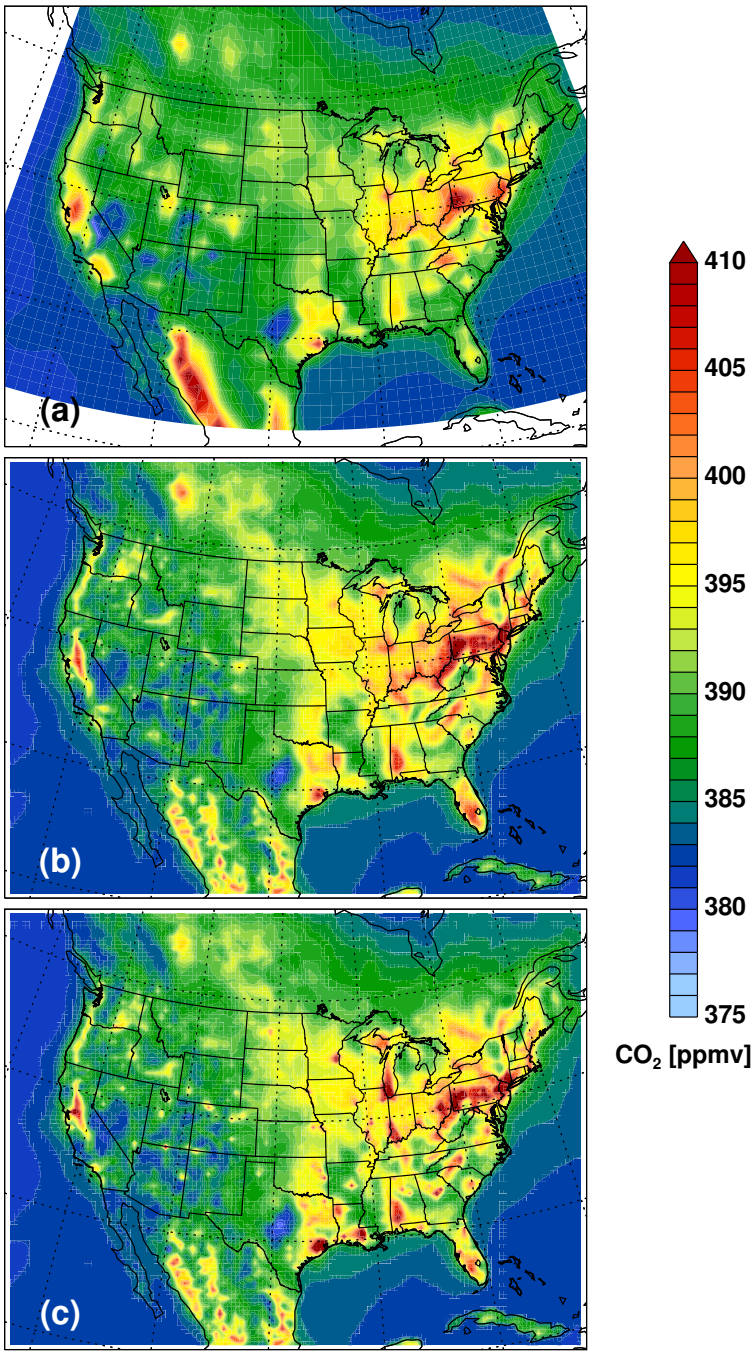


Figure 5. Monthly mean CO_2 concentrations near the surface in October 2007 simulated by (a) CT2011, (b) CMCCT using NEE from CT2011 and fossil-fuel emissions from CDIAC for the entire domain, and (c) CMVCT using NEE from CT2011, and fossil-fuel emissions from CDIAC and Vulcan for model grids outside and inside the U. S., respectively.

side the contiguous U.S., CMAQ simulates a relatively uniform background CO_2 of 380-383 ppmv over most places in the country in October 2007, with discernable gradients in the northeastern and western U.S. In Fig. 6b, the biosphere, as a net source (positive NEE) of CO_2 in this month, is responsible for prevalent CO_2 enhancement in the U.S. on top of the background in Fig. 6a. Weaker CO_2 enhancement is seen in the southwest due to lower NEE fluxes, and depletion of CO_2 (up to more than 4 ppmv) occurs in the central Texas and, to a lesser degree, in the LA Basin. In comparison, the fossil-fuel component in Fig. 6c exhibits a slightly different spatial pattern from the biosphere component. Numerous domes of CO_2 (> 16 ppmv) form near large emission sources (as shown in Fig. 3a). Dispersion of CO_2 from these domes and those scatter smaller emission sources creates a 2-4 ppmv of CO_2 superimposed on the background. Comparing Figs. 6b and 6c suggests that (1) in areas far away from large fossil-fuel emission sources, the biosphere component is similar to or even higher than the fossil fuel component, and (2) the biosphere component in the majority of cities cannot be regarded as negligible, with possible exception of some urban areas, e.g., the LA Basin in October 2007. The decomposition of biosphere and fossil-fuel components also facilitates the interpretation of CO_2 distribution shown in Fig. 5c. For example, Fig. 6b clearly shows that NEE is the main contributor to the high CO_2 in central Pennsylvania. It is important to note that, the biosphere CO_2 component is expected to vary significantly over different seasons, and thus its contribution to atmospheric CO_2 is expected to change with seasons as well.

Sensitivity of CO_2 Spatial Distribution to Uncertainty in NEE

All the results discussed above are from simulations using CT2011 NEE as input. In this section, we examine the impact of uncertainty of NEE on the spatial patterns of CO_2. Figure 7 compares monthly mean spatial distributions of CO_2 near the surface simulated by CMVCT, CMVCS, and CMVLM, which use the same fossil-fuel emission input but different NEE (Table 2). A comparison among Figs. 7a-7c reveals considerable model discrepancies as a result of differences in NEE inputs, consistent with a recent inverse modeling study using synthetic data at a few tower sites over the contiguous U.S. [93]. The regional mean concentration of CO_2 is lower in CMVLM (Fig. 7c) than in the other two models, consistent with its overall lower NEE (Table 1). CO_2 domes in the southeastern U.S. shown in CMVCT (Fig.7a) diminish in both CMVCS and CMVLM (Fig. 7b and Fig. 7c). In contrast, CO_2 concentrations along the West Coast do not vary much with different NEE inputs, implying less severe interference by NEE uncertainty. Such regionally varied situations imply that conclusions drawn from studies at one locale (e.g., in the LA Basin) need to be reexamined when looking at another locale (e.g., Washington D. C. or Atlanta). The uncertainty of NEE as reflected by inter-model differences has been found to depend on seasons (e.g. [93]). Therefore, the conclusions drawn here from one month of simulation cannot be seen as representative for all seasons. More comprehensive model comparisons for all seasons using NEE outputs from a larger group of TBMs are needed in future to better understand the issue of NEE uncertainty and biospheric interference.

Comparison with Tall-Tower Measurements

In this section, we examine the model results against observations at the six NOAA ESRL tall-tower sites listed in Table A.2, to evaluate and understand CMAQ-simulated temporal variations

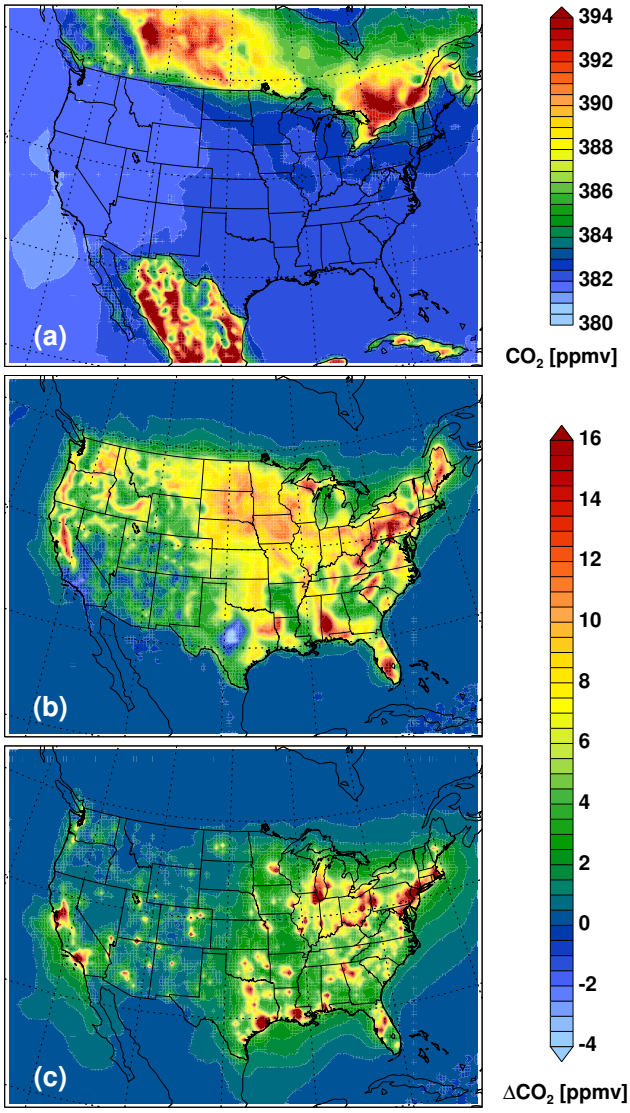


Figure 6. (a) Background, (b) biosphere, and (c) fossil-fuel components of CO_2 near the surface over the contiguous U.S. in October 2007 simulated by CMAQ. The background CO_2 component is simulated by the background run (CMBG) with fossil-fuel emissions and NEE fluxes turned off within the U.S.; The biosphere CO_2 component is calculated by subtracting CO_2 simulated by the background run (CMBG) from that by the biosphere run (CMBIO), for which fossil-fuel emissions are turned off within the U.S.; Fossil-fuel CO_2 component is calculated by subtracting CO_2 simulated by the background run CMBG from that by the fossil-fuel run (CMFF), in which NEE is turned off within the U.S.

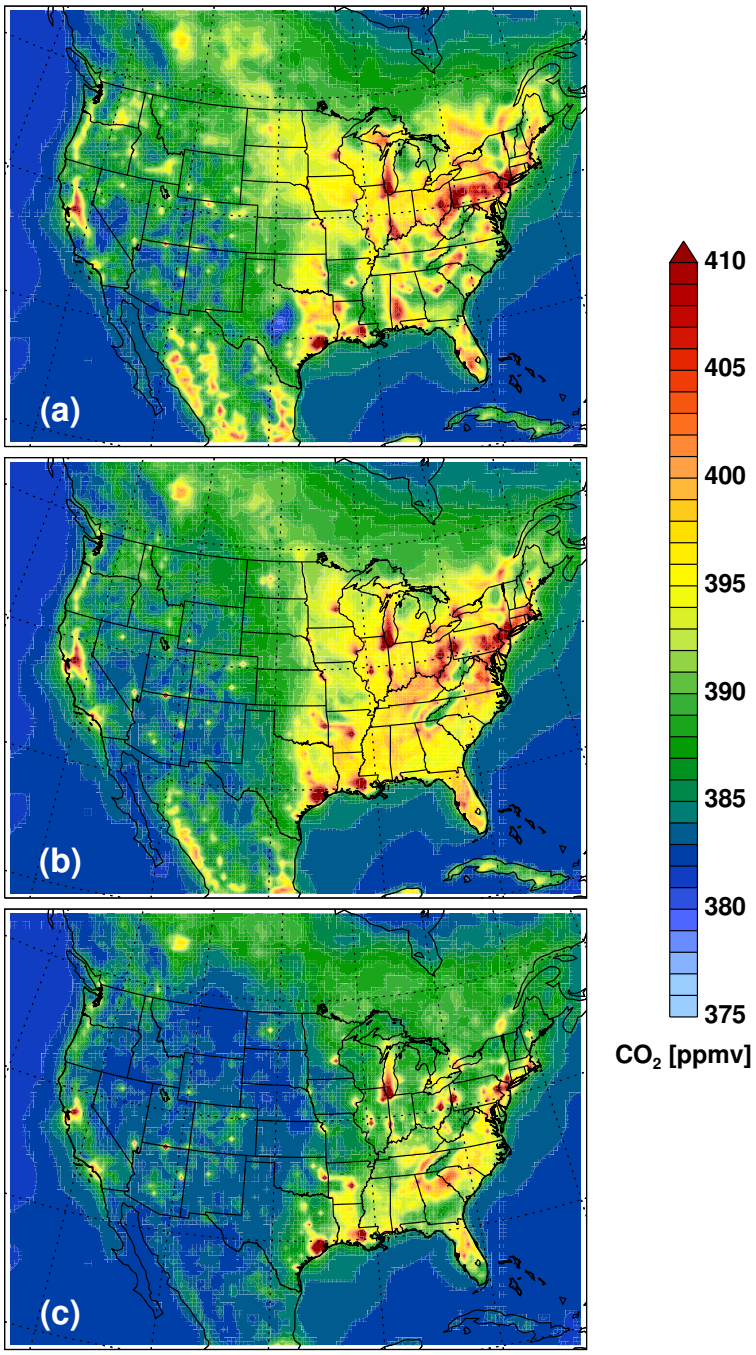


Figure 7. Monthly mean CO_2 concentrations near the surface simulated for October 2007 by (a) CMVCT that uses Vulcan fossil-fuel emissions and CT2011 NEE, (b) CMVCS that uses Vulcan fossil-fuel emissions and CASA NEE, and (c) CMVLM that uses Vulcan fossil-fuel emissions and CLM4VIC-BG1 NEE. For model grids outside the U.S., Vulcan has no values and CDIAC emissions are used instead.

of CO_2. Figures A.1 through A.6 show the CO_2 time series and mean diurnal profiles from CMAQ and CT2011 simulations, compared to their compartments from observations at the six sites. One needs to bear in mind that CT2011 has perturbed NEE fluxes to match CO_2 observations at five of the six sites (except for BAO). In general, CMAQ models using CASA-derived NEE fluxes (CMCCT, CMVCT, and CMVCS) show better performance than the model using VLM4VIC NEE (CMVLM). Compared to CT2011, CMAQ models using CASA-derived NEE can simulate better monthly mean concentrations (as suggested by the reduced mean biases) and resolve more high-frequency variability (as suggested by the closer-to-unity ratios of standard deviations) at most of the sites. However, these CMAQ models do not always show higher correlations with observations or lower RMSE than CT2011, suggesting that switching to new transport and fluxes at higher resolution also introduces more model-data mismatches. The mean diurnal profiles in Figs. A.1-A.6 show that almost all the CMAQ models have a low bias at night through early morning compared to observation. Possible reasons include but are not restricted to errors in model transport in the boundary layer and emission temporal profiles.

As mentioned earlier, BAO is unique and of higher interest compared to other five sites because (1) it receives fresh fossil fuel emissions from Denver, and (2) the observations were not assimilated by CT2011 in this month. It can be seen from Fig. A.1 that all the CMAQ models show improved performance than CT2011 in reproducing the 3-hourly observed CO_2 concentrations at BAO. CT2011-simulated CO_2 roughly tracks the observed background (lowest observed level) and shows negligible diurnal variability, possibly due to the smoothed local topography with the coarse grid of TM5 (the global model used for CT2011) and diluted emissions in CDIAC near BAO. CMAQ-simulated diurnal profiles are in general stronger than CT2011, but also vary with different NEE and/or fossil fuel emissions. The standard model (CMVCT) with CT2011 optimized NEE and Vulcan fossil-fuel emissions shows the best agreement with observation in terms of both 3-hourly and mean diurnal variability, but also has a low bias in early morning, as found for all other CMAQ models at all sites. CMCCT, which uses the same fluxes as CT2011 but higher-resolution meteorology, simulates a slightly stronger diurnal variability than CT2011. Using the hourly-varying Vulcan emissions (CMVCT, CMVCS, and CMVLM), which resolve the morning rush-hour emission peak, helps to capture the observed morning peak around 8:00 am. By switching to different NEE inputs, CMVCS and CMVLM simulate lower CO_2 concentrations in general than CMVCT. These results suggest that both time-varying emissions and biospheric fluxes are important drivers of the 3-hourly and diurnal variability of CO_2 at BAO. The importance of time-varying emissions was recently demonstrated by a recent modeling study, which shows that diurnal and weekly variations of emissions could result in up to 8 ppmv of perturbations of CO_2 near the surface [11].

It is very important to understand the causes for the model-data mismatch to guide subsequent inverse modeling. Since inverse modeling essentially seeks to match observations by perturbing selected model fluxes, an incorrect attribution of the model-data mismatch would directly lead to erroneous inversion results (e.g., [71]). For instance, attributing the early-morning low bias to errors in fossil fuel emissions, or NEE, or model transport would lead to drastically different conclusions (i.e., scaling up emissions in the first versus no scaling in the latter two). Indeed, our model results illustrate that emission verification is confounded by factors that affect CO_2 concentrations simultaneously with emissions, such as transport and biospheric fluxes. Alternatively,

fossil-fuel emissions could be isolated using certain tracer or proxy techniques.

Spatial Correlations between CO_2 and Traditionally Regulated Pollutants

Another question that can be conveniently addressed with CMAQ is the feasibility of using traditionally regulated air pollutants, such as CO, NO_x and SO_2, to provide constraints for fossil fuel CO_2 (e.g., [20]). For these traditionally regulated air pollutants, there are abundant long-term ground-based and satellite monitoring data that contain valuable information about historical trends and spatial patterns of emissions. Effective and economical strategies for emission verification are urgently needed to ensure the success of near-term emission reductions [16, 2]. It is thus worthwhile to consider approaches that (1) take full advantage of currently available observational networks and experiences in air pollutant emission monitoring and reduction, and (2) combine state-of-the-art atmospheric transport and emission modeling techniques. It has been shown recently that concurrent measurements of CO_2, CO, NO_y and SO_2 can be used to derive a top-down estimate of CO_2 emissions from a city (e.g. [20]). The emission trend of CO_2 over China was recently inferred from satellite NO_2 columns [98]. CO:CO_2 correlation slopes from aircraft observations during TRACE-P were used for understanding model-data mismatches and constraining emission fluxes [99].

Figure 9 compares the CMAQ-simulated monthly mean spatial patterns of CO_2, SO_2, NO_x and CO. Fossil-fuel combustion is the largest source for all four compounds, although emission factors for their common source sectors are different, and each of them has its unique sources and sinks. In terms of spatial distribution, CO_2 correlates better with NO_x (R=0.63) and CO (R=0.61) than with SO_2 (R=0.38). These correlations of concentrations are slightly better than the correlations of emissions of these compounds (R=0.43 for CO_2:NO_x, R=0.4 for CO_2:CO and R=0.24 for CO_2:SO_2), implying that similar emission sources and transport processes both contribute to the similar spatial patterns observed in Fig. 9. We note that using emissions of CO_2 and tracer species for different years and regridding the Vulcan inventory to the model resolution might have degraded the correlations between CO_2 and the three pollutants. A consistent processing procedure for emissions of CO_2 and its tracers is necessary for future studies. CO_2 hotspots (e.g., the CO_2 hotspot in central Pennsylvania and the broad high CO_2 region in the central and northern U. S.) arising primarily from biospheric fluxes can be readily identified with the assistance of the three tracers that are mainly emitted in urban areas (with the possible exception of Electricity Generating Units which can be located in rural areas). A quantitative understanding of such correlations and their utility to CO_2 emission inference needs to be attained by taking into account (1) emission factors and activities for each individual source (e.g., [21, 98]), and (2) model simulated transport ([100]) and model errors [66]. The tracer correlation problem has been studied extensively in the stratosphere (e.g., [101, 102]). Modeling and observations need to be combined to in future work to better understand the characteristics of tracer correlations in the troposphere. Correlations on different dimensions, e.g., a 1-dimensional (1-D) temporal correlation from a single ground site (e.g., [103]), a 2-D spatial correlation shown in Fig. 8, or a 4-D spatiotemporal correlation from aircraft measurements (e.g., [20, 21]) convey different physical meanings and should be examined and used carefully. Incorporating CO_2 into a Positive Matrix Factorization (PMF) analysis with multiple tracers (possibly including both gaseous species and PM2.5 components) could also be

considered to aid the tracking of CO_2 from different sources. Another possible direction is to explore a joint CO_2:tracer flux inversion [66] that makes use of the correlation of model errors between CO_2 and a tracer.

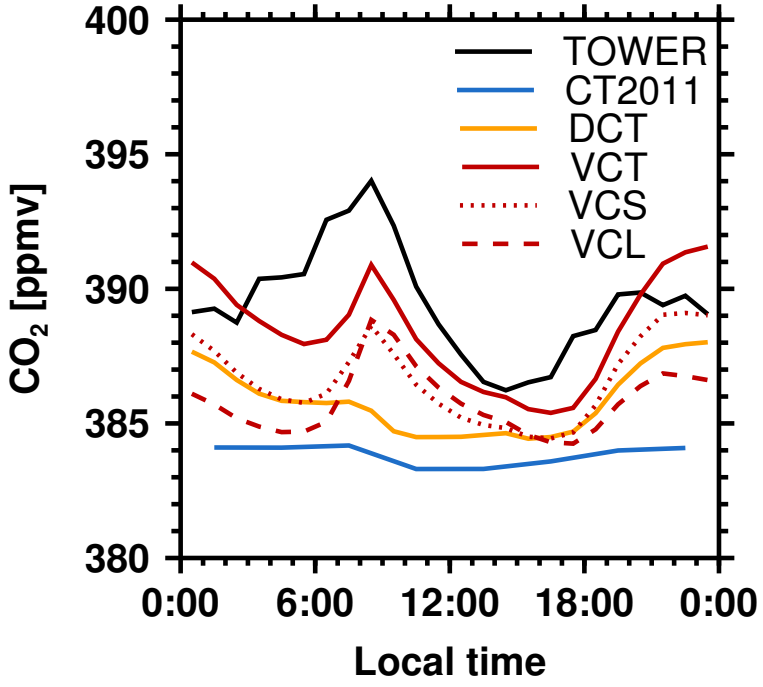


Figure 8. Monthly mean diurnal profiles of CO_2 in October 2007 observed at Boulder Atmospheric Observatory (BAO) (TOWER) and simulated by CT2011 and CMAQ with different configurations. CMCCT uses CDIAC fossil-fuel emissions and CT2011 NEE; CMVCT uses Vulcan fossil-fuel emissions and CT2011 NEE; CMVCS uses Vulcan fossil-fuel emissions and CASA NEE; and CMVLM uses Vulcan fossil-fuel emissions and CLM4VIC NEE. For model grids outside the U.S., Vulcan has no values and CDIAC is used instead.

2.4 Summary

In this work, we have described the motivation and methods for simulating CO_2 with CMAQ, and have presented initial modeling results for the contiguous U.S. in October 2007 to examine the potential of using CMAQ to characterize CO_2 spatial and temporal variability.

- By decomposing CMAQ-simulated CO_2 into background, biosphere, and fossil-fuel components, we have found that biospheric fluxes and fossil-fuel emissions are comparably im-

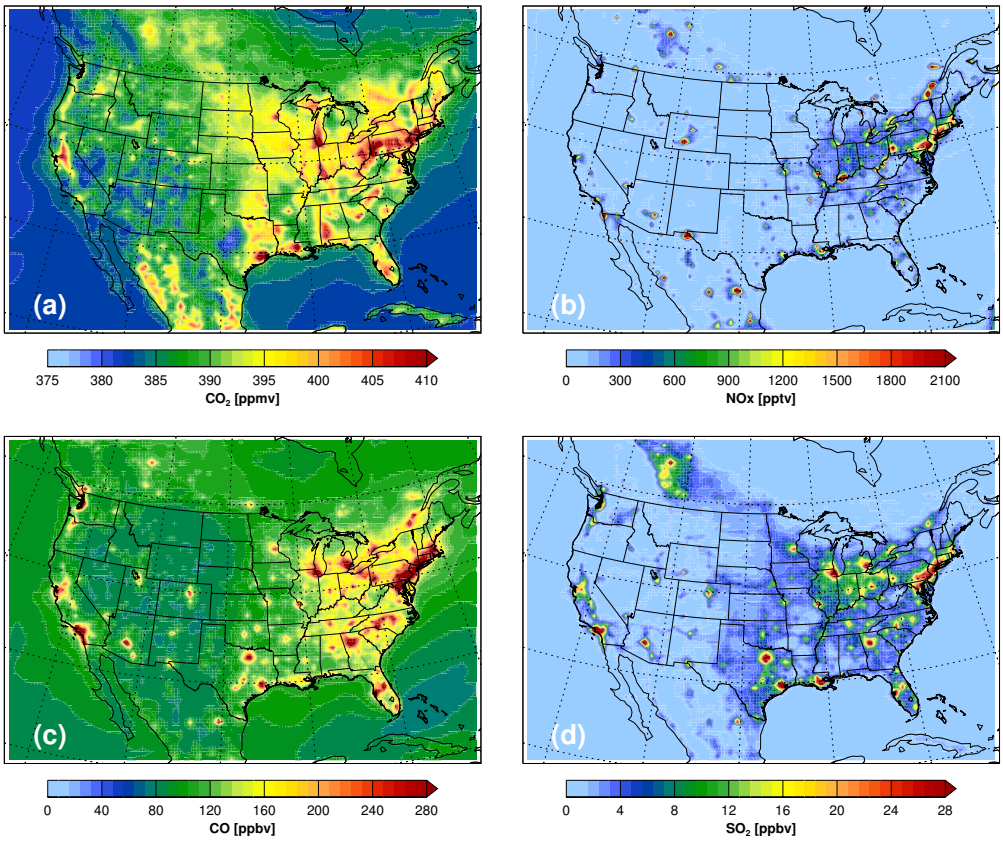


Figure 9. Monthly mean concentrations of (a) CO_2, (b) NO_x, (c) CO and (d) SO_2 near the surface simulated by CMAQ for October 2007. SO_2 is simulated by CMVCT, which uses Vulcan fossil-fuel emissions in the U.S. and CT2011 NEE.

portant in shaping spatial distributions CO_2 near the surface over the contiguous U.S. during October 2007, with each of them showing its unique characteristics.

- By using three different sets of NEE as inputs, we have shown that the uncertainty of NEE estimates has considerable impact on model-simulated atmospheric CO_2 concentrations near the surface, a finding that is consistent with previous studies (e.g., [93]). While only three sets of NEE inputs from two different TBMs are used here, more different TBM outputs from model inter-comparison projects, e.g., the ongoing MsTMIP project, will be used in future work to comprehensively address the issue of NEE uncertainty.

- By comparing the model results with observations from six tall-tower sites in the NOAA ESRL network, we have evaluated the model-simulated 3-hourly and diurnal temporal variability of CO_2. In particular, at BAO near Denver, CO, the model using the Vulcan emissions and CT2011 NEE shows the best performance in matching the observed mean diurnal profile, although with a low bias in the early morning. Using different NEE inputs would degrade the model-data agreement. More work is needed to better understand the model-data mismatch to inform subsequent inverse modeling.

- The model-simulated spatial pattern of CO_2 near the surface shows varying degrees of correlations with NO_x, CO and SO_2, as a result of their similar emission sources and common transport processes. Future work will explore the utility of these tracers for constraining fossil-fuel CO_2 emissions.

Findings from this work serve as a proof-of-concept and suggest that a regional CTM such as CMAQ has the potential to facilitate interpretation of CO_2 observations and emission verification. Future work will improve CMAQ CO_2 simulations in the following aspects: (1) increasing the model spatial resolution to better resolve urban and point sources; (2) processing gridded CO_2 emissions using SMOKE; (3) developing inline simulation of bidirectional biospheric fluxes; and (4) comprehensively evaluating the model performance using observations from ground networks, aircrafts, and satellites for all seasons of a year. Other greenhouse gases can also be studied using CMAQ in a similar manner to that shown here.

3 Bias-Enhanced Bayesian Inference of Atmospheric Trace Gas Sources and Sinks

3.1 Introduction

Inverse modeling is a formal approach to derive 'top-down' estimates of trace gas fluxes at Earth's surface using atmospheric concentration measurements. The main input-output relationship is the chemical transport model used to establish the relationship between measured atmospheric concentrations and strengths of sources and sinks at the surface. For the case of nonreactive constituents such as CO_2, which will be the focus of this study, the relationship is linear by virtue of the nature of atmospheric transport. The inverse problem is challenged by potential non-Gaussian structure of the measurement errors, as well as by the errors due to simplifying assumptions in the transport models, i.e. the structural error. Both data and model errors, if not properly characterized or quantified, will bias the inferred strengths of the sources and sinks away from their true values. The inverse problem may also be severely ill-conditioned because of the lack of data compared to a a potentially high number of sources/sinks.

In the literature, the CO_2 flux inversion problem has been commonly formulated to minimize the following objective function based on the least-squares criterion[104]:

$$J = \frac{1}{2}(\mathbf{d}-\mathbf{Rs})^T\mathbf{C_d}^{-1}(\mathbf{d}-\mathbf{Rs}) + \frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^T\mathbf{C_s}^{-1}(\mathbf{s}-\mathbf{s}_0) \tag{23}$$

in which $\mathbf{s}$ is the vector of fluxes to be estimated, $\mathbf{s}_0$ is a vector of *prior* estimate of fluxes that are usually output from a terrestrial biosphere model (TBM), $\mathbf{r}$ is the matrix of linear concentration-flux response functions established by a deterministic transport model, $\mathbf{d}$ is the vector of observed concentrations, $\mathbf{C_d}$ and $\mathbf{C}_s$ are error covariance matrices for the model-data misfit and the prior fluxes, respectively. Eq. (23) is appealing because an analytical solution of the minimization problem

$$\mathbf{s} = \arg\min_{\mathbf{s}} J \tag{24}$$

exists, allowing for efficient computation even when there are large number of fluxes to be determined, i.e. for high-dimensional problems. Various CO_2 flux inversion schemes have been developed based on the objective function (23). For instance, Bayesian synthesis inversion [105] and Ensemble Kalman Filter (EnKF) [106] use the analytical solution of Eq (23); the 4D-Var method finds the minimum of J iteratively [107]; geostatistical inversion minimizes a modified objective function slightly different from Eq. (23) by not explicitly imposing the prior flux constraint $\mathbf{s}_0$ [108].

From a statistical inversion point of view, the least-squares criterion is well known to be tied with the assumption that both model-data misfit and the prior fluxes are Gaussian [109]. As commonly seen in other applications, the least-squares criterion is adopted to solve the source inversion problem mainly for its convenience, whereas the justification of the underlying Gaussian assumption has not been well investigated [110]. Among the very few existing studies, [111] showed

evidence of non-Gaussian prior flux error distribution. Nor does information exist for the probability distribution functions (PDFs) of the model-data mismatch, the lack of robustness of the least-squares criterion for being sensitive to outliers [109] actually makes it not the best justified choice of statistical estimator to describe model-data mismatch in CO_2 flux inversion [105]. There is a clear need for a robust statistical mechanism of inversion that is independent of unjustified Gaussian assumptions. The second major motivation for this study is to build a statistical inversion framework that appropriately treats the *model errors*, i.e. the errors associated with the linearized transport model structural deficiencies. Indeed, the estimated fluxes may be strongly biased because of the fact that the model is imperfect. Most, if not all, studies of transport model inversion have inherently assumed that the model itself is a perfect representation of reality. This assumption may lead to biased estimates of the fluxes that try to compensate for the model deficiencies. Moreover, with increasing volume of observations the biased estimates will have smaller uncertainties around the wrong values. We will develop a strategy of inversion that handles the uncertainties associated with such errors. Having said that, the role of transport model uncertainty has been investigated by many studies in the literature, however most of the techniques that deal with model errors stem from ad-hoc assumptions and problem-specific adjustments. As a representative example, the TransCom3 Experiment studied the sensitivity of Bayesian synthesis inversions to transport process differences using 16 different transport models and model variants under the same input data and protocol [112, 113] with iterative, ad-hoc adjustments to the assumed data covariance structure to ensure different models lead to reasonably similar outcomes.

In this work, we revisit the TransCom3 Experiment using Bayesian inference to assess the impact of statistical assumptions on inversion solutions of CO_2 sources and sinks.

3.2 TransCom3 Inversion Formalism

The Level 1 control inversion in the TransCom3 Experiment seeks to solve for 5-yr mean aggregated biosphere and ocean fluxes over $M = 22$ regions using 5-yr mean observations at $N = 77$ sites around the world.

Following the notations used by [114] and [113], let $D(\mathbf{x_i})$ be the steady state concentration of CO_2 observed at the ith site $\mathbf{x_i} = (x_i, y_i, z_i)$. A decomposition of $D(\mathbf{x_i})$ based on mass balance, assuming no measurement errors for simplicity, gives

$$D(\mathbf{x_i}) = D_{FF}(\mathbf{x_i}) + D_{BB}(\mathbf{x_i}) + \sum_{j=1}^{M} D_j(\mathbf{x_i}), \tag{25}$$

in which $D_{FF}(\mathbf{x_i})$ and $D_{BB}(\mathbf{x_i})$ are the fraction of CO_2 concentration resulting from fossil-fuel CO_2 emissions and annually balanced net ecosystem exchange (NEE) fluxes, respectively; $D_j(\mathbf{x_i})$, the so-called *response function*, represents the concentration due to the net flux from the jth region and is obtained from

$$D_j(\mathbf{x_i}) = s_j r_j(\mathbf{x_i}) \tag{26}$$

wherein $r_j(\mathbf{x_i})$ is the *basis function* that represents CO_2 concentration at the ith site, $\mathbf{x_i}$, caused by unit CO_2 flux from the jth region, s_j.

The basis functions evaluated at the locations of interest form the response matrix $R_{ij} = r_j(\mathbf{x}_i)$, and Eq. (25) can now be written as

$$\mathbf{d} = \mathbf{R}\mathbf{s}, \tag{27}$$

where the 'effective' observations $\mathbf{d}$ are obtained by removing the 'background' CO_2 concentrations $\mathbf{d}_i = D(\mathbf{x_i}) - D_{FF}(\mathbf{x_i}) - D_{BB}(\mathbf{x_i})$. In the TransCom3 Experiment, the background quantity $D_{FF}(\mathbf{x_i}) + D_{BB}(\mathbf{x_i})$ is pre-calculated using transport simulations and is considered a known quantity.

The inversion task - and the main focus of our studies - now becomes finding $\mathbf{s}$ given a measurement vector $\mathbf{d}$. This is a *linear* problem as the right-hand-side is linear with respect to the object of inference $\mathbf{s}$. While there are various methods to solve for $\mathbf{s}$, e.g., see [114], we will argue for Bayesian techniques and employ Bayesian inference framework that is well suited for handling problems with various sources of errors, and is able to seamlessly incorporate prior information with the available measurement data [115].

3.3 Bayesian Inference

Bayesian machinery relies on the Bayes' formula, which in this context reads as

$$\underbrace{p(\mathbf{s}|\mathbf{d})}_{posterior} \propto \underbrace{p(\mathbf{d}|\mathbf{s})}_{likelihood}\ \underbrace{p(\mathbf{s})}_{prior} \tag{28}$$

Here the set of all measurements $\mathbf{d} = (d_1, \ldots, d_M)$ is considered data, and the fluxes for all regions, $\mathbf{s} = (s_1, \ldots, s_N)$ are the object of inference in Bayesian formulation.

Bayes' formula (28) relates the prior probability distribution $p(\mathbf{s})$ of the fluxes to the posterior one, in light of data, using the likelihood function

$$L(\mathbf{s}) = p(\mathbf{d}|\mathbf{s}), \tag{29}$$

which is a measure of the goodness-of-fit of the observed data to the model predictions stemming from the value $\mathbf{s}$.

While generally the exact computation of the posterior distribution is challenging for high-dimensional (i.e. for large N) problems, one often relies on Markov chain Monte Carlo algorithms to sample from the posterior distribution. With both prior and likelihood functions in place, we then employ adaptive MCMC (AMCMC) [57] algorithm in order to sample values of $\mathbf{s}$ according to the posterior distribution $p(\mathbf{s}|\mathbf{d})$. Three commonly used summaries of the posterior distribution are the mean $\mathbf{s}_{\text{mean}}$, standard deviation $\mathbf{s}_{\text{std}}$ and the maximum a posteriori (MAP) value $\mathbf{s}_{\text{MAP}} = \arg\max_{\mathbf{s}} p(\mathbf{s}|\mathbf{d})$. Often, when only the latter is of interest, one can proceed to posterior maximization via standard optimization methods, without invoking MCMC.

The likelihood function construction is the key step in Bayesian methods. It is intended to incorporate probabilistic representations of all sources of discrepancies between observed data and

the proposed model. The simplest, and most typical formulation relies on explicitly assuming statistical structure for the discrepancy vector

$$\boldsymbol{\varepsilon}_{\mathbf{d}} = \mathbf{d} - \mathbf{R}\mathbf{s}. \tag{30}$$

The most common scenario is to assume $\boldsymbol{\varepsilon}_{\mathbf{d}}$ is a multivariate normal random variable with vanishing mean, i.e. no measurement bias, and data covariance matrix $\mathbf{C}_{\mathbf{d}}$. This almost necessarily implies that model deficiencies are not captured and the only source of discrepancy is the measurement error. In fact, often the data covariance matrix $\mathbf{C}_{\mathbf{d}}$ is taken to be a diagonal one as the measurements are expected to have uncorrelated errors.

Furthermore, the Gaussian assumption is also typical for the prior distribution of the fluxes, since studies often report a mean value and a standard deviation for each flux. Note that, the so-called uninformative prior that is uniform over a wide range of possibilities can be thought of as a Gaussian in the limit of large variance. Such Gaussian assumptions for both the likelihood and the prior lead to analytical formulae for the posterior distribution and help avoiding the use of potentially expensive sampling-based methods, such as MCMC.

3.4 Analytical Solution for Gaussian case

Consider Gaussian *prior* $p(\mathbf{s})$ with mean $\mathbf{s}_0$ and covariance $\mathbf{C}_{\mathbf{s}}$ as well as a Gaussian additive data error term with zero mean and covariance $\mathbf{C}_{\mathbf{d}}$, leading to prior and likelihood formulae,

$$\begin{aligned} p(\mathbf{s}) &= ((2\pi)^n |\mathbf{C}_{\mathbf{s}}|)^{-1/2} \exp[-\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^T \mathbf{C}_s^{-1}(\mathbf{s}-\mathbf{s}_0)] \\ L(\mathbf{s}) = p(\mathbf{d}|\mathbf{s}) &= ((2\pi)^n |\mathbf{C}_{\mathbf{d}}|)^{-1/2} \exp[-\frac{1}{2}(\mathbf{d}-\mathbf{R}\mathbf{s})^T \mathbf{C}_{\mathbf{d}}^{-1}(\mathbf{d}-\mathbf{R}\mathbf{s})] \end{aligned} \tag{31}$$

With these Gaussian assumptions, the posterior $p(\mathbf{s}|\mathbf{d})$ can be proven to be Gaussian, too

$$p(\mathbf{s}|\mathbf{d}) = ((2\pi)^n |\mathbf{C}_p|)^{-1/2} \exp[-\frac{1}{2}(\mathbf{s}-\mathbf{s}_{\mathrm{MAP}})^T \mathbf{C}_p^{-1}(\mathbf{s}-\mathbf{s}_{\mathrm{MAP}})] \tag{32}$$

with mean $\mathbf{s}_{\mathrm{MAP}}$ and covariance $\mathbf{C}_p$ given by

$$\mathbf{s}_{\mathrm{MAP}} = \mathbf{s}_0 + \mathbf{C}_p \mathbf{R}^T \mathbf{C}_{\mathbf{d}}^{-1}(\mathbf{d} - \mathbf{R}\mathbf{s}_0) \tag{33}$$

and

$$\mathbf{C}_p = (\mathbf{R}^T \mathbf{C}_{\mathbf{d}}^{-1} \mathbf{R} + \mathbf{C}_s^{-1})^{-1} \tag{34}$$

Since the posterior is Gaussian, the mean and the maximum a posteriori (MAP) estimates coincide. Furthermore, the MAP estimate $\mathbf{s}_{\mathrm{MAP}}$ is the classical least-squares solution with cost function (23), since the latter is an additive constant away from the negative logarithm of the posterior.

3.5 Implications of Gaussian Data Error Assumption

The key assumption in the results described above is that of Gaussianity of the discrepancy between observational data and model's response. While this allows analytical posterior calculations, we will challenge this assumption below by producing synthetic data with a different noise characteristics. Specifically, let us produce synthetic data with a true response matrix $\mathbf{R}_{\text{true}}$ and true flux $\mathbf{s}_{\text{true}}$ as

$$\mathbf{d} = \mathbf{R}_{\text{true}}\mathbf{s}_{\text{true}} + \underbrace{\boldsymbol{\varepsilon}_{\mathbf{d}} + \mathbf{e}_{\mathbf{d}}}_{\text{total obs. error } \boldsymbol{\eta}_{\mathbf{d}}} \tag{35}$$

where $\boldsymbol{\varepsilon}_{\mathbf{d}}$ obeys the same i.i.d. Gaussian character with vanishing mean and standard deviation $\sigma = 0.1$, while there is another observational error term $\mathbf{e}_{\mathbf{d}}$ that we will assume follows an i.i.d. exponential distribution with mean $\beta = 2$. The total error $\boldsymbol{\eta}_{\mathbf{d}}$ now follows the so-called Exponentially-modified Gaussian (EMG) distribution with a probability density function $f(x;\mu,\sigma,\beta) = \frac{1}{2\beta}\exp\left[\frac{2\mu+\sigma^2/\beta-2x}{2\beta}\right]\text{erfc}\left[\frac{\mu+\sigma^2/\beta-x}{\sqrt{2}\sigma}\right]$, where $\text{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^{+\infty} e^{-t^2}dt$ is the complimentary error function. The probability density functions of both the Gaussian only and the EMG error term are illustrated in Figure 10.

Now we compare two scenarios of inferring $\mathbf{s}$ from the relation $\mathbf{d} = \mathbf{R}_{\text{true}}\mathbf{s} + \boldsymbol{\eta}_{\mathbf{d}}$: one with the classical, Gaussian likelihood function with analytically tractable posterior, and the other with the EMG likelihood that the synthetic data is generated with, albeit having to resort to MCMC and optimization routines to find the MAP values for the fluxes $\mathbf{s}$ and its posterior density. Figure 11 illustrates a result of this comparison for data amount $M = 100$ and the number of inferred fluxes $N = 22$. It compares the true flux $\mathbf{s}_{\text{true}}$ to the MAP values stemming from the two scenarios $\mathbf{s}^{\text{G}}_{\text{MAP}}$ and $\mathbf{s}^{\text{EMG}}_{\text{MAP}}$. Clearly, using more appropriate likelihood function improves the accuracy of the resulting fluxes. Furthermore, we note that the accuracy of the resulting fluxes depends on the amount of information, i.e. the amount of data observations M. Figure 12 illustrates this effect. Namely, it compares the two likelihood scenarios and extract the committed error in the fluxes with respect to increasing amount of data, i.e. increasing M. Again, the more appropriate, skewed EMG likelihood is more accurate.

3.6 Bayesian Formulation Accounting for Model Error

The linearized model $\mathbf{R}$ is generally an approximation of the true relationship between the fluxes and the observations. The conventional formulation from Eq. (30) does not take this model error into account. This leads to a few concerns regarding flux estimation and further predictions based on those. First of all, the estimated fluxes will be somewhat biased trying to compensate for the structural deficiencies of the model, i.e. the model error induced by the linear approximation of the true relationship. Moreover, increasing the amount of observations will not help. It will reduce the posterior density towards a delta-function centered near a single, best flux vector $\mathbf{s}$, under the wrong assumption that the model is perfectly replicating the truth. This flux vector will be biased away from the true vector as the model error has not been accounted for. Thus, not only the mean flux is incorrect, but also the estimated posterior uncertainty does not reflect the true uncertainty

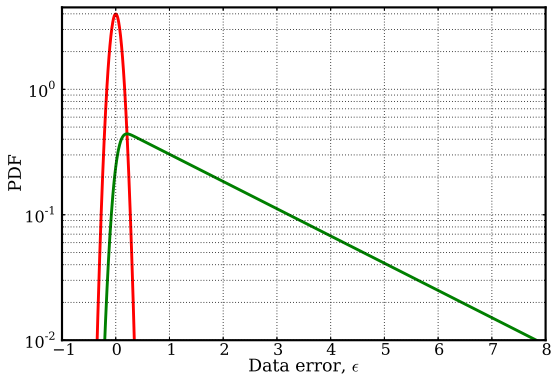


Figure 10. The probability density functions of Gaussian and the associated EMG variables. The parameters are set to $\mu = 0$, $\sigma = 0.1$ and $\beta = 2$.

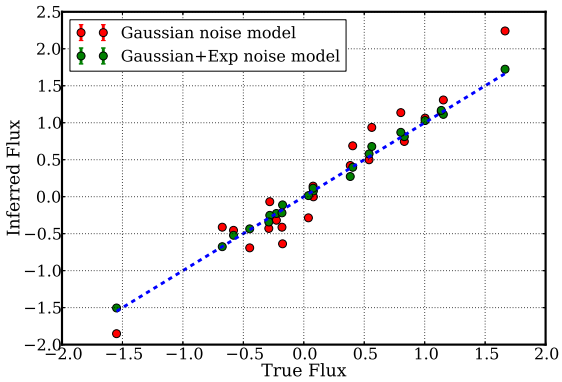


Figure 11. Comparison of the true fluxes with the best values inferred using a) Gaussian likelihood and analytical formula, and b) EMG likelihood and optimization algorithm.

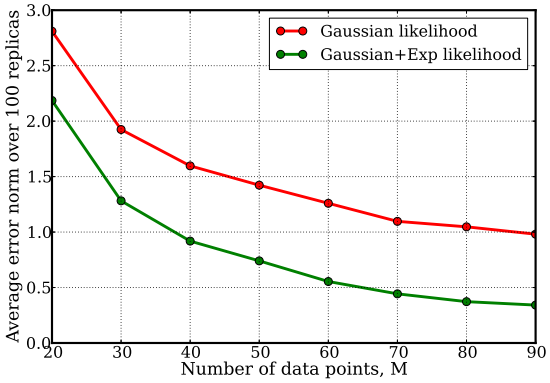


Figure 12. Convergence of the average error norm, over 100 replica runs, of the MAP value of fluxes, as the amount of data M, grows. Two scenarios are compared, a) Gaussian likelihood and analytical formula, and b) EMG likelihood and optimization algorithm.

associated with the found mean flux. This will subsequently lead to predictions that are wrong both in terms of mean and in terms of the uncertainty around it.

We strive to formulate an inverse problem where the resulting uncertainties take into account both the model bias and data errors. We would like to build-in uncertainties in the result that do not vanish when the amount of observations increases or the observational error vanishes. In order to do so, we cast the flux vector as a random vector, i.e. allow variability in input parameters that will lead to output variability consistent with observations. For simplicity, and in order to restrict the number of unknown parameters, we assume independent Gaussian distributions for each flux, i.e.

$$s_i = \mu_i + \sigma_i \xi_i, \text{ for } i = 1, \ldots, N, \tag{36}$$

where ξ_i are standard normal random variables. In a vector form, this can be written as

$$\mathbf{s} = \mu + \mathbf{I}\sigma\xi, \tag{37}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Within a Bayesian formulation, the objects of inference now are the pairs of vectors $\mu = (\mu_1, \ldots, \mu_N)$ and $\sigma = (\sigma_1, \ldots, \sigma_N)$. The Bayes formula in this case reads

$$p(\mu, \sigma|\mathbf{d}) \propto p(\mathbf{d}|\mu, \sigma) p(\mu, \sigma). \tag{38}$$

The likelihood function $L_{\mathbf{d}}(\mu, \sigma) = p(\mathbf{d}|\mu, \sigma)$ now measures the probability of obtaining the specific data set $\mathbf{d}$ given the probabilistic flux description defined as independent-component Gaussian vector with mean μ and standard deviations σ. One can write the relation between observations and the model as

$$\mathbf{d} = \mathbf{R}\mathbf{s} + \varepsilon_{\mathbf{d}}, \tag{39}$$

where the measurement error $\varepsilon_{\mathbf{d}}$ is assumed to be a multivariate normal with vanishing mean and diagonal covariance matrix $\mathbf{C_d}$ indicating independence of its components. With the MVN

characterization of the fluxes $\mathbf{s}$ from Eq. (37), this implies

$$\mathbf{d} = \mathbf{R}\mu + \mathbf{R}\sigma\xi + \varepsilon_{\mathbf{d}}, \text{ and } \mathbf{d} \sim \text{MVN}(\mathbf{R}\mu, \underbrace{\mathbf{R}\sigma\sigma^T\mathbf{R}^T + \mathbf{C_d}}_{\mathbf{G}}), \tag{40}$$

where ξ is a vector of i.i.d. standard normal variables, leading to a likelihood function

$$L_{\mathbf{d}}(\mu,\sigma) = p(\mathbf{d}|\mu,\sigma) = (2\pi)^{-\frac{N}{2}} \det(\mathbf{G})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{R}\mu)^T \mathbf{G}^{-1}(\mathbf{d}-\mathbf{R}\mu)\right] \tag{41}$$

Note that when there is no data noise, i.e. $\mathbf{C_d} = \mathbf{0}$, the $M \times M$ covariance matrix $\mathbf{G}$ has a rank at most N, which is typically smaller than M, since one expects to have more observations than flux sources that are being sought for. Indeed, if there is no data noise, generally there can not be a set of N values $\mathbf{s}$ that lead to an exact match of $M > N$ values $\mathbf{d} = \mathbf{R}\mathbf{s}$. Therefore, for vanishing or small enough data noise, the likelihood (41) is degenerate. Only if there is large enough data noise characterized by the matrix $\mathbf{C_d}$, the covariance $\mathbf{G}$ has full rank and can be inverted.

One can draw parallels with the formulation of the Kennedy-O'Hagan approach [116], where the model error δ_m is explicitly written as a Gaussian with a known, usually square-exponential covariance with respect to the underlying spatial distance measure. That is

$$\mathbf{d} = \mathbf{R}\mu + \underbrace{\delta}_{\mathbf{E}\xi} + \varepsilon_{\mathbf{d}}, \tag{42}$$

and $(\mathbf{E}\mathbf{E}^T)_{ij} = A\exp\left(-(x_i - x_j)/l^2\right)$, for a predefined correlation length l and variance magnitude A, while the underlying spatial locations $\mathbf{x}$ correspond to the observations $\mathbf{d}$. In our case, instead of building an explicit spatially correlated Gaussian model error term, we force the spatial correlations to be consistent with the model itself by embedding the variability of the fluxes in the model and obtaining $\mathbf{E} = \mathbf{R}\sigma$.

In order to avoid the likelihood degeneracy, one can generally resort to Approximate Bayesian Computation (ABC) which employs likelihoods that are based on matching moments or other statistics of the data $\mathbf{d}$ and the model predictions $\mathbf{R}\mathbf{s}$ [117, 118]. In this work, we will employ *marginalized* likelihood, which is essentially an approximation to the full likelihood (41) by replacing the covariance matrix $\mathbf{G}$ by its diagonal approximation $\tilde{\mathbf{G}} = diag(\mathbf{G})\mathbf{I}$. The likelihood function then is written as

$$\tilde{L}_{\mathbf{d}}(\mu,\sigma) = (2\pi)^{-\frac{N}{2}} \det(\tilde{\mathbf{G}})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{d}-\mathbf{R}\mu)^T \tilde{\mathbf{G}}^{-1}(\mathbf{d}-\mathbf{R}\mu)\right], \tag{43}$$

and it constrains each observation to the marginal distribution of the model prediction independently.

The prior function can be split into independent components

$$p(\mu,\sigma) = \prod_{i=1}^{N} p(\mu_i)p(\sigma_i) \tag{44}$$

We have used uninformative, uniform priors for the means $p(\mu_i) = \text{const}$, and Jeffrey's priors [119] for the standard deviations $p(\sigma_i) \propto 1/\sigma_i$. In practice, we employ $\log \sigma_i$ to enforce positivity of the standard deviation, and the Jeffrey's prior is equivalent to a uniform prior on the logarithm, i.e. $p(\log \sigma_i) = \text{const}$.

While one should ideally proceed with sampling from the posterior using MCMC methods, here we will mainly focus on the best estimates of μ and σ, i.e. we search the values that maximize the posterior distribution using Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm. The calibration problem is effectively reformulated as a *density estimation* problem, since the object of interest are the mean μ and the variance σ of the cast multivariate density of the flux vector $\mathbf{s}$. For details of such reformulation of calibration in order to properly take into account model errors, see [120].

Numerical tests

Consider synthetic generation of data from 'true' fluxes and a 'true' $M \times N$ response matrix $\mathbf{R}_{\text{true}}$ with small measurement noise $\varepsilon_{\mathbf{d}}$, in order to focus solely on the model error, leading to

$$\mathbf{d} = \mathbf{R}_{\text{true}}\mathbf{s}_{\text{true}} + \varepsilon_{\mathbf{d}} \tag{45}$$

Now, we infer fluxes $\mathbf{s}$ using a response matrix $\mathbf{R}$ that is generally biased, i.e. $\mathbf{R} \neq \mathbf{R}_{\text{true}}$, and we seek $\mathbf{s}$ such that

$$\mathbf{d} \approx \mathbf{R}\mathbf{s}. \tag{46}$$

Clearly no matter how much data is used, this will lead to the best estimate of $\mathbf{s}$ that is biased, i.e. $\mathbf{s} \neq \mathbf{s}_{\text{true}}$, in order to compensate for inadequacy of the response matrix. Indeed, Figure 13 illustrates a case with $N = 77$ data points and $M = 22$ flux sources. We have randomly selected a 'correct' response matrix $\mathbf{R}_{\text{true}}$ and inferred the fluxes $\mathbf{s}$ with both the correct matrix $\mathbf{R}_{\text{true}}$ and a slightly perturbed, 'wrong' one $\mathbf{R}$. The posterior width is very small, together with the data measurement error magnitude, leading to very small errorbars in both the inferred model predictions and the inferred flux values. This confirms the deficiency of the conventional inference approach out lied above. That is, the prediction errorbars are misleadingly small, and the posterior errorbars can be made small with large enough amount of data, thus failing to handle the clearly-present modeling bias.

On the other hand, with the proposed density estimation reformulation, we cast the flux vector $\mathbf{s}$ as a random quantity, and infer its component-wise means μ and standard deviations σ. In this work, we focused on the MAP estimates of these quantities, μ_{MAP} and σ_{MAP}. As Figure 14 demonstrates, the standard deviations σ_{MAP} do not vanish irrespective of the amount of data, therefore leading to estimates of fluxes with errorbars that are consistent with the true, unknown flux values. This in turn leads to prediction uncertainties that are not negligible and are consistent with the observational data.

Let us now turn to the TransCom3 study. We will generally investigate 14 models summarized in [13], i.e. 14 different response matrices $\mathbf{R}$. These $M \times N$ matrix-models are summarized in

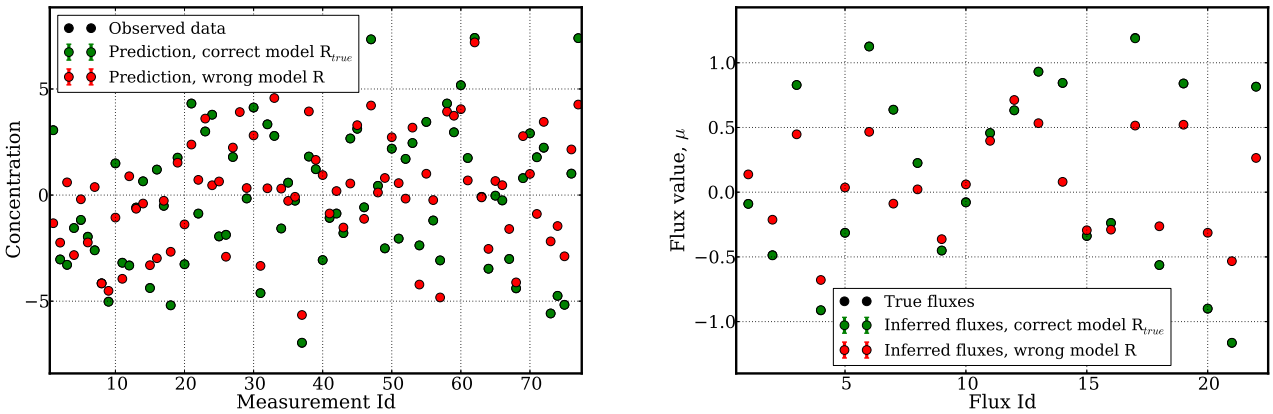


Figure 13. Results of conventional Bayesian inference with Gaussian likelihoods and priors. Two inference scenarios are studied: one with the 'correct' response matrix, and the other with a biased one. On the left plot, synthetically generated observational data is shown together with the predictions from two inference tests. If one uses the correct response matrix, the predictions perfectly match the data, while the wrong model leads to biased predictions with small errorbars that are not consistent with the commited error. On the right plot, the fluxes are shown in both scenarios. Again, the inferred fluxes are biased away from the true ones, if one uses the perturbed model for the inference.

Figure 15. The observational CO_2 concentration data is collected at $M = 77$ sites, and the goal is to infer fluxes at $N = 22$ selected regions. The observed data, depending on the location latitude, is illustrated in Figure 16. For the density-estimation formulation, we have used flat priors for μ, and somewhat informative, lognormal priors for components of σ. The parameters of the lognormal prior were chosen to enforce mean of 1.0 and mode of 0.5 approximately reflecting the expected discrepancy that we would commit in inferring the fluxes. As Figures 17 and 18 illustrate, the conventional inference approach leads to strongly varying flux values from model to model. Moreover, the posterior errorbars, mainly reflecting the amount of data, fail to cover the model-to-model variability. On the other hand, the density estimation method introduced here, allows more uniform results across all the models. In fact, the relatively 'flat' mean fluxes for most of the ocean regions simply indicate the fact that the observations do not constrain well the flux values in these locations of interest. This is a considerably more robust result compared to the results from Gaussian methodology that tend to overfit and compensate for model deficiencies by driving the flux values away from truth or from prior knowledge.

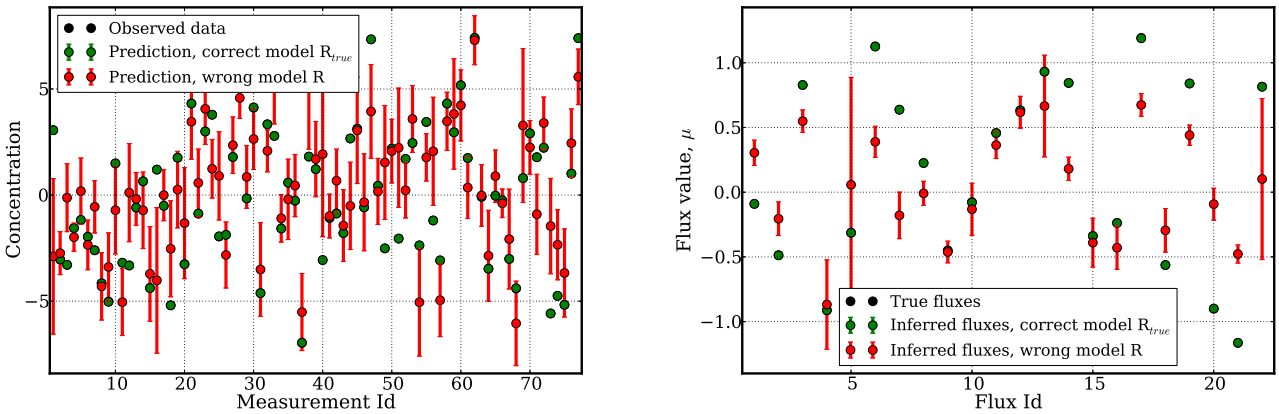


Figure 14. The same scenario as in Figure 13, only with the proposed density-estimation framework. The inferred MAP values for the standard deviations lead to consistent errorbars in both flux estimation and concentration predictions.

3.7 Summary

In this section, we have discussed challenges associated with the conventional Bayesian inference strategy for inferring sources and sinks in linear transport models given measurements of CO_2 concentrations. We have focused on challenges associated with a) data bias, and b) model bias. In particular, we demonstrated that for non-Gaussian, biased measurement errors the usual Gaussian likelihood may lead to biased results, and one should be careful in choosing an appropriate likelihood function. Having said that, the Gaussian likelihood construction allows efficient, analytical computations of the posterior distributions, while for non-conventional likelihoods one almost always needs to resort to MCMC or optimization algorithms. Furthermore, we have demonstrated that the conventional Bayesian methodology does not properly address the *model errors*, i.e. the biases that stem from models being imperfect or inadequate. In order to alleviate this issue, we introduced a reformulated inverse problem that is one of density estimation. We demonstrated how inferred values are biased and overfit in order to accommodate model errors, in a synthetic case. Finally, we used 14 different models from the TransCom3 experiment and demonstrated how the conventional inference approach leads to strong model-to-model variability for the inferred fluxes, while our proposed density estimation approach results in robust answers with quantified uncertainties that are associated with model errors.

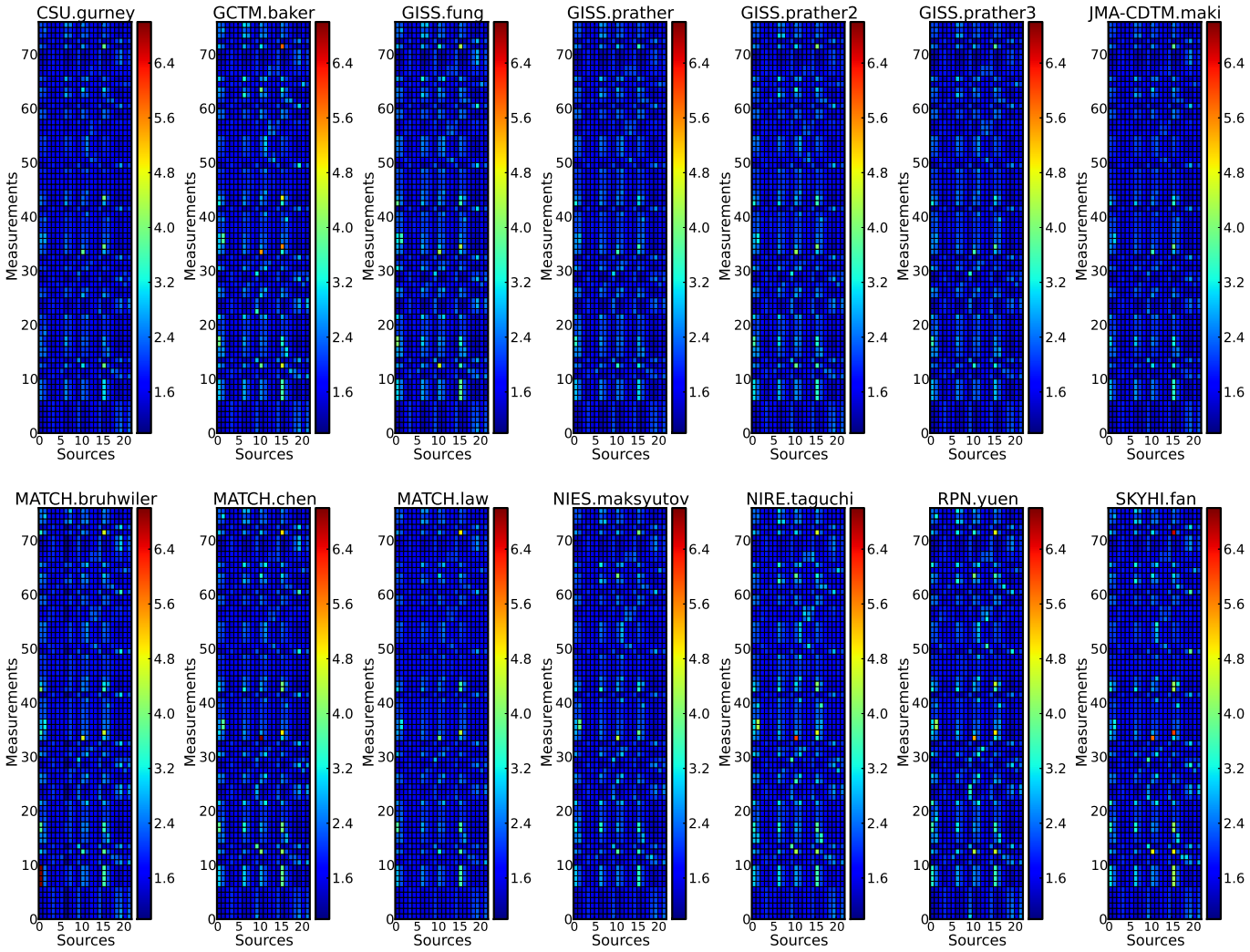


Figure 15. Matrices corresponding to the 14 linear response models under consideration.

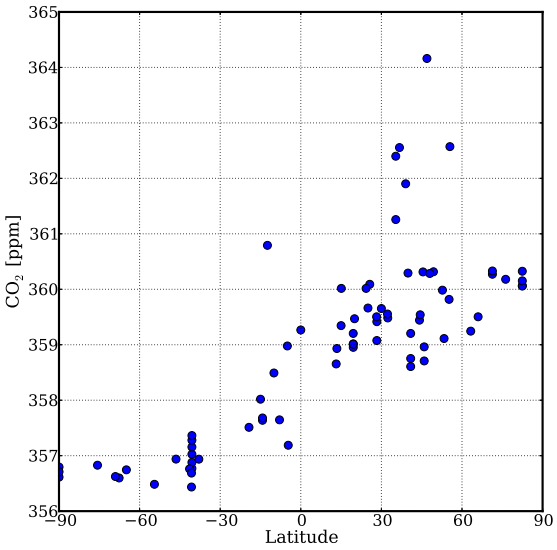


Figure 16. Observed concentrations with respect to location latitudes.

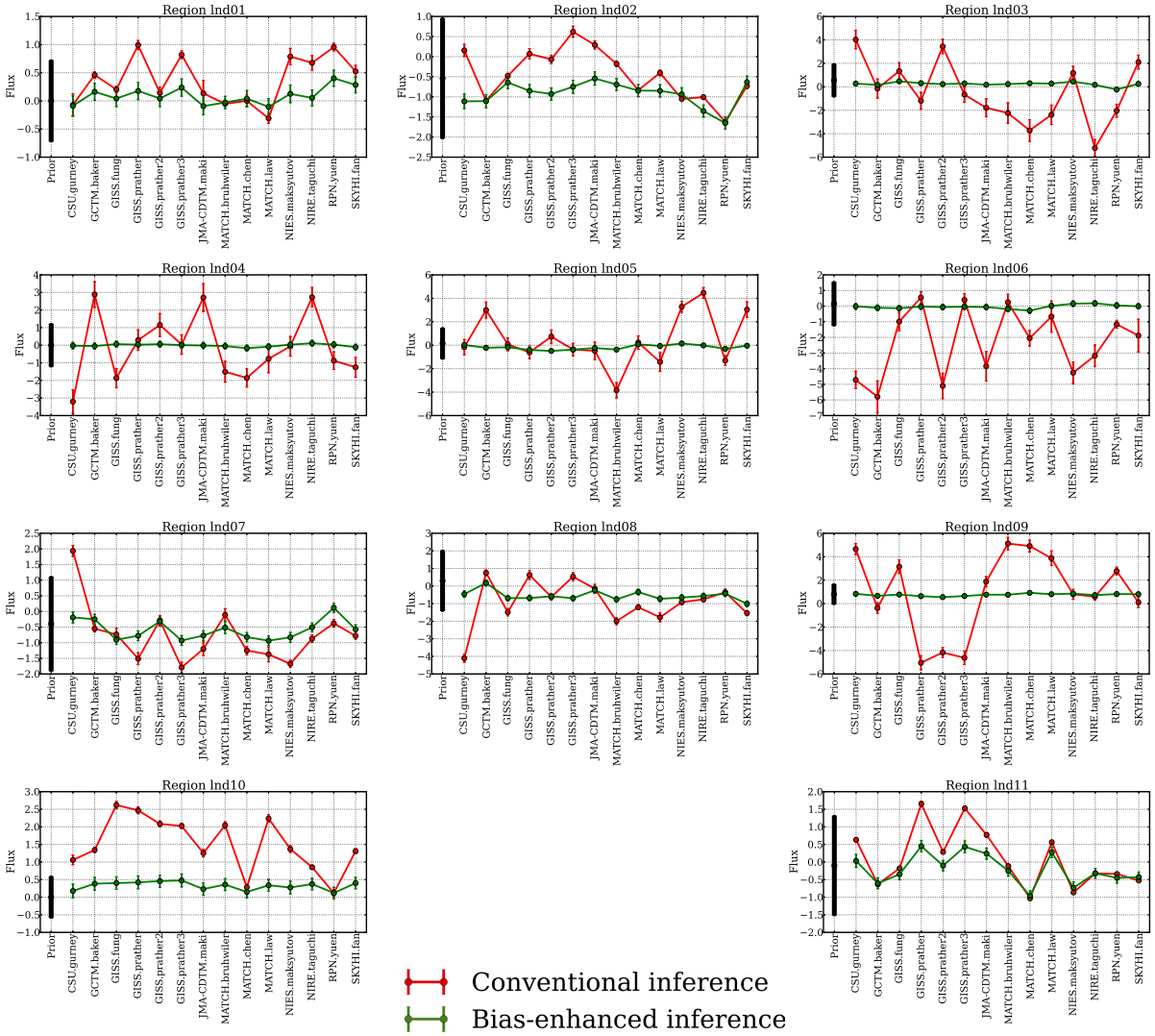


Figure 17. Illustration of the posterior fluxes in the land regions inferred by two methods, using 14 different models. The prior flux and its standard deviation is also depicted.

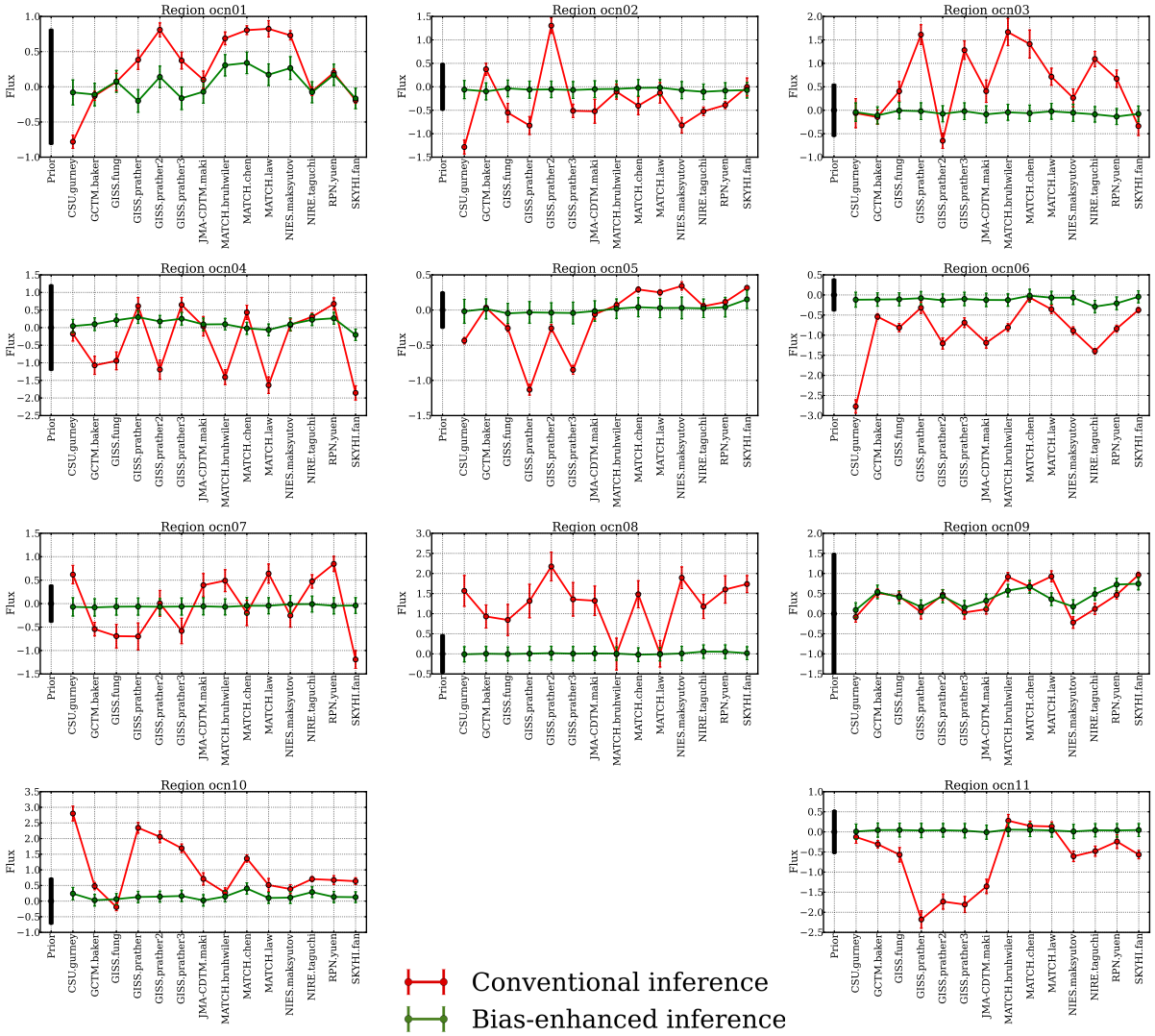


Figure 18. Illustration of the posterior fluxes in the ocean regions inferred by two methods, using 14 different models. The prior flux and its standard deviation is also depicted.

4 Source Inversion using Regional Transport Models and Passive Scalar Transport

4.1 Simulations using the Weather Research and Forecasting Model

The Weather Research and Forecasting (WRF) model simulated wind fields for a period of seven days, starting on Oct. 22, 2010 at 12:00am. The simulation includes 3 nested mesh levels. Figure 19 shows the topology of the computational domain. The coarsest mesh has 151 computational cells in the East-West direction and 201 North-South, and covers most of the Norh-American continent. The intermediate and finest mesh have 151×199 grid cells each and are centered on OK and KS. The lower left corner of the intermediate mesh correspond to cell $(52, 68)$ in the coarse grid, while the fine mesh is anchored at the same cell number in the intermediate grid. We used a

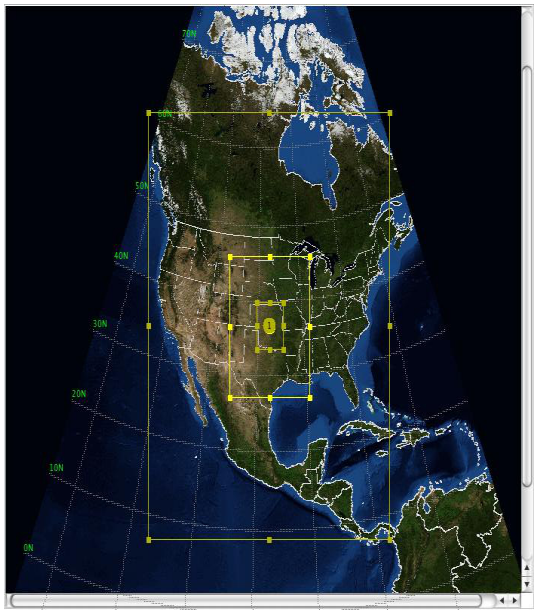


Figure 19. Topology of the nested grids centered around OK and KS.

refinement factor of 3 between successive mesh blocks, starting a cell size of 30km in the coarse mesh and continuing with 10km and 3.33km for the intermediate and fine mesh cells, respectively.

The simulation was set up to output intermediate solution files every hour, and restart files every 24h at 12:00am. Figures 20 and 21. Show 2D velocity vectors at a height of 10m. This data is directly available in the WRF netcdf output file.

4.2 Passive Scalar Transport

The 2D velocity at a height of 10m, with components U_{10} and V_{10} in the East-West and North-South directions, respectively, is extracted from the WRF output files and then used to drive the advection-diffusion of a passive scalar field, modeled as

$$\frac{\partial c}{\partial t} + (u \cdot c) = D\nabla^2 c + S_x(x)S_t(t) \tag{47}$$

Here c is the concentration of a generic scalar, u is the velocity vector, D is the scalar diffusivity and S_x and S_t are functions that control the spatial and temporal profiles for scalar sources that will be placed in the computational domain.

Eq. 47 is discretized on a computational grid that corresponds to the finest mesh in the WRF simulations. The spatial derivatives are computed using a second-order upwind scheme, while the time advancement is done using a second-order TVD Runge-Kutta scheme [121]. The velocity field is linearly interpolated in time to generate velocity values for each time step. Spatial interpolation is not necessary since the computational mesh coincides with the WRF grid.

Setup of Scalar Sources and Sensors

The spatial profile component $S_x(x)$ of the source allowed to have the scalar emitted from one particular locations in the computational domain. The results presented in this report correspond to a singular scalar source located in cell # $(70, 20)$. The time profile $S_t(t)$ consists of a sequence of periodic puffs shown in Fig. 22. The amplitude A_0 of the puffs is assumed to be known since the scalar concentrations depend linearly on this value. The duration, s_1, and the interval between them, s_2, are the unkown parameters which will be inferred based on concentration measurements at select sensor locations.

Figure 22 shows instantanous scalar concentration contour plots for a simulation with $s_1 = s_2 =$ 3.6h. The source location is shown with a filled black circle. The filled squares show the locations of the numerical "sensors" that record the concentration values as a function of time.

Figure 24 shows time histories of scalar concentrations measured at the sensor locations shown with black squares in Fig. 23, numbered 1 through 4 starting from the leftmost location. The signal is highly nonlinear due to the shifting patterns in the windfield.

4.3 Bayesian Inference of the Source Characteristics

Bayes formula

$$p(\vec{s}|\mathscr{D}) \propto L_{\mathscr{D}}(\vec{s})p(\vec{s}) \tag{48}$$

relates the prior distribution $p(\vec{s})$ of source para-meters $\vec{s}$ to the posterior $p(\vec{s}|\mathscr{D})$, where the data $\mathscr{D}$ is the set of measurements at various sites around the source.

The likelihood $L_{\mathscr{D}}(\vec{s})$ accounts for the discrepancy between the data $\mathscr{D}$ and the model $f(\vec{s})$.

$$L_{\mathscr{D}}(\vec{s}) \propto \exp\left(-\sum_{i=1}^{N} \frac{(f(\vec{s})-\vec{y}_i)^2}{2\sigma^2}\right) \tag{49}$$

Here N is the number of measurement sites, $f(\vec{s})$ are pollutant concentrations at the measurement site i computed using a transport model of choice (e.g., WRF+scalar transport), and $\vec{y}_i$ are the experimental values. The standard deviation σ includes both the instrument error as well as any model discrepancy error (initial and boundary conditions, sub-grid models, numerical approximations) introduced by f.

Given the likelihood $L_{\mathscr{D}}(\vec{s})$ and the prior $p(\vec{s})$, we then draw samples from the posterior distribution $p(\vec{s}|\mathscr{D})$ via Markov Chain Monte Carlo (MCMC) sampling. MCMC is a class of techniques that allows sampling from a posterior distribution by constructing a Markov Chain that has the posterior as its stationary distribution [57].

Surrogate Model Construction

The computational expense of the WRF and scalar transport simulations, typically associated with a large number of MCMC samples, will be circumvented by employing surrogate models, which are used instead of the forward model $f(s)$ in the MCMC.

For this study the surrogate models are based on polynomial chaos (PC) expansions [122, 123] and are used to represent quantities of interest, e.g., scalar concentration at specific locations, as functions of source and model parameterizations.

In order to use PC representations we interpret input parameters $\vec{s}$ as random variables, which can be represented via their cumulative distribution function (CDF) $F(\cdot)$, such that, with $\xi_i \sim$ Uniform$[-1,1]$, we have:

$$s_i = F_{s_i}^{-1}\left(\frac{\xi_i+1}{2}\right), \qquad \text{for } i = 1,2,\ldots. \tag{50}$$

The forward model output for the scalar dispersion given by $f(\cdot)$ can be represented as a PC expansion:

$$f(\vec{s}) = Z \approx \sum_{k=0}^{K} Z_k \Psi_k(\xi) \tag{51}$$

$\Psi_k(\cdot)$ are standard Legendre polynomials of independent, random variables ξ, orthogonal w.r.t. uniform pdf $p_\xi(\xi)$, i.e.,

$$\langle \Psi_i(\xi)\Psi_j(\xi)\rangle \equiv \int \Psi_i(\xi)\Psi_j(\xi)p_\xi(\xi)d\xi = \delta_{ij}\langle \Psi_i(\xi)^2\rangle \tag{52}$$

The coefficients Z_k are computed by Galerkin (orthogonal) projection

$$Z_k = \frac{\langle f(\vec{s}(\xi))\Psi_k(\xi)\rangle}{\langle \Psi_k^2(\xi)\rangle} \tag{53}$$

Here, the projection integrals are computed by quadrature

$$\langle f(\vec{s}(\xi))\Psi_k(\xi)\rangle = \sum_{l=1}^{N_{quad}} w_l f(\vec{s}(\xi_l))\Psi_k(\xi_l) \tag{54}$$

Figures 25 and 26 show comparisons between model values and surrogate model values corresponding to sensors #1 and #3. In this study the "model" is the cumulative scalar concentrations obtained by integrating in time the time series values shown in Fig. 24. The surrogate model values, computed using eq. 51 are based on 7-th order polynomials using Legendre basis functions. The polynomial coefficients were constructed by quadrature (eq. 54) using 11 quadrature points per dimension for a total of 121 simulations. The surrogate models show a maximum discrepancy of about 4% compared to results from full model simulations.

Results

We considered two synthetic source scenarios. For scenario A the source parameters were set to $(s_1, s_2) = (0.082, 0.075)$, while for B the values were set to $(s_1, s_2) = (0.062, 0.082)$. Here the time values were normalized with respect to the total measurement time of 48h. To generate the data we ran the transport model for these sets of parameters, recorded the concentrations at the 4 sensor locations and then perturbed the values with multiplicative Gaussian noise to simulate the measurement error.

$$y_{i,\text{meas}} = y_{i,\text{model}}(1 + N(0,\sigma)) \tag{55}$$

For both scenarios we used $\sigma = 0.1$.

Figures 27 and 28 compare the inference results for the two scenarios. For this particular setup the MCMC sampling of the posterior distributions for the source parameters can be compared with analytical solutions since the problem is two-dimensional only.

In both cases the joint PDF obtained through MCMC sampling agrees well with the analytical values. In the first scenario the inference process detected a multi-modal distribution with one of the modes centered around the "truth". For this scenario the information available from the measurements is not sufficient to pin-point the source characteristics. In the second scenario the joint PDF is centered around the expected values.

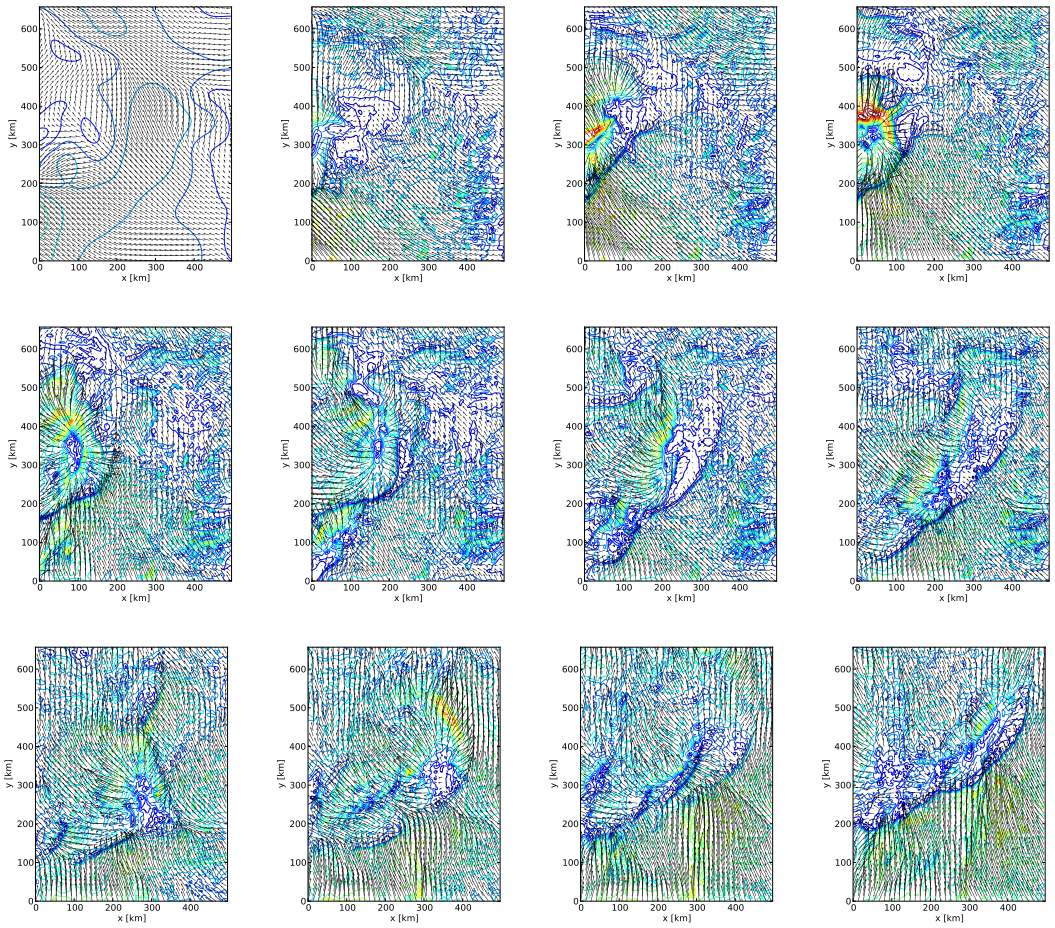


Figure 20. Sample 2D velocity fields at a height of 10m. The contour line correspond to the velocity magnitude, changing from blue for small values to red for a magnitude of 15m/s. The frames, left to right and top to bottom, correspond to 2h increments starting on 10/22/2010 at 12:00am GMT.

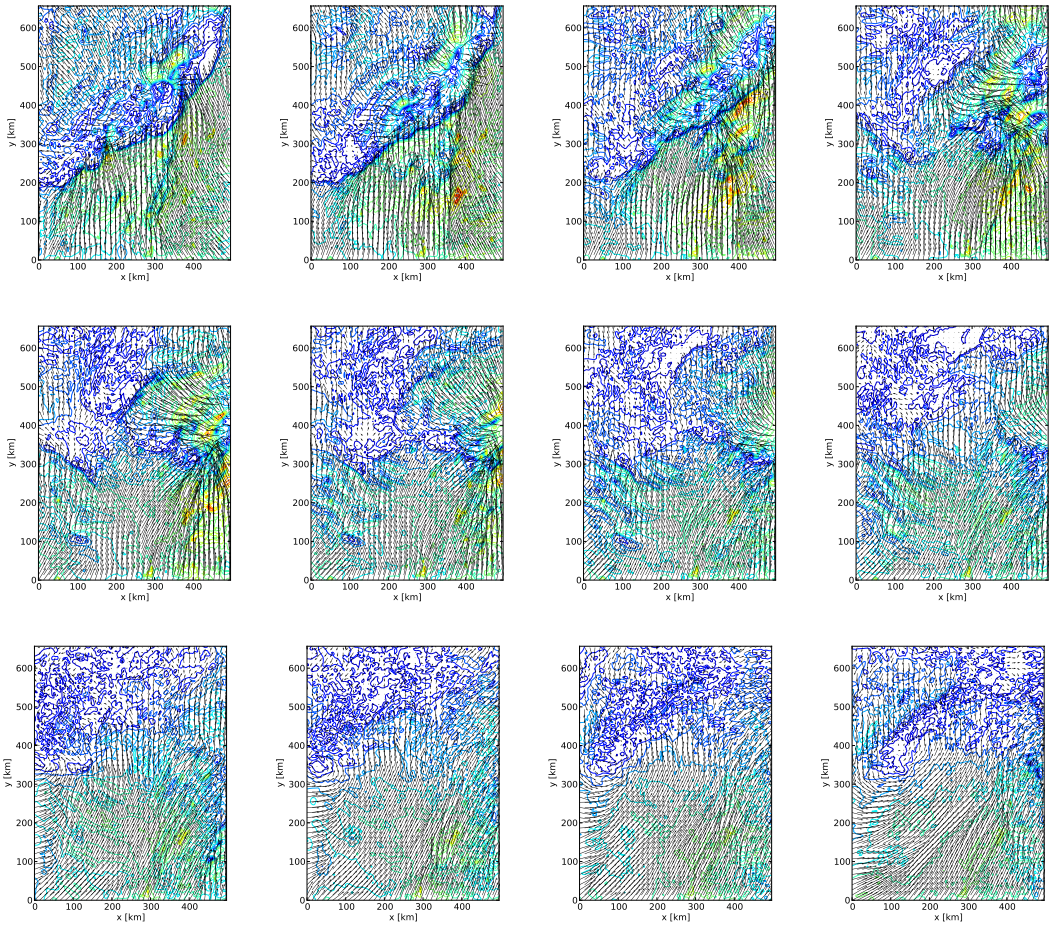


Figure 21. Sample 2D velocity fields at a height of 10m. The contour line correspond to the velocity magnitude, changing from blue for small values to red for a magnitude of 15m/s. The frames, left to right and top to bottom, correspond to two hour increments continuing from Fig. 20.

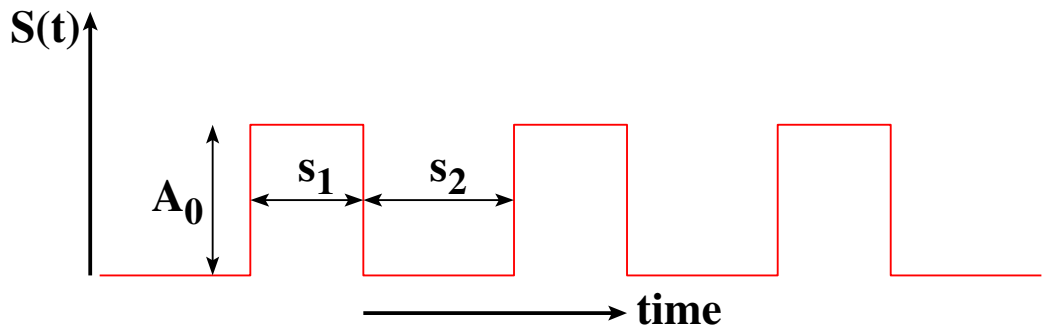


Figure 22. Time profile function $S_t(t)$ showing a sequence of periodic puffs.

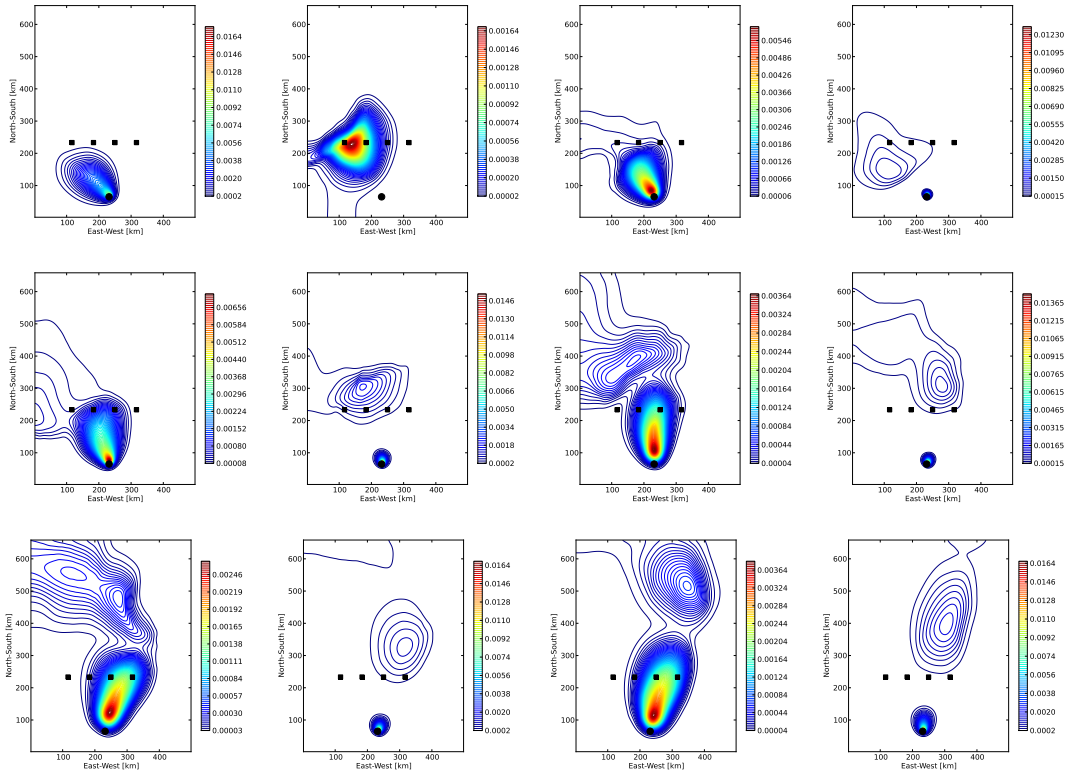


Figure 23. Contour plots of scalar concentrations corresponding to a simulation with $s_1 = s_2 = 3.6$h. The frames, left to right and top to bottom, correspond to 8h increments starting 8h from the beginning of the simulatin.

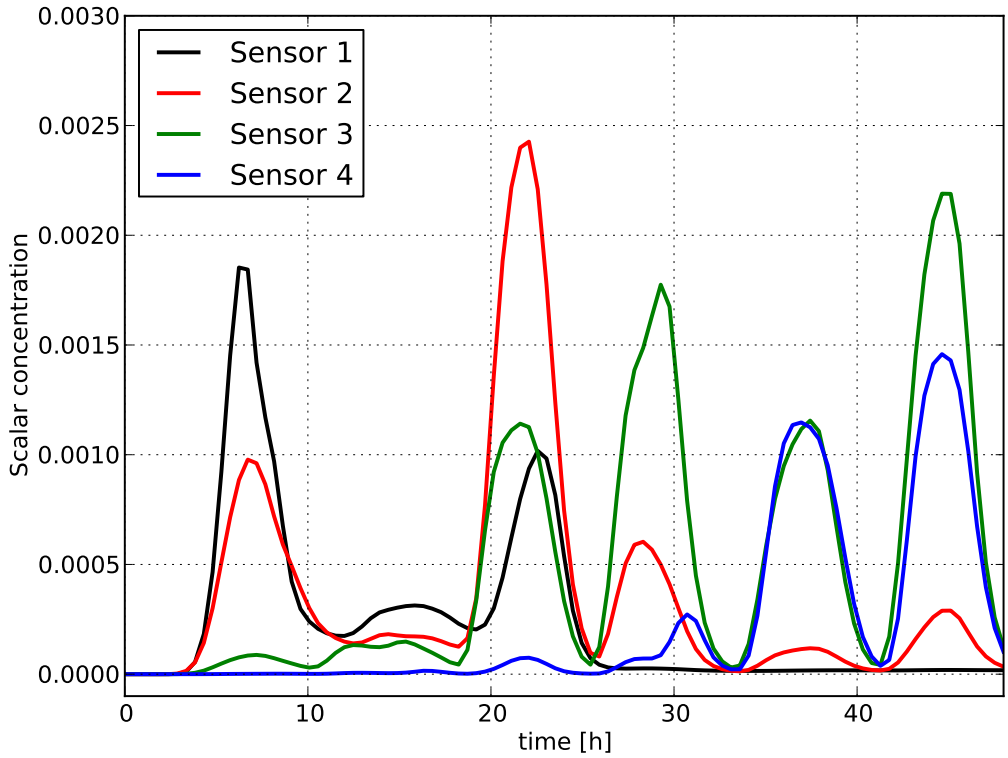


Figure 24. Time histories of scalar concentrations measured at the sensor locations shown in Fig. 23.

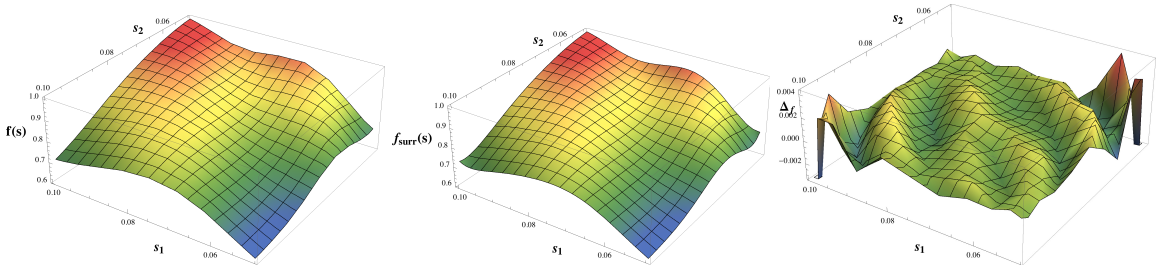


Figure 25. Sensor 1 data as a function of the source parameters s_1 and s_2. The left frame shows transport model data, the middle corresponds to the surrogate model, and the right frame show the discrepancy between the two.

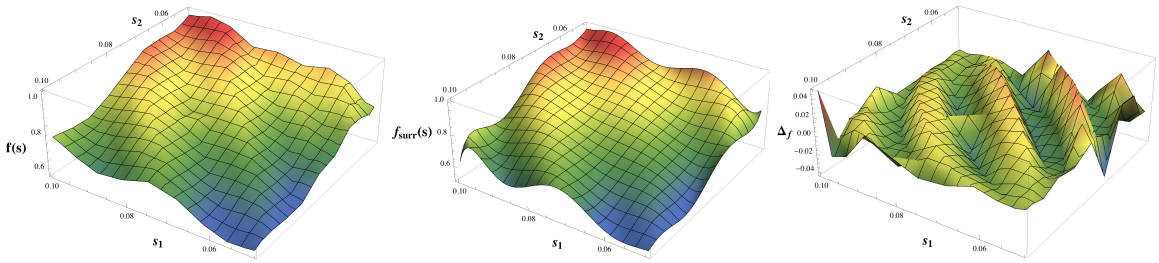


Figure 26. Sensor 3 data as a function of the source parameters s_1 and s_2. The left frame shows transport model data, the middle corresponds to the surrogate model, and the right frame show the discrepancy between the two.

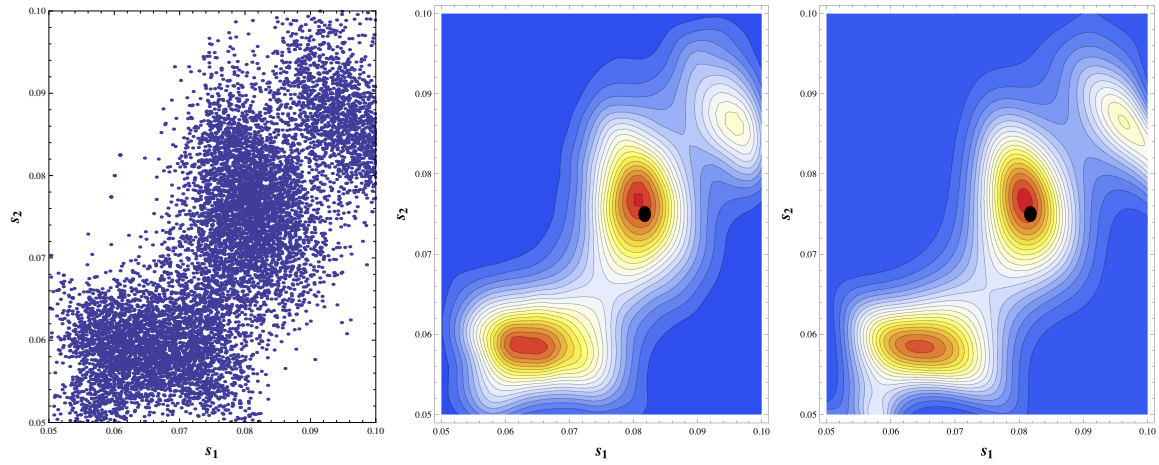


Figure 27. Inference of source parameters for Scenario A. Left frame shows the MCMC samples, the middle frame shows the posterior density of (s_1, s_2) based on these samples, and the right frame shows the analytical solution.

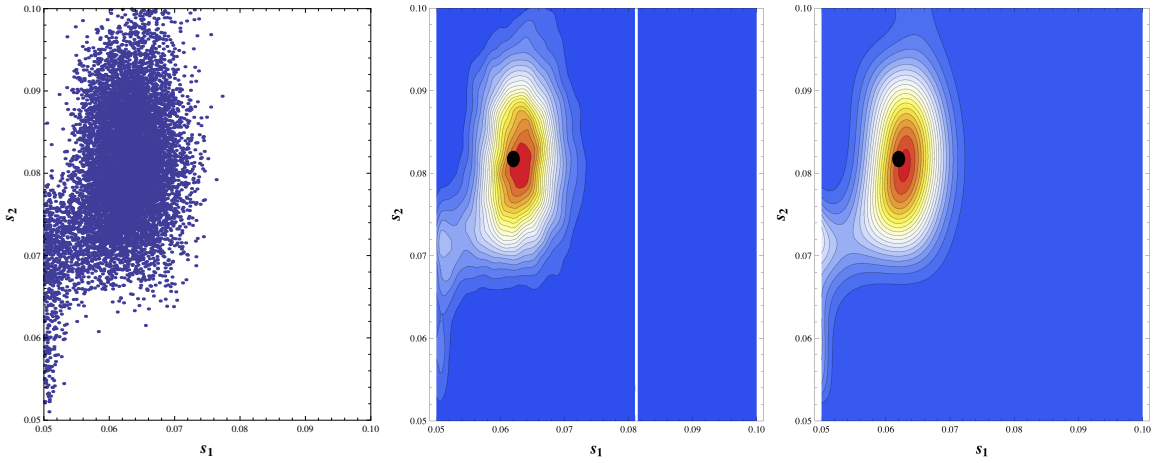


Figure 28. Inference of source parameters for Scenario B. Left frame shows the MCMC samples, the middle frame shows the posterior density of (s_1, s_2) based on these samples, and the right frame shows the analytical solution.

5 Inversion Under Uncertainty for Trace-Gases using Convection-Diffusion-Reaction

5.1 Introduction

The characterization of trace-gas sources is important to help control pollutants in the atmosphere. Carbon-dioxide is one of several species that has been linked to the increase of average global temperatures and understanding the overall dynamics of these trace-gases depends on knowing the spatial distribution and magnitudes of carbon-dioxide fluxes. Exact characterizations could for instance support a Green House Gas Information System (GHGIS), which would be responsible for monitoring and managing the overall production of greenhouse gases. The determination of locations and characters of sources is complicated by multiple factors: the multiple spatial distributions, extreme sparsity of the measurements, temporal variations, uncertainty of natural CO_2 source and sinks, and the many uncertainties associated with data and model parameters (in particular the velocity field that needs to be calculated from atmospheric/weather models). In this work, we investigate two critical aspects of this inversion. First we explore an efficient inversion under uncertainty scheme that leverages concepts from stochastic optimization. We account for uncertainties assosicated with the velocity fields. Second, we investigate the inversion of trace gas source terms by considering multiple trace gase measurements.

The inversion of source terms motivates a large optimization problem in which the goal is to reconcile the differences between sparse observations and numerical predictions of convection-diffusion-reaction dynamics by manipulating magnitudes of source terms as target inversion parameters. To eventually develop methodolgies that can reconstruct source terms in sufficient detail, many inversion variables need to be considered, potentially at every computational discretization point. Black box approaches in which gradients of the objective function are determined through finite difference methods or local interrogation (non-gradient based) methods quickly become computationally intractable in addition to suffereing from quality issues as a result of for instance selecting an appropriate finite difference step. To address both the computational expense and the accuracy of the gradient, adjoint-based sensitivities need to be implemented. This however poses several implementation challenges associated with parallelism and stabilized finite element discretization. Furthermore, first order optimization methods, such as steepest descent and non-linear conjugate gradient (CG), are not efficient and potentially not sufficently accurate. Second order approaches, such as Newton and Quasi-Newton methods, may be reqruied, which may introduce additional implementation challenges.

The determination of accurate trace-gas dynamics in atmospheric flows introduces uncertainties ranging from inaccurate velocity fields at fine spatial scales to the variability of the temporal signals for both anthropogenic and biological source terms. Efficient methods must be considered to manage uncertainties without compromising our ability to invert for large number of source terms while managing stochastic model parameters.

In this work, we first present a large scale optimization approach that leverages adjoint-based sensitivities. Our optimization methods are implemented in a separate package, called Rapid Op-

timization Library (ROL) within the Trilinos framework, which features a range of algorithms including first and second order methods, line search and trust region globalization, and the ability to accomodate inexact gradient and objective function evaluations. The finite element approach is used to discretize convection-diffusion-reaction physics. For high Peclet numbers, a Stabilized Upwind Petrov Galerkin (SUPG) stabilization method is implemented in both the forward and adjoint operators. The Jacobians in the forward simulator are calculated with automatic differentiation through C^{++} template overloading. Parallelization is achieved through the Epetra package in Trilinos. We leverage concepts from stochastic optimization to manage model uncertainties and strive to derive robust solutions. A "risk measure" is introduced in the objective function and then discretized with collocation methods. Although a range of risk measures can be considered, we limit our approach to several popular ones and to risk measures that can be easily mapped from the financial to the engineering world. In particular we consider an expected value and a coherent value at risk, which are related to risk-neutral and probabilty-of-failure measures, respectively. This approach was first developed in an optimal control problem, and although one might prefer a stochastic inverse solution, the formulation is identical and extends to inverse solution with the computational advantages of the large scale deterministic methods.

Mathematical Formulation

This section describes the optimization problem that we solve in inferring for model parameters from data, and derives the formula by which the gradient of the objective function is calculated. Several types of risk measures are described, and the reasoning for choosing one is explained.

Optimization problem formulation

The physics of the test cases are described by the convection-diffusion-reaction equations for two species with concentration states ϕ_1 and ϕ_2, denoted together by $\bar{\phi}$. Although our target is to invert for source terms f, we also consider inversion of the diffusion coefficients μ. Among other unknowns in the model, the velocity field $\vec{v}(\zeta)$ is one of the more important sources of uncertainty. We do not try to infer it but instead assume a stochastic description is available where the velocity term is a function of a random variables ζ with an appropriate distribution. Our mathematical formulation for the inverse problem is given as follows:

$$\min_d \mathscr{J}(\phi,d) = \sigma\left[\frac{1}{2}\int_T\int_\Omega \left(\bar{\phi}-\phi^\star\right)^2 \delta(x-x^*,t-t^*)\mathrm{d}\Omega\mathrm{d}t - \frac{\beta}{2}\int_\Omega \|d\|^2\right]$$

$$\text{where } \phi \text{ solves } F(\phi,d) = \frac{\partial\phi}{\partial t} - \nabla\cdot\mu(x))\nabla\phi) + \vec{v}(\zeta)\cdot\nabla\phi - r(\phi) - f(x) = 0,$$

σ is the risk measure, β is the Tikhonov regularization parameter, which controls the magnitude of the penalty term and depends on the quantity and quality of data. σ is the risk measure, and in the case of a risk-neutral measure, it can be replaced with an expected value. The optimization

parameter vector d can be either μ or f. To solve this optimization problem, a trust-region method is used, with the use of a truncated conjugate-gradient method to solve the trust-region subproblem; the gradient required by this method is calculated using an adjoint approach, described in [124].

The gradient can be calculated by differentiating the objective function and making use of the chain rule:

$$\frac{D\mathscr{J}}{Dd} = \frac{\partial \mathscr{J}}{\partial \phi}\frac{\partial \phi}{\partial d} + \frac{\partial \mathscr{J}}{\partial d}.$$

Since ϕ and d are constrained by $F(\phi, d) = 0$, the direct sensitivity matrix can be expressed as

$$\frac{\partial \phi}{\partial d} = -\frac{\partial F}{\partial \phi}^{-1}\frac{\partial F}{\partial d},$$

which when placed in the gradient equation gives

$$\frac{D\mathscr{J}}{Dd} = -\frac{\partial \mathscr{J}}{\partial \phi}\frac{\partial F}{\partial \phi}^{-1}\frac{\partial F}{\partial d} + \frac{\partial \mathscr{J}}{\partial d}.$$

To avoid solving for the direct sensitivity matrix, which for n_ϕ states requires solving a linear system with Jacobian $\frac{\partial F}{\partial \phi}$ for each of the n_ϕ columns of $\frac{\partial F}{\partial d}$, we reorder the calculation:

$$\frac{D\mathscr{J}}{Dd} = -\frac{\partial F}{\partial \phi}^{-T}\frac{\partial \mathscr{J}}{\partial \phi}\frac{\partial F}{\partial d} + \frac{\partial \mathscr{J}}{\partial d}.$$

where an adjoint solution arises:

$$\frac{\partial F}{\partial \phi}^{T}\lambda = \frac{\partial \mathscr{J}}{\partial \phi}^{T}.$$

The gradient can then be calculated:

$$\frac{D\mathscr{J}}{Dd} = -\lambda\frac{\partial F}{\partial d} + \frac{\partial \mathscr{J}}{\partial d}.$$

Risk Measures

The motivation for augmenting the objective function with a risk measure is to account for some model based uncertainty in an attempt to provide a robust solution. The risk measure is a concept from stochastic optimization and often applied to the management of financial portfolios. The risk measure allows for a mechanism to achieve a range of objectives given the unknown future of the economy. For uncertain market conditions a bank may use risk measures to decide how much currency to keep in reserve, or a business may use them to decide how much to produce.

Generally, a risk measure is a mapping from a set of random variables to the real numbers. In actual applications the risk measure is applied to a probability distribution of losses. Given a loss distribution, the measure should encapsulate the risk associated with it. The choice of risk measures depends on what is considered risky; perhaps a risky investment is one with great variation in its possible returns, or perhaps it is one with very great loss expected in the worst case scenarios. In the case of inverse problems we consider, the "loss" as the observational mismatch combined with the regularization terms.

Risk measures from the financial world are not all easily mapped to engineering applications but there are a few common ones that can be described in the context of engineering targets, including expected value, standard deviation or variance, and Conditional Value-at-Risk (CVaR), The expected value risk measure for function of a random variable $f(X)$, where X has probability distribution $\rho(X)$, is

$$\sigma_{EV}(f(X)) = \mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(X)\rho(X)\mathrm{d}X.$$

In finance, the expected value is considered "risk-neutral"; a decision maker who is risk neutral cares only about the expected returns or losses. In engineering, the expected value is used when creating a design that is robust to uncertainties[125] and has a favorable mean response. In our context, the loss distribution would correspond to a distribution of penalized errors, composed from observation mismatch and regularization penalties; to minimize the expected loss would be to find a parameter estimate that is robust to uncertainties in the model form.

If neutral to risk is not desired, one might use the standard deviation and variance risk measures, which are defined

$$\sigma_{SD}(f(X)) = \mathbb{E}[f(X)] + w\sqrt[p]{\int_{-\infty}^{\infty} |f(X) - \mathbb{E}[f(X)]|^p \rho(X)\mathrm{d}X},$$

$$\sigma_{V}(f(X)) = \mathbb{E}[f(X)] + w\int_{-\infty}^{\infty} |f(X) - \mathbb{E}[f(X)]|^p \rho(X)\mathrm{d}X,$$

where usually $p = 2$. These risk measures are used with the assumption that the higher the variance of a variable, such as a portfolio, the more risky. Semideviation and semivariance risk measures, defined by

$$\sigma_{SemiD}(f(X)) = \mathbb{E}[f(X)] + w\sqrt[p]{\int_{-\infty}^{\infty} (\max\{0, f(X) - \mathbb{E}[f(X)]\})^p \rho(X)\mathrm{d}X},$$

$$\sigma_{SemiV}(f(X)) = \mathbb{E}[f(X)] + w\int_{-\infty}^{\infty} (\max\{0, f(X) - \mathbb{E}[f(X)]\})^p \rho(X)\mathrm{d}X,$$

can be used if only deviation towards the worse side (in the definition above, the more positive side) of the mean is considered risky. For a portfolio, these risk measures might represent a trade-off between expected returns and the risk one associates with the uncertainty in these returns; in engineering, they might represent a tradeoff between expected performance and uncertainty in the performance that can actually be achieved. The choice of weighting, however, reflects one's own personal aversion to variability.

If one is instead concerned about worst-case scenarios, then one might use the CVaR risk meansure. For a chosen confidence level $0 \leq \alpha \leq 1$ and continuous distribution $\rho(X)$, the CVaR risk measure is defined by

$$\sigma_{CVaR}(f(X),\alpha) = \frac{1}{1-\alpha}\int_{Q_\alpha}^{\infty} f(X)\rho(X)\mathrm{d}X,$$

where $Q_\alpha = \sup\{x \in \mathbb{R} | P(X \leq x) \leq \alpha\}$ is the α-quantile. The CVaR risk measure represents the expected loss in the worst $100(1-\alpha)\%$ of the distribution, and as a mean is less sensitive to sampling error than the quantile risk measure Q_α by itself.[126] The CVaR risk measure is related to that used in reliability-based design formulations, in which one wishes to minimize the probability of exceeding a certain threshold. When there is less of a clear line between an acceptable outcome and catastrophic failure, the CVaR risk measure can be used instead to minimize the expected outcome of the worst-case scenarios. What value of α is chosen to represent the most extreme outcomes is again a matter of one's personal degree of aversion to this risk.

In this study, the expected value risk measure is chosen to obtain a parameter estimate that is robust to uncertainties in the model form and least dependent on any personal preference for or against risk.

5.2 Implementation

To solve the optimization problem, we use a trust-region method; the trust-region subproblem is solved using the truncated conjugate-gradient method. Each of the source coefficient parameters is bounded above and below to prevent the sources from becoming sinks and to keep the concentration states from becoming so large that the forward solve does not converge. The risk measure is evaluated by sampling the three-dimensional stochastic space using a seven-point sparse grid generated from the Gauss-Hermite quadrature rule.

The forward model PDE is solved using a Galerkin finite element method, with backwards Euler for timestepping and a Newton method to solve the nonlinear system at each timestep. Since a convection-dominated problem solved by the standard Galerkin finite element method can produce eroneous oscillatory solutions if the Peclet number is too high, the streamline upwind Petrov Galerkin (SUPG) method is used to stabilize the solution. For weight function w, the local residual for the weak form of the convection-diffusion-reaction problem is

$$R = \frac{\partial \phi}{\partial t} + \mu\nabla\phi\cdot\nabla w + (\vec{v}\cdot\nabla\phi)w - r(\phi)w - fw.$$

With SUPG stabilization, the local residual becomes

$$R = \frac{\partial \phi}{\partial t} + \mu\nabla\phi\cdot\nabla w + (\vec{v}\cdot\nabla\phi)w - r(\phi)w - fw + \tau(\vec{v}\cdot\nabla\phi - f)(\vec{v}\cdot\nabla w),$$

where

$$\tau = \left(\frac{C_1 k}{h^2} + \frac{C_2\|\vec{v}\|}{h}\right)^{-1}$$

and h is the size of the element with $C_1 = 4.0$ and $C_2 = 2.0$. The stabilized problem is no longer adjoint consistent, so taking the adjoint (transpose) of the discretized forward system (discretize-then-optimize) is no longer equivalent to discretizing the continuous adjoint system (optimize-then-discretize)[127]. The discretize-then-optimize approach was selected for ease of implementation and ultimately the optimization requires the discrete form of the adjoint.

5.3 Numerical Results

This section presents numerical experimentation results to demonstrate our algorithmic approach. We first consider the deterministic inverse problem when data are available from many sensors, for progressively more complex inferences, then add uncertainty to the model and reduce the number of sensors.

We start by presenting estimation results for diffusivity and source parameters, individually and simultaneously, for a linear convection-diffusion model with data from numerous sensors are given. In these cases, the data is sufficiently informative to accurately recover the true parameter values. The simultaneous estimation of diffusivity and source parameters is repeated for a nonlinear two-species convection-diffusion-reaction model, with data about either one or both species available. For this more complex physics, having a large amoung of data about just one species is not quite enough to recover all the source terms; this can be remedied by adding data about the second species.

We then consider the case where there is uncertainty in the model, and a simultaneous estimation of diffusivity and source parameters for a nonlinear two-species convection-diffusion-reaction model is again performed, for different levels of uncertainty; more uncertainty in the model resulted in parameter estimates that deviated further from the truth values. Lastly, source parameters are estimated given data from sparse sensors; it is shown that adding data about one species can help estimate the source terms of the other.

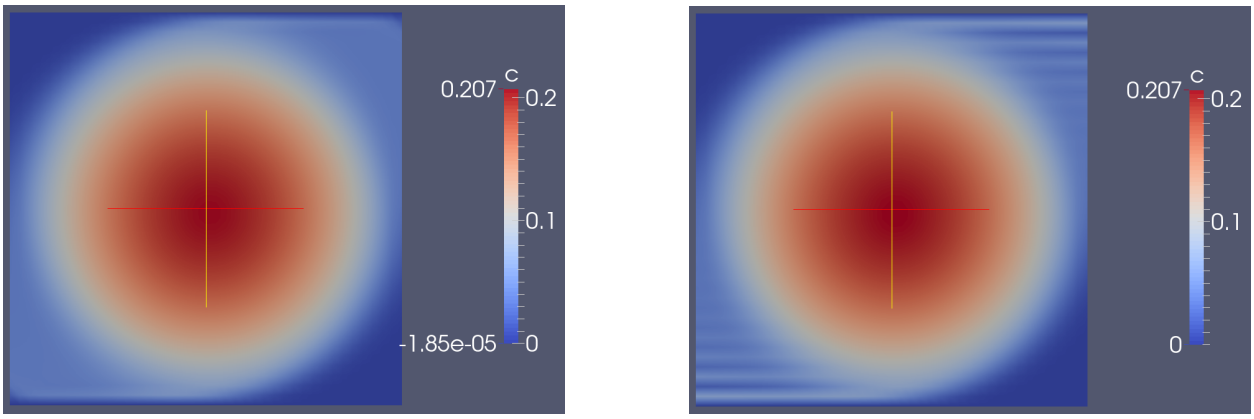
In all these cases, the computational domain is $\Omega = [0,1] \times [0,1]$ with zero intial conditions, discretized by 40×40 elements with linear nodal bases. To avoid an inversion crime, data is generated from a finer mesh 80×80 and then contaminated with Gaussian white noise.

Deterministic Convection-Diffusion

In this section, results are given for the estimation of various parameters given numerous sensors and a convection-diffusion model

$$\frac{\partial \phi}{\partial t} - \nabla \cdot (\mu \nabla \phi) + \vec{v} \cdot \nabla \phi = f$$

for a single species ϕ. The species is allowed to evolve over 32 timesteps from $t_0 = 0.0$ to $t_f = 1.0$ with homogeneous Dirichlet boundary conditions imposed along the top and bottom of the domain. Data is obtained from 64 sensors and placed in a square grid throughout the domain,



(a) with stabilization

(b) without stabilization

Figure 29. Final state, with and without SUPG

taking measurements every $\Delta t = 0.1875$. Since data were available from a large number of sensors at frequent measurements, it was qualitatively decided that regularization was unecessary for the estimation of a handful of parameters in this case with linear physics.

Diffusivity Coefficient Inversion

As an initial phase, data from 64 sensors is used to estimate just a single parameter: a constant diffusivity coefficient μ. The known source f is described by

$$f = 10\exp(-10((x-0.25)^2 + (y-0.25)^2)).$$

The known velocity $\vec{v} = (u, v)$ is an irrotational vortex described by

$$u = -1000(y-0.5),\ v = 1000(x-0.5).$$

For the given velocity field and element size, and the range of diffusivity coefficients considered, the Peclet number is high enough to warrent the use of a stabilization method to avoid oscillations in the simulated concentration field, as shown in Figure 29.

To avoid an inversion crime, Gaussian white noise with standard deviation $\sigma = 10^{-3}$ is added to the data. Although there is a large amount of data available for the estimation of a single parameter, the standard deviation of the noise is only an order of magnitude smaller than the pure measurements, and it would be expected that this relatively high noise level would interfere with what should otherwise be a very accurate estimate of the parameter. This expectation is borne out in the resulting parameter estimate which, as shown in Table 3, is much closer to the truth than the initial guess but not as close as might be expected from such a large amount of data.

Initial guess	Estimated	Truth
2.0	0.9826992	1.0

Table 3. Estimated diffusivity coefficient - convection-diffusion, no model uncertainty

Source Inversion

Next we consider a case where more parameters need to be estimated from the same number of data points; although the resulting optimization problem is simpler than the previous in that it is convex quadratic it is more complex because it is more inversion parameters.

In this case, the diffusivity $k = 1$ is known and the parameters $\vec{C}$ are to be estimated from an algebraic parameterization of the source terms:

$$\begin{aligned} f &= C_1 \exp(-20((x-0.15)^2+(y-0.85)^2)) + C_2 \exp(-20((x-0.5)^2+(y-0.85)^2)) \\ &\quad + C_3 \exp(-20((x-0.85)^2+(y-0.85)^2)) + C_4 \exp(-20((x-0.15)^2+(y-0.5)^2)) \\ &\quad + C_5 \exp(-20((x-0.5)^2+(y-0.5)^2)) + C_6 \exp(-20((x-0.85)^2+(y-0.5)^2)) \\ &\quad + C_7 \exp(-20((x-0.15)^2+(y-0.15)^2)) + C_8 \exp(-20((x-0.5)^2+(y-0.15)^2)) \\ &\quad + C_9 \exp(-20((x-0.85)^2+(y-0.15)^2)). \end{aligned}$$

The known velocity $\vec{v} = (u, v)$ is an irrotational vortex described by

$$u = -42(y-0.5),\ v = 42(x-0.5).$$

Again, Gaussian white noise with standard deviation $\sigma = 10^{-3}$ is added to the data, and the estimated source terms are shown in Table 4 and Figure 30. As in the previous case, the estimated parameters are much closer to the truth than the initial guesses were, but the noise level in the data limited the accuracy of the parameter inversion.

Simultaneous Source and Diffusivity Inversion

Here we consider an inference problem that combines the difficulties of the nonlinear optimality conditions of the first case with the larger parameter space of the second. Neither the diffusivity nor the source strengths are known. The diffusivity field is modeled as piecewise constant, with the first four parameters representing the diffusivty in four quadrants of the domain. The remaining five parameters describe the source term

$$\begin{aligned} f &= C_5 \exp(-20((x-0.5)^2+(y-0.75)^2)) + C_6 \exp(-20((x-0.75)^2+(y-0.75)^2)) \\ &\quad + C_7 \exp(-20((x-0.5)^2+(y-0.5)^2)) + C_8 \exp(-20((x-0.75)^2+(y-0.5)^2)) \\ &\quad + C_9 \exp(-20((x-0.25)^2+(y-0.25)^2)). \end{aligned}$$

Parameter	Initial guess	Estimated	Truth
C_1	2.0	0.9164218	1.0
C_2	2.0	0.9727920	1.0
C_3	2.0	0.9596691	1.0
C_4	2.0	1.037823	1.0
C_5	2.0	1.014904	1.0
C_6	2.0	1.037821	1.0
C_7	2.0	0.9596682	1.0
C_8	2.0	0.9727950	1.0
C_9	2.0	0.9164226	1.0

Table 4. Estimated source coefficients - convection-diffusion, no model uncertainty

The velocity field is the same as in the previous case. The standard deviation of the Gaussian white noise added to the data is reduced to $\sigma = 10^{-4}$, and the estimated parameters are shown in Table 5. Compared to the previous case, the accuracy of the inferred parameters is improved, reflecting the reduced noise level in the data.

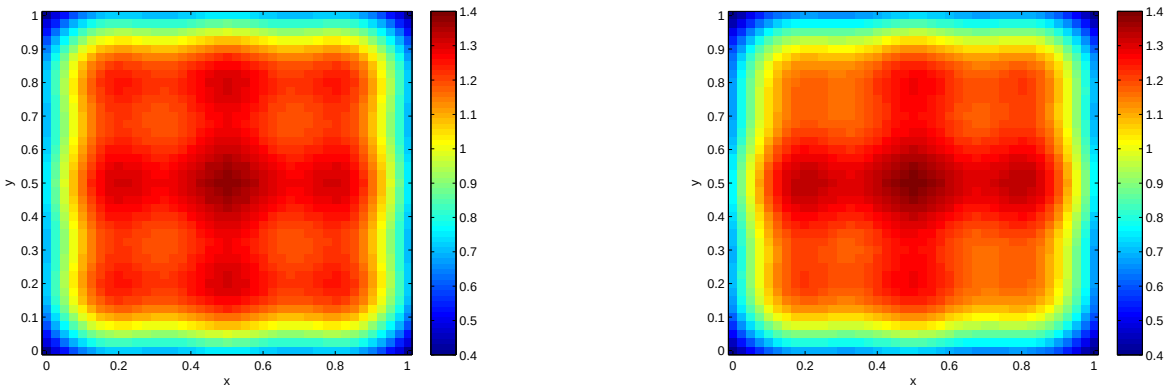
Parameter	Initial guess	Estimated	Truth
C_1	2.0	1.000415	1.0
C_2	2.0	0.9965526	1.0
C_3	2.0	1.002004	1.0
C_4	2.0	0.9990992	1.0
C_5	2.0	0.9859356	1.0
C_6	2.0	1.008036	1.0
C_7	2.0	1.007769	1.0
C_8	2.0	0.9917184	1.0
C_9	2.0	0.9968621	1.0

Table 5. Estimated parameters - convection-diffusion, no model uncertainty

5.4 Deterministic Convection-Diffusion-Reaction

In this case, diffusivity and source coefficients are again simultaneously estimated, but with a nonlinear two-species convection-diffusion-reaction model. The state equations are

$$\frac{\partial \phi_1}{\partial t} - \nabla \cdot (\mu \nabla \phi_1) + \vec{v} \cdot \nabla \phi_1 = \alpha \phi_2 + f_1$$



(a) true f **(b)** inferred f

Figure 30. True and inferred sources - convection-diffusion, no model uncertainty

$$\frac{\partial \phi_2}{\partial t} - \nabla \cdot (\mu \nabla \phi_2) + \vec{v} \cdot \nabla \phi_2 = \alpha \phi_1 + f_2$$

where ϕ_1 and ϕ_2 are the concentration states of the two species and $\alpha = 1.0$ is the reaction coefficient. The model is run from $t_0 = 0.0$ to $t_f = 1.0$ in 32 timesteps, and homogeneous Dirichlet boundary conditions are imposed along the top and bottom of the domain. There are nine parameters to estimate, the first being the constant diffusivity and the rest describing the source terms

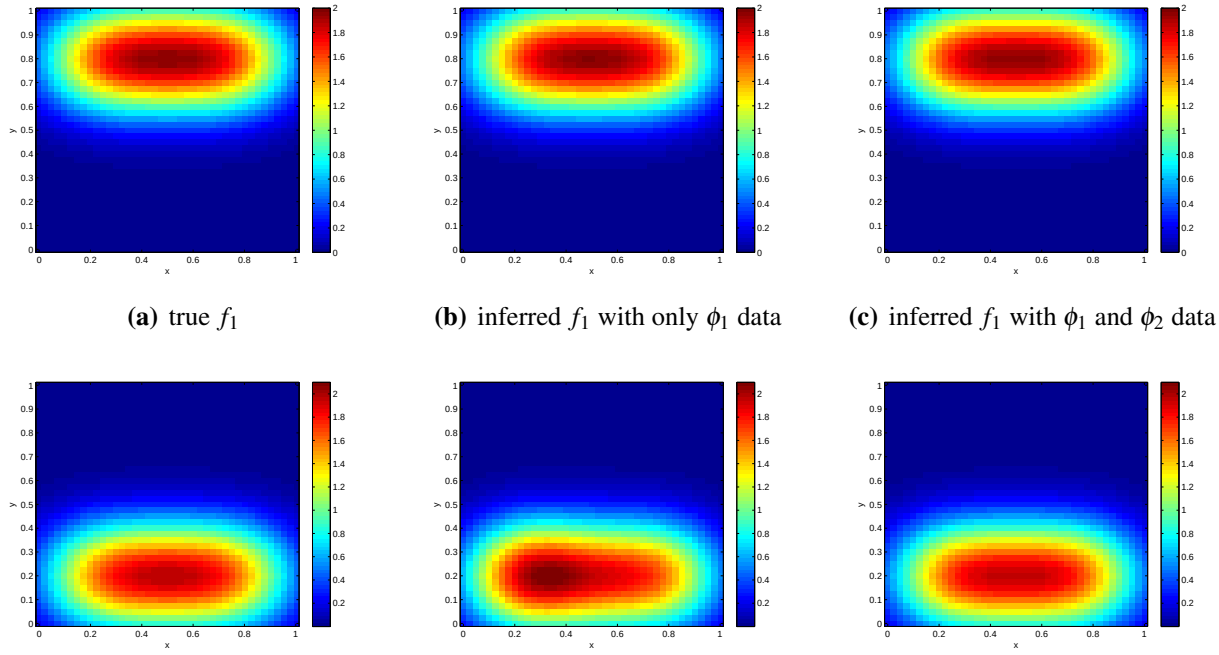
$$\begin{aligned} f_1 &= C_2 \exp(-20((x-0.2)^2+(y-0.8)^2)) + C_3 \exp(-20((x-0.4)^2+(y-0.8)^2)) \\ &\quad + C_4 \exp(-20((x-0.6)^2+(y-0.8)^2)) + C_5 \exp(-20((x-0.8)^2+(y-0.8)^2)) \\ f_2 &= + C_6 \exp(-20((x-0.2)^2+(y-0.2)^2)) + C_7 \exp(-20((x-0.4)^2+(y-0.2)^2)) \\ &\quad + C_8 \exp(-20((x-0.6)^2+(y-0.2)^2)) + C_9 \exp(-20((x-0.8)^2+(y-0.2)^2)). \end{aligned}$$

The known velocity $\vec{v} = (u, v)$ is an irrotational vortex described by

$$u = -42(y-0.5),\ v = 42(x-0.5),$$

and since the diffusivity varies as the parameters space is explored, SUPG stabilization is used to avoid possible oscillations. Data contaminated with Gaussian white noise with standard deviation $\sigma = 10^{-3}$ is taken from 64 sensors every $\Delta t = 0.1875$. The estimated parameters are shown in Table 6 and the estimated source terms are shown in Figure 31.

The data is sufficient to obtain close estimates of the diffusivity and source parameters for the first species, and although data of only the first species is available, the interaction of the two species through the reaction term, along with the large number of sensors, allows for a close estimate of two of the four source parameters for the second species as well. Of course, if each sensor could provide data for both species, the parameter estimate is much improved.



(a) true f_1 **(b)** inferred f_1 with only ϕ_1 data **(c)** inferred f_1 with ϕ_1 and ϕ_2 data

(d) true f_2 **(e)** inferred f_2 with only ϕ_1 data **(f)** inferred f_2 with ϕ_1 and ϕ_2 data

Figure 31. True and inferred sources - convection-diffusion-reaction, no model uncertainty

Parameter	Initial guess	Data from ϕ_1 only	Data from ϕ_1 and ϕ_2	Truth
C_1	2.0	1.001678	0.9952705	1.0
C_2	2.0	1.014302	1.035935	1.0
C_3	2.0	1.002739	0.9887638	1.0
C_4	2.0	0.9935702	0.9863631	1.0
C_5	2.0	1.014242	1.038581	1.0
C_6	2.0	1.359563	1.038581	1.0
C_7	2.0	1.022204	0.9863532	1.0
C_8	2.0	0.8746213	0.9887834	1.0
C_9	2.0	1.097095	1.035925	1.0

Table 6. Estimated parameters - convection-diffusion-reaction, no model uncertainty

5.5 Convection-Diffusion-Reaction with model uncertainty

In this section, uncertainty is added to the convection-diffusion-reaction model and the traditional deterministic objective function augmented with the expected value risk measure. First, a simultaneous estimation of diffusivity and source paramaters is performed with data from numerous sensors available and in the presence of different degrees of uncertainty. Then a case is examined in which source parameters are estimated given sparse sensors.

Numerous Sensors

The optimization problem can be formulated by:

$$\min_d \mathscr{J}(\phi, d) = \mathbb{E}\left[\frac{1}{2}\int_T \int_\Omega \left(\bar{\phi} - \phi^\star\right)^2 \delta(x - x^*, t - t^*) \mathrm{d}\Omega \mathrm{d}t - \frac{\beta}{2}\int_\Omega \|d\|^2\right]$$

where ϕ solves:

$$\frac{\partial \phi_1}{\partial t} - \nabla \cdot (\mu \nabla \phi_1) + \vec{v} \cdot \nabla \phi_1 = \alpha \phi_1^2 + f_1$$

$$\frac{\partial \phi_2}{\partial t} - \nabla \cdot (\mu \nabla \phi_2) + \vec{v} \cdot \nabla \phi_2 = \alpha \phi_1^2 + f_2$$

with $\alpha = 1.0$; the reaction term is based on the reaction rate $r = k[NO_2]^2$ of the reaction

$$CO_{(g)} + NO_{2(g)} \rightarrow CO_{2(g)} + NO_{(g)},$$

with $\phi_1 = [NO_2]$ and $\phi_2 = [CO]$. The same timesteps and boundary conditions are used as in the previous case. There are three parameters to estimate, the first being the constant diffusivity and the other two describing the source terms for the first and second species, respectively:

$$f_1 = C_2 \exp(-20((x-0.3)^2 + (y-0.5)^2))$$
$$f_2 = C_3 \exp(-20((x-0.7)^2 + (y-0.5)^2)).$$

The velocity field $\vec{v} = (u, v)$ is again an irrotational vortex, but there is uncertainty in its magnitude described by

$$u = -20\zeta(y-0.5),\ v = 20\zeta(x-0.5),$$

where $\zeta \sim \mathcal{N}(1, \sigma_\zeta^2)$ and the mean was used to generate the data. Data with Gaussian white noise with standard deviation $\sigma = 10^{-4}$ is available from 36 sensors taking measurements of both species every $\Delta t = 0.1875$, so no regularization is used. As in the previous case, since the diffusivity varies as the parameters space is explored, stabilization is used to avoid possible oscillations. The stochastic space was sampled using a sparse grid built form the Gauss-Hermite quadrature rules.

The parameter estimates obtained in the presence of an uncertain velocity field are shown in Table 7, for $\sigma_\zeta = 0.005$ and $\sigma_\zeta = 0.5$. The data is sufficient to obtain a good estimate the parameter values when the velocity field is known, but as the uncertainty in the velocity field increases, the estimates increasingly deviate from the truth, as would be expected.

Parameter	Initial guess	$\sigma_\zeta = 0.5$	$\sigma_\zeta = 0.005$	Truth
C_1	2.0	1.1582	1.02741	1.0
C_2	2.0	1.4644	1.33455	1.3
C_3	2.0	1.8022	1.63228	1.6

Table 7. Estimated parameters - convection-diffusion-reaction with model uncertainty

Sparse Sensors

The physics of interest are described by the convection-diffusion-reaction equations for two species since the diffusivity varies as the parameters space is explored.

$$\frac{\partial \phi_1}{\partial t} - \nabla \cdot (k \nabla \phi_1) + \vec{v} \cdot \nabla \phi_1 = \frac{\alpha}{2} \phi_1^2 \phi_2 + f_1$$

$$\frac{\partial \phi_2}{\partial t} - \nabla \cdot (k \nabla \phi_2) + \vec{v} \cdot \nabla \phi_2 = \alpha \phi_1^2 \phi_2 + f_2$$

where ϕ_1 and ϕ_2 are the concentration states of the two species, $k = 0.01$ is the diffusivity coefficient, and $\alpha = 2.0$ is the reaction coefficient. The reaction term is based on the reaction rate $r = \alpha[NO]^2[Cl_2]$ of the reaction

$$Cl_{2(g)} + 2NO_{(g)} \rightarrow 2NOCl_{(g)},$$

with $\phi_1 = [NO]$ and $\phi_2 = [Cl_2]$. Nitric oxide is a byproduct of combustion in the presence of nitrogen, which is the main component of air, and chlorine gas has commercial and industrial applications as a disinfectant and for water treatment.

The species concentrations are allowed to evolve over 32 timesteps from $t_0 = 0.0$ to $t_f = 1.0$ with natural boundary conditions. In this simple test case, we assume that we know the locations of the sources producing species 1, but wish to invert for their magnitudes; the source f_1 is described by

$$\begin{aligned} f_1 \;=\; & C_1 \exp(-20((x-0.1)^2 + (y-0.6)^2)) \\ & + C_2 \exp(-20((x-0.25)^2 + (y-0.7)^2)) \\ & + C_3 \exp(-20((x-0.5)^2 + (y-0.8)^2)) \\ & + C_4 \exp(-20((x-0.7)^2 + (y-0.85)^2)) \\ & + C_5 \exp(-20((x-0.8)^2 + (y-0.9)^2)), \end{aligned}$$

where $\vec{C} = (C_1, C_2, C_3, C_4, C_5)$ are the parameters we try to estimate from the data. The source f_2 also has a known location, but its magnitude is treated as an uncertainty in the model form rather than a parameter:

$$f_2 = \zeta_3 \exp(-10((x-0.6)^2 + (y-0.4)^2)),$$

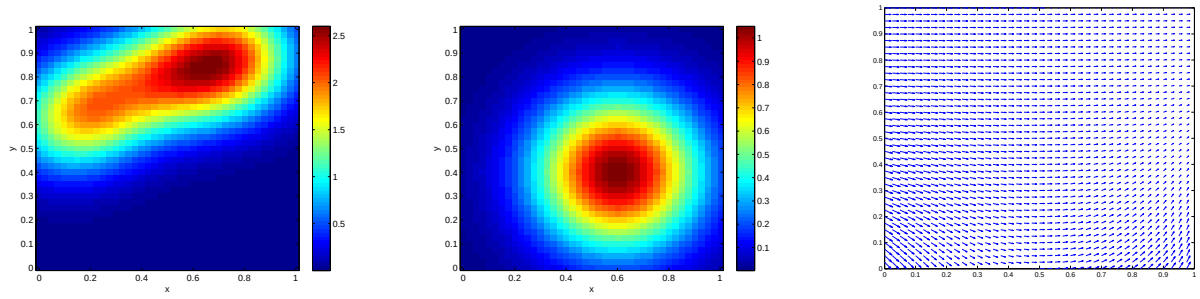
where $\zeta_3 \sim \mathcal{N}(1, (0.01)^2)$ is a random variable. The velocity field $\vec{v} = (u, v)$,

$$u = 1.0 + \zeta_1 \frac{y^2 - (x+0.5)^2}{y^2 + (x+0.5)^2} - \zeta_2 \frac{y^2 - (x-1.5)^2}{y^2 + (x-1.5)^2}$$

$$v = -\zeta_1 \frac{y^2 - (x+0.5)^2}{y^2 + (x+0.5)^2} + \zeta_2 \frac{y^2 - (x-1.5)^2}{y^2 + (x-1.5)^2},$$

is also a source of uncertainty in the model form, with $\zeta_1, \zeta_2 \sim \mathcal{N}(1, (0.1)^2)$. The "true" values of these random variables that are used to produce synthetic data are $\zeta^\star = (1.05, 1.05, 1.05)$; the true source coefficients are $C^\star = (1.0, 1.2, 1.4, 1.2, 1.0)$. The data is perturbed by normally distributed white noise with standard deviation $\sigma = 10^{-4}$. The true sources and velocity field are shown in Figure 32.

The data comes from two sensors placed at $(0.3, 0.2)$ and $(0.3, 0.7)$, one near the sources and one at a location that the first species was expected to be convected through, based on the mean velocity field; each sensor took measurements of ϕ_1 every $\Delta t = 0.1875$. Given the sparse sensors, Tikhonov regularization with $\beta = 10^{-4}$ is used. The resulting estimated source coefficients are compared with that obtained if additional data is available from a sensor at $(0.75, 0.25)$, taking



(a) true f_1 **(b)** true f_2 **(c)** true velocity

Figure 32. True sources and velocity field

Parameter	Initial guess	2 ϕ_1 sensors 0 ϕ_2 sensors	2 ϕ_1 sensors 1 ϕ_2 sensor	2 ϕ_1 sensors 3 ϕ_2 sensors	Truth
C_1	3.0	0.966013	0.990649	0.992926	1.0
C_2	3.0	1.31042	1.22505	1.22334	1.2
C_3	3.0	1.01494	1.27215	1.23333	1.4
C_4	3.0	0.065891	0.276508	0.93828	1.2
C_5	3.0	0.00815815	0.111768	0.700139	1.0

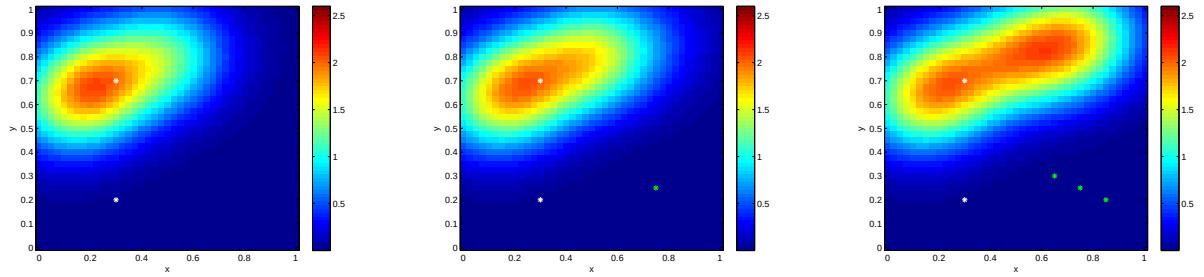
Table 8. Estimated source coefficients

measurements of ϕ_2 at the same timesteps; this additional sensor is located near the center of f_2 and thus where ϕ_2 was expected to be high.

The estimated source coefficient parameters are summarized in Table 8. Using only the measurements of ϕ_1 from two sensors gives an estimate of f_1 shown in Figure 33(a). The two sensors are only able to provide enough information for a fair estimate of the source components they are closest to; the ones further away are mostly informed by the regularization term. Using measurements of ϕ_2 from an additional sensor provides information on the state and thus source of the first species, improving the estimates of the source coefficients. The improved source estimate is shown in Figure 33(b). Including data from two more sensors of ϕ_2 at $(0.65, 0.3)$ and $(0.85, 0.2)$, also located near the center of f_2 and thus where ϕ_2 is expected to be higher and the reaction term larger, further improves the source estimate, shown in Figure 33(c).

5.6 Conclusions

We present the efficient solution of a parameter estimation problem that is robust to model uncertainties, taking advantage of stochastic optimization algorithms. Both the inversion of diffusivity



(a) two ϕ_1 sensors (white)

(b) two ϕ_1 sensors (white) and one ϕ_2 sensor (green)

(c) two ϕ_1 sensors (white) and three ϕ_2 sensors (green)

Figure 33. Inferred source - convection-diffusion-reaction, with model uncertainty

and source coefficients are investigated for numerous and sparse sensors, utilizing Thikonov regularization for the latter. Convection-diffusion-reaction physics with SUPG stabilization is used to evaluate the use of two reacting species in the presence of uncertainty in the velocity field. It is shown that additional information from another trace-gas species improves the reconstruction of the other trace-gas source term coefficient. Furthermore, robust inversion solutions are obtained in the face of incertainty, expoiting an expected value in the objective function to reflect a risk neutral measure.

6 Experimental Setup for GHG Observations

A mobile atmospheric monitoring facility (the ATML) has been deployed in Livermore, CA (Lon: -121.71°, Lat: 37.67°), where observations of various greenhouse gases, related gas phase tracers, and meteorological parameters have been ongoing since May of 2013. The primary measurements of interest are CO_2 and CH_4, measured by a Picarro $CO_2/CH_4/H_2O$ analyzer (Picarro, Model 1301), which has been operational virtually without interruption since May 2013. In support of these observations, a standard suite of meteorological measurements have been ongoing, a Vaisala ceilometer (CL51) has been measuring aerosol backscatter vertical profiles to determine mixing layer depth, and an Ecotech air quality analyzer has been measuring trace gases related to air quality and combustion emissions: CO, O_3, SO_2, and NO_x. Most recently, a proton transfer reaction mass spectrometer, PTR-MS, (Ionicon Analytik, PTR-QMS-HS) was deployed during the months of May and June, 2014 to measure a number of different volatile organic compounds in the atmosphere. NO_x The Picarro analyzer samples air from an inlet that extends to 27 m a.g.l. up a tower that is adjacent to the ATML. Air is continuously pumped through the inlet lines at a rate of 20 liters per minute and sub-sampled at approximately 0.4 SLPM into the Picarro ring down cavity. Prior to entering the Picarro, the air passes through a Nafion dryer (Perma Pure Corp.) that removes water vapor, thereby minimizing the impact of a water interference on the absorption lines of CO_2 and CH_4. Automated calibrations of the Picarro CO_2 and CH_4 signals have been conducted daily (every 23 hours) since late January 2014 and have indicated that the measurements are very stable over this time period (discussed in detail below). The calibrations are performed using whole air samples from pressurized cylinders that are referenced to the internationally recognized NOAA/WMO scales for both CO_2 and CH_4. Three different whole air samples with varying CH_4 and CO_2 concentrations are used to define a calibration slope and intercept, which is then applied to bias-correct the raw data. Water vapor (for CO_2 and CH_4) and isotopic corrections (for CO_2 only) are applied to the data, post-calibration, according to Chen *et.al.* [128] and Nara *et.al.* [129].

Figure 34 shows the deviation from the mean bias correction for each calibration performed through June 26, 2014, providing an indication of the stability of the measurements. The bias is defined as the resulting correction applied to a 400 ppm CO_2 sample and a 1.97 ppm CH_4 sample based on a given calibration slope and intercept of a three point linear regression. The variability in the bias can provide an indication of the pre-calibration precision of the measurements. The shaded areas in the figure show the 1σ (dark grey) and 2σ (light grey) standard deviations in the measured bias from the mean for the ensemble of calibrations. For CO_2 the 1σ stability is 0.031 ppm and for CH_4 it is 0.23 ppb. There are trends apparent in the bias, however, especially for CH_4, which declines relatively sharply during the month of May 2014, underscoring the importance of performing regular calibrations to catch such drifts in instrument performance and to minimize as much as possible the uncertainty due to instrumental drifts.

The mean bias during the 5 months when calibrations were performed was applied across the full data set (May 2013 to June 2014). Unless significant instrumental drifts occurred prior to January 2014, the variability in the instrument bias provides limits on the uncertainty resulting from the lack of calibrations for the first 8 months of the data period, which are small relative to

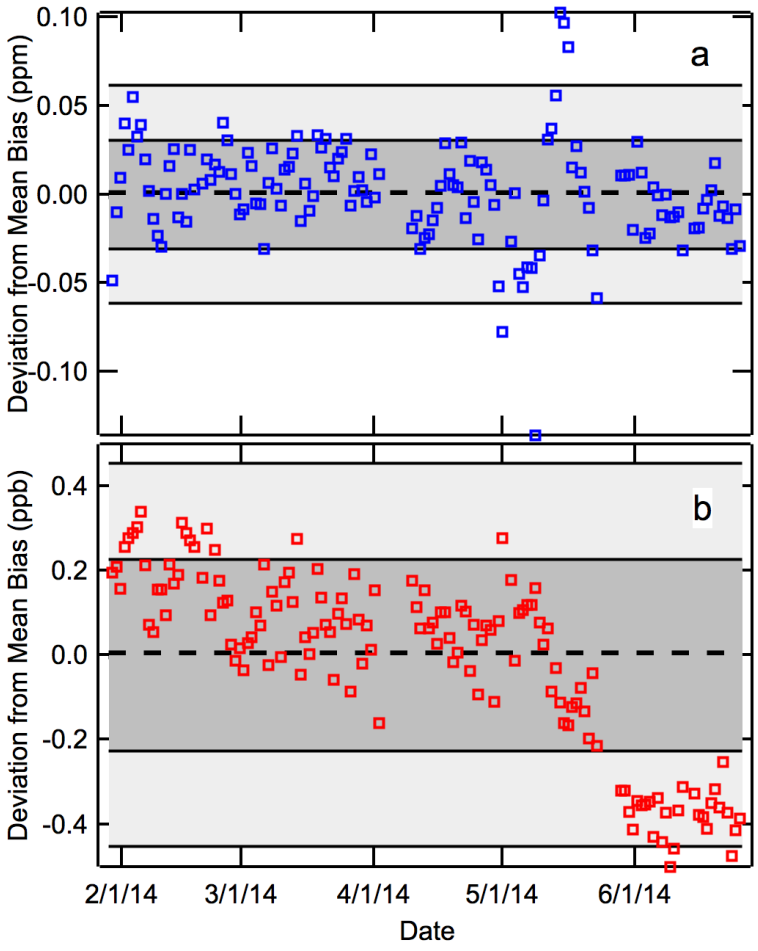


Figure 34. Deviation from the mean bias correction for each calibration performed through June 26, 2014, for (a) CO_2 and (b) CH_4.

other likely sources of uncertainty, such as pressure broadening, isotope corrections, and water corrections [128, 129, 130]. For example, prior to the inclusion of the Nafion drier (late December, 2013), the water vapor concentrations typically ranged from 0.5-1.5%, where uncertainties in the H_2O correction translate to uncertainties of 0.05 ppm and 0.5 ppb for CO_2 and CH_4, respectively [128]. As calibrations continue to be processed the instrumental drift will continue to be evaluated to provide better constraints on the uncertainty for data collected before calibrations commenced.

The time series for all CO_2 and CH_4 data collected through late June 2014 are shown in Fig. 35 as hourly averages. The CO_2 time series shows significant variability due to local emissions on top of a seasonally varying hemispheric background characterized by uptake during the summer and release during the winter. CH_4 also shows strong variability due to local sources on top of a seasonally varying background that is influenced primarily by a summertime photochemical sink. The seasonal variation of CH_4 is similar in phase to CO_2 but different in amplitude. Our interest is in the interpretation of the variability due to local emissions and uptake of these two gases, therefore it is useful to express the CO_2 and CH_4 mole fractions as excess values (ΔCO_2 and ΔCH_4) where the observations from a clean air background site are subtracted. In this case we use observations from aircraft flying in the free troposphere over Trinidad Head, CA, thereby removing variability in the data set due to hemispheric scale processes so that we may focus on interpreting the impact of local processes on the observations.

A significant portion of the variability in ΔCO_2 and ΔCH_4 is due to boundary layer dynamics operating on diurnal, synoptic, and seasonal scales. During so-called stable boundary layer conditions, typical at nighttime and during the winter months, mixing is weak because of minimal convective heating at the surface, and the observations are more sensitive to local emissions. On the other hand, during the daytime and the summer months, surface heating results in strong convective mixing and deeper boundary layers, and concentrations are lower for equivalent emission rates. The seasonality of this effect can be seen in the time series of CO_2 and CH_4, both of which exhibit much higher variability on average during the winter months than during the summer, more so than would be expected from shifts in the background concentrations alone. It can also be seen on diurnal scales, as shown in Fig. 36, which shows the diurnal averages for the entire data set for both ΔCO_2 and ΔCH_4. On average, concentrations are higher during the nighttime and early morning hours than during the afternoons, and this effect is expected because of stronger convective mixing during the daytime. Synoptic scale processes are harder to discern, however, and this results in day-to-day and week-to-week variability in the mixing dynamics that is much more difficult to predict. This motivates the use of our ceilometer observations to provide real-time information about mixing layer dynamics from vertical profiles of aerosol backscatter. Fig. 37 shows an example of the aerosol backscatter for a typical summertime day and the inferred mixing layer depth that is calculated from an algorithm that identifies sharp gradients in aerosol backscatter as a function of altitude. Mixing layer depth can also be simulated using the Weather Research and Forecasting model, and the ceilometer observations provide an observational constraint for the model. Knowledge of the mixing layer height allows us to appropriately analyze the CO_2 and CH_4 data set and discern variability due to mixing dynamics from variability due to changes in local emissions.

Another useful strategy for interpreting variability in the observations to provide information

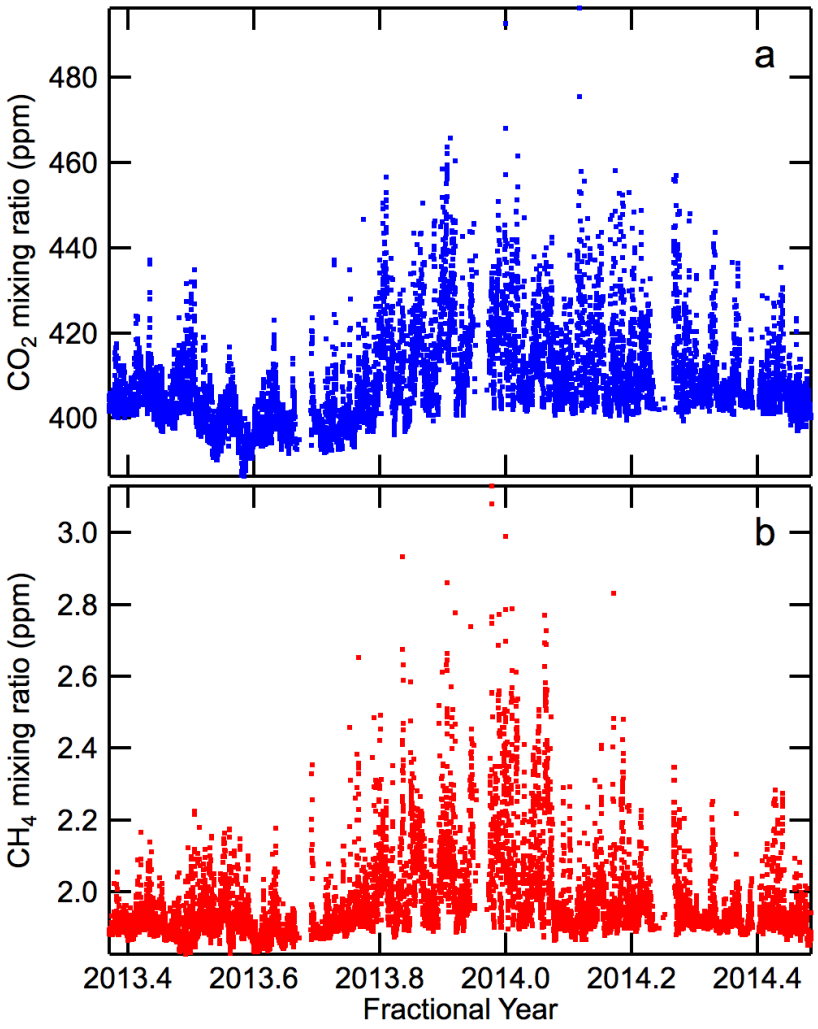


Figure 35. Concentrations time series for (a) CO_2 and (b) CH_4 data collected through late June 2014.

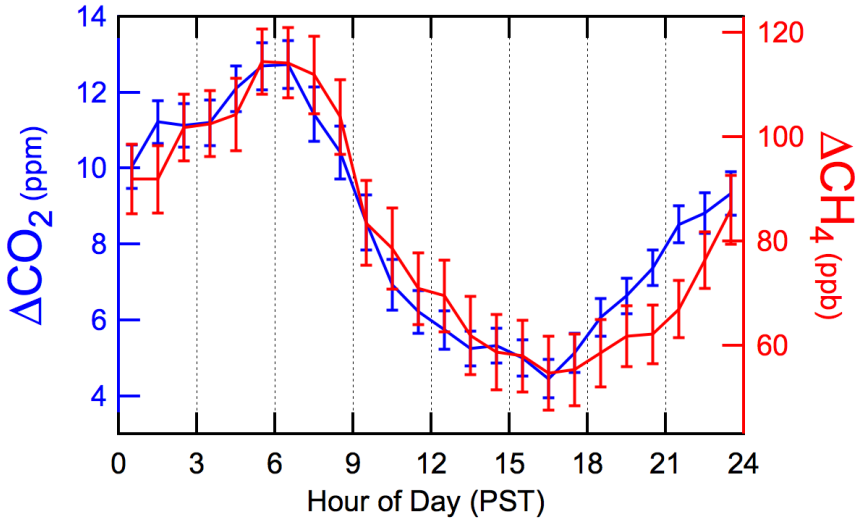


Figure 36. Diurnal averages for the entire data set for both ΔCO_2 and ΔCH_4.

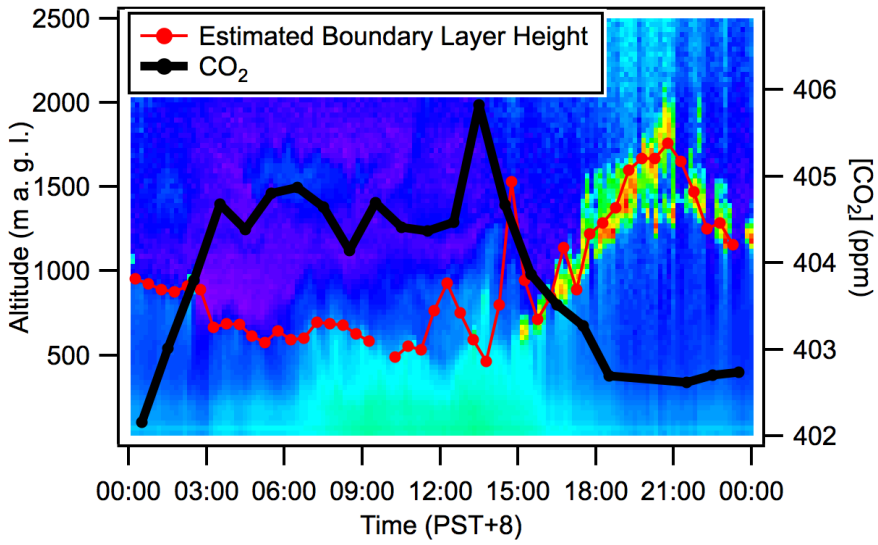


Figure 37. An example of the aerosol backscatter for a typical summertime day and the inferred mixing layer depth.

on local emission sources is to study tracer-tracer relationships between CO_2 and CH_4 and other trace gases of interest, such as CO, NO_x, and VOCs. This approach tends to minimize the impact of changing mixing dynamics because the tracers are subject to the same physical processes in the atmosphere. Thus, the ratio of two different tracers will be driven primarily by changes in the type of emissions sampled rather than mixing. For example, CO and NO_x are both primarily emitted by on-road vehicles, with some contribution from off-road vehicles. CO is emitted in larger amounts by gasoline vehicles while NO_x is emitted more so from diesel vehicles. CO_2, of course, is emitted significantly from on-road and off-road vehicles, but also from a wide range of other combustion sectors (utility generation, residential and commercial heating, etc). The behavior of CO and NO_x, relative to CO_2, should reflect the differences in timing of on-road vehicle traffic, with respect to the other combustion sources that emit CO_2. It is instructive, therefore, to examine the diurnal behavior of CO and NO_x in Fig. 38 along with that of CO_2 in Fig. 35. In contrast to the CO_2 and CH_4 plots, the CO and NO_x data used in the diurnal profile is from October-November 2013 only, so there may be some seasonal variability that is not reflected in these data. Nevertheless, the diurnal behavior for CO and NO_x is qualitatively different from that of CO_2, and the opposite of what would be expected by a constant emissions source and a diurnally varying mixing layer height. In contrast to CO_2 and CH_4, CO and NO_x are higher in concentration on average during the daytime hours, peaking during the morning and evening rush hours, indicating that the timing of on-road emission sources is a strong factor in driving the variability of these tracers, and that boundary layer dynamics have less of an impact. The primary CH_4 and CO_2 sources in the region, therefore, appear to be decoupled from traffic patterns, suggesting that the transportation sector is not the primary component of the sources contributing to both gases. As an extension of this analysis, the PTR-MS measurements will allow for analogous studies of a whole suite of VOC tracers in a multi-variable regression analysis to fingerprint more specific sources of CO_2 and CH_4.

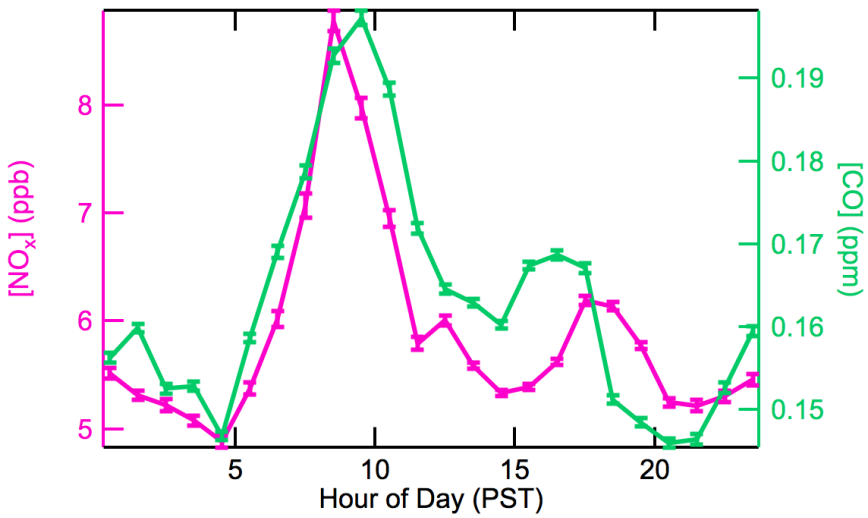


Figure 38. Diurnal behavior of CO and NO_x from October-November 2013.

The sampling infrastructure developed as part of this project has enabled multiple collaborations with other institutions interested in the general area of greenhouse gas attribution studies, and who can bring a wide variety of tools to approach the problem. One such collaboration with Lawrence Livermore National Lab and University of California at Merced is ongoing and brings analysis and modeling capabilities in radiocarbon in CO_2 ($^{14}CO_2$) and carbonyl sulfide (COS) in order to partition the observed CO_2 mole fraction into its fossil and biospheric components. Another collaboration involves a statewide effort lead by researchers at Imperial College London/Scripps Institute of Oceanography in which our Livermore site is part of a network of towers across California where an identical set of observations are being made, including CO, CO_2, CH_4, and $^{14}CO_2$. The goal of this study is to provide improvements to emissions inventories in CA for fossil CO_2 and CH_4 using the observations within an inversion framework.

7 Conclusion

In this project, we have developed atmospheric measurement capabilities and a suite of atmospheric modeling and analysis tools that are ready to be used for verifying emissions of greenhouse gases (GHGs) on an urban-through-regional scale.

- We have for the first time applied the Community Multiscale Air Quality (CMAQ) model to simulate atmospheric CO_2. This will allow for the examination of regional-scale transport and distribution of CO_2 along with those traditionally studied air pollutants at relatively high spatial and temporal resolution, with the goal of leveraging emissions verification efforts for air quality and climate.

- We have developed a bias-enhanced Bayesian inference approach that can remedy the well-known problem of transport model errors in atmospheric CO_2 inversions. We have tested the approach using the well-documented data and model outputs from the TransCom3 global CO_2 inversion comparison project.

- We have also performed two prototyping studies on inversion approaches in the generalized convection-diffusion context. One of these approaches explored the use of Polynomial Chaos Expansion in accelerating the evaluation of a regional transport model to enable efficient Markov Chain Monte Carlo sampling of the posterior for Bayesian inference. The other approach aims at a deterministic inversion of convection-diffusion-reaction system at the presence of uncertainty. These approaches should in principle be applied to realistic atmospheric problems with moderate adaptation.

- We outline a regional greenhouse gas source inference system. As shown in Figure 39, the system integrates (1) two approaches of atmospheric dispersion simulation and (2) a class of Bayesian inference and uncertainty quantification algorithms. We use two different and in principle complementary approaches to simulate atmospheric dispersion. Specifically, a Eulerian chemical transport model CMAQ and a Lagrangian Particle Dispersion Model - FLEXPART-WRF are used. These two models share the same WRF assimilated meteorology fields, making it possible to perform a hybrid simulation, in which the Eulerian model (CMAQ) can be used to compute the initial condition needed by the Lagrangian model, while the source-receptor relationships for a large state vector can be efficiently computed using the Lagrangian model in its backward mode. In addition, CMAQ has a complete treatment of atmospheric chemistry of a suite of traditional air pollutants, many of which could help attribute GHGs from different sources. The inference of emissions sources using atmospheric observations is casted as a Bayesian model calibration problem, which is solved using a class of Bayesian inference techniques, such as the bias-enhanced Bayesian inference algorithm that account for the intrinsic model deficiency, Polynomial Chaos Expansion to accelerate model evaluation and Markov Chain Monte Carlo sampling, and Karhunen-Loève (KL) Expansion to reduce the dimensionality of the state space.

- We have established an atmospheric measurement site in Livermore, CA and are collecting continuous measurements of CO_2, CH_4 and other species that are typically co-emitted

with these GHGs. Measurements of co-emitted species can assist in attributing the GHGs to different emissions sectors. Automatic calibrations using traceable standards are performed routinely for the gas-phase measurements. We are also collecting standard meteorological data at the Livermore site as well as planetary boundary height measurements using a ceilometer. The location of the measurement site is well suited to sample air transported between the San Francisco Bay area and the California Central Valley.

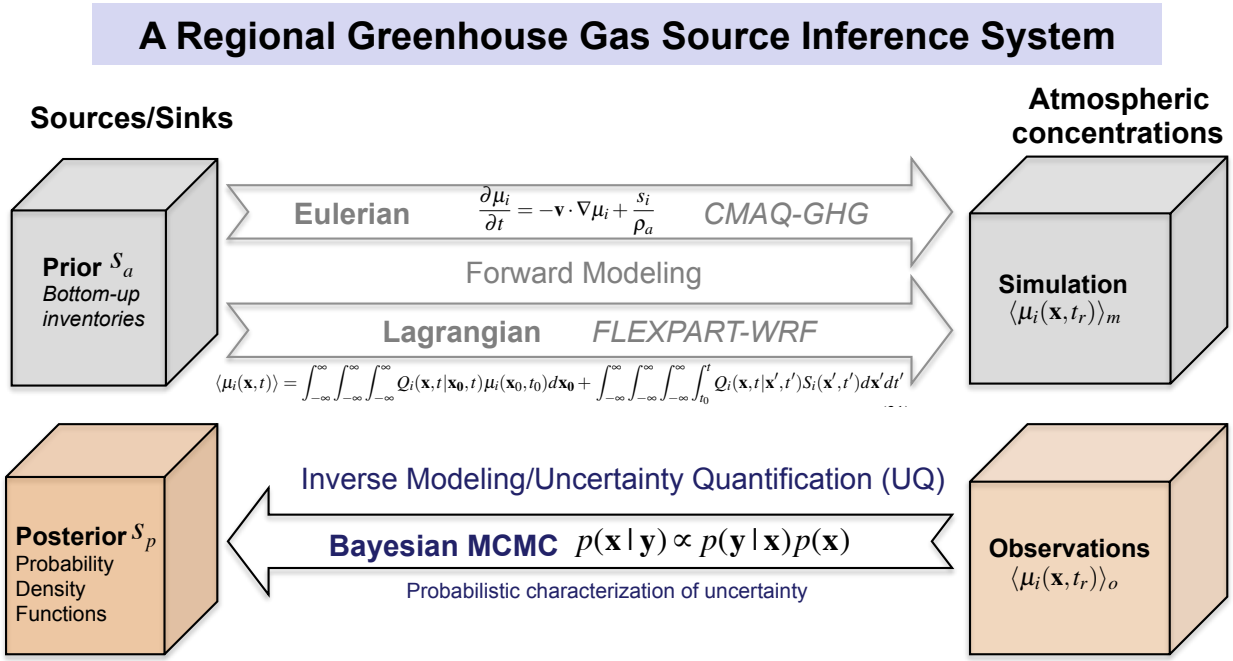


Figure 39. Schematic of the Regional GHGs Source Inference System

References

[1] Josep G. Canadell, Corinne Le Qur, Michael R. Raupach, Christopher B. Field, Erik T. Buitenhuis, Philippe Ciais, Thomas J. Conway, Nathan P. Gillett, R. A. Houghton, and Gregg Marland. Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proceedings of the National Academy of Sciences*, 104(47):18866–18870, 2007.

[2] Committee on Methods for Estimating Greenhouse Gas Emissions; National Research Council. *Verifying Greenhouse Gas Emissions:Methods to Support International Climate Agreements*. The National Academies Press, 2010.

[3] R. J. Andres, T. A. Boden, F.-M. Bréon, P. Ciais, S. Davis, D. Erickson, J. S. Gregg, A. Jacobson, G. Marland, J. Miller, T. Oda, J. G. J. Olivier, M. R. Raupach, P. Rayner, and K. Treanton. A synthesis of carbon dioxide emissions from fossil-fuel combustion. *Biogeosciences*, 9(5):1845–1871, 2012.

[4] D. Guan, Z. Liu, Y. Geng, S. Lindner, and K. Hubacek. The gigatonne gap in china/'s carbon dioxide inventories. *Nature Clim. Change*, 2:672–675, 2012.

[5] Gregg Marland. Uncertainties in Accounting for CO2 From Fossil Fuels. *Journal of Industrial Ecology*, 12(2):136–139, 2008.

[6] V. Matthias, I. Bewersdorff, A. Aulinger, and M. Quante. The contribution of ship emissions to air pollution in the north sea regions. *Environmental Pollution*, 158(6):2241 – 2250, 2010.

[7] Corinne Le Quere, Michael R. Raupach, Josep G. Canadell, and Gregg Marland et al. Trends in the sources and sinks of carbon dioxide. *Nature Geosci*, 2(12):831–836, Dec 2009.

[8] T.A. Boden, G. Marland, and R. J. Andres. Global, Regional, and National Fossil-Fuel CO2 Emissions. Technical report, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A., 2011.

[9] Kevin R. Gurney, Daniel L. Mendoza, Yuyu Zhou, Marc L. Fischer, Chris C. Miller, Sarath Geethakumar, and Stephane de la Rue du Can. High Resolution Fossil Fuel Combustion CO2 Emission Fluxes for the United States. *Environmental Science & Technology*, 43(14):5535–5541, 2009.

[10] P. Peylin, S. Houweling, M. C. Krol, U. Karstens, C. Rödenbeck, C. Geels, A. Vermeulen, B. Badawy, C. Aulagnier, T. Pregger, F. Delage, G. Pieterse, P. Ciais, and M. Heimann. Importance of fossil fuel emission uncertainties over europe for co_2 modeling: model intercomparison. *Atmospheric Chemistry and Physics*, 11(13):6607–6622, 2011.

[11] Ray Nassar, Louis Napier-Linton, Kevin R. Gurney, Robert J. Andres, Tomohiro Oda, Felix R. Vogel, and Feng Deng. Improving the temporal and spatial distribution of CO2 emissions from global fossil fuel emission data sets. *Journal of Geophysical Research: Atmospheres*, 118(2):917–933, 2013.

[12] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

[13] Kevin Robert Gurney, Rachel M. Law, A. Scott Denning, Peter J. Rayner, David Baker, Philippe Bousquet, Lori Bruhwiler, Yu-Han Chen, Philippe Ciais, Songmiao Fan, Inez Y. Fung, Manuel Gloor, Martin Heimann, Kaz Higuchi, Jasmin John, Takashi Maki, Shamil Maksyutov, Ken Masarie, Philippe Peylin, Michael Prather, Bernard C. Pak, James Randerson, Jorge Sarmiento, Shoichi Taguchi, Taro Takahashi, and Chiu-Wai Yuen. Towards robust regional estimates of CO2 sources and sinks using atmospheric transport models. *Nature*, 415(6872):626–630, Feb 2002.

[14] Michael J. Prather, Xin Zhu, Susan E. Strahan, Stephen D. Steenrod, and Jose M. Rodriguez. Quantifying errors in trace species transport modeling. *Proceedings of the National Academy of Sciences*, 105(50):19617–19621, 2008.

[15] D. Wunch, P. O. Wennberg, G. C. Toon, G. Keppel-Aleks, and Y. G. Yavin. Emissions of greenhouse gases from a north american megacity. *Geophysical Research Letters*, 36(15), 2009.

[16] Kathryn McKain, Steven C. Wofsy, Thomas Nehrkorn, Janusz Eluszkiewicz, James R. Ehleringer, and Britton B. Stephens. Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region. *Proceedings of the National Academy of Sciences*, 109(22):8423–8428, 2012.

[17] Kelly L. Mays, Paul B. Shepson, Brian H. Stirm, Anna Karion, Colm Sweeney, and Kevin R. Gurney. Aircraft-based measurements of the carbon footprint of indianapolis. *Environmental Science & Technology*, 43(20):7816–7823, 2009.

[18] Eric A. Kort, Christian Frankenberg, Charles E. Miller, and Tom Oda. Space-based observations of megacity carbon dioxide. *Geophysical Research Letters*, 39(17), 2012.

[19] J. C. Turnbull, J. B. Miller, S. J. Lehman, P. P. Tans, R. J. Sparks, and J. Southon. Comparison of 14CO2, CO, and SF6 as tracers for recently added fossil fuel CO2 in the atmosphere and implications for biological CO2 exchange. *Geophysical Research Letters*, 33(1), 2006.

[20] J. Brioude, G. Petron, G. J. Frost, R. Ahmadov, W. M. Angevine, E.-Y. Hsie, S.-W. Kim, S.-H. Lee, S. A. McKeen, M. Trainer, F. C. Fehsenfeld, J. S. Holloway, J. Peischl, T. B. Ryerson, and K. R. Gurney. A new inversion method to calculate emission inventories without a prior at mesoscale: Application to the anthropogenic CO2 emission from Houston, Texas. *Journal of Geophysical Research: Atmospheres*, 117(D5), 2012.

[21] Paul I. Palmer, Parvadha Suntharalingam, Dylan B. A. Jones, Daniel J. Jacob, David G. Streets, Qingyan Fu, Stephanie A. Vay, and Glen W. Sachse. Using CO2:CO correlations to improve inverse analyses of carbon fluxes. *Journal of Geophysical Research: Atmospheres*, 111(D12), 2006.

[22] L. Rivier, P. Ciais, D. A. Hauglustaine, P. Bakwin, P. Bousquet, P. Peylin, and A. Klonecki. Evaluation of SF6, C2Cl4, and CO to approximate fossil fuel CO2 in the Northern Hemisphere using a chemistry transport model. *Journal of Geophysical Research: Atmospheres*, 111(D16), 2006.

[23] P. P. Tans, I. Y. Fung, and T. Takahashi. Observational Constraints on the Global Atmospheric CO_2 Budget. *Science*, 247(4949):1431–1438, 1990.

[24] M. Buchwitz, R. de Beek, J. P. Burrows, H. Bovensmann, T. Warneke, J. Notholt, J. F. Meirink, A. P. H. Goede, P. Bergamaschi, S. Körner, M. Heimann, and A. Schulz. Atmospheric methane and carbon dioxide from sciamachy satellite data: initial comparison with chemistry and transport models. *Atmospheric Chemistry and Physics*, 5(4):941–962, 2005.

[25] M. T. Chahine, Luke Chen, Paul Dimotakis, Xun Jiang, Qinbin Li, Edward T. Olsen, Thomas Pagano, James Randerson, and Yuk L. Yung. Satellite remote sounding of mid-tropospheric co2. *Geophysical Research Letters*, 35(17):n/a–n/a, 2008.

[26] S. S. Kulawik, D. B. A. Jones, R. Nassar, F. W. Irion, J. R. Worden, K. W. Bowman, T. Machida, H. Matsueda, Y. Sawa, S. C. Biraud, M. L. Fischer, and A. R. Jacobson. Characterization of tropospheric emission spectrometer (tes) co_2 for carbon cycle science. *Atmospheric Chemistry and Physics*, 10(12):5601–5623, 2010.

[27] T. Yokota, Y. Yoshida, N. Eguchi, Y. Ota, T. Tanaka, H. Watanabe, and S. Maksyutov. Global concentrations of co¡sub¿2¡/sub¿ and ch¡sub¿4¡/sub¿ retrieved from gosat: First preliminary results. *SOLA*, 5:160–163, 2009.

[28] P. J. Rayner, M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann. Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (ccdas). *Global Biogeochemical Cycles*, 19(2), 2005.

[29] R. G. Prinn. *Inverse Methods in Global Biogeochemical Cycles*, volume 114 of *Geophys. Monogr. Ser.*, chapter Measurement equation for trace chemicals in fluids and solution of its inverse. Washington, DC, 2000.

[30] P. Seibert and A. Frank. Source-receptor matrix calculation with a lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics*, 4(1):51–63, 2004.

[31] I. G. Enting. *Inverse Problems in Atmospheric Constituent Transport*. Cambridge University Press, 2002.

[32] Song-Miao Fan, Jorge L. Sarmiento, Manuel Gloor, and Stephen W. Pacala. On the use of regularization techniques in the inverse modeling of atmospheric carbon dioxide. *Journal of Geophysical Research: Atmospheres*, 104(D17):21503–21512, 1999.

[33] KEVIN ROBERT GURNEY, RACHEL M. LAW, A. SCOTT DENNING, PETER J. RAYNER, DAVID BAKER, PHILIPPE BOUSQUET, LORI BRUHWILER, YU-HAN CHEN, PHILIPPE CIAIS, SONGMIAO FAN, INEZ Y. FUNG, MANUEL GLOOR, MARTIN HEIMANN, KAZ HIGUCHI, JASMIN JOHN, EVA KOWALCZYK,

TAKASHI MAKI, SHAMIL MAKSYUTOV, PHILIPPE PEYLIN, MICHAEL PRATHER, BERNARD C. PAK, JORGE SARMIENTO, SHOICHI TAGUCHI, TARO TAKAHASHI, and CHIU-WAI YUEN. Transcom 3 co2 inversion intercomparison: 1. annual mean control results and sensitivity to transport and prior flux information. *Tellus B*, 55(2):555–579, 2003.

[34] Thomas Kaminski, Peter J. Rayner, Martin Heimann, and Ian G. Enting. On aggregation errors in atmospheric transport inversions. *Journal of Geophysical Research: Atmospheres*, 106(D5):4703–4715, 2001.

[35] W. Peters, J. B. Miller, J. Whitaker, A. S. Denning, A. Hirsch, M. C. Krol, D. Zupanski, L. Bruhwiler, and P. P. Tans. An ensemble data assimilation system to estimate co2 surface fluxes from atmospheric trace gas observations. *Journal of Geophysical Research: Atmospheres*, 110(D24), 2005.

[36] Abhishek Chatterjee, Anna M. Michalak, Jeffrey L. Anderson, Kim L. Mueller, and Vineet Yadav. Toward reliable ensemble kalman filter estimates of co2 fluxes. *Journal of Geophysical Research: Atmospheres*, 117(D22):n/a–n/a, 2012.

[37] L. F. Tolk, A. J. Dolman, A. G. C. A. Meesters, and W. Peters. A comparison of different inverse carbon flux estimation approaches for application on a regional domain. *Atmospheric Chemistry and Physics*, 11(20):10349–10365, 2011.

[38] C. D. Rodgers. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Publishing Company, Inc., 2000.

[39] Frdric Chevallier, Nicolas Viovy, Markus Reichstein, and Philippe Ciais. On the assignment of prior errors in bayesian inversions of co2 surface fluxes. *Geophysical Research Letters*, 33(13), 2006.

[40] Frdric Chevallier, Tao Wang, Philippe Ciais, Fabienne Maignan, Marc Bocquet, M. Altaf Arain, Alessandro Cescatti, Jiquan Chen, A. Johannes Dolman, Beverly E. Law, Hank A. Margolis, Leonardo Montagnani, and Eddy J. Moors. What eddy-covariance measurements tell us about prior land flux errors in co2-flux inversion schemes. *Global Biogeochemical Cycles*, 26(1), 2012.

[41] D. F. Baker, R. M. Law, K. R. Gurney, P. Rayner, P. Peylin, A. S. Denning, P. Bousquet, L. Bruhwiler, Y.-H. Chen, P. Ciais, I. Y. Fung, M. Heimann, J. John, T. Maki, S. Maksyutov, K. Masarie, M. Prather, B. Pak, S. Taguchi, and Z. Zhu. Transcom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional co2 fluxes, 19882003. *Global Biogeochemical Cycles*, 20(1), 2006.

[42] T. Lauvaux, O. Pannekoucke, C. Sarrat, F. Chevallier, P. Ciais, J. Noilhan, and P. J. Rayner. Structure of the transport uncertainty in mesoscale inversions of co2 sources and sinks using ensemble model simulations. *Biogeosciences*, 6(6):1089–1102, 2009.

[43] J. C. Lin and C. Gerbig. Accounting for the effect of transport errors on tracer inversions. *Geophysical Research Letters*, 32(1), 2005.

[44] A. A. Alkhaled, A. M. Michalak, and S. R. Kawa. Using co2 spatial variability to quantify representation errors of satellite co2 retrievals. *Geophysical Research Letters*, 35(16), 2008.

[45] I. G. Enting, P. J. Rayner, and P. Ciais. Carbon cycle uncertainty in regional carbon cycle assessment and processes (reccap). *Biogeosciences*, 9(8):2889–2904, 2012.

[46] Daniel M. Ricciuto, Kenneth J. Davis, and Klaus Keller. A bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty. *Global Biogeochemical Cycles*, 22(2), 2008.

[47] Andrew D. Richardson, Miguel D. Mahecha, Eva Falge, Jens Kattge, Antje M. Moffat, Dario Papale, Markus Reichstein, Vanessa J. Stauch, Bobby H. Braswell, Galina Churkina, Bart Kruijt, and David Y. Hollinger. Statistical properties of random CO2 flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, 148(1):38 – 50, 2008.

[48] Kevin Robert Gurney, Yu-Han Chen, Takashi Maki, S. Randy Kawa, Arlyn Andrews, and Zhengxin Zhu. Sensitivity of atmospheric co2 inversions to seasonal and interannual variations in fossil fuel emissions. *Journal of Geophysical Research: Atmospheres*, 110(D10), 2005.

[49] Marc Bocquet. Inverse modelling of atmospheric tracers: non-Gaussian methods and second-order sensitivity analysis. *Nonlinear Processes in Geophysics*, 15(1):127–143, 2008.

[50] Marc Bocquet, Carlos A. Pires, and Lin Wu. Beyond gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138(8):2997–3023, 2010.

[51] S. J. Fletcher and M. Zupanski. A data assimilation method for log-normally distributed observational errors. *Quarterly Journal of the Royal Meteorological Society*, 132(621):2505–2519, 2006.

[52] J. Brioude, S.-W. Kim, W. M. Angevine, G. J. Frost, S.-H. Lee, S. A. McKeen, M. Trainer, F. C. Fehsenfeld, J. S. Holloway, T. B. Ryerson, E. J. Williams, G. Petron, and J. D. Fast. Top-down estimate of anthropogenic emission inventories and their interannual variability in houston using a mesoscale inverse modeling technique. *Journal of Geophysical Research: Atmospheres*, 116(D20), 2011.

[53] MILIND KANDLIKAR. Bayesian inversion for reconciling uncertainties in global mass balances. *Tellus B*, 49(2):123–135, 1997.

[54] D. G. Partridge, J. A. Vrugt, P. Tunved, A. M. L. Ekman, H. Struthers, and A. Sorooshian. Inverse modelling of cloud-aerosol interactions part 2: Sensitivity tests on liquid phase clouds using a markov chain monte carlo based simulation approach. *Atmospheric Chemistry and Physics*, 12(6):2823–2847, 2012.

[55] S. M. Burrows, P. J. Rayner, T. Butler, and M. G. Lawrence. Estimating bacteria emissions from inversion of atmospheric transport: sensitivity to modelled particle characteristics. *Atmospheric Chemistry and Physics*, 13(11):5473–5488, 2013.

[56] D. Gamerman and H.F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2006.

[57] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[58] Malte Meinshausen, Nicolai Meinshausen, William Hare, Sarah C. B. Raper, Katja Frieler, Reto Knutti, David J. Frame, and Myles R. Allen. Greenhouse-gas emission targets for limiting global warming to 2 degrees C. *NATURE*, 458(7242):1158–U96, APR 30 2009.

[59] Kari Karhunen. ber lineare methoden in der wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37.

[60] M Loève. *Probability Theory I*. Comprehensive Manuals of Surgical Specialties. Springer, 1977.

[61] D.S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press. Academic Press, 2011.

[62] R. Zhuravlev, B. Khattatov, B. Kiryushov, and S. Maksyutov. Technical note: A novel approach to estimation of time-variable surface sources and sinks of carbon dioxide using empirical orthogonal functions and the kalman filter. *Atmospheric Chemistry and Physics*, 11(20):10305–10315, 2011.

[63] Karl K. Jonietz, Paul E. Dimotakis, Douglas A. Roman, and Bruce C. Walker. *A greenhouse-gas information system monitoring and validating emissions reporting and mitigation*. 2011.

[64] G. Keppel-Aleks, P. O. Wennberg, C. W. O'Dell, and D. Wunch. Towards constraints on fossil fuel emissions from total column carbon dioxide. *Atmospheric Chemistry and Physics*, 13(8):4349–4357, 2013.

[65] Ingeborg Levin, Bernd Kromer, Martina Schmidt, and Hartmut Sartorius. A novel approach for independent budgeting of fossil fuel co2 over europe by 14co2 observations. *Geophysical Research Letters*, 30(23):n/a–n/a, 2003.

[66] H. Wang, D. J. Jacob, M. Kopacz, D. B. A. Jones, P. Suntharalingam, J. A. Fisher, R. Nassar, S. Pawson, and J. E. Nielsen. Error correlation between co_2 and co as constraint for co_2 flux inversions using satellite data. *Atmospheric Chemistry and Physics*, 9(19):7313–7323, 2009.

[67] H. Bovensmann, M. Buchwitz, J. P. Burrows, M. Reuter, T. Krings, K. Gerilowski, O. Schneising, J. Heymann, A. Tretner, and J. Erzinger. A remote sensing technique for global monitoring of power plant co2 emissions from space and related applications. *Atmos. Meas. Tech.*, 3(4):781–811, 2010.

[68] V. A. Velazco, M. Buchwitz, H. Bovensmann, M. Reuter, O. Schneising, J. Heymann, T. Krings, K. Gerilowski, and J. P. Burrows. Towards space based verification of co_2 emissions from strong localized sources: fossil fuel power plant emissions as seen by a carbonsat constellation. *Atmospheric Measurement Techniques*, 4(12):2809–2822, 2011.

[69] A. S. Denning, I. Y. Fung, and D. Randall. Latitudinal gradient of atmospheric co2 due to seasonal exchange with land biota. *Nature*, 376(6537):240–243, 1995.

[70] Kazuyuki Miyazaki, Prabir K. Patra, Masayuki Takigawa, Toshiki Iwasaki, and Takakiyo Nakazawa. Global-scale transport of carbon dioxide in the troposphere. *Journal of Geophysical Research: Atmospheres*, 113(D15), 2008.

[71] R. Ahmadov, C. Gerbig, R. Kretschmer, S. Koerner, B. Neininger, A. J. Dolman, and C. Sarrat. Mesoscale covariance of transport and CO2 fluxes: Evidence from observations and simulations using the WRF-VPRM coupled atmosphere-biosphere model. *Journal of Geophysical Research: Atmospheres*, 112(D22), 2007.

[72] M. K. van der Molen and A. J. Dolman. Regional carbon fluxes and the effect of topography on the variability of atmospheric CO2. *Journal of Geophysical Research: Atmospheres*, 112(D1), 2007.

[73] A. Chevillard, U. Karstens, P. Ciais, S. Lafont, and M. Heimann. Simulation of atmospheric CO2 over Europe and western Siberia using the regional scale model REMO. *Tellus B*, 54(5), 2002.

[74] C. Sarrat, J. Noilhan, P. Lacarrre, S. Donier, C. Lac, J. C. Calvet, A. J. Dolman, C. Gerbig, B. Neininger, P. Ciais, J. D. Paris, F. Boumard, M. Ramonet, and A. Butet. Atmospheric CO2 modeling at the regional scale: Application to the CarboEurope Regional Experiment. *Journal of Geophysical Research: Atmospheres*, 112(D12), 2007.

[75] R. Ahmadov, C. Gerbig, R. Kretschmer, S. Körner, C. Rödenbeck, P. Bousquet, and M. Ramonet. Comparing high resolution wrf-vprm simulations and two global co_2 transport models with coastal tower measurements of co_2. *Biogeosciences*, 6(5):807–817, 2009.

[76] Mark Z. Jacobson. On the causal link between carbon dioxide and air pollution mortality. *Geophysical Research Letters*, 35(3), 2008.

[77] Mark Z. Jacobson. Enhancement of Local Air Pollution by Urban CO2 Domes. *Environmental Science & Technology*, 44(7):2497–2502, 2010.

[78] D. Pillai, C. Gerbig, R. Kretschmer, V. Beck, U. Karstens, B. Neininger, and M. Heimann. Comparing lagrangian and eulerian models for co_2 transport a step towards bayesian inverse modeling using wrf/stilt-vprm. *Atmospheric Chemistry and Physics*, 12(19):8979–8991, 2012.

[79] D. Byun and K. L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (cmaq) modeling system. *Applied Mechanics Reviews*, 59(2):51–77, 2006.

[80] Yongtao Hu, M. Talat Odman, and Armistead G. Russell. Top-down analysis of the elemental carbon emissions inventory in the united states by inverse modeling using community multiscale air quality model with decoupled direct method (cmaq-ddm). *Journal of Geophysical Research: Atmospheres*, 114(D24), 2009.

[81] Daniel S. Cohan, Amir Hakami, Yongtao Hu, and Armistead G. Russell. Nonlinear response of ozone to emissions: source apportionment and sensitivity analysis. *Environmental Science & Technology*, 39(17):6739–6748, 2005.

[82] Alan M. Dunker. The decoupled direct method for calculating sensitivity coefficients in chemical kinetics. *The Journal of Chemical Physics*, 81(5):2385–2393, 1984.

[83] Amir Hakami, Mehmet T. Odman, and Armistead G. Russell. Nonlinearity in atmospheric response: A direct sensitivity analysis approach. *Journal of Geophysical Research: Atmospheres*, 109(D15), 2004.

[84] Amir Hakami, Daven K. Henze, John H. Seinfeld, Kumaresh Singh, Adrian Sandu, Soontae Kim, Byun, and Qinbin Li. The adjoint of cmaq. *Environmental Science & Technology*, 41(22):7807–7817, 2007.

[85] J. Pleim and L. Ran. Surface flux modeling for air quality applications. *Atmosphere*, 2:271–302, 2011.

[86] D. N. Huntzinger, C. Schwalm, A. M. Michalak, K. Schaefer, A. W. King, Y. Wei, A. Jacobson, S. Liu, R. B. Cook, W. M. Post, G. Berthier, D. Hayes, M. Huang, A. Ito, H. Lei, C. Lu, J. Mao, C. H. Peng, S. Peng, B. Poulter, D. Riccuito, X. Shi, H. Tian, W. Wang, N. Zeng, F. Zhao, and Q. Zhu. The north american carbon program multi-scale synthesis and terrestrial model intercomparison project - part 1: Overview and experimental design. *Geoscientific Model Development*, 6(6):2121–2133, 2013.

[87] D.N. Huntzinger, W.M. Post, Y. Wei, A.M. Michalak, T.O. West, A.R. Jacobson, I.T. Baker, J.M. Chen, K.J. Davis, D.J. Hayes, F.M. Hoffman, A.K. Jain, S. Liu, A.D. McGuire, R.P. Neilson, Chris Potter, B. Poulter, David Price, B.M. Raczka, H.Q. Tian, P. Thornton, E. Tomelleri, N. Viovy, J. Xiao, W. Yuan, N. Zeng, M. Zhao, and R. Cook. North american carbon program (nacp) regional interim synthesis: Terrestrial biospheric model intercomparison. *Ecological Modelling*, 232(0):144 – 157, 2012.

[88] Kevin Schaefer, Christopher R. Schwalm, Chris Williams, M. Altaf Arain, Alan Barr, Jing M. Chen, Kenneth J. Davis, Dimitre Dimitrov, Timothy W. Hilton, David Y. Hollinger, Elyn Humphreys, Benjamin Poulter, Brett M. Raczka, Andrew D. Richardson, Alok Sahoo, Peter Thornton, Rodrigo Vargas, Hans Verbeeck, Ryan Anderson, Ian Baker, T. Andrew Black, Paul Bolstad, Jiquan Chen, Peter S. Curtis, Ankur R. Desai, Michael Dietze, Danilo Dragoni, Christopher Gough, Robert F. Grant, Lianhong Gu, Atul Jain, Chris Kucharik, Beverly Law, Shuguang Liu, Erandathie Lokipitiya, Hank A. Margolis, Roser Matamala, J. Harry McCaughey, Russ Monson, J. William Munger, Walter Oechel, Changhui Peng, David T. Price, Dan Ricciuto, William J. Riley, Nigel Roulet, Hanqin Tian, Christina Tonitto, Margaret Torn, Ensheng Weng, and Xiaolu Zhou. A model-data comparison of gross primary productivity: Results from the north american carbon program site synthesis. *Journal of Geophysical Research: Biogeosciences*, 117(G3), 2012.

[89] Christopher R. Schwalm, Christopher A. Williams, Kevin Schaefer, Ryan Anderson, M. Altaf Arain, Ian Baker, Alan Barr, T. Andrew Black, Guangsheng Chen, Jing Ming

Chen, Philippe Ciais, Kenneth J. Davis, Ankur Desai, Michael Dietze, Danilo Dragoni, Marc L. Fischer, Lawrence B. Flanagan, Robert Grant, Lianhong Gu, David Hollinger, R. Csar Izaurralde, Chris Kucharik, Peter Lafleur, Beverly E. Law, Longhui Li, Zhengpeng Li, Shuguang Liu, Erandathie Lokupitiya, Yiqi Luo, Siyan Ma, Hank Margolis, Roser Matamala, Harry McCaughey, Russell K. Monson, Walter C. Oechel, Changhui Peng, Benjamin Poulter, David T. Price, Dan M. Riciutto, William Riley, Alok Kumar Sahoo, Michael Sprintsin, Jianfeng Sun, Hanqin Tian, Christina Tonitto, Hans Verbeeck, and Shashi B. Verma. A model-data intercomparison of CO2 exchange across North America: Results from the North American Carbon Program site synthesis. *Journal of Geophysical Research: Biogeosciences*, 115(G3), 2010.

[90] Wouter Peters, Andrew R. Jacobson, Colm Sweeney, Arlyn E. Andrews, Thomas J. Conway, Kenneth Masarie, John B. Miller, Lori M. P. Bruhwiler, Gabrielle Ptron, Adam I. Hirsch, Douglas E. J. Worthy, Guido R. van der Werf, James T. Randerson, Paul O. Wennberg, Maarten C. Krol, and Pieter P. Tans. An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences*, 104(48):18925–18930, 2007.

[91] Hongyi Li, Maoyi Huang, Mark S. Wigmosta, Yinghai Ke, Andr M. Coleman, L. Ruby Leung, Aihui Wang, and Daniel M. Ricciuto. Evaluating runoff simulations from the community land model 4.0 using observations from flux towers and a mountainous watershed. *Journal of Geophysical Research: Atmospheres*, 116(D24), 2011.

[92] Y. Wei, S. Liu, D. N. Huntzinger, A. M. Michalak, N. Viovy, W. M. Post, C. R. Schwalm, K. Schaefer, A. R. Jacobson, C. Lu, H. Tian, D. M. Ricciuto, R. B. Cook, J. Mao, and X. Shi. The north american carbon program multi-scale synthesis and terrestrial model intercomparison project - part 2: Environmental driver data. *Geoscientific Model Development Discussions*, 6(4):5375–5422, 2013.

[93] Deborah N. Huntzinger, Sharon M. Gourdji, Kimberly L. Mueller, and Anna M. Michalak. The utility of continuous atmospheric measurements for identifying biospheric CO2 flux variability. *Journal of Geophysical Research: Atmospheres*, 116(D6), 2011.

[94] G. Keppel-Aleks, P. O. Wennberg, R. A. Washenfelder, D. Wunch, T. Schneider, G. C. Toon, R. J. Andres, J.-F. Blavier, B. Connor, K. J. Davis, A. R. Desai, J. Messerschmidt, J. Notholt, C. M. Roehl, V. Sherlock, B. B. Stephens, S. A. Vay, and S. C. Wofsy. The imprint of surface fluxes and transport on variations in total column carbon dioxide. *Biogeosciences*, 9(3):875–891, 2012.

[95] C. J. Eyers, P. Norman, J. Middel, M. Plohr, S. Michot, K. Atkinson, and R. A. Christo. Aero2k global aviation emissions inventories for 2002 and 2025. Technical Report QINETIQ/04/01113, Center for Air Transport and the Environment, 2004.

[96] R. Andres, J. Gregg, L. Losey, G. Marland, and T. Boden. Monthly, global emissions of carbon dioxide from fossil fuel consumption. *Tellus B*, 63(3), 2011.

[97] A.E. Andrews, J. Kofler, P. S. Bakwin, C. Zhao, and P. Tans. Carbon dioxide and carbon monoxide dry air mole fractions from the noaa esrl tall tower network. 1992-2009. version: 2011-08-31. ftp://ftp.cmdl.noaa.gov/ccg/towers/, 2009. Accessed: 2012-01-14.

[98] E. V. Berezin, I. B. Konovalov, P. Ciais, A. Richter, S. Tao, G. Janssens-Maenhout, M. Beekmann, and E.-D. Schulze. Multiannual changes of co_2 emissions in china: indirect estimates derived from satellite measurements of tropospheric no_2 columns. *Atmospheric Chemistry and Physics Discussions*, 13(1):255–309, 2013.

[99] Parvadha Suntharalingam, Daniel J. Jacob, Paul I. Palmer, Jennifer A. Logan, Robert M. Yantosca, Yaping Xiao, Mathew J. Evans, David G. Streets, Stephanie L. Vay, and Glen W. Sachse. Improved quantification of Chinese carbon fluxes using CO2/CO correlations in Asian outflow. *Journal of Geophysical Research: Atmospheres*, 109(D18), 2004.

[100] Yuhang Wang and Tao Zeng. On tracer correlations in the troposphere: The case of ethane and propane. *Journal of Geophysical Research: Atmospheres*, 109(D24), 2004.

[101] H. A. Michelsen, G. L. Manney, M. R. Gunson, C. P. Rinsland, and R. Zander. Correlations of stratospheric abundances of ch4 and n2o derived from atmos measurements. *Geophysical Research Letters*, 25(15):2777–2780, 1998.

[102] R. A. Plumb. Tracer interrelationships in the stratosphere. *Reviews of Geophysics*, 45(4), 2007.

[103] Felix Vogel, Samuel Hammer, Axel Steinhof, Bernd Kromer, and Ingeborg Levin. Implication of weekly and diurnal 14 c calibration on hourly estimates of co-based fossil fuel co 2 at a moderately polluted site in southwestern germany. *Tellus B*, 62(5), 2011.

[104] P. Peylin, R. M. Law, K. R. Gurney, F. Chevallier, A. R. Jacobson, T. Maki, Y. Niwa, P. K. Patra, W. Peters, P. J. Rayner, C. Rödenbeck, and X. Zhang. Global atmospheric carbon budget: results from an ensemble of atmospheric co_2 inversions. *Biogeosciences Discussions*, 10(3):5301–5360, 2013.

[105] I.G. Enting. *Inverse Problems in Atmospheric Constituent Transport*. Cambridge Atmospheric and Space Science Series. Cambridge University Press, 2002.

[106] Wouter Peters, Andrew R. Jacobson, Colm Sweeney, Arlyn E. Andrews, Thomas J. Conway, Kenneth Masarie, John B. Miller, Lori M. P. Bruhwiler, Gabrielle Ptron, Adam I. Hirsch, Douglas E. J. Worthy, Guido R. van der Werf, James T. Randerson, Paul O. Wennberg, Maarten C. Krol, and Pieter P. Tans. An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences*, 104(48):18925–18930, 2007.

[107] P. J. Rayner, M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann. Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (ccdas). *Global Biogeochemical Cycles*, 19(2):n/a–n/a, 2005.

[108] Anna M. Michalak, Lori Bruhwiler, and Pieter P. Tans. A geostatistical approach to surface flux estimation of atmospheric trace gases. *Journal of Geophysical Research: Atmospheres*, 109(D14):n/a–n/a, 2004.

[109] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 2005.

[110] Frdric Chevallier and Christopher W. O'Dell. Error statistics of bayesian co2 flux inversion schemes as seen from gosat. *Geophysical Research Letters*, 40(6):1252–1256, 2013.

[111] Frdric Chevallier, Nicolas Viovy, Markus Reichstein, and Philippe Ciais. On the assignment of prior errors in bayesian inversions of co2 surface fluxes. *Geophysical Research Letters*, 33(13):n/a–n/a, 2006.

[112] Kevin Robert Gurney, Rachel M Law, A Scott Denning, Peter J Rayner, David Baker, Philippe Bousquet, Lori Bruhwiler, Yu-Han Chen, Philippe Ciais, Songmiao Fan, et al. Towards robust regional estimates of co2 sources and sinks using atmospheric transport models. *Nature*, 415(6872):626–630, 2002.

[113] Kevin Robert Gurney, Rachel M. Law, A. Scott Denning, Peter J. Rayner, David Baker, Philippe Bousquet, Lori Bruhwiler, Yu-Han Chen, Philippe Ciais, Songmiao Fan, Inez Y. Fung, Manuel Gloor, Martin Heimann, Kaz Higuchi, Jasmin John, Eva Kowalczyk, Takashi Maki, Shamil Maksyutov, Philippe Peylin, Michael Prather, Bernard C. Pak, Jorge Sarmiento, Shoichi Taguchi, Taro Takahashi, and Chiu-Wai Yuen. Transcom 3 co2 inversion intercomparison: 1. annual mean control results and sensitivity to transport and prior flux information. *Tellus B*, 55(2):555–579, 2003.

[114] Song-Miao Fan, Jorge L. Sarmiento, Manuel Gloor, and Stephen W. Pacala. On the use of regularization techniques in the inverse modeling of atmospheric carbon dioxide. *Journal of Geophysical Research: Atmospheres*, 104(D17):21503–21512, 1999.

[115] D. S. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial, Second Edition*. Oxford University Press, 2006.

[116] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.

[117] M. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[118] Scott A Sisson and Yanan Fan. Likelihood-free Markov chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, 2010.

[119] H. Jeffreys. *Theory of Probability*. Oxford: Clarendon Press, 3^{rd} edition, 1939.

[120] K. Sargsyan, H. Najm, and R. Ghanem. On the statistical calibration of physical models. *International Journal for Chemical Kinetics*, submitted, 2014.

[121] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77(2):439 – 471, 1988.

[122] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer Verlag, New York, 1991.

[123] O.P. Le Maître and O.M. Knio. *Spectral Methods for Uncertainty Quantification*. Springer, New York, NY, 2010.

[124] S. S. Collis, R. A. Bartlett, T. M. Smith, M. Heinkenschloss, L. C. Wilcox, J. C. Hill, O. Ghattas, M. O. Berggren, V. Akcelik, C. C. Ober, B. G. van Bloemen Waanders, and E. R. Keiter. Sensitivity technologies for large scale simulation. Technical Report SAND2004-6574, Sandia National Laboratories, 2005.

[125] A. Giunta, M. Eldred, L. Swiler, T. Trucano, and S. Wojtkiewicz. Perspectives in optimization under uncertainty: Algorithms and applications. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, number 2004-4451. AIAA, 2004.

[126] M. R. Hardy. An introduction to risk measures for actuarial applications. Technical Report C-25-07, Education and Examination Committee of the Society of Actuaries, 2006.

[127] S. S. Collis and M. Heinkenschloss. Analysis of the streamline upwind/petrov galerkin method applied to the solution of optimal control problems. Technical Report TR02-01, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2002.

[128] H. Chen, J. Winderlich, C. Gerbig, A. Hoefer, C. W. Rella, E. R. Crosson, A. D. Van Pelt, J. Steinbach, O. Kolle, V. Beck, B. C. Daube, E. W. Gottlieb, V. Y. Chow, G. W. Santoni, and S. C. Wofsy. High-accuracy continuous airborne measurements of greenhouse gases (co_2 and ch_4) using the cavity ring-down spectroscopy (crds) technique. *Atmospheric Measurement Techniques*, 3(2):375–386, 2010.

[129] H. Nara, H. Tanimoto, Y. Tohjima, H. Mukai, Y. Nojiri, K. Katsumata, and C. W. Rella. Effect of air composition (n_2, o_2, ar, and h_2o) on co_2 and ch_4 measurement by wavelength-scanned cavity ring-down spectroscopy: calibration and measurement strategy. *Atmospheric Measurement Techniques*, 5(11):2689–2701, 2012.

[130] A. E. Andrews, J. D. Kofler, M. E. Trudeau, J. C. Williams, D. H. Neff, K. A. Masarie, D. Y. Chao, D. R. Kitzis, P. C. Novelli, C. L. Zhao, E. J. Dlugokencky, P. M. Lang, M. J. Crotwell, M. L. Fischer, M. J. Parker, J. T. Lee, D. D. Baumann, A. R. Desai, C. O. Stanier, S. F. J. De Wekker, D. E. Wolfe, J. W. Munger, and P. P. Tans. Co_2, co, and ch_4 measurements from tall towers in the noaa earth system research laboratory's global greenhouse gas reference network: instrumentation, uncertainty analysis, and recommendations for future high-accuracy greenhouse gas monitoring efforts. *Atmospheric Measurement Techniques*, 7(2):647–687, 2014.

[131] A. E. Andrews, J. D. Kofler, M. E. Trudeau, J. C. Williams, D. H. Neff, K. A. Masarie, D. Y. Chao, D. R. Kitzis, P. C. Novelli, C. L. Zhao, E. J. Dlugokencky, P. M. Lang, M. J.

Crotwell, M. L. Fischer, M. J. Parker, J. T. Lee, D. D. Baumann, A. R. Desai, C. O. Stanier, S. F. J. de Wekker, D. E. Wolfe, J. W. Munger, and P. P. Tans. Co_2, co and ch_4 measurements from the noaa earth system research laboratory's tall tower greenhouse gas observing network: instrumentation, uncertainty analysis and recommendations for future high-accuracy greenhouse gas monitoring efforts. *Atmospheric Measurement Techniques Discussions*, 6(1):1461–1553, 2013.

[132] J. E. Campbell, G. R. Carmichael, T. Chai, M. Mena-Carrasco, Y. Tang, D. R. Blake, N. J. Blake, S. A. Vay, G. J. Collatz, I. Baker, J. A. Berry, S. A. Montzka, C. Sweeney, J. L. Schnoor, and C. O. Stanier. Photosynthetic control of atmospheric carbonyl sulfide during the growing season. *Science*, 322(5904):1085–1088, 2008.

[133] Changsub Shim, Yuhang Wang, Hanwant B. Singh, Donald R. Blake, and Alex B. Guenther. Source characteristics of oxygenated volatile organic compounds and hydrogen cyanide. *Journal of Geophysical Research: Atmospheres*, 112(D10), 2007.

[134] M. D. Gunzburger. *Perspectives in Flow Control and Optimization*. Society for Industrial and Applied Mathematics, 2002.

[135] H. Bovensmann, J. P. Burrows, M. Buchwitz, J. Frerick, S. Noel, V. V. Rozanov, K. V. Chance, and A. P. H. Goede. SCIAMACHY: Mission Objectives and Measurement Modes. *Journal of the Atmospheric Sciences*, 56(2):127–150, 1999.

[136] C.D. Rodgers. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. Series on Atmospheric Oceanic and Planetary Physics, Volume 2. World Scientific Publishing Company, Incorporated, 2000.

[137] Thomas Kaminski and Martin Heimann. Inverse modeling of atmospheric carbon dioxide fluxes. *Science*, 294(5541):259, 2001.

[138] Britton B. Stephens, Kevin R. Gurney, Pieter P. Tans, Colm Sweeney, Wouter Peters, Lori Bruhwiler, Philippe Ciais, Michel Ramonet, Philippe Bousquet, Takakiyo Nakazawa, Shuji Aoki, Toshinobu Machida, Gen Inoue, Nikolay Vinnichenko, Jon Lloyd, Armin Jordan, Martin Heimann, Olga Shibistova, Ray L. Langenfelds, L. Paul Steele, Roger J. Francey, and A. Scott Denning. Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric co2. *Science*, 316(5832):1732–1735, 2007.

A Detailed Descriptions of NEE From the Community Land Model (CLM4VIC) Simulation

Under MsTMIP, CLM4VIC was configured and run following the protocol described in [86], with driver datasets provided by MsTMIP as described by Wei et al. [92]. The CLM4VIC-based NEE used in this study was from the baseline global simulation (i.e., BG1 simulation), in which the model was driven by a 110 year (1901 - 2010) atmospheric forcing dataset, with annual variations in atmospheric nitrogen deposition, CO2 concentration, and land-use change. The carbon-nitrogen biogeochemistry in the model was turned on, allowing for simulating vegetation dynamics in response to a changing environment, including prognostic estimates of emissions due to wild fires under appropriate environmental conditions. For a detailed description of the configuration and model setup of CLM4VIC simulations from MsTMIP, interested readers are referred to Huntzinger et al. [86].

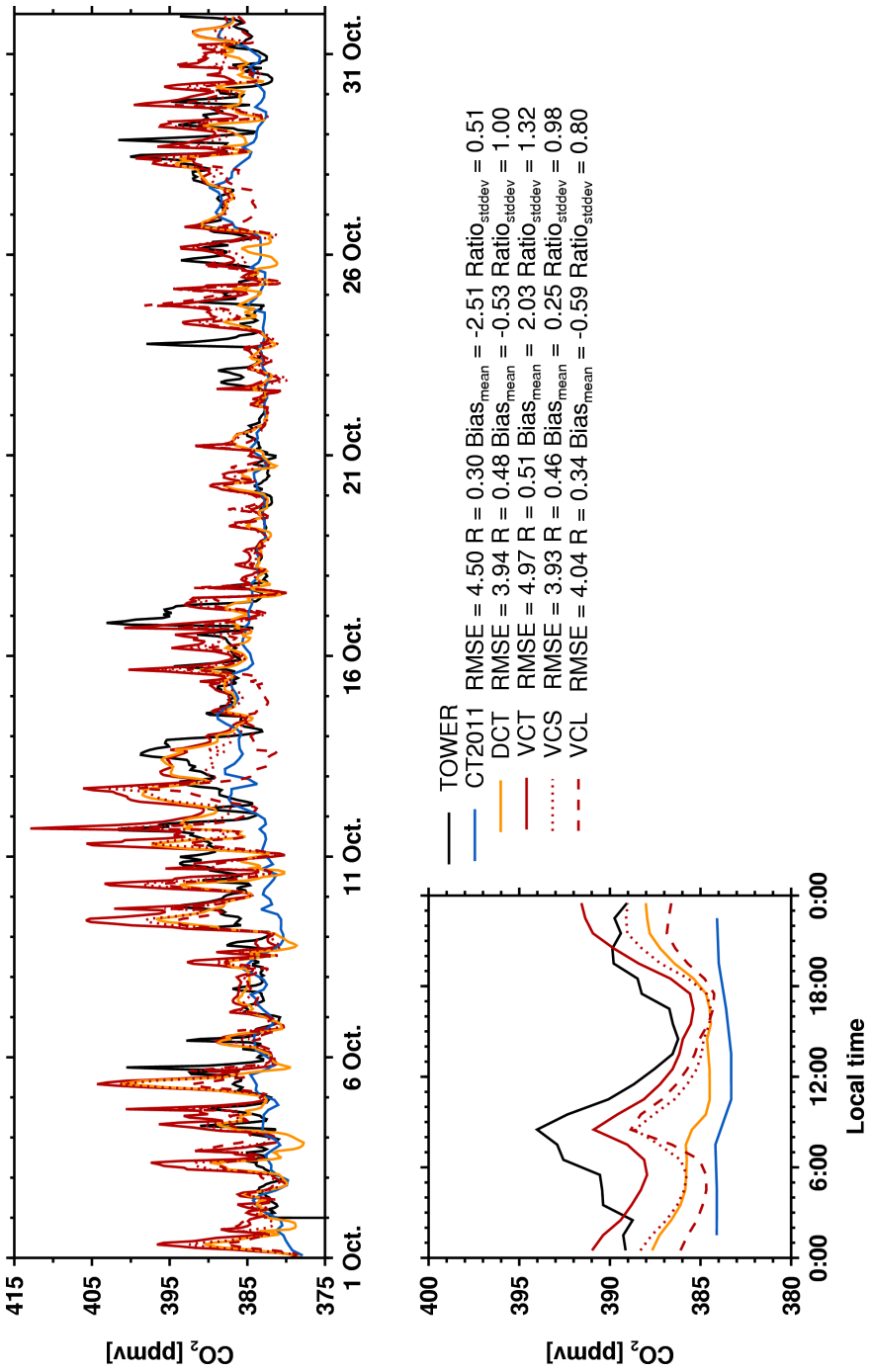


Figure A.1. Hourly time series and average diurnal cycle of CO2 observed and simulated at BAO. For the time series, model root mean square error (RMSE), correlation coefficient (R), model mean bias (meanmodel meanobservation), and ratio of standard deviations (stddevmodel/stddevobservation) are shown for both CMAQ and CT2011 after aggregating CMAQ hourly outputs to 3-hourly time series (to match the time resolution of CT2011).

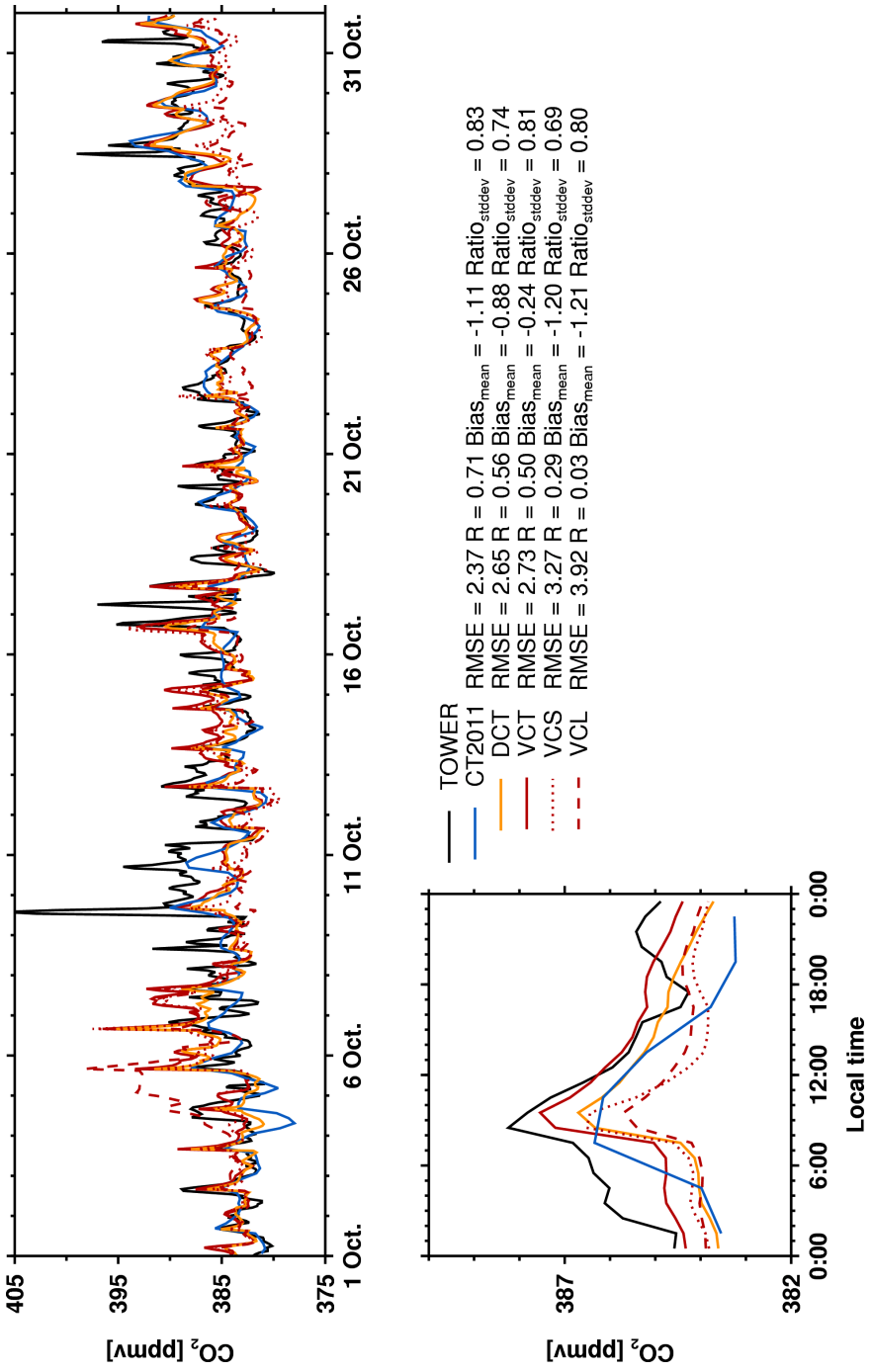


Figure A.2. Same as Fig. A.1, but for WKT.

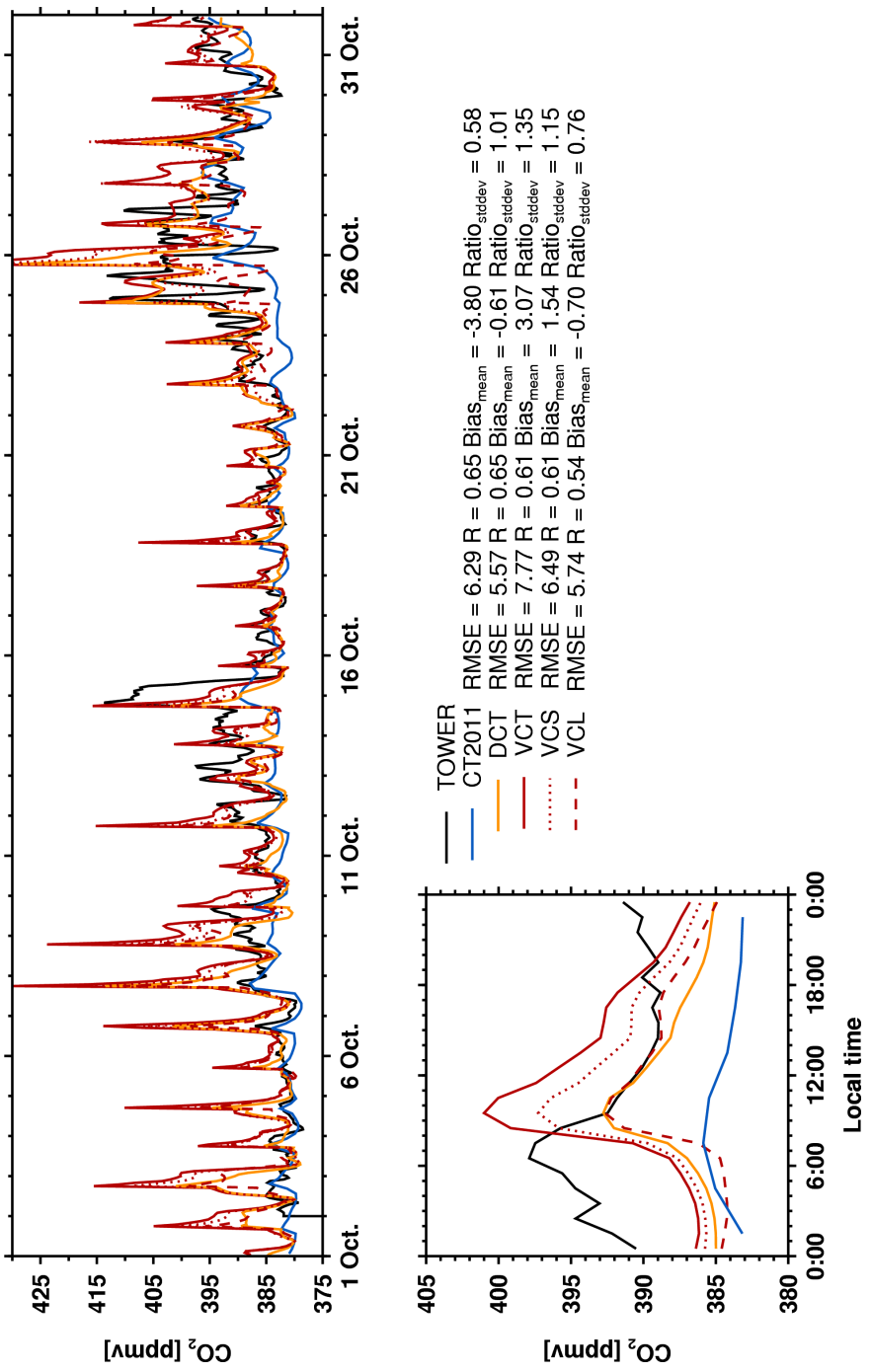


Figure A.3. Same as Fig. A.1, but for WGC.

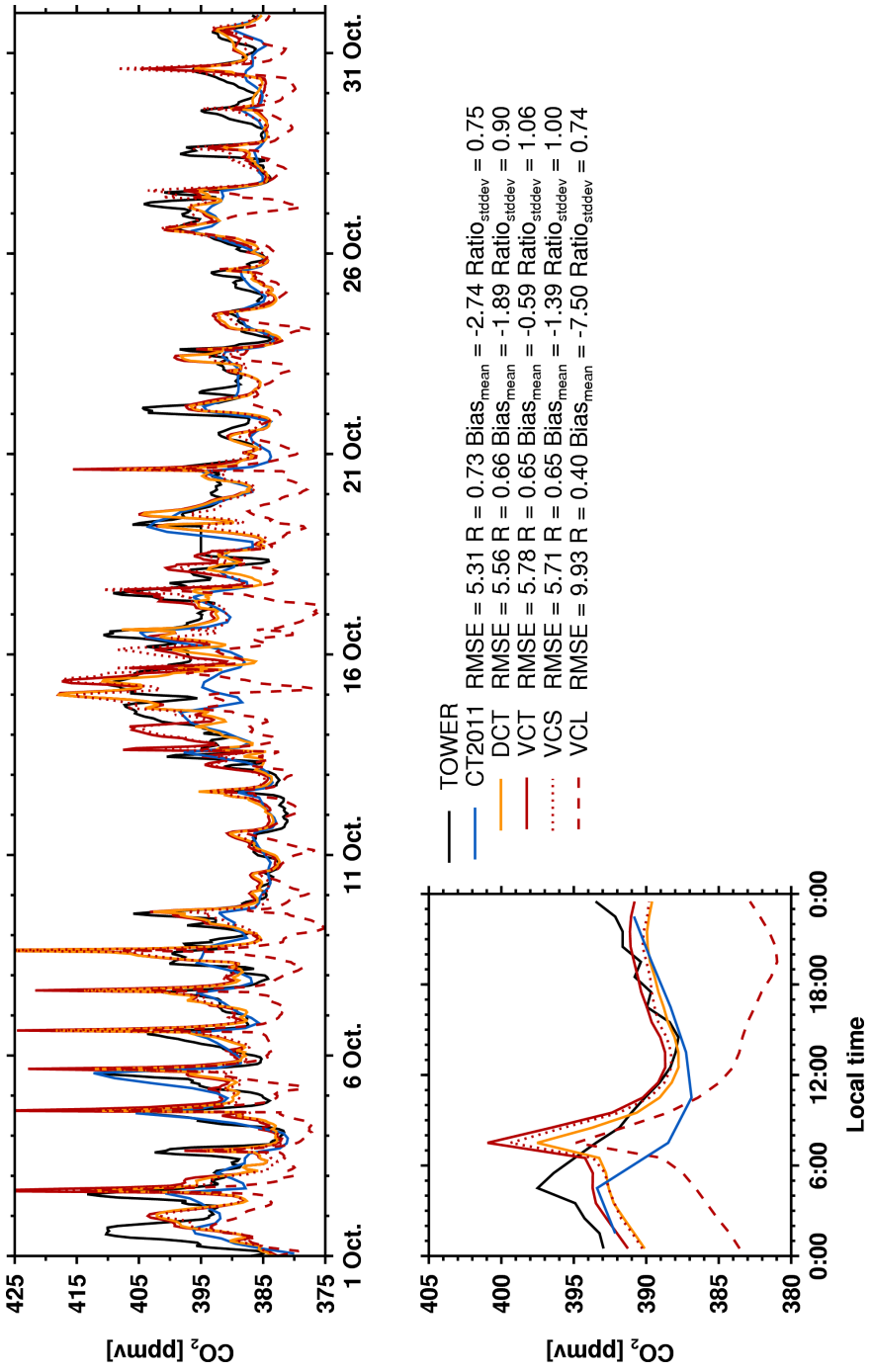


Figure A.4. Same as Fig. A.1, but for WBI.

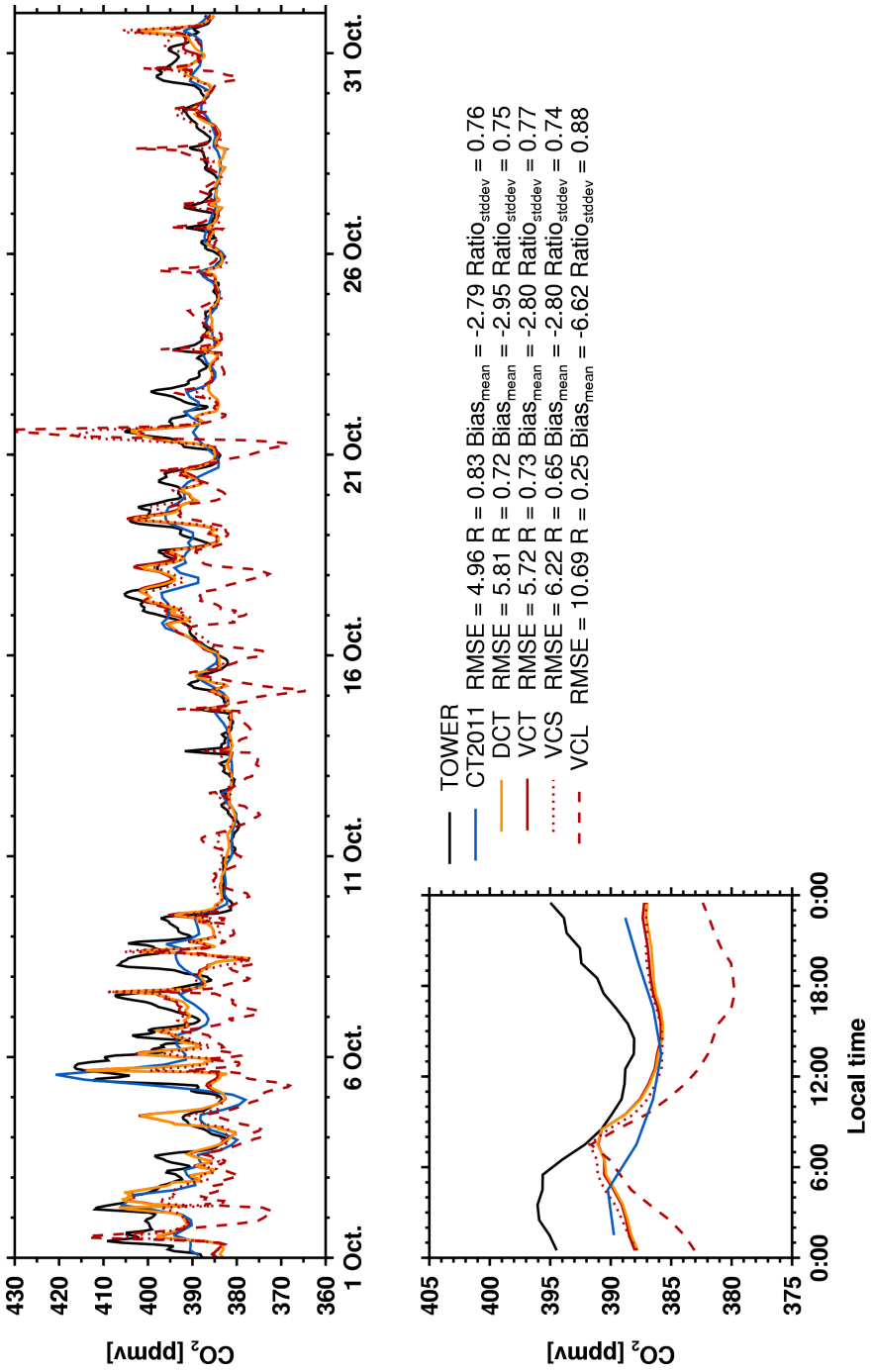


Figure A.5. Same as Fig. A.1, but for LEF.

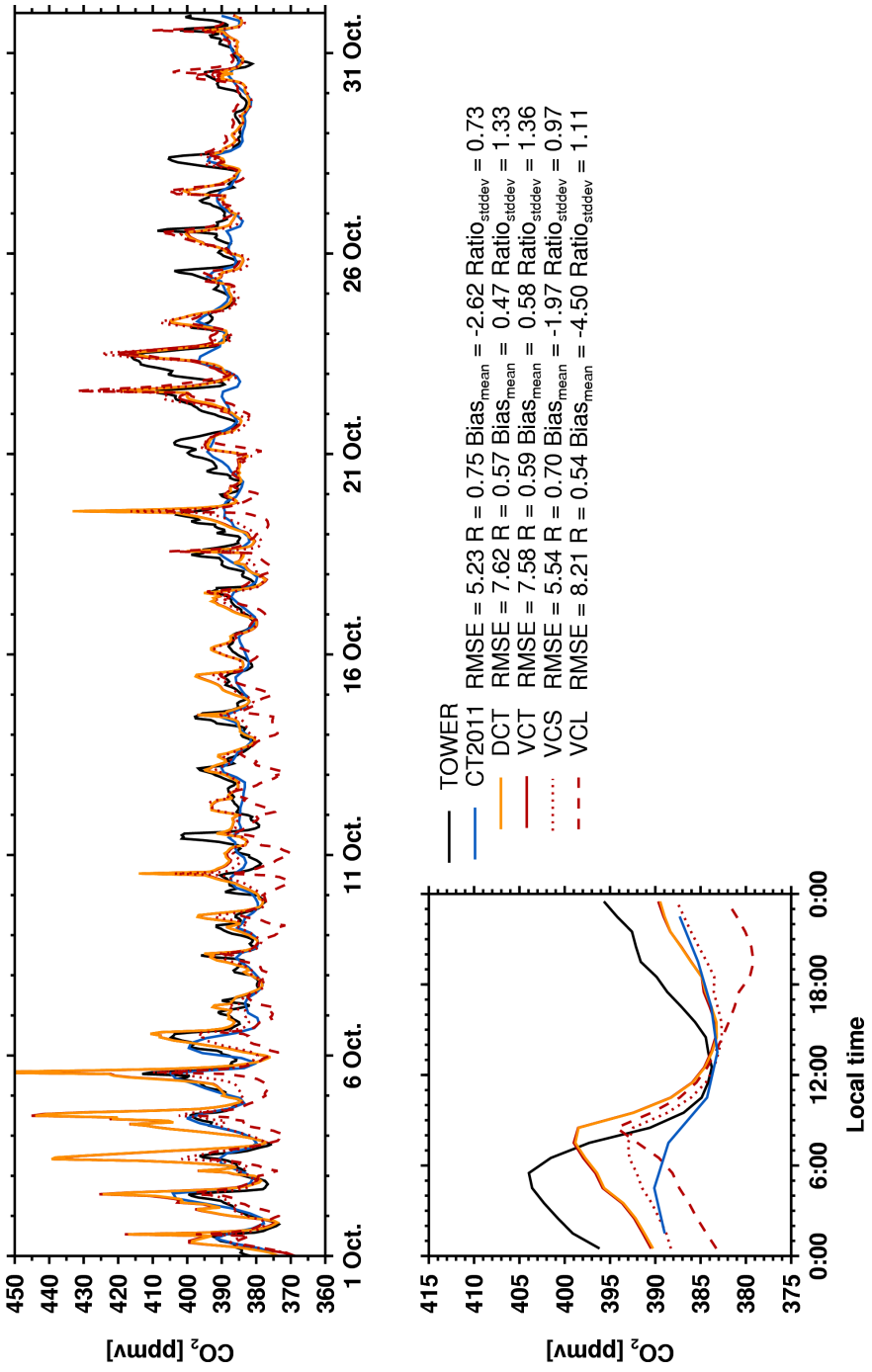


Figure A.6. Same as Fig. A.1, but for AMT.

Table A.1. Configurations of WRF and CMAQ

WRF	
Microphysics	Morrison
Cumulus	Kain-Fritsch
Surface	Pleim-Xiu
Radiation (longwave and shortwave)	RRTMG
Others	vertical velocity damping; 6th-order diffusion
	For more details in the WRF simulation, including model setup and evaluation against observations, see the SEMAP project report for WRF `http://sesarm.aer.com/static/pages/v0.9/SESARM-Final-Report-20111219.pdf`
CMAQ	
Gas phase chemistry	Carbon Bond 05 (CB05) with updated toluene chemistry
Aerosols	5^{th}-generation modal CMAQ aerosol model (AERO5)
Cloud	Asymmetric Convective Method (ACM)
Vertical diffusion	ACM2
Horizontal diffusion	multiscale scheme based on local wind deformation
Vertical advection	WRF
Horizontal advection	Yamo
Dry deposition	In-line for non-CO2 species
Emissions	Anthropogenic emissions: SESARM regional inventory (2007) for SESARM states; NEI 2005 v5 for non-SESARM states Fire: SESARM regional inventory and Blue Sky inventory for non-SESARM states Biogenic: BEIS3 For more details, see the SouthEastern Modeling, Analysis and Planning (SEMAP) modeling protocol `http://airqualitymodeling.org/semapwiki/index.php?title=SEMAP_Modeling_Protocol`; accessed January 14, 2013

Table A.2. Configurations of WRF and CMAQ

Tower	Location	Elevation (meters above sea level)	Intake height (meters above ground)
Argyle, Maine (AMT)	45.03°N 68.68°W	50	107
Boulder Atmospheric Observatory (BAO)	40.05°N 105.01°W	1584	300
Park Falls, Wisconsin (LEF)	45.95°N 90.27°W	472	122
West Branch, Iowa (WBI)	41.73°N 91.35°W	242	99
Walnut Grove, California (WGC)	38.27°N 121.49°W	0	483
Waco Killeen, Texas (WKT)	31.32°N 97.33°W	251	457

This page intentionally left blank.

Sandia National Laboratories