SANDIA REPORT

SAND2013-2919
Unlimited Release
Printed April 2013

A multiresolution spatial parametrization for the estimation of fossil-fuel carbon dioxide emissions via atmospheric inversions

J. Ray, V. Yadav, J. Lee, A. M. Michalak, S. Lefantzi, B. van Bloemen Waanders and S. A. McKenna

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

SAND2013-2919
Unlimited Release
Printed April 2013

A multiresolution spatial parametrization for the estimation of fossil-fuel carbon dioxide emissions via atmospheric inversions

J. Ray, J. Lee and S. Lefantzi
Sandia National Laboratories, P. O. Box 969, Livermore CA 94551
V. Yadav and A. M. Michalak,
Carnegie Institution for Science, Stanford, CA 94305
B. van Bloemen Waanders,
Sandia National Laboratories, P. O. Box 5800, Albuquerque NM 87185-0751
S. A. McKenna,
IBM Research, Smarter Cities Technology Centre, Bldg 3
Damastown Industrial Estate, Mulhuddart, Dublin 15, Ireland.

{jairay,jlee3,slefant,bartv@sandia.gov}, {vineety,michalak@stanford.edu},
{seanmcke@ie.ibm.com}

Abstract

The estimation of fossil-fuel CO_2 emissions ($ffCO_2$) from limited ground-based and satellite measurements of CO_2 concentrations will form a key component of the monitoring of treaties aimed at the abatement of greenhouse gas emissions. To that end, we construct a multiresolution spatial parametrization for fossil-fuel CO_2 emissions ($ffCO_2$), to be used in atmospheric inversions. Such a parametrization does not currently exist. The parametrization uses wavelets to accurately capture the multiscale, non-stationary nature of $ffCO_2$ emissions and employs proxies of human habitation, e.g., images of lights at night and maps of built-up areas to reduce the dimensionality of the multiresolution parametrization. The parametrization is used in a synthetic data inversion to test its suitability for use in atmospheric inverse problem. This linear inverse problem is predicated on observations of $ffCO_2$ concentrations collected at measurement towers. We adapt a convex optimization technique, commonly used in the reconstruction of compressively sensed images, to perform sparse reconstruction of the time-variant $ffCO_2$ emission field. We also borrow concepts from compressive sensing to impose boundary conditions i.e., to limit $ffCO_2$ emissions within an irregularly shaped region (the United States, in our case). We find that the optimization algorithm performs a data-driven sparsification of the spatial parametrization and retains only of those wavelets whose weights could be estimated from the observations. Further, our method for the imposition of boundary conditions leads to a $10\times$ computational saving over conventional means of doing so. We conclude with a discussion of the accuracy of the estimated emissions and the suitability of the spatial parametrization for use in inverse problems with a significant degree of regularization.

Acknowledgment

Thanks are due to Sharon Gourdji of Stanford University, and Kim Mueller and Abhishek Chatterjee of University of Michigan, Ann Arbor for fruitful discussions regarding algorithms, the challenges of working with field observations as well as some of the data used in this work. This work was supported by Sandia National Laboratories' LDRD (Laboratory Directed Research and Development) funds, sponsored by the Geosciences Investment Area. 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.

Contents

Appendix

Figures

Tables

1 Introduction

The potential role of anthropogenic greenhouse gas emission in climate change has led to a host of policies that seek to reduce them e.g., the UN-REDD program [5] and the UNFCCC [6]. Emissions of CO_2 from the burning of fossil fuels (for power generation, transportation, industrial and residential use) form the largest component of these greenhouse gas emissions and any policies aimed at the abatement of fossil-fuel ($ffCO_2$) emissions are expected to have significant socioeconomic impacts. Monitoring the ability of such policies to reduce greenhouse gas emissions is a complex endeavor and is expected to receive increasing attention in the future. To this end, The United States National Academy of Sciences considered the design of a Greenhouse Gas Information System that would combine measurements of CO_2 concentrations and fluxes, with reports of national inventories of fossil fuels, to estimate $ffCO_2$ emissions [7].

In this report, we present a method for estimating $ffCO_2$ emissions from CO_2 concentration measurements at observation towers. We develop a spatial parameterization for the emissions and an inversion scheme, based on sparse reconstruction, to calibrate the parametrization. The method is demonstrated on $ffCO_2$ in the United States using synthetic data. Its extension to global scale inversion is conceptually simple. It will, naturally, require more data. It will also be significantly more computationally demanding, thus necessitating the development of scalable counterparts of the sparse reconstruction methods described in this report.

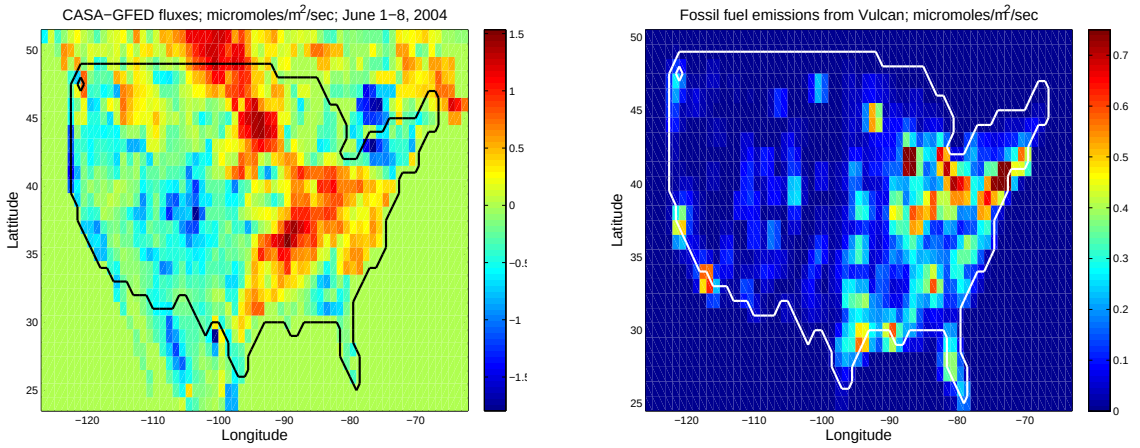
Currently, the best data on $ffCO_2$ emissions are obtained from inventories/databases that are constructed from national reporting of the consumption of fossil fuels. Some inventories [8, 9, 10] start with national or provincial estimates of fossil-fuel consumption and disaggregate them using proxies of human habitation (e.g., population density), whereas others aggregate them from local information of fossil-fuel use patterns [4]. Their accuracy depends on the data used for constructing the inventories, as well as the method; see [11, 12] for a discussion on their uncertainties and revisions when more data becomes available. In some cases, these revisions can be large [13].

An alternate way of estimating CO_2 emissions is via atmospheric inversion. In this process, time-varying CO_2 fluxes $\mathbf{f}$ are estimated from measurements of time-varying CO_2 concentrations at certain ground-based locations (towers) or satellite soundings of column-integrated CO_2 concentrations ($\mathbf{y}^{obs}$). The CO_2 fluxes $\mathbf{f}$ are defined on a gridded domain and related to $\mathbf{y}^{obs}$ via an atmospheric transport model. Since the magnitude of $\mathbf{f}$ does not impact the wind velocity field and dispersion, CO_2 is modeled as a passive scalar and a linear relationship holds

$$\mathbf{y}^{obs} = \mathbf{y} + \varepsilon = \mathbf{H}\mathbf{f} + \varepsilon, \tag{1}$$

where $\mathbf{H}$ is the transport or sensitivity matrix, obtained from a transport model like Weather Research and Forecasting Model (WRF, [14]). $\mathbf{y}$ is the CO_2 concentration predicted by the atmospheric model which differs from its measured counterpart by an error ε. Atmospheric inversion has become a routine tool for estimating *biospheric* CO_2 fluxes, conditioned on both satellite and ground-based observations [15, 16, 17, 18, 19]. The inverse problem is set up as an optimization between the need to reproduce observations while adhering to a prior belief regarding the spatio-temporal distribution of fluxes. Since biospheric CO_2 fluxes vary smoothly in space (see Fig. 1), the prior modeling of CO_2 fluxes is performed with a stationary, multivariate Gaussian field. The prior covariance provides the regularization in the inverse problem and allows high-dimensional gridded emissions to be estimated from relatively few measurements, at the cost of smoothing out fine-scale spatial structures in the estimated fluxes.

There are two methods for calculating these emission estimates : (1) deterministic methods, based on a variational formulation (which provide a "mean" or a "best-fit" estimation and (2) Bayesian methods that



(a) Biospheric fluxes

(b) ffCO$_2$ emissions

Figure 1: Differences in the nature of spatial distribution of biospheric (left) and fossil-fuel (right) CO$_2$ fluxes. The biospheric fluxes are stationary, whereas ffCO$_2$ emissions are multiresolution (non-stationary) and correlated with human habitation. The biospheric fluxes are for June 1 - June 8, 2004, obtained from CASA-GFED [1]. The post-processing steps to obtain the fluxes as plotted are described in [2]. The units of fluxes/emissions are μmoles s^{-1} m^{-2} of C. The ffCO$_2$ emissions are obtained from the Vulcan database [3, 4]. Note the different colormaps; ffCO$_2$ emissions can assume only non-negative values.

provide a measure of the uncertainty in the estimate in addition. These methods have been reviewed in [20]. Kalman filters, too, have been used in these inversions [21]. All these methods are strongly influenced by the choice of the prior model, and the *geostatistical inversion* method [22] was developed to lessen the dependence.

Fig. 1 contrasts the spatial distribution of biospheric CO$_2$ fluxes against $\overline{\mathbf{f}_V}$, the ffCO$_2$ emissions from the Vulcan database [3, 4]. Vulcan provides emissions for the lower 48 states of the US, on a 0.1° grid and at an hourly resolution. The emissions are coarsened to a 1° resolution and averaged over the entire year, to obtain $\overline{\mathbf{f}_V}$. It is clear that the spatial distribution of ffCO$_2$ is correlated with human habitation, is strongly multiscale, and may be difficult to represent with a variogram. Instead, we seek a different spatial parameterization that can represent its strongly multiresolution character. Further, the spatial parametrization must necessarily be low-dimensional (i.e., have few free/independent parameters) so that it can be used for estimation within the context of an inverse problem with sparse observations.

In this paper, we construct a spatial parametrization for ffCO$_2$ based on wavelets. We will refer to it as the Multiscale Random Field (MsRF) model. Wavelets have compact support, form an orthogonal basis set and are widely used to model non-stationary fields e.g. images [23, 24]. We will reduce the dimensionality of the MsRF using an easily-observed proxy of human habitation e.g., images of lights at night (henceforth called nightlights), maps of built-up areas etc. The MsRF so formed will be used in a synthetic atmospheric inversion test using a novel, sparsity-enforcing optimization method. The inversion assumes a pure ffCO$_2$ signal i.e., we can measure the fossil-fuel contribution to the CO$_2$ concentration in a sample, as done in [25, 26]. This procedure will identify the subset of wavelets in the MsRF that can be actually estimated from the synthetic observations, while "turning off" the rest. In doing so, it will ensure that the MsRF, as designed, has sufficient flexibility to extract the information on ffCO$_2$ in the observations. We conclude with a discussion

on the efficiency of the inversion/ffCO_2 estimation. The Vulcan inventory will serve as the 'ground-truth" in our synthetic data inversions.

The paper is structured as follows. In Sec. 2, we review existing literature on the construction of ffCO_2 inventories, with emphasis on how proxies of human habitation (population density, nightlights etc) are used to disaggregate national and provincial-level fossil-fuel emissions to finer resolution. We will also review existing literature on atmospheric inversions (both biospheric and fossil-fuel CO_2 fluxes) and *compressive sensing*, a wavelet-based image-processing technique that we adapt to our inversion problem. In Sec. 3, we construct two MsRF models, based on nightlights and maps of built-up areas. Sec. 4 contains the formulation of the inverse problem and the algorithm to solve it. In Sec. 5, we perform tests with synthetic data, discuss the quality of the competing MsRFs and examine the impact of various features of the inverse problem (e.g. number of measurements, errors in measurements etc.). Conclusions are in Sec. 6.

This page intentionally left blank

2 Background

In this section, we review how $ffCO_2$ emission inventories are currently calculated and the shortcomings in them. This is followed by a discussion of CO_2 flux (both fossil-fuel and biospheric) estimation via atmospheric inversion. We identify the essential difference between fossil-fuel and biospheric CO_2 fluxes and motivate the need for a spatial parameterization for $ffCO_2$ emissions. We then review existing literature on compressive sensing, which provides the kernel of the spatial parametrization, as well as the optimization techniques which exploit the regularization such random field models can provide.

2.1 Estimation of CO_2 fluxes

$ffCO_2$ emissions, calculated from the consumption of fossil fuels, are reported by most countries and published by the International Energy Agency (http://www.iea.org) as well the United Nations Statistics Division [27]. They are usually published after aggregation to the national (and sometimes provincial/state) levels. $ffCO_2$ emissions can be thought of as a combination of intense point-sources (e.g., electricity generation and cement production) and diffuse spatially distributed emissions associated with transportation, residential and commercial activities. Emissions from individual, geotagged power plants can be obtained from CARMA [28] whereas [29] provides cement production data; both can be subtracted from national inventories. The remainder, constituting the diffuse sources, can be disaggregated onto a grid based on a number of easily-observed proxies of human activity. Emission inventories are generally employed in their gridded form.

Images of lights at night [30] have been correlated to many socioeconomic parameters [31] and are a common tool for disaggregating national or regional $ffCO_2$ emission onto a grid. This is discussed in [8], where gridded inventory with a kilometer resolution was constructed. In contrast, EDGAR (Emission Database for Global Atmospheric Research, [32]) and CDIAC (Carbon Dioxide Information Analysis Center, [33]) use population density to perform the disaggregation [10, 34]. In [9] both population density and nightlights are jointly assimilated to perform the disaggregation. The Vulcan inventory [3, 4] and its follow-on, Hestia [35, 36], follow a more complex method, consisting of both aggregation from census-tract data, as well as disaggregation to obtain their gridded inventory. They are not directly dependent on nightlights or population density. Nightlights and population density differ in their spatial distribution at the small (10 kilometer) scale and therefore the two disaggregation methods produce different results (see [9, 8] for discussion); further, the inventories are ultimately dependent on the accuracy of national reporting of emissions, which raises its own set of uncertainties [37, 12, 11].

$ffCO_2$ emissions or source strengths can also be back-calculated from measurements of CO_2 concentrations using a CO_2 transport model. Such an approach requires one to measure the concentration of $ffCO_2$, rather than CO_2 , in a given sample. This can performed by either measuring $\Delta^{14}CO_2$ or measuring CO and estimating $ffCO_2$ from CO/CO_2 ratios observed in incomplete combustion; see [25] for an example of the estimation of emissions from Sacramento, California, using airborne measurements. Alternatively, in conjunction with an atmospheric transport model, one can scale the emissions from an inventory, e.g., Vulcan, to reproduce observations in the vicinity of an urban area to estimate its emissions, as performed for Salt Lake City in [26]. In both cases a single variable was evaluated from the data - the source strength in case of Sacramento and the scaling factor for Salt Lake City. There was no attempt to back-calculate a spatially variable quantity.

Spatially variable *biospheric* CO_2 fluxes are routinely obtained via atmospheric inversions, using both

ground-based and satellite measurements. The spatial domain (a region or the entire Earth) is discretized with a grid; each grid-cell hosts a CO_2 flux source. A Bayesian inverse problem is posed using a transport model that linearly relates CO_2 concentration measurements at a set of locations to the strength of each of the sources. The resolution of the grid and the temporal discretization employed for temporally-varying – diurnal and seasonal– biospheric CO_2 fluxes result in more unknowns that can be constrained by the observations and regularization is used to reduce the effective dimensionality of the problem. The "true" fluxes are assumed to be distributed around a "prior"/guess (often obtained from process-based models of biospheric CO_2 fluxes e.g. CASA [38]); the discrepancy between the "true" and prior fluxes is modeled as a multivariate Gaussian field, whose covariance is calculated beforehand/offline. The inverse problem thereafter reduces to finding a CO_2 flux distribution that is a compromise between reproducing CO_2 observations and deviation from the prior fluxes. The multivariate Gaussian field, along with the prior fluxes, provide the regularization. Since the inverse problem is linear, the posterior distribution of the fluxes is also a multivariate Gaussian field. A review of inversion methods can be found in [20]. This approach has been successfully used with satellite retrievals [15] and ground-based measurements [39]. A scalable variational approach has successfully obtained global estimates of CO_2 fluxes at high resolution using satellite measurements [18, 40], and jointly with ground data [19]. Inversion methods based on ensemble Kalman filters and its derivatives have also been successfully used [17, 41, 21].

Given the paucity of observations, the prior fluxes influence the inferred CO_2 fluxes strongly [15]. This is especially true when inversions are required at a fine spatial resolution. The *geostatistical inverse* method was constructed to lessen its effect [22]. In [42, 43, 16], the prior fluxes were constructed as a linear model which included some easily observed proxies of biospheric and fossil-fuel fluxes (e.g., leaf-area index, population density, per capita GDP etc) with the sensitivities treated as unknowns to be estimated from data. The inversions inferred spatial patterns in CO_2 emissions which reflected both biospheric and anthropogenic contributions.

To summarize, biospheric CO_2 fluxes are routinely estimated from ground and satellite measurements using a variety of methods. Spatial resolutions finer than $1^\circ \times 1^\circ$ and temporal ones that resolve the diurnal cycle, can be achieved. This capability is due, in large part, to the smooth spatial variation of biospheric fluxes, which allows them to be modeled as multivariate Gaussian fields, and serves as a regularization in the inverse problem. In contrast, no such spatial parameterization exists for ffCO_2 emissions, limiting current work to estimating single parameters. In this paper, we will construct a MsRF model for ffCO_2 emissions, as a first step towards enabling their inference via atmospheric inversion. The spatial patterns in nightlights (and other proxies of human activity) will serve as the source of regularization. The approach will be tested in a synthetic data problem, where Vulcan [3, 4], which does not use these proxies, will supply the "ground truth" emissions.

2.2 Wavelet modeling

Wavelets are a family of orthogonal bases with compact support. They are generated using a scaling function ϕ' which obeys the recursive relationship

$$\phi'(x) = \sum_i c_i \phi'(2x - i).$$

A wavelet ϕ is generated from the scaling function by taking differences in the following manner:

$$\phi(x) = \sum_i (-1)^i c_{1-i} \phi'(2x - i).$$

Note that the wavelet has compact support. The choice of the filter coefficients c_i and ϕ' determine the type of the resulting wavelets. Wavelets with a large number of non-zero filter coefficients have larger support and are smoother. The simplest are the Haars, which are symmetrical in shape, but not smooth and not differentiable. They have only their first 2 moments equal to zero. Daubechies order 4 and order 6 (Daubechies 4 and 6) are smoother, with up to $4^{\text{th}}-$ (respectively 6^{th})-order moments that are zero. They have wider support and are asymmetrical in shape. Symlet wavelets are a modified form of Daubechies wavelets, which are more symmetrical. Wavelets can be shrunk and translated to model functions

$$\phi_{s,i} = 2^{\frac{s}{2}}\phi(2^s x - i)$$

where s is the dilation scale and i refers to translation (location). For each increment in scale, the support of the wavelet halves.

Wavelets are defined on dyadic (power-of-two) hierarchical or multi-resolution grids. Consider a domain of size D, discretized by a hierarchy of meshes with resolutions $\Delta D/D = \{1, 1/2, 1/2^2, \ldots 1/2^M\}$. The coarsest mesh, with $\Delta D = D$ is called the scale (or level) $s = 0$, whereas the finest mesh, with $\Delta D = D/2^M$ is the scale M mesh. Wavelets are defined on each of the levels of the hierarchical mesh. Their support spans the same number of grid-points on each scale i.e., wavelets' supports halve every level as we traverse down the hierarchical mesh. Wavelets can be translated or positioned at any grid-cell $i, 0 \leq i \leq 2^s - 1$, on any scale s of the hierarchical mesh. Thus a wavelet $\phi_{s,i}(x)$ requires the specification of s and i to denote its position in the hierarchical grid.

Consider a 1D function $g(x)$ defined on the hierarchical grid. It can be represented as

$$g(x) = w'\phi'(x) + \sum_{s=1}^{M}\sum_{i=0}^{2^s-1} w_{s,i}\phi_{l,i}(x).$$

The coefficients (or weights) $w_{s,i}$ and w' are obtained by taking projections of $f(x)$,

$$w_{s,i} = \int_{-\infty}^{\infty} g(x)\phi\left(\frac{x-b}{a}\right)dx$$

where $b = iL/2^s$, $a = 2^s$ and $\phi(\xi)$ is the wavelet on the coarsest level. The compact support of $\phi_{s,i}(x)$ and the large number of coefficients $w_{s,i}$ allow one to represent arbitrary $f(x)$ accurately, though not necessarily efficiently (i.e., with few non-zero $w_{s,i}$). Fast methods to compute the projection, called wavelet transforms, exist.

Wavelet transforms of a function are performed using Mallat's pyramid [44]. Wavelet coefficients c_i are determined by calculating weighted averages and differences, as we proceed recursively from the finest scale (grid resolution) to the coarsest (where the grid is reduced to 1 cell). The transform of a function $g(x)$, discretized on the mesh, can be represented by $\mathbf{w} = \mathcal{W}\mathbf{g}$, though $\mathcal{W}$ is never actually formed. $\mathbf{w}$ are the wavelet coefficients or weights. $\mathbf{w}$ contains 2^M wavelet coefficients. Details on wavelet decompositions and transforms are in [45, 23].

An arbitrary 2D field e.g., an image, can, in general, be expressed as a linear superposition of wavelets [45, 23]. A function $f(x,y)$, defined on a $D \times D$ domain and discretized on a hierarchical $2^M \times 2^M$ mesh, can be wavelet transformed by applying 1D wavelet transforms repeatedly, e.g., first by rows and then by columns. In 2D, wavelets of scale s have a support $2^{M-s} \times 2^{M-s}, 0 \leq s \leq M$. A wavelet of scale s can also be "translated" in space and positioned (in 2D space) at location $(i, j), 0 \leq (i, j) < 2^s$. Compact support and the ability to translate wavelets allow them to model complex, non-stationary fields. A 2D wavelet transform results

in $2^M \times 2^M$ wavelet coefficients. If the type of wavelets is chosen judiciously, many $w_{s,i,j}$ may be small and can be approximated as zero. In such a case we obtain an approximate, but *sparse* representation of $f(x,y)$ on the mesh.

2.3 Sparse reconstruction

Compressive Sensing (CS) [46, 47] is an efficient means of encoding sparse images. Consider an image $\mathbf{g}$ of size N. Assume, too, that it can be represented sparsely using $L \ll N$ wavelets. CS asserts that a sparse image may be sampled compactly by projecting on to a set of random vectors ψ_i, to obtain a measurement $\mathbf{g}'$, of size N_m, $L < N_m \ll N$,

$$\mathbf{g}' = \mathbf{\Psi g} = \mathbf{\Psi \Phi w}, \tag{2}$$

where the rows of $\mathbf{\Psi}$ consist of the random vectors ψ_j, and columns of $\mathbf{\Phi}$ consist of the orthonormal basis vectors ϕ_i. $\mathbf{\Phi}$ is a $N \times N$ matrix while $\mathbf{\Psi}$ is $N_m \times N$. The bulk of the theory was established in [48, 49, 50]. Sparsity expresses the idea that the "information rate" of a signal may be smaller than its bandwidth and this "information sparsity" may find a reflection in its concise representation in a suitable set of bases $\mathbf{\Phi}$ (see Eq. 2). $\mathbf{\Psi}$ and $\mathbf{\Phi}$ should be "incoherent". Incoherence relates to the fact that a concise signal in $\mathbf{\Phi}$ must be measured in a space $\mathbf{\Psi}$ where it is spread out, so that a few measurements may suffice to capture its behavior. Further, while sparsity is assumed, the support – the identity of basis vectors with (appreciably) non-zero weights – is not known *a priori* and is inferred along with the values of $\mathbf{w}$. The incoherence between $\mathbf{\Psi}$ and $\mathbf{\Phi}$ is ensured by choosing some well known wavelets bases (Haars, Daubechies 4 and 8) for $\mathbf{\Phi}$ and random vectors like noiselets [51], random sign ensembles, uniform spherical ensembles, partial Fourier and Hadamard ensembles [52] for $\mathbf{\Psi}$. In [53] the authors establish the degree to which compression might be performed

$$M \geq CL\log(N/L). \tag{3}$$

C is generally 3-5.

In CS, the reconstruction of $\mathbf{g}$ (alternatively, $\mathbf{w}$) can be performed using a number of methods. It can be posed as an optimization of the cost function

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left[\frac{1}{2} ||\mathbf{g}' - \mathbf{Aw}||_2 + \lambda ||\mathbf{w}||_1 \right]$$

which is a trade-off between the ℓ_1 norm of $\mathbf{w}$ (to enforce sparsity) and the ℓ_2 norm of the misfit between $\mathbf{g}'$ and the signal reconstructed from $\mathbf{w}$. $\mathbf{A} = \mathbf{\Psi \Phi}$. This can be solved using Basis Pursuit (BP) [54] or LASSO (Least Absolute-Shrinkage and Selection Operator) [55]. A slightly different formulation,

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^N}{\text{minimize}} \quad & ||\mathbf{w}||_1 \\ \text{subject to} \quad & ||\mathbf{g}' - \mathbf{Aw}||_2 < \varepsilon_2, \end{aligned} \tag{4}$$

which achieves the same effect, can be addressed using Matching Pursuit (MP) [56], Orthogonal Matching Pursuit (OMP, [57]) and Stagewise OMP [58] (StOMP). StOMP is much faster and offers many of the theoretical guarantees of BP. Bayesian equivalents also exist [59, 60], where Laplace priors are used to enforce sparseness in the inferred $\mathbf{w}$. All these algorithms are general and do not exploit any particular structure in the signal (except sparsity).

One may also create a model of the wavelet distribution and use it along with sparsity-enforcement when constraining $\mathbf{w}$. In a typical wavelet transform, the coefficients of the wavelets at different scales and translations can be stored in a binary tree. The tree can be pruned, *a priori*, based on a learning set of images

and thereafter employed with MP or OMP [61, 62, 63]. A Bayesian modification removed the need for a learning set [64]. Comparison with "regular" methods that did not exploit the tree-structure of wavelet coefficients showed a marked improvement in the quality and speed of reconstruction.

To summarize, CS has some striking parallels with the problem of atmospheric inversion. Fig. 1 (right) reveals that $ffCO_2$ emissions are clustered near the coasts and there are vast areas with little emissions. This indicates that a sparse wavelet representation of $ffCO_2$ emissions may be possible. Further, in Eq. 1 the transport matrix $\mathbf{H}$ serves much the same function as the measurement matrix $\mathbf{\Psi}$ in Eq. 2 – it samples the $\mathbf{f}$ and aggregates the contribution of each grid cell to the $ffCO_2$ concentration $\mathbf{y}^{obs}$ measured at the sensing locations. While $\mathbf{H}$ is not random, wind flow patterns over the United States are unlikely to align along wavelets and the incoherence requirement might be met. Given the limited number of measurement sites, we may only be able to reconstruct a very sparse subset of the wavelet coefficients required to represent $ffCO_2$ emissions accurately. Sparse reconstruction methods that allow us to do so were reviewed above.

This page intentionally left blank

3 Constructing a multiscale random field model

Here, we examine wavelet models of $ffCO_2$ emissions. We seek an approximate representation, which is low dimensional or sparse i.e., many of the $w_{s,i,k}$ may be set to zero. We subject $\overline{\mathbf{f}_V}$ to wavelet analysis. The emissions are described on a $2^M \times 2^M$ grid, $M = 6$ i.e., our hierarchical mesh has 6 levels. The spatial resolution is $1^\circ \times 1^\circ$ and the rectangular domain extents are given by the corners (24.5N,-63.5W) and (87.5N, -126.5W). $ffCO_2$ emissions are restricted to $\mathcal{R}$ the lower 48 states of the US.

We select a wavelet type e.g., Daubechies 4, and perform a wavelet transform of $\overline{\mathbf{f}_V}$. At each scale s, we identify the wavelet coefficient $w_{max,s}$ with the largest magnitude and set all wavelets $|w_{s,i,j}| < w_{max,s}/1000$ equal to zero. In Fig. 2 (left) we plot the fraction of non-zero wavelet coefficients, on each scale s, for each of the wavelet types. We see a significant decrease in the number of wavelet coefficients when using Haar wavelets, whereas the others show a decrease only at scale 4 and higher. Clearly Haars provide the sparsest representation of $ffCO_2$ emissions due to their non-smooth distribution in space. In Fig 2 (right), we plot the average and standard deviation of the non-zero wavelet coefficients. Most of the wavelet coefficients at the finer scales are set to zero, and the means are small, regardless of the wavelet type. We see that the means and standard deviations shrink, especially after scale $s = 3$; further, the distributions of wavelet coefficients arising from the different wavelet types begin to resemble each other. This arises from the fact that there are sharp boundaries around the areas where $ffCO_2$ emissions occur; when subjected to a wavelet transform, the region in the vicinity of a sharp boundary gives rise to large wavelet coefficients down to the finest scale. Thus the few non-zero wavelet coefficients at the finer scales assume similar values, irrespective of the wavelet type.

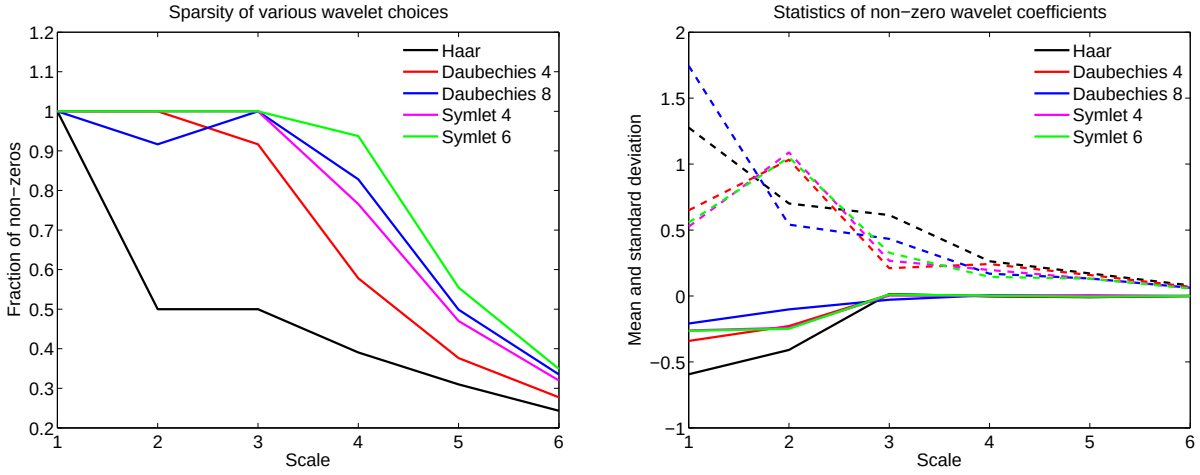


Figure 2: Left: The fraction of wavelet coefficients which are non-zero on each scale l, when subjected to a wavelet transform using Haars, Daubechies 4 and 6 and Symlet 4 and 6 wavelets. We see that Haar wavelets provide the sparsest representation. Right: We plot the average value of the non-zero coefficients (solid lines) and their standard deviation (dashed line). We find that while Haars may provide the sparsest representation, the non-zero values tend to be large and distinct.

Henceforth, we will proceed with Haar wavelets as the basis set of choice for representing $ffCO_2$ emissions,

since they provide the sparsest representation. We seek a spatial parametrization for ffCO$_2$ emissions, of the form

$$\mathbf{f} = w'\phi' + \sum_{s=1}^{M}\sum_{i,j} w_{s,i,j}\phi_{s,i,j}, \quad \{s,i,j\} \in W^{(s)} \tag{5}$$

where $W^{(s)}$ is a set containing a small number of Haar bases. We will select the components of $W^{(s)}$ using an easily observed proxy $\mathbf{X}$ of human activity (which correlates with ffCO$_2$ emissions). We will use radiance calibrated nightlights [65, 30, 66, 67] for the proxy. However, nightlight radiances are also affected by economic factors [68], and we will explore maps of built-up area [69, 70] as an alternative. As mentioned in [69], the map of built-up areas uses nightlight radiances in its computations, and so these are *not* independent proxies; however the built-up area map also includes information from IGBP (International Geosphere-Biosphere Programme [71]) land-cover map. The two choices for $\mathbf{X}$ will be compared with respect to (1) sparsity, i.e., the size of $W^{(s)}$ compared to 4^M, size of the full complement of wavelet bases on a $2^M \times 2^M$ grid, (2) the correlation between $\mathbf{X}$ and $\overline{\mathbf{f}_V}$ and (3) the ability of $W^{(s)}$ to capture $\overline{\mathbf{f}_V}$.

In Fig. 3 (top row), we plot maps of the two proxies, coarsened to 1° resolution. Comparing with Fig. 1 (right), we see that they bear a strong resemblance to $\overline{\mathbf{f}_V}$. We then subject $\mathbf{X}$ to a wavelet transform and set all wavelet coefficients $|w_{s,i,j}| < \delta$ to zero, where δ is a user-defined threshold. The bases with non-zero coefficients are selected to constitute $W^{(s)}$. We reconstitute a "sparsified" proxy, $\mathbf{X}^{(s)}$, using just the bases in $W^{(s)}$, and compute the correlation between $\mathbf{X}^{(s)}$ and $\overline{\mathbf{f}_V}$. Finally, we project $\overline{\mathbf{f}_V}$ onto $W^{(s)}$, obtain its "sparsified" form $\overline{\mathbf{f}_V}^{(s)}$, and compute the error $\varepsilon_f = ||\overline{\mathbf{f}_V}^{(s)} - \overline{\mathbf{f}_V}||_2 / ||\overline{\mathbf{f}_V}||_2$, where $||\quad||_2$ denotes the ℓ_2 norm. In Fig. 3 (middle row), we plot the sparsity, correlation and ε_f for various values of δ, for both nightlights and built-up areas. For nightlights, we achieve a sparsity of around 0.25 for $\delta < 10^{-2}$ i.e., we need retain only a quarter of the Haar bases to represent nightlights. The nightlights so represented bear a correlation of around 0.7 with $\overline{\mathbf{f}_V}$, and achieve an error ε_f of around 0.1. Note that this measure of error reflects the inability of the MsRF to represent fine-scale details i.e., if we were interested only in spatially-aggregated quantities, the error using the sparsified representation could be far less. In contrast, using built-up area as a proxy, we see that while the sparsity achieved is similar, the correlation between $\mathbf{X}^{(s)}$ and $\overline{\mathbf{f}_V}$ is slightly higher. The behavior of ε_f is similar, except the error increases faster with δ, as compared to nightlights. However both nightlights and built-up area maps show significant correlation with $\overline{\mathbf{f}_V}$ and the sparsified set of Haar bases that they (i.e., the proxies) provide (using $\delta = 10^{-2}$ in both the cases) allow us to construct a a low dimensional parametrization of ffCO$_2$ emissions.

Finally, we use $\mathbf{X}^{(s)}$ to create a "prior model" $\mathbf{f}_{pr} = c\mathbf{X}^{(s)}$ for ffCO$_2$ emissions, $\mathbf{f}$. c is computed such that

$$\int_{\mathcal{R}} \overline{\mathbf{f}_V} dA = \int_{\mathcal{R}} \mathbf{f}_{pr} dA = c \int_{\mathcal{R}} \left(w'^{,(X)}\phi' + \sum_{l,i,j} w^{(X)}_{s,i,j}\phi_{l,i,j} \right) dA, \quad \{l,i,j\} \in W^{(s)} \tag{6}$$

where $\mathcal{R}$ denotes the Lower 48 states of USA and $w^{(X)}_{l,i,j}$ are coefficients from a wavelet transform of $\mathbf{X}$. This implies that c is calculated such that both $\overline{\mathbf{f}_V}$ and $\mathbf{f}_{pr}$ provide the same value for the total emissions for the US. In Fig. 3 (bottom row), we plot the error $(\mathbf{f}_{pr} - \overline{\mathbf{f}_V})$. We see that neither nightlights nor the built-up area map provide a $\mathbf{f}_{pr}$ that is an accurate representation of $\overline{\mathbf{f}_V}$, though they share similar spatial patterns i.e., $\mathbf{f}_{pr}$ may be used to provide regularization for $\mathbf{f}$ in an inverse problem, but, by itself, is a poor predictor, regardless of the proxy $\mathbf{X}$ used to create it.

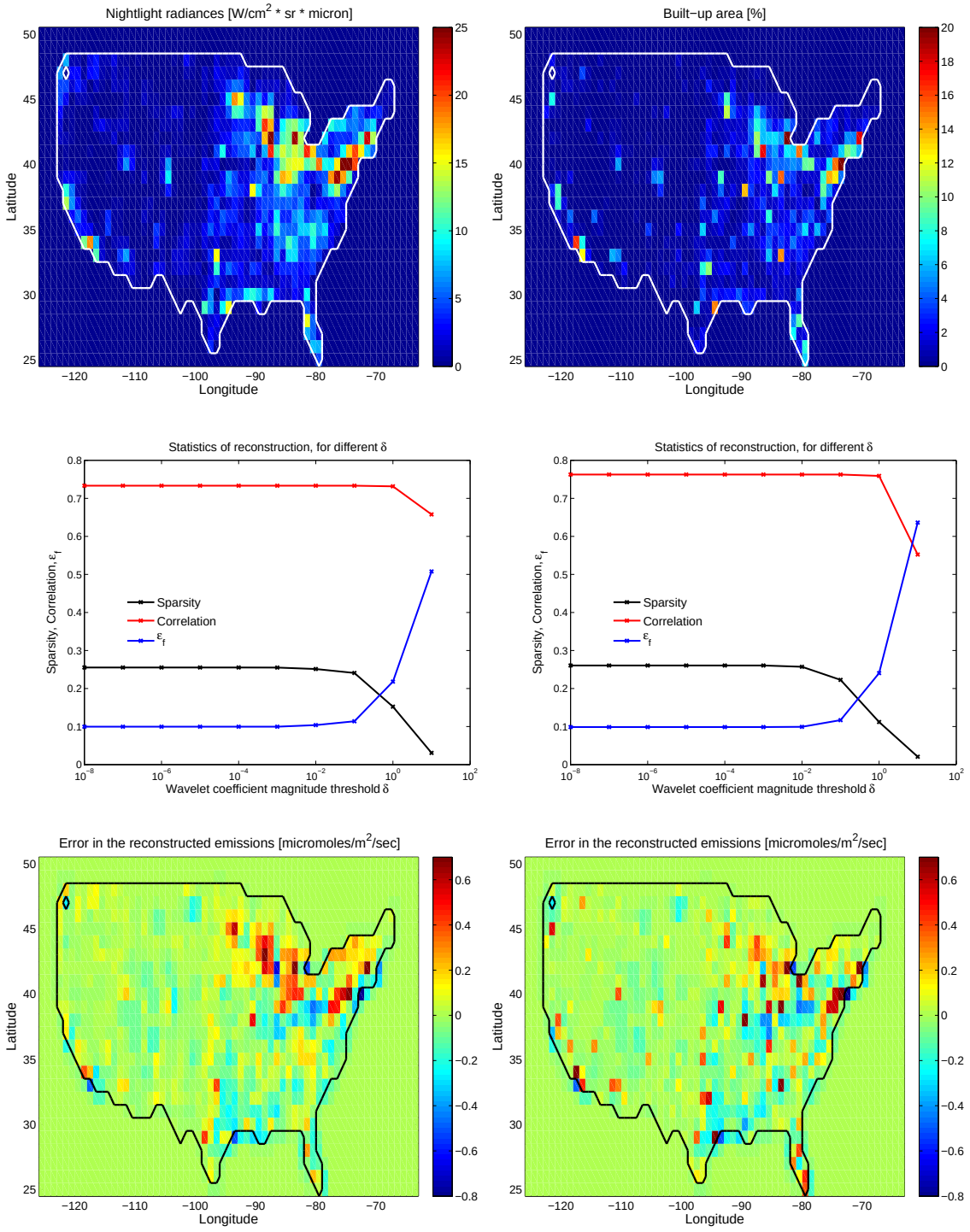


Figure 3: Top row: Maps of nightlight radiances (left) and built-up area percentage (right), for the US. Middle row: The sparsity of representation, the correlation between $\mathbf{X}$ and $\overline{\mathbf{f}_V}$ and the normalized error ε_f between the Vulcan emissions $\overline{\mathbf{f}_V}$ and the sparsified form obtained by projecting it on $\mathbf{X}$. These values are plotted for nightlights (left) and the built-up area maps (right). Bottom row: Plots of $(\mathbf{f}_{pr} - \overline{\mathbf{f}_V})$ obtained from nightlights (left) and built-up area maps (right).

This page intentionally left blank

4 Formulation of the estimation problem

In this section, we pose the inverse problem to estimate $ffCO_2$ emissions from limited measurements and describe a method to solve it. Fig. 4 provides a flowchart of the estimation procedure. Sec. 3 addressed the top half of the figure where we (1) selected Haar wavelets for modeling $ffCO_2$ emissions and (2) derived "prior" models for emissions based on nightlights and built-up area maps. The lower half of Fig. 4 illustrates the issues addressed in this section. We will use MsRF constructed from Haars to create a model for the (unknown) $ffCO_2$ fluxes and concentrations (which will also involve an atmospheric transport model). The fluxes will be constrained by observations of $ffCO_2$ concentrations. We then use sparse reconstruction, along with the "prior" model to estimate the $ffCO_2$ emissions.

Below, we briefly describe the transport model used to link observations at a set of measurement towers to emissions on a numerical grid. Thereafter, we pose the estimation problem, followed by a review of the method used to solve it. Finally, we describe how we enforce non-negativity in the estimated $ffCO_2$ emission estimates.

4.1 Transport model

The transport of CO_2 in the atmosphere is modeled using Eq. 1. The calculation of the sensitivities $\mathbf{H}$ is described in detail in the supplement to [2], and we provide a summary below.

The elements of the $\mathbf{H}$ matrix are calculated using the Stochastic Time-Inverted Lagrangian Transport Model (STILT) model [72]. STILT has already been used in a number of inversion studies for estimating biospheric CO_2 fluxes [73, 2]. It represents air arriving at the observation locations as an ensemble of particles, and transports them in reverse using wind fields. These wind fields are generated by numerical weather prediction models, in this particular case, the Weather Research & Forecasting (WRF) model [74] version 2.2. WRF has been customized for STILT and other transport models [75]. The details of the submodels used in the WRF-STILT calculations are in the supplementary material of [2]. The grid used for generating $\mathbf{H}$ was three-level, with a 2 kilometer resolution level around the 3 tallest towers (LEF, AMT and WKT; see Sec. B for details). The 2 km grid was embedded in a 10 km resolution grid over northern the Midwest and Gulf Coast region and New England, extending to approximately 105° W. An outermost, 40 km resolution, grid covered the rest of the domain (see Fig. 1 in [73] for a diagram). At each measurement location, 500 particles were released every hour and their 10-day back-trajectories calculated using meteorology from 2008.

Concentration footprints (or sensitivities) were calculated at 3-hour intervals by integrating the trajectories over the North American $1^\circ \times 1^\circ$ grid as described in [72]. The sensitivity of the CO_2 concentration at each observation location due to the flux at each grid-cell (the "footprint") is calculated in units of ppmv / μmol m^{-2} s^{-1} (ppmv: parts per million by volume). The footprint, which is the adjoint of the transport field, is calculated by counting the number of particles in a surface-influenced region (defined as a fraction of the estimated planetary boundary layer height at that grid-cell) for a given time-period spent in the region (for details, see [72]). When multiplied by a flux field, the sensitivities provide the contribution of fluxes to the mixing ratios (ppmv) at the measurement location/receptor. These sensitivities, calculated for 2008, were used in the estimation of biospheric fluxes [76, 21]. The sensitivities for CO_2 fluxes with an 8-day resolution (i.e., the ones used in this work) were obtained, from the 3-hour sensitivities described above by simply adding the $8 \times 24/3 = 64$ sensitivities that spanned the 8-day period. Thereafter, the grid-cells outside $\mathcal{R}$ were removed to obtain the $\mathbf{H}$ matrix used in this study. The size of the $\mathbf{H}$ matrix is $(K_s N_s) \times (N_{\mathcal{R}} K)$, where

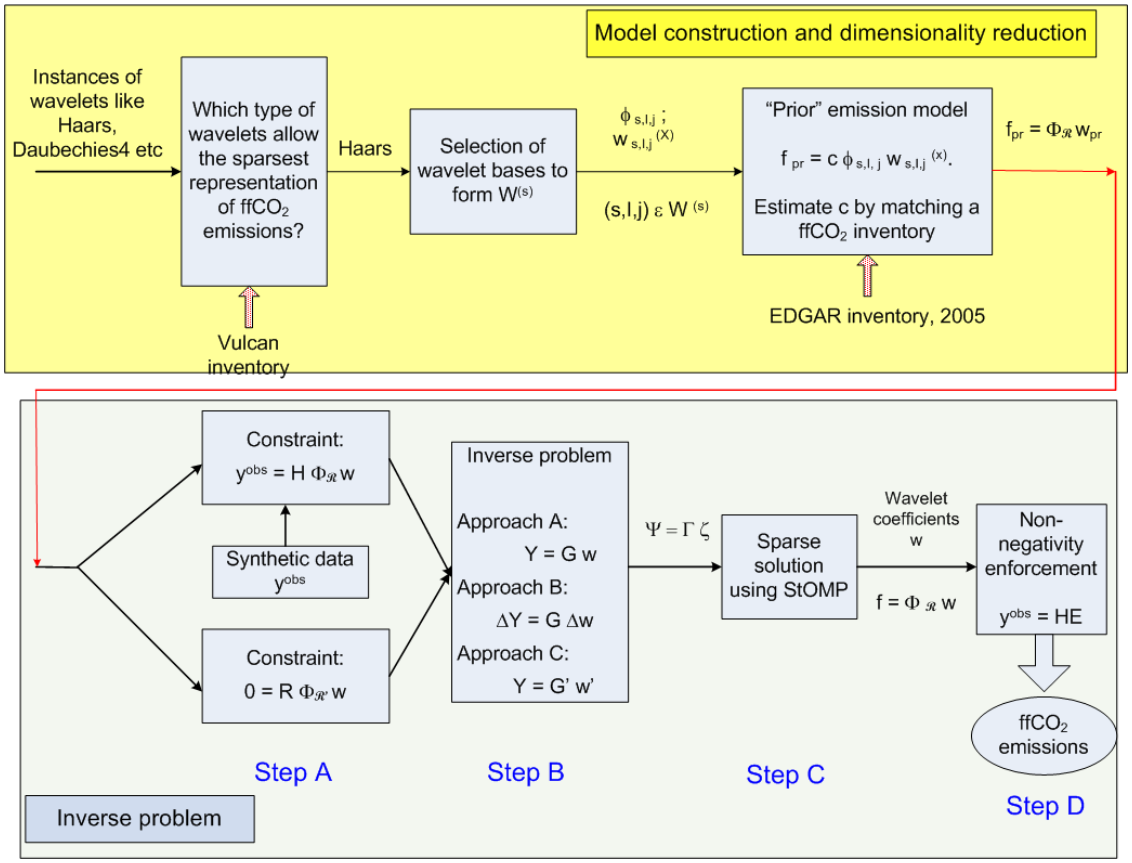


Figure 4: Flowchart of the procedure for performing the inversion of $ffCO_2$ emissions. The top half of the figure was addressed in Sec. 3. We start with a collection of wavelet families that *could* be used for modeling the emissions, and find that Haars provide the sparsest representation. Thereafter, we use the Haars to create a "prior" model $f_{pr} = c\widetilde{\mathbf{\Phi}}'_{\mathcal{R}} w$ for the emissions. In the lower half of the figure, we illustrate the posing and solution of the inverse problem. Step A, the posing of the inverse problem is described in Sec. 4.2. Three slightly different formulations (Step B), based on how the "prior" is incorporated into the inverse problem, are also described in the same section. Step C, the StOMP algorithm, is described in Sec. 4.3. Sec. 4.4, the description of non-negativity on the $ffCO_2$ emissions, corresponds to Step D.

K_s is the number of tower measurements every year ($24/3 \times 45 = 2880$), N_s is the number of sensors/towers, $N_{\mathcal{R}} = 816$ is the number of grid-cells in $\mathcal{R}$, the part of the domain covered by the lower 48 states of the US and $K = 45$ is the number of 8-day periods in 360 days (approximately a year).

4.2 The inverse problem

We seek to estimate $ffCO_2$ emissions over the region constituting the lower 48 states of the US, averaged over 8-day periods. The spatial distribution of emissions during an arbitrary 8-day period k is denoted by $\mathbf{f}_k$. We are interested in estimating emissions over an entire year i.e., we seek $\mathbf{F} = \{\mathbf{f}_k\}, k = 1 \ldots K$.

We will model emissions on the $2^M \times 2^M, M = 6$ mesh with wavelets:

$$\begin{aligned} \mathbf{f}_k &= w'_k \phi' + \sum_{s=1}^{M} \sum_{i,j} w_{s,i,j,k} \phi_{s,i,j}, \quad \{s,i,j\} \in W^{(s)} \\ &= \mathbf{\Phi} \mathbf{w}_k. \end{aligned} \tag{7}$$

Note that $\mathbf{\Phi}$ comprises of only those wavelets selected using $\mathbf{X}$ and contained in $W^{(s)}$, and not the 4^M wavelets that the grid can support. For the entire year, the expression for emissions becomes

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_K \end{pmatrix} = \text{diag}(\mathbf{\Phi}, \mathbf{\Phi}, \ldots, \mathbf{\Phi}) \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_K \end{pmatrix} = \widetilde{\mathbf{\Phi}} \mathbf{w}.$$

Since $\mathbf{\Phi} \mathbf{w}_k$ models the emissions over all grid-cells, and not just $\mathcal{R}$, $\mathbf{F}$ contains emissions over the lower 48 states, as well as the region outside it (where we have assumed that the emissions are non-existent). We separate out the two fluxes by permuting the rows of $\widetilde{\mathbf{\Phi}}$

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_{\mathcal{R}} \\ \mathbf{F}_{\mathcal{R}'} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{\Phi}}_{\mathcal{R}} \\ \widetilde{\mathbf{\Phi}}_{\mathcal{R}'} \end{pmatrix} \mathbf{w},$$

where $\widetilde{\mathbf{\Phi}}_{\mathcal{R}}$ and $\widetilde{\mathbf{\Phi}}_{\mathcal{R}'}$ are $(N_{\mathcal{R}} K) \times (LK)$ and $(N_{\mathcal{R}'} K) \times (LK)$ matrices respectively. Here L is the number of wavelets in $W^{(s)}$ and $N_{\mathcal{R}'} = 4096 - 816 = 3280$ is the number of grid-cells in $\mathcal{R}'$. The modeled concentrations at the measurement towers, caused by $\mathbf{F}_{\mathcal{R}}$, can be written as $\mathbf{y} = \mathbf{H}\mathbf{F}_{\mathcal{R}}$. For arbitrary $\mathbf{w}$, $\mathbf{F}_{\mathcal{R}'}$, the emissions in the region outside $\mathcal{R}$, are not zero. Consequently, it will be necessary to specify $\mathbf{F}_{\mathcal{R}'} = 0$ as a constraint in the inverse problem.

Specifying the constraint $\mathbf{F}_{\mathcal{R}'} = 0$ directly is not very efficient since it leads to $N_{\mathcal{R}'} K$ constraints. In a global inversion, or at resolutions higher than $1^\circ \times 1^\circ$, this could get very large. Consequently, we adapt an approach from compressive sensing to enforce this constraint approximately. Consider a $M_{cs} \times (N_{\mathcal{R}'} K)$ matrix $\mathbf{R}$, whose rows are direction cosines of random points on the surface of $N_{\mathcal{R}'} K$-dimensional unit sphere. This is called a uniform spherical ensemble and is used in compressive sampling [52]. The projection of the emission field $\mathbf{F}_{\mathcal{R}'}$ on $\mathbf{R}$ i.e., $\mathbf{R}\mathbf{F}_{\mathcal{R}'}$ compressively samples $\mathbf{F}_{\mathcal{R}'}$. Setting this projection to zero during inversion allows us to enforce zero emissions outside $\mathcal{R}$. However, to do so, we add only M_{cs} constraint equations rather than $N_{\mathcal{R}'} K$, which would be the case if we set the emission in each grid-cell in $\mathcal{R}'$ to zero. In Sec. 5.5, we will investigate the degree of computational saving afforded by imposing the $\mathbf{F}_{\mathcal{R}'} = 0$ constraint in this manner.

The optimization problem can be written as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}^{obs} \\ 0 \end{pmatrix} \approx \begin{pmatrix} \mathbf{H}\,\widetilde{\mathbf{\Phi}}_{\mathcal{R}} \\ \mathbf{R}\,\widetilde{\mathbf{\Phi}}_{\mathcal{R}'} \end{pmatrix} \mathbf{w} = \mathbf{G}\mathbf{w}. \tag{8}$$

In this equation, $\mathbf{G}$ is akin to $\mathbf{A}$ discussed in Sec. 2.3. It is a "sampling" of a basis set. The $\mathbf{R}\,\widetilde{\mathbf{\Phi}}_{\mathcal{R}'}$ component of $\mathbf{G}$ is formally a random sampling. The $\mathbf{H}\,\widetilde{\mathbf{\Phi}}_{\mathcal{R}}$ component of $\mathbf{G}$, on the other hand, "samples" the emission field as guided by the transport processes encoded in $\mathbf{H}$. The left hand side $\mathbf{Y}$ is approximately equal to $\mathbf{G}\,\mathbf{w}$ since the observations $\mathbf{y}^{obs}$ contain measurement errors that cannot be modeled with $\mathbf{H}$. This completes Step A in Fig. 4.

The wavelet coefficients $\mathbf{w}$ in Eq. 8 are not normalized and usually display a large range of magnitudes. The wavelets in $W^{(s)}$ at finer scales i.e. those with a small support / "footprint" tend to have coefficients with a large magnitude. Their small footprint cause the fine-scale wavelets to impact only neighboring measurement towers. In contrast, wavelets at the coarser scales have large "footprints" that span multiple measurement locations. Total emissions in $\mathcal{R}$, as well as $\mathbf{y}^{obs}$, are very sensitive to their coefficients. Solving Eq. 8, as-is, incorporates no information from $\mathbf{X}$ beyond the selection of wavelets to be included in $\widetilde{\mathbf{\Phi}}$. We explore the incorporation of $\mathbf{X}$ in the estimation of $\mathbf{w}$ using three different approaches:

Approach A : This is the baseline approach and solves Eq. 8 as-is. The lack of normalization of $\mathbf{w}$, in conjunction with the optimization procedure described below in Sec. 4.3, leads to artifacts which will be described in Sec. 5.1.

Approach B : In this formulation, $\mathbf{X}$ is included in the optimization as a "prior". We write the emissions as $\mathbf{F} = \mathbf{f}_{pr} + \Delta\mathbf{F}$. Substituting into Eq. 8, we get $\mathbf{Y} \approx \mathbf{H}\mathbf{f}_{pr} + \mathbf{G}\Delta\mathbf{w}$, where $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{(X)}$. Here, $\mathbf{w}_{(X)} = c\{w'^{,(X)}, w^{(X)}_{s,i,j}\}$, $\{s,i,j\} \in W^{(s)}$, where c is obtained from Eq. 6. It provides a wavelet decomposition of $\mathbf{f}_{pr}$ using the bases in $W^{(s)}$. Simplifying, we get

$$\Delta\mathbf{Y} = \mathbf{Y} - \mathbf{H}\mathbf{f}_{pr} \approx \mathbf{G}\Delta\mathbf{w}, \tag{9}$$

The term "prior" model is used somewhat loosely since this is not a Bayesian method. However, $\mathbf{f}_{pr}$ serves a similar function by providing regularization in the inverse problem.

Approach C : The incorporation of the spatial patterns in $\mathbf{X}$ into the estimation procedure can be performed in an alternative manner. We note that $\mathbf{w}_{(X)}$ can be used to normalize $\mathbf{w}$. We rewrite Eq. 8 as

$$\mathbf{Y} \approx \mathbf{G}\ \mathrm{diag}(\mathbf{w}_{(X)})\ \mathrm{diag}(\mathbf{w}_{(X)}^{-1})\mathbf{w} = \mathbf{G}'\mathbf{w}' = \begin{pmatrix} \mathbf{H}\ \widetilde{\mathbf{\Phi}}'_{\mathcal{R}} \\ \mathbf{R}\ \widetilde{\mathbf{\Phi}}'_{\mathcal{R}'} \end{pmatrix} \mathbf{w}', \tag{10}$$

where $\mathbf{w}' = \{w_{s,i,j}/w_{(X),s,i,j}\}, \{s,i,j\} \in W^{(s)}$, is the normalized set of wavelet coefficients, $\widetilde{\mathbf{\Phi}}'_{\mathcal{R}} = \widetilde{\mathbf{\Phi}}_{\mathcal{R}}\ \mathrm{diag}(\mathbf{w}_{(X)})$ and $\widetilde{\mathbf{\Phi}}'_{\mathcal{R}'} = \widetilde{\mathbf{\Phi}}_{\mathcal{R}'}\ \mathrm{diag}(\mathbf{w}_{(X)})$.

In all the three cases, we obtain an underdetermined set of linear equations of the form

$$\mathbf{\Upsilon} \approx \mathbf{\Gamma}\boldsymbol{\zeta}. \tag{11}$$

This completes Step B in Fig. 4. We outline the solution procedure in Sec. 4.3 and compare the performance of the three formulations in Sec. 5.1.

4.3 Solving the inverse problem

Eq. 11 is solved using an optimization procedure. Since we obtain our observations from a set of measurement locations that were sited with an eye towards biospheric CO_2 fluxes (see Appendix B), it is unlikely that they will constrain all the elements of $\boldsymbol{\zeta}$. Here we fall back on our multiscale model of $\mathbf{F}_{\mathcal{R}}$, which explicitly parametrizes orthogonal spatial patterns (the wavelets) of different scales. Further, only those patterns that were observed in the $\mathbf{X}$ field were retained in $W^{(s)}$. $\mathbf{y}^{obs}$ is probably sufficient to estimate the larger patterns, whereas it may not be possible to constrain the finer details. However, *a priori*, we do not know the identity of these "un-constrainable" details.

In our solution of Eq. 11, we will attempt to identify the wavelets coefficients that cannot be estimated from $\boldsymbol{\Upsilon}$ and drive them to zero i.e, the solution method will impose sparsity on $\boldsymbol{\zeta}$. This will provide us with a reconstruction of $\mathbf{F}_{\mathcal{R}}$, containing only those features/patterns that are supported by data. The sparse solution of Eq. 11 is performed using Stagewise Orthogonal Matching Pursuit (StOMP, [58]). Eq. 11 is recast similar to Eq. 4

$$\begin{aligned} \underset{\boldsymbol{\zeta}\in\mathbb{R}^N}{\text{minimize}} \quad & ||\boldsymbol{\zeta}||_1 \\ \text{subject to} \quad & ||\boldsymbol{\Upsilon}-\boldsymbol{\Gamma}\boldsymbol{\zeta}||_2^2 < \varepsilon_2. \end{aligned} \tag{12}$$

$||\boldsymbol{\zeta}||_1$ is minimized by setting as many elements of $\boldsymbol{\zeta}$ to zero as possible, thus enforcing sparsity. Meanwhile, the constraint $||\boldsymbol{\Upsilon}-\boldsymbol{\Gamma}\boldsymbol{\zeta}||_2$ ensures that the solutions being proposed by the optimization procedure provide a good reproduction of the observations. While the details of StOMP are in [58], we summarize the algorithm below.

StOMP operates iteratively via S stages, building up a sequence of approximations $\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \ldots$ by removing detected structure from a sequence of residual vectors $r_1, r_2, \ldots$. We start with an initial guess $\boldsymbol{\zeta}_0 = 0$ and initial residual $r_0 = \boldsymbol{\Upsilon}$. The stage counter is initialized $s = 1$. The algorithm maintains a sequence of estimates $I_1, I_2, \ldots$ of the locations of non-zeros in $\boldsymbol{\zeta}$.

In the l^{th} stage, we obtain a vector of residual correlation

$$c_l = \boldsymbol{\Gamma}^T r_{l-1}$$

which we consider to be a vector with a few significant non-zeros and the rest containing Gaussian noise of small magnitude. Based on a threshold calculated from the assumption that c_l is mostly i.i.d. Gaussian noise, we identify a small set J_l of "large" coordinates:

$$J_l = \{j : |c_l(j)| > t_l \sigma_l\}$$

where $\sigma_l = ||r_l||_2/\sqrt{|r_l|}$, $2 \leq t_l \leq 3$ and $|r_l|$ is the length of the vector r_l. The newly detected locations of non-zeros are added to the running list of the support of $\boldsymbol{\zeta}$

$$I_l = I_{l-1} \cup J_l.$$

We then project $\boldsymbol{\Upsilon}$ on the columns of $\boldsymbol{\Gamma}$ contained in the enlarged support. Let $\boldsymbol{\Gamma}_I$ denote the $(K_s N_s + M_{cs}) \times |I|$ matrix constructed with the columns chosen using the index set I. The new approximation $\boldsymbol{\zeta}_l$, with support I_l is given by

$$(\boldsymbol{\zeta}_l)_{I_l} = \left(\boldsymbol{\Gamma}_{I_l}^T \boldsymbol{\Gamma}_{I_l}\right)^{-1} \boldsymbol{\Gamma}_{I_l}^T \boldsymbol{\Upsilon}$$

and the updated residual is

$$r_l = \boldsymbol{\Upsilon} - \boldsymbol{\Gamma}\boldsymbol{\zeta}_l.$$

We check a stopping criterion, and proceed to the next iteration after setting $:= l + 1$.

This completes Step C in Fig. 4.

4.4 Enforcing non-negativity of $\mathbf{F}_{\mathcal{R}}$

Estimates of $\mathbf{w}$ calculated by StOMP do not necessarily provide $\mathbf{F}_{\mathcal{R}} = \mathbf{H}\,\widetilde{\boldsymbol{\Phi}}_{\mathcal{R}}\,\mathbf{w}$ that are non-negative. In practice the negative ffCO_2 emissions occur in only a few grid-cells and are usually small in magnitude. A

large fraction of elements of $\mathbf{F}_{\mathcal{R}}$ are set to zero by StOMP. Having identified the sparsity pattern, i.e., the spatial scales that can and cannot be estimated from $\mathbf{Y}$, we devise an iterative procedure for enforcing non-negativity on $\mathbf{F}_{\mathcal{R}}$. We discard $\mathbf{F}_{\mathcal{R}'}$ and manipulate the emissions in $\mathcal{R}$ directly, rather than via the wavelet coefficients.

We seek non-negative ffCO$_2$ emissions $\mathbf{E} = \{E_i\}, i = 1 \ldots Q, Q = (N_{\mathcal{R}} K)$ such that

$$\frac{||\mathbf{y}^{obs} - \mathbf{H}\mathbf{E}||_2}{||\mathbf{y}^{obs}||_2} \leq \varepsilon_3. \tag{13}$$

$\mathbf{E}$ is constructed iteratively through a sequence $\mathbf{E}_1, \mathbf{E}_2, \ldots$. $\mathbf{E}_0$ is initialized by using the absolute values of $\mathbf{F}_{\mathcal{R}}$ calculated by solving Eq. 12.

At each iteration m, we seek a correction $\xi = \{\xi_i\}, i = 1 \ldots Q$, where $|\xi_i| \leq 1$, such that

$$\begin{aligned}
\mathbf{E}^{(m)} &= \text{diag}(\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_Q))\mathbf{E}^{(m-1)} \\
&\approx \text{diag}(1+\xi_1, 1+\xi_2, \ldots, 1+\xi_Q)\mathbf{E}^{(m-1)} \\
&= \mathbf{E}^{(m-1)} + \Delta\mathbf{E}^{(m-1)}, \quad \text{where } \Delta\mathbf{E}^{(m-1)} = \xi^T \mathbf{E}^{(m-1)}.
\end{aligned}$$

Since the emissions must satisfy $\mathbf{y}^{obs} \approx \mathbf{H}\mathbf{E}^{(m)}$, we get

$$\mathbf{y}^{obs} - \mathbf{H}\mathbf{E}^{(m-1)} = \Delta\mathbf{y} \approx \mathbf{H}\Delta\mathbf{E}^{(m-1)} \tag{14}$$

This is an underconstrained problem, and we seek the sparsest set of updates $\Delta\mathbf{E}^{(m-1)}$ using StOMP.

The corrections are calculated, and the emissions updated as

$$\begin{aligned}
\xi_i &= \text{sgn}\left(\frac{\Delta E_i^{(m-1)}}{E_i^{(m)}}\right) \max\left(1, \left|\frac{\Delta E_i^{(m-1)}}{E_i^{(m)}}\right|\right), \\
E_i^{(m)} &= E_i^{(m-1)} \exp(\xi_i),
\end{aligned} \tag{15}$$

to obtain $\mathbf{E}^{(m)}$.

The convergence requirement Eq. 13 is checked with $\mathbf{E}^{(m)}$, and if not met, the iteration count is updated $m := m+1$ and Eq. 14 is solved again. This completes Step D in Fig. 4.

To summarize, the solution of the inverse problem proceeds in two steps:

1. *Step I – Sparse estimation of $\mathbf{F}_{\mathcal{R}}$ using the wavelet model:* This is performed by StOMP, which explicitly identifies small scales that cannot be constrained by $\mathbf{y}^{obs}$ and drives them to zero. The estimation of $\mathbf{F}_{\mathcal{R}}$ is performed by calculating the wavelet coefficients $\mathbf{w}$. We obtain an approximation to the emissions, $\mathbf{F}_{\mathcal{R}}$, which may be negative in certain grid-cells.

2. *Step II – Enforcement of non-negativity:* Non-negative emissions $\mathbf{E}$ are calculated by iteratively updating $|\mathbf{F}_{\mathcal{R}}|$ with a multiplicative correction. Updates are performed on fluxes in grid-cells, not the wavelet coefficients. Each iteration includes an invocation of StOMP.

5 Tests

In this section, we test the multiscale parameterization for $ffCO_2$ emissions, developed in Sec. 3, using the sparse estimation technique in Sec. 4.

Generating synthetic observations: The true $ffCO_2$ emissions are obtained, for 2002, from the Vulcan inventory [3, 4]. Vulcan provides emissions in $\mathcal{R}$ and we assume that there are no emissions elsewhere. Hourly Vulcan fluxes are coarsened from 0.1° resolution to 1°, and averaged to 8-day periods. These fluxes are multiplied by $\mathbf{H}$ to obtain $ffCO_2$ concentrations at $K_s = 35$ measurement towers (see Appendix B). Observations are available every 3 hours and span a full year. A measurement error $\varepsilon \sim N(0, \sigma^2)$ is added to the concentrations to obtain $\mathbf{y}^{obs}$, as used in Eq. 8. The same σ is used for all towers and is set to 0.01 ppm (approximately, 1% of the average concentration observed at the 35 towers).

5.1 Comparison of optimization formulations

We solve Eq. 8 using the StOMP method (Sec. 4.3) and enforce non-negativity on $\mathbf{F}_{\mathcal{R}}$ to obtain $\mathbf{E}$. The coefficients $\mathbf{w}_{(\mathbf{X})}$ used in Eq. 9 and Eq. 10 are obtained from a wavelet decomposition of $\mathbf{f}_{pr}$ based on nightlights (Sec. 3). The constant c in Eq. 6 is obtained by using fluxes from the EDGAR inventory [32] for 2005 i.e., instead of using emissions from Vulcan to calculate $\overline{\mathbf{f}_V}$, we use EDGAR. We thus ensure that Vulcan is not used in any capacity during the inverse modeling (except to generate the synthetic observations). The inversion is performed for $k = 1 \ldots 45$, for the entire year. The following parameters are used in the inversion process (Sec. 4.3 & 4.4): $\varepsilon_2 = 10^{-5}, \varepsilon_3 = 5.0 \times 10^{-4}, M_{cs} = 13{,}500$ i.e., 300 random projections for each 8-day period.

In Fig. 5 we plot the estimated emissions during the 31^{st} 8-day period, as calculated using Approaches A, B and C. The true emissions are also plotted for reference. Four quadrants are also plotted for easier comparison and reference. The distribution of towers is very uneven, with most of the towers being concentrated in the Northeast quadrant. We see that Approach A (Fig. 5, top right) provides estimates that have large areas in the Northwest (NW) and Southwest (SW) quadrants with low levels of $ffCO_2$ emissions. In contrast, the true emissions (Fig. 5, top left) are mostly empty, which is also borne out by nightlight and built-up area maps in Fig. 3. Thus we see that the minimization of $||\boldsymbol{\zeta}||_1$ (alternatively $||\mathbf{w}||_1$) drives the wavelet coefficients to small values, but not identically to zero. In Fig. 5 (bottom left), Approach B provides estimates that show much structure in the Eastern quadrants, and the patterns seen in nightlights (Fig. 3) are clearly reproduced. The reason is as follows. While $\mathbf{f}_{pr}$ captures the broad, coarse scale patterns of $ffCO_2$ emissions, it incurs significant errors at the finer scales. Eq. 9 essentially seeks to rectify these errors. However, as mentioned in Sec. 4.2, fine-scale wavelets tend to have large wavelet coefficients and the minimization of $||\boldsymbol{\zeta}||_1$ (alternatively $||\Delta\mathbf{w}||_1$) removes them since the constraint $||\boldsymbol{\Upsilon} - \boldsymbol{\Gamma}\boldsymbol{\zeta}||_2^2 < \varepsilon_2$ is not very sensitive to individual wavelets at the fine scale (they only affect neighboring towers significantly). The inability to rectify the fine-scale discrepancies lead to a final $ffCO_2$ estimate that resembles $\mathbf{f}_{pr}$ in the finer details. Fig. 5 (bottom right) plots the estimates obtained using Approach C, which uses normalized wavelet coefficients $\mathbf{w}'$. The estimates from Approach C show large areas of little or no emissions in the Western quadrants, similar to the true emissions in the top left figure. In the Eastern quadrants, the emissions show less spatial structure than the true emissions as well as those obtained using Approach A.

In Fig. 6 (left) we evaluate the accuracy of the reconstruction quantitatively. We total the emissions in $\mathcal{R}$ to obtain the country-level $ffCO_2$ emissions and compare that with the emissions from Vulcan. We plot a

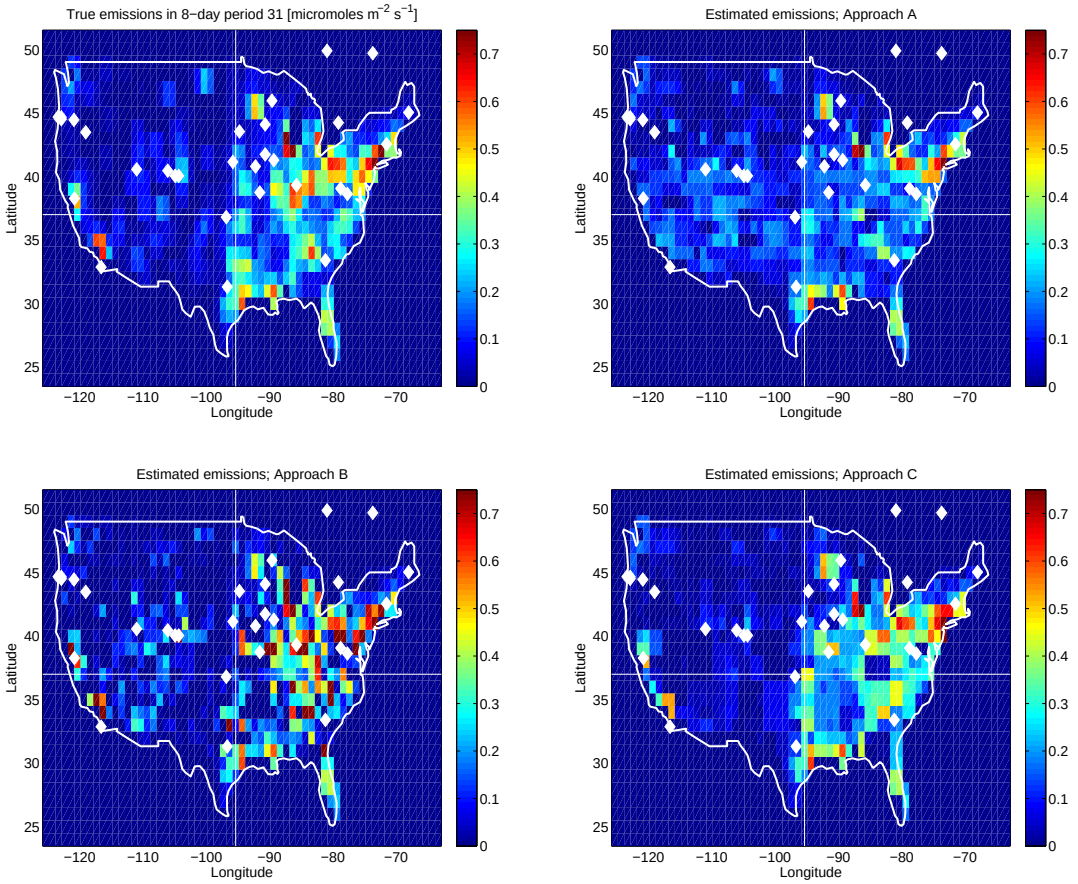


Figure 5: Plots of ffCO$_2$ emissions during the 31st 8-day period. Top left, we plot true emissions from the Vulcan inventory. Top right, the estimates from Approach A. Bottom left and right figures contain the estimates obtained from Approaches B and C respectively. Each figure contains the measurement towers as white diamonds. Each figure is also divided into quadrants. We see that Approach A, unconstrained by $\mathbf{f}_{pr}$ provides low levels of (erroneous) emissions in large swathes of the Western quadrants. Approach B reflects $\mathbf{f}_{pr}$ very strongly. Approach C provides a balance between the influence of $\mathbf{f}_{pr}$ and the information in $\mathbf{y}^{obs}$.

time-series of errors defined as a percentage of total, country-level Vulcan emissions

$$\text{Error}_k\ (\%) = \frac{100}{K}\sum_{k=1}^{K}\frac{E_k - E_{V,k}}{E_{V,k}}, \quad \text{where} \quad E_k = \int_{\mathcal{R}} \mathbf{E}_k\, dA \quad \text{and} \quad E_{V,k} = \int_{\mathcal{R}} \mathbf{f}_{V,k}\, dA. \tag{16}$$

Here, $\mathbf{f}_{V,k}$ are Vulcan emissions averaged over the k^{th} 8-day period and $\mathbf{E}_k$ are the non-negativity enforced emission estimates in the same time period. A positive error denotes an overestimation by the inverse problem. In Fig. 6 (right) we plot the correlation between the true and reconstructed emissions over the same duration. This was done by calculating the Pearson correlation coefficient between the vector of grid-cells covering the United States. It is clear that Approach B provides the worst reconstructions, with the largest errors and smallest correlations. Approach C tends to over-predict emissions, whereas Approach B tends to under-predict them, often by approximately the same (small) amount. The correlations of the reconstructed emissions from these two emissions are about the same.

In Fig 7 we see the essential difference between Approach A and C. We plot the reconstruction error (left

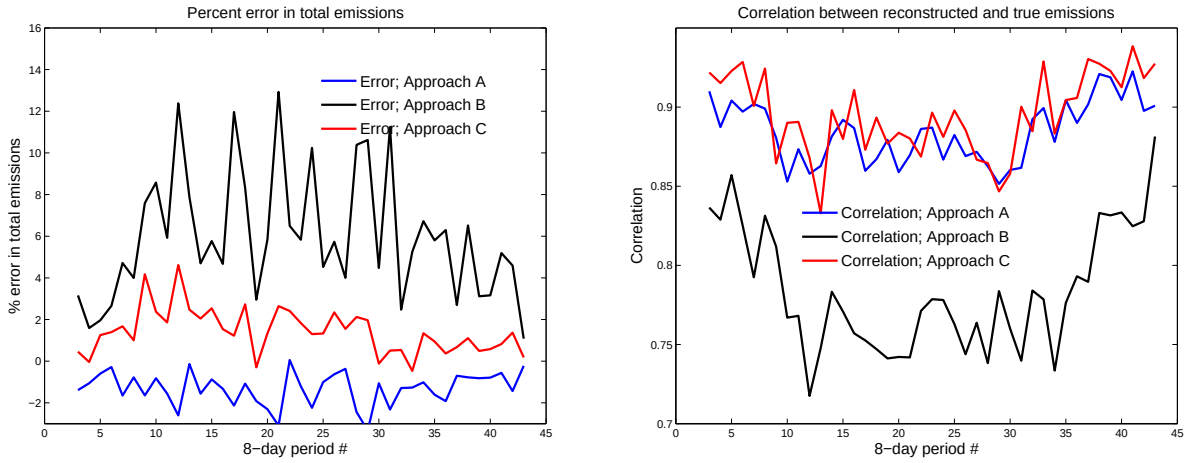


Figure 6: Comparison of estimation error (left) and the correlation between true and estimated emissions (right) using Approaches A, B and C. It is clear that Approach B is inferior to the others.

figure) and correlation between true and reconstructed emissions (right figure) in the Northeast (NE) and Northwest (NW) quadrants. Errors in the emissions are represented as a percentage of the total (true) emissions in that quadrant. We see the Approach C has smaller errors in both the quadrants. It also provides higher correlation in the NW quadrant which does not have many measurement towers (white diamonds in Fig. 5). Thus normalization using $\mathbf{w}_{(X)}$ and minimization of $||\boldsymbol{\zeta}||_1$ (alternatively $||\mathbf{w}'||_1$) prevents large departures from $\mathbf{f}_{pr}$ and also rectifies the tendency to remove large wavelet coefficients belonging to the finer wavelets. Approach C therefore provides a formulation that is more accurate and robust at the quadrant scale. Note, however, that both Approach A and C have similar fidelity at the scale of $\mathcal{R}$.

In the tests below, we will restrict ourselves to Approach C.

5.2 Evaluating formulation using compressive sensing metrics

In this section we compare Approach A versus C in terms of certain compressive sensing metrics to explain the lower accuracy of the estimates developed using Approach A. Note that in Approach A, sparsity is the sole source of regularization.

As specified in Eq. 2, the process of acquiring compressive measurements $\mathbf{g}'$ involves project wavelet bases $\boldsymbol{\Phi}$ on the rows of $\boldsymbol{\Psi}$. Since the rows $\psi_{i,\cdot}$ of $\boldsymbol{\Psi}$ are random unit vectors, they are neither aligned with nor orthogonal to the bases $\phi_{\cdot,j}$ and consequently the elements of $\mathbf{A}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}\boldsymbol{\Phi}$ tend to assume a distribution of magnitudes which are generally not very close to 0 or 1. This *coherence* $\mu(\boldsymbol{\Psi},\boldsymbol{\Phi})$ is defined as [77]

$$\mu(\boldsymbol{\Psi},\boldsymbol{\Phi}) = \sqrt{N} \max_{1\le i,j\le N} | < \psi_{i,\cdot}, \phi_{\cdot,j} > |, \qquad 1 \le \mu(\boldsymbol{\Psi},\boldsymbol{\Phi}) \le \sqrt{N} \tag{17}$$

Values of μ near 1 (alternative $\max | < \psi_{i,\cdot}, \phi_{\cdot,j} > |$ near $1/\sqrt{N}$) indicate incoherence between $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ and more efficient sampling. Note that the lower bound assumes that $\boldsymbol{\Psi}$ is a random matrix.

In compressive sensing, random matrices such as Gaussians, Hadamard, Circulant/Toeplitz or functions such as noiselets [52, 78, 79, 80] serve as $\boldsymbol{\Psi}$. In Fig. 8, we plot the distribution of $\log_{10}(|A_{i,j}|)$, the elements

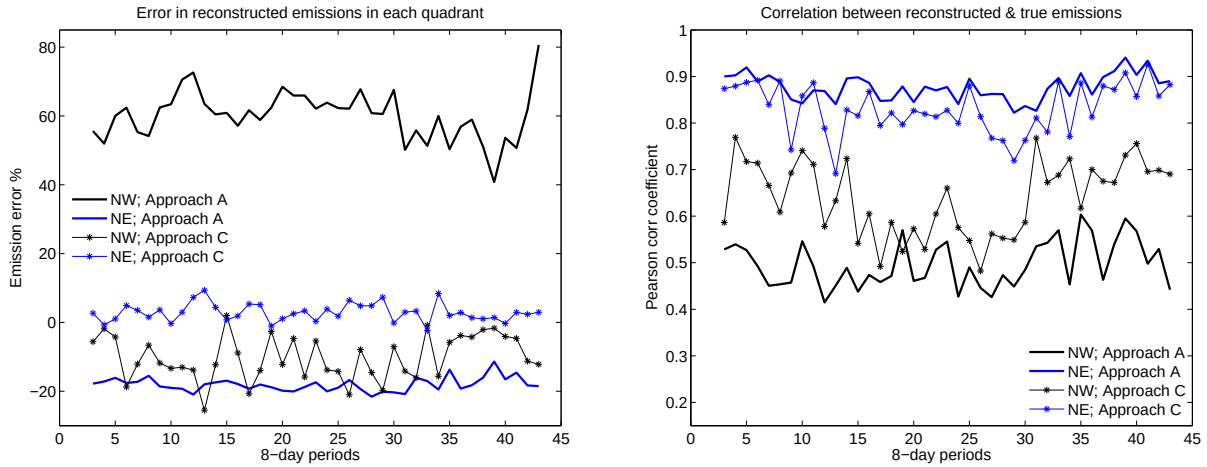


Figure 7: Reconstruction error (left) and correlation between the true and estimated emissions, using Approaches A and C, for the Northeast (NE) and Northwest (NW) quadrants. We see that Approach C, which includes information from $\mathbf{f}_{pr}$, leads to lower errors in both the quadrants and better correlations in the less instrumented NW quadrant.

of $\mathbf{A}_{\mathbf{\Psi}}$ for some "standard" sampling matrices. $\mathbf{\Phi}$ contains the wavelets used in Sec. 5.1 i.e., those selected using nightlights. We see that $\log_{10}(|A_{i,j}|)$ may assume continuous (Gaussian and circulant sampling matrices) or discrete (Hadamard, scrambled-block Hadamard and Noiselets) distributions, and generally lie between -3 and -1. The samples collected by these projection schemes are sufficiently informative, due to the incoherence between $\mathbf{\Psi}$ and $\mathbf{\Phi}$, to allow reconstruction of the original image with sparsity as the sole regularization.

In Eq. 8, $\mathbf{H}$ serves a similar sampling purpose. $\mathbf{H}$ is determined by atmospheric transport processes and is not a random matrix. However, the efficiency of sampling depends on the incoherence between $\mathbf{H}$ and $\mathbf{\Phi}$. We construct a new $\mathbf{H}'$ by picking the rows of $\mathbf{H}$ corresponding to 2 towers and for the 21$^{\text{st}}$ and 22$^{\text{nd}}$ 8-day periods. We compute $\mathbf{A}_{\mathbf{H}'} = \mathbf{H}'\mathbf{\Phi}$, and in Fig. 8, plot the log-transformed magnitudes of the elements of $\mathbf{A}_{\mathbf{H}'}$. The distributions for the two towers are almost identical. We clearly see that, unlike $\mathbf{A}_{\mathbf{\Psi}}$, $\mathbf{A}_{\mathbf{H}'}$ contains a significant number of elements that are close to 1, and a large number of elements which are close to 0 (e.g. near 10^{-6}). This is a consequence of the rows of $\mathbf{H}'$ being approximately aligned to some of the columns of $\mathbf{\Phi}$ and consequently, nearly orthogonal to others. The small values in $\mathbf{A}_{\mathbf{H}'}$ indicate that the CO_2 concentration prediction $\mathbf{y}$ at the two selected towers are insensitive to many of the wavelets i.e., to many scales and locations. In fact, the dominance of near-field CO_2 fluxes on tower measurements is well known [81] and is responsible for the particular structure of $\mathbf{A}_{\mathbf{H}'}$.

Further, the coherence $\mu(H', \mathbf{\Phi})$ is larger than $\mu(\mathbf{\Psi}, \mathbf{\Phi})$, indicating that the sampling efficiency of $\mathbf{H}'$ is inferior to what is generally achieved in compressive sensing. Consequently a combination of sparsity and a "prior" emission model $\mathbf{f}_{pr}$ were required to regularize the problem and enhance the accuracy of the emission estimates in Approach C.

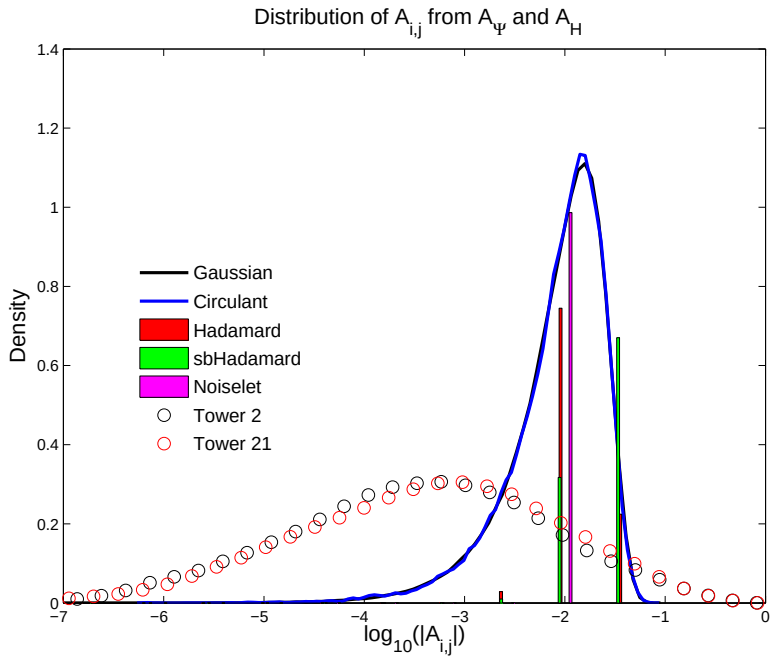


Figure 8: Comparison of the distribution of the elements of $\mathbf{A_\Psi}$ and $\mathbf{A_\Phi}$. We see that Gaussian and circulant random matrices lead to continuous distributions whereas Hadamard, scrambled-block Hadamard (sbHadamard) and noiselets serving as sampling matrices lead to $\mathbf{A_\Psi}$ where the elements assume discrete values. In contrast, the elements of $\mathbf{A_{H'}}$ assume values which are spread over a far larger range, some of which are quite close to 1 while others are very close to zero.

5.3 Inversions with nightlights as proxy

In Fig. 9, we plot the true and reconstructed emissions for 3 8-day periods ($k = 9, 33, 42$). We see that due to the abundance of measurement towers in the NE quadrant, the reconstruction there is accurate. On the other hand, the tower density on the West coast is sparse and gives rise to significant inaccuracies. For example, in 8-day periods 33 and 42, we see that the Los Angeles–San Diego region (Southwest quadrant) is underestimated, since the only tower is in La Jolla (32.87N, -117.26W). The estimated emissions in the center of the country (Continental Divide and Great Plains, in the Western quadrants) show some underestimation, and far less structure than the true ffCO$_2$ emissions. This is due to the presence of just 3 towers, which are also clustered in a relatively small region. The region around the Gulf of Mexico is not well estimated since there are few towers there.

In Fig. 10 (left) we plot a time-series of errors defined as a percentage of total, country-level Vulcan emissions. Percent errors in reconstructed emissions are calculated using Eq. 16. The "prior" errors are computed as

$$\text{Error}_{pr,k}\ (\%) = \frac{100}{K}\sum_{k=1}^{K}\frac{E_{pr}-E_{V,k}}{E_{V,k}} \quad \text{where} \quad E_{pr} = \int_{\mathcal{R}} \mathbf{f}_{pr}\, dA.$$

We see 25% errors in $\mathbf{f}_{pr}$. This is a consequence not only of the disagreement between EDGAR (in 2005) and Vulcan (in 2002), but also the manner in which they account for emissions. Since we are only interested in obtaining a rough guess of US emissions with $\mathbf{f}_{pr}$, we did not perform a careful analysis and comparison. As

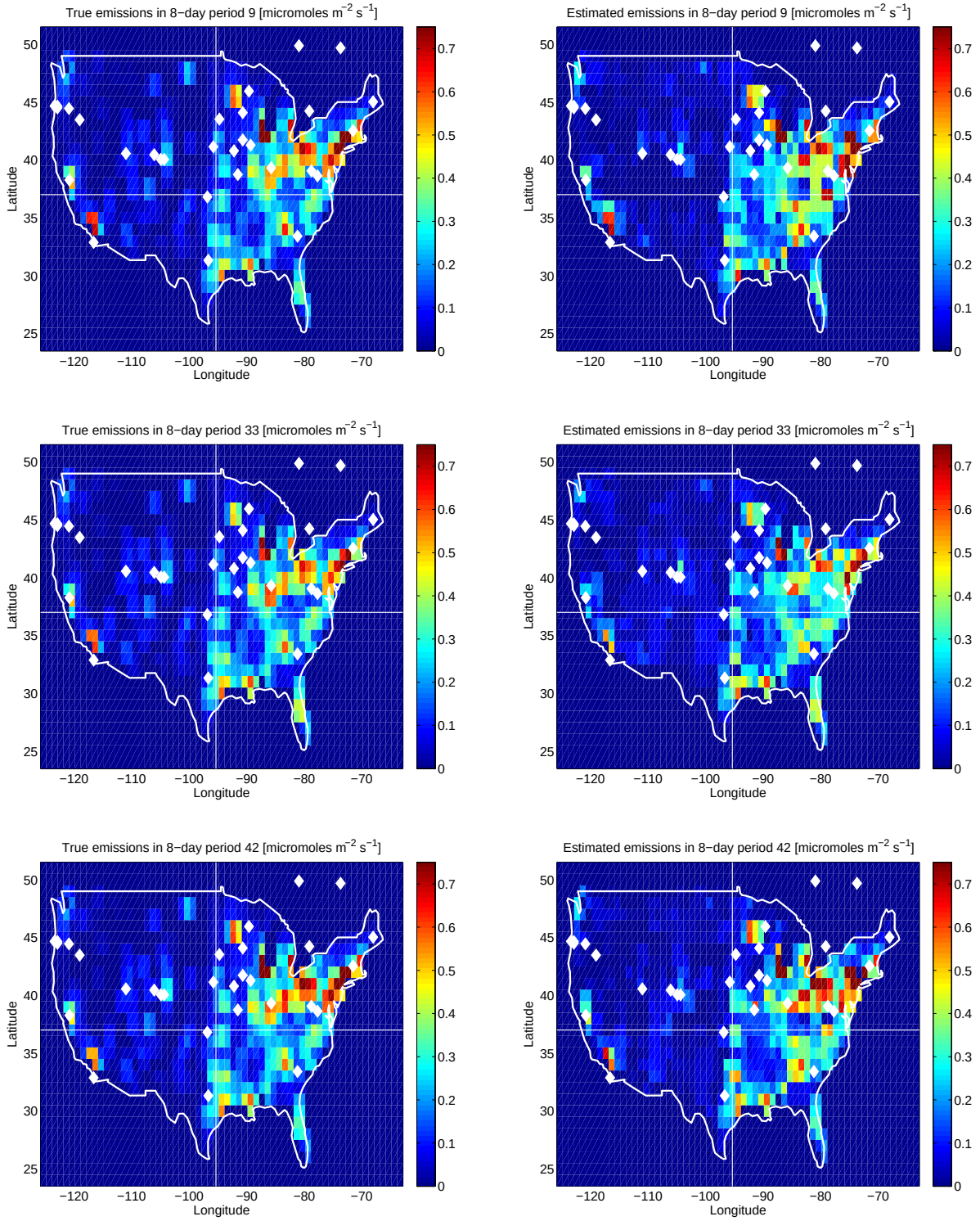


Figure 9: Reconstruction of the ffCO$_2$ emissions from the 35 towers (plotted as diamonds) over one year. The true emissions are on the left and the reconstructions on the right. We see that the large scale structure of the emissions have been captured, as seen by the comparisons for 8-day periods number 9 (mid-March), 33 (end of August) and 42 (early November). The west coast of the US has few towers near heavily populated regions and thus is not very well estimated. In the middle and bottom rows, we see that the emissions in the Los Angeles-San Diego region (32.87N, -117.26W) is underestimated. On the other hand, due to the higher density of towers in the Northeast, the true and estimated emissions are qualitatively similar. Emissions have units of μmol m^{-2} s^{-1} of C (not CO$_2$).

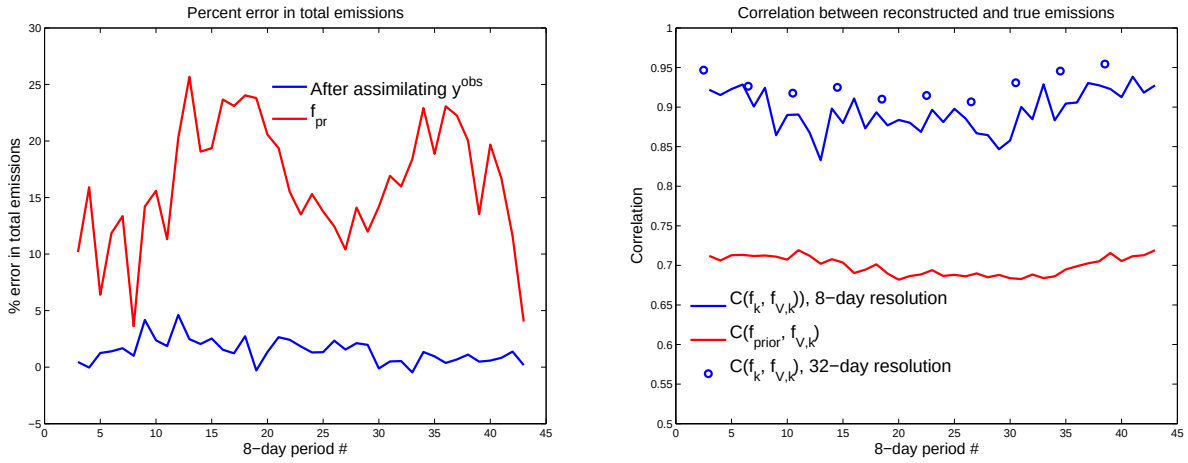


Figure 10: Comparison of reconstruction error and correlations. Left: We plot the error between the reconstructed and true (Vulcan) emissions in blue. In red, we plot the error between $\mathbf{f}_{pr}$ using nightlights and Vulcan emissions. We see that assimilation of $\mathbf{y}^{obs}$ leads to significantly improved accuracy. Right: We plot the accuracy of the spatial distribution of the reconstructed emissions. The Pearson correlations $C(\mathbf{E}_k, \mathbf{f}_{V,k})$ and $C(\mathbf{f}_{pr}, \mathbf{f}_{V,k})$ show that incorporating $\mathbf{y}^{obs}$ improves the spatial agreement of estimated emissions versus the true one from 0.7 to around 0.9. If the emissions are averaged over 32-day periods, rather than 8-day periods, the correlation with true (Vulcan) emissions rises even higher.

can be seen, assimilation of $\mathbf{y}^{obs}$ reduces the error to a maximum of 4%. The least accurate reconstructions are during spring. In order to check the accuracy of the spatial distribution of $\mathbf{E}_k$, we calculate the Pearson correlations $C(\mathbf{E}_k, \mathbf{f}_{V,k})$ and $C(\mathbf{f}_{pr}, \mathbf{f}_{V,k})$. We see that data assimilation increases the correlation from around 0.7 to 0.85-0.95, with a mean around 0.9. When the emissions are aggregated/averaged over 32-day periods, the correlation increases to 0.9-0.95, with a mean around 0.93. Thus, the ffCO$_2$ emissions obtained using a nightlight proxy are substantially improved by the incorporation of $\mathbf{y}^{obs}$.

Next, we address the impact of tower density. As was clear in Fig. 9, the reconstruction in the Northeast (NE) quadrant is more accurate compared to the reconstruction elsewhere. We compute the percent error between reconstructed and true (Vulcan) emissions on a quadrant basis, and plot them in Fig. 11. The correlation between the reconstructed and Vulcan emissions are plotted too. In Fig. 11 (left), we see that the reconstruction error in the NE quadrant is far smaller than elsewhere. Further, the reconstruction error in individual quadrants is far higher than in the country as a whole (Fig. 10), where the total reconstruction error never exceeded 4%. Thus, while large scale structures (e.g., at the scale of the US) are being estimated quite accurately, significant errors are seen when we address regions the size of the individual quadrants. The consequences of not being able to constrain quadrant-scale structures is seen in Fig. 11, where the NE quadrant displays better correlation with Vulcan than the rest. The most inaccurate spatial patterns are seen in the Northwest (NW) quadrant that includes the Great Plains and the Continental Divide. Here, the lack of strong emissions, nightlights and towers make the reconstruction particularly poor.

We now address some of the numerical aspects of the solution. In Fig. 12 (left) we plot $\mathbf{y}$ predicted by the reconstructed emissions at 3 towers. We see that the ffCO$_2$ concentrations are well reproduced by the estimated emissions. Thus, the lack of fidelity at the smaller scales (seen in Fig. 11) do not substantially impact the measurements. In Fig. 12 (right) we plot the wavelet coefficients obtained by projecting the emissions (both the true and reconstructed) on the wavelet bases. The wavelet coefficient values have been

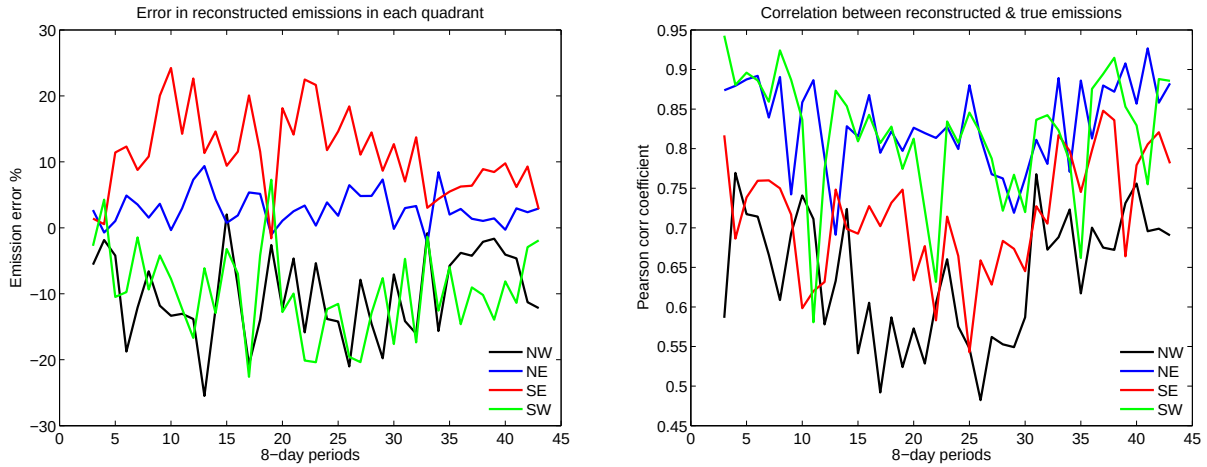


Figure 11: Plot of the error in, and the correlation between reconstructed emissions and true (Vulcan) ones. Left: We see that the reconstruction error in the NE quadrant is small compared to the others. However, the error in each of the quadrants is much larger than the error at the country level (see Fig. 10). Right: Correlation in each of the quadrants. The NE quadrant is substantially better than the others due to the higher density of towers.

subjected to a hyperbolic tangent transformation for ease of plotting. The true wavelet coefficients with a magnitude above 0.01 are plotted with red symbols. The true (Vulcan) emissions have a large number of coefficients with small magnitude; these are usually for small-scale features i.e., have coefficient indices in the right half of the range (Fig. 12, right; red symbols). During the sparse reconstruction from sparse data, these coefficients are set to zero (blue line in Fig. 12, right). The coefficients corresponding to the low-index coefficients, which represent large structures are estimated accurately. This explains the good correlation between true and estimated emissions at the country scale and a far poorer one at the quadrant scale (Fig. 11). The explicit separation of scales is thus leveraged into ignoring unimportant, fine-scale details (which are difficult to constrain with data) and focusing model-fitting effort on the scales that can be resolved by the data. Sparse reconstruction achieves this in an automatic, purely data-driven manner, rather than via a pre-processing, scale-selection step.

Finally we see the impact of the enforcement of non-negativity. In Fig. 13 we plot the cumulative distribution function (CDF) of the ffCO$_2$ emissions in $\mathcal{R}$ before and after the imposition of non-negativity, as described in Sec. 4.4. The emissions are from the 31$^{\text{st}}$ 8-day period (depicted in Fig. 12). We see from the CDF that before the imposition of non-negativity, the number of grid-cells with negative emissions is small; further, the negative emissions are small in magnitude. Thus the sparse reconstruction (Step I in Sec. 4.4) provides a very good approximation to the final estimated fluxes, by detecting the spatial patterns/wavelets that could be constrained by observations. Given a good approximation, the non-negativity enforcement converges quickly.

5.4 Inversions with built-up area maps as proxy

We investigate the effect of built-up area (BUA) maps, instead of images of nightlights, as the proxy. Changing the proxy results in a different set of wavelets being chosen (nightlights resulted in a $W^{(s)}$ of 1031

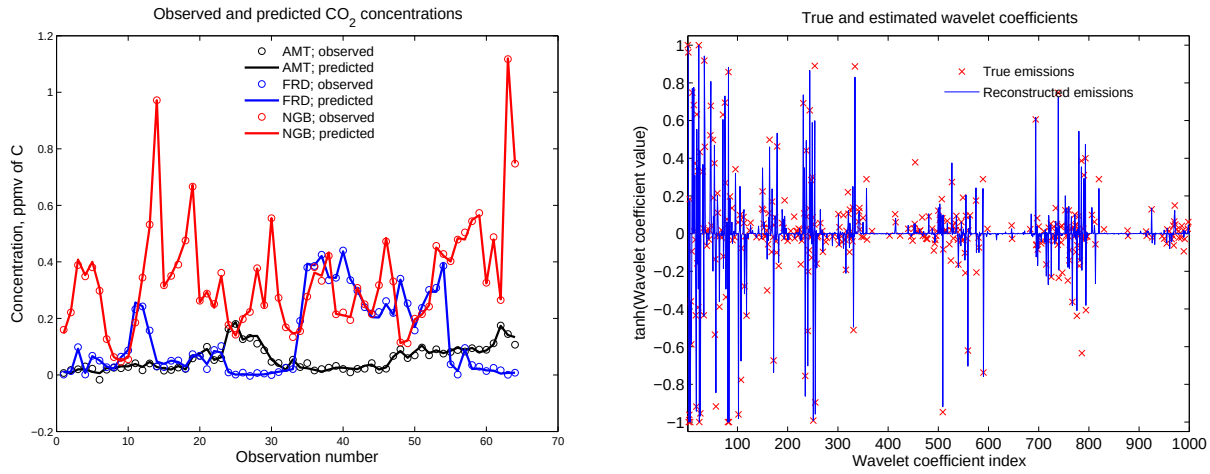


Figure 12: Left: Prediction of ffCO$_2$ concentrations at 3 measurement locations, using the true (Vulcan) and reconstructed emissions (blue lines) over an 8-day period (Period no. 31). Observations occur every 3 hours. We see that the concentrations are accurately reproduced by the estimated emissions. Right: Projection of the true and estimated emissions on the wavelet bases for the same period. Coarse wavelets have lower indices, and they progressively get finer with the index number. We see that the true emissions have a large number of wavelets with small, but not zero, coefficients. In the reconstruction (plotted in blue), a number of wavelet coefficients are set to very small values (almost zero) by the sparse reconstruction. The larger scales are estimated accurately.

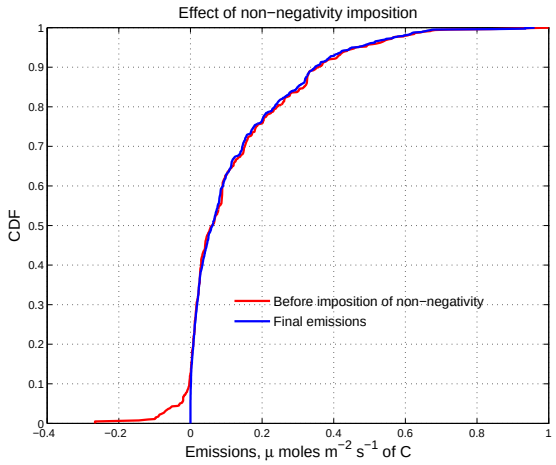


Figure 13: CDF of emissions in $\mathcal{R}$, before and after the imposition of non-negativity, as described in Sec. 4.4. We see that the CDF of the emissions without non-negativity imposed contains a few grid-cells with negative fluxes; further, the magnitude of the negative emissions is small. Thus the spatial parameterization, with sparse reconstruction provides a good approximation of the final, non-negative emissions.

wavelets; the corresponding number for BUA was 1049); further one was *not* a strict subset of the other. It also results in a different normalization in Eq. 10. The inversion was performed in a manner identical to that in Sec. 5.3. In Fig. 14 (left) we see that emissions reconstructed from BUA as the proxy are more variable; however, averaged over the entire 45 8-day periods, the reconstruction errors are not very different. In Fig. 14 (right), we see that the correlation between estimated and true emissions is better in case of nightlight-based reconstruction (when performed at a 8-day temporal resolution). However, when averaged to a 32-day temporal resolution, the difference between the two reconstructions (nightlights- versus BUA-based) is minimal. In Fig. 15 we investigate the differences between the nightlight- and BUA-based reconstructions at the quadrant level. We see in Fig. 15 (left) that the difference between nightlight- and BUA-based reconstruction errors in the NE quadrant are smaller than those for the NW quadrant. Thus, while the prior emissions from nightlights and BUA are quite different (see the last row of Fig. 3), the estimated emissions are well constrained by $\mathbf{y}^{obs}$ in the NE quadrant and the impact of the priors/proxies is small. This is not the case for the NW quadrant, where not only are the errors high, the two $\mathbf{f}_{pr}$ lead to reconstruction errors of different signs. Thus the region is badly constrained, which is not surprising given the paucity of towers (see Fig. 9). However, on the whole, BUA-based reconstruction is a less accurate one. In Fig. 15 (right) we plot the correlation of the reconstructed and true emissions in the NE and NW quadrants. We see that the correlations for the nightlight- and BUA-based estimated emissions are somewhat similar, though BUA-based estimation is more variable, and has lower correlation as a whole. Thus, while Fig 3 (middle row) showed that BUA-had a slightly better correlation with true (Vulcan) emissions, its larger errors, as seen in Fig. 3 (bottom row) lead to a less accurate reconstruction. This result is also a testament to the inadequacy of $\mathbf{y}^{obs}$ over the whole country; had there been sufficient data to constrain $\mathbf{E}$, the impact of $\mathbf{f}_{pr}$ would have been minimal.

In Fig. 16 we compare the estimated emissions developed from the two competing prior models. In the top row we plot the estimated emissions for the 34^{th} 8-day period using the nightlight prior (left) and built-up area maps (right). The difference between the two estimates, plotted in the bottom left subfigure, shows differences spread over a large area, though their magnitudes are not very big. Thus the "prior" model has a measurable impact on the spatial distribution of the emissions. Bottom right, we plot the degree to which the measurements update/change the prior model. On the horizontal axis, we plot the emissions predicted by the "prior" model, while the vertical axis represents the estimated emission in the corresponding grid-box. We see some correlation between the "prior" and estimated emission when the emissions are large (more than 0.25 μmol m^{-2} s^{-1} of C).

5.5 Impact of M_{cs}

In Sec. 4.2 we had used M_{cs} random projections of $\mathbf{F}_{\mathcal{R}'}$ to implement a $\mathbf{F}_{\mathcal{R}'} = 0$ constraint efficiently. Since Eq. 11 is solved approximately, and due to the small number of wavelets in $W^{(s)}$ that span $\mathcal{R}'$, the constraint $\mathbf{F}_{\mathcal{R}'} = 0$ is not satisfied exactly and $\mathbf{F}_{\mathcal{R}'}$ is usually small. This error varies with M_{cs}; a larger number of random projections result in a closer realization of the constraint. However, they are never driven to zero, primarily because the wavelets used to model $\mathbf{F}_{\mathcal{R}'}$ and $\mathbf{F}_{\mathcal{R}}$ were chosen using $\mathbf{X}$ (and thus may not form a complete basis set for ffCO$_2$ emissions). Errors in the enforcement of the $\mathbf{F}_{\mathcal{R}'}$ constraint lead to commensurate errors in $\mathbf{F}_{\mathcal{R}}$. In practice, this affects only Step I of the emission estimation procedure, where a sparse approximation of $\mathbf{F}_{\mathcal{R}}$ is calculated; thereafter it is used as a guess in Step II, the enforcement of non-negativity of emissions. However, a good estimate of $\mathbf{F}_{\mathcal{R}}$ accelerates Step II.

In Fig. 17, we plot the impact of M_{cs} on the reconstruction. We perform Step I of the emission estimation

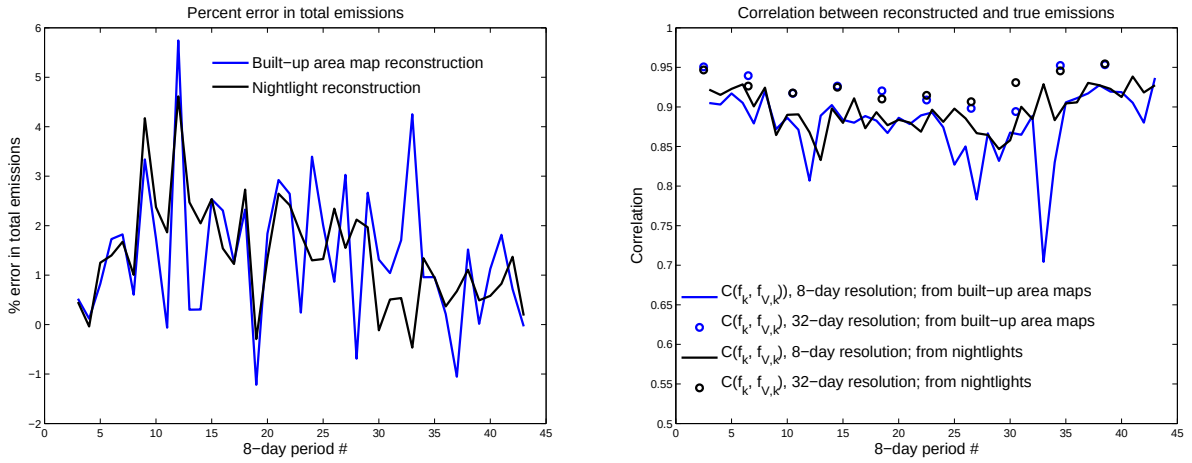


Figure 14: Left: Plot of the error in the reconstruction, as performed with built-up area maps (blue line) and nightlights (black line) as the proxies. Right: Correlation between the true and reconstructed emissions, as performed with built-up area maps and nightlights as the proxies. The nightlights-based reconstruction has slightly less reconstruction error and better correlation with the true emissions, when compared at 8-day temporal resolution. When averaged to a 32-day temporal resolution, emissions obtained via the two methods are very similar.

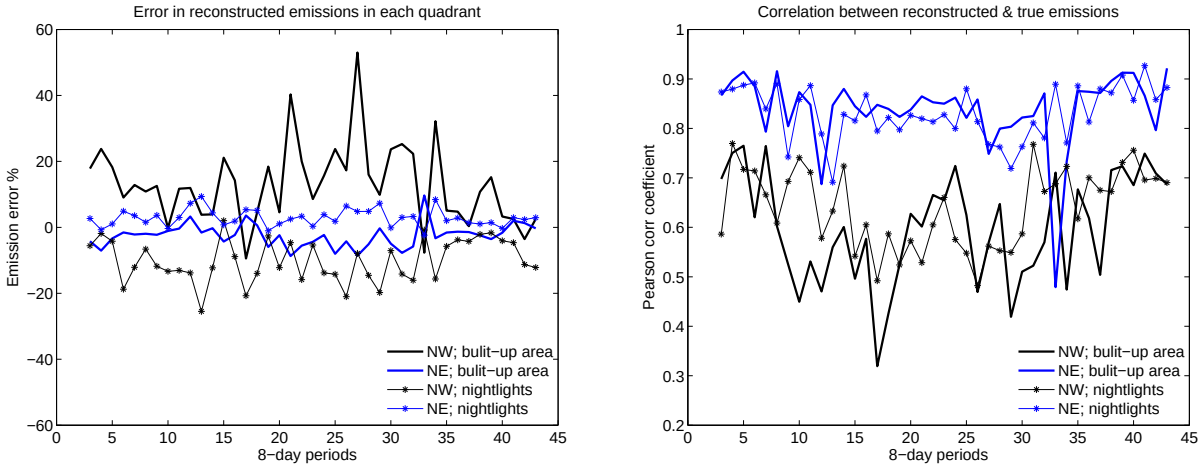


Figure 15: Left: Emission reconstruction error in the NE (blue) and NW (black) quadrants, when performed with BUA (line) and nightlights (symbols) as proxies. We see that the NW quadrant is very badly constrained and a change in proxies changes the sign of the error. This is also seen in the NE quadrant; however, the error magnitudes are far smaller. Right: The comparison of correlations between true and reconstructed emissions shows similar trends; nightlights-based estimation produces better reconstructions.

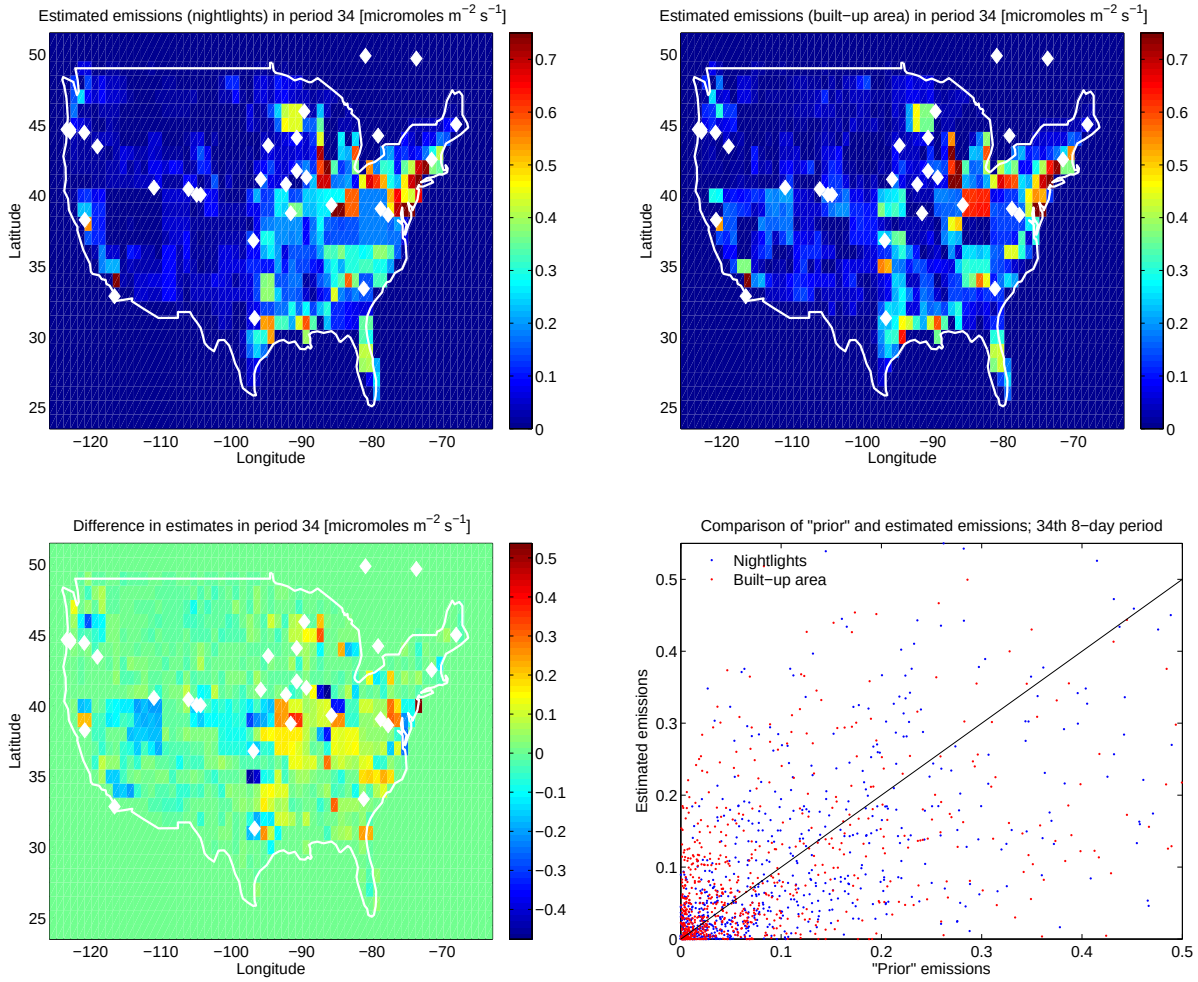


Figure 16: Comparison of emission estimates developed using $\mathbf{f}_{pr}$ constructed from nightlight radiances and built-up area maps. Top: Estimated ffCO$_2$ emissions for the 34$^{\text{th}}$ 8-day period developed using nightlight prior (left) and the prior from built-up areas maps (right). Bottom left: we plot the difference between the two estimates. Bottom right: We plot a scatter plot between the estimated and prior emissions, for the two prior models.

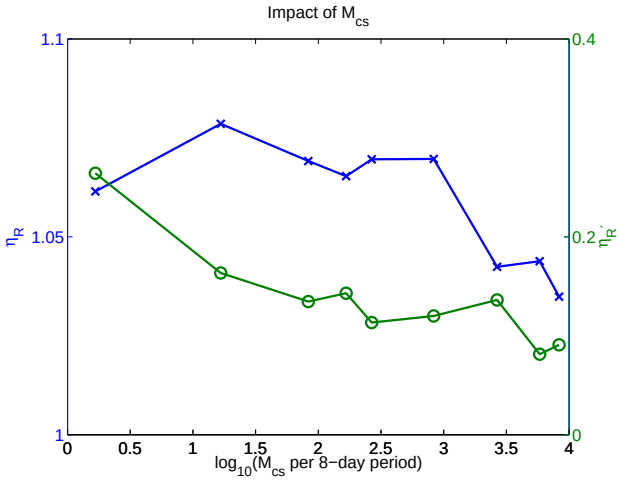


Figure 17: The impact of the number of compressive samples M_{cs} on the reconstruction of $\mathbf{F}_{\mathcal{R}}$ ($\eta_{\mathcal{R}}$) and $\mathbf{F}_{\mathcal{R}'}$ ($\eta_{\mathcal{R}'}$). $\eta_{\mathcal{R}}$ and $\eta_{\mathcal{R}'}$ are plotted on the Y1 and Y2 axes respectively. Results are plotted for the 31^{st} 8-day period. We see that $M_{cs} > 10^3$ does not result in an appreciable increase in reconstruction quality. Also, $M_{cs} < 10^2$ shows a marked degradation in $\eta_{\mathcal{R}'}$.

procedure (see Sec. 4), for the 31^{st} and 32^{nd} 8-day periods and compute the ratios

$$\eta_{\mathcal{R}} = \frac{||\mathbf{f}_{k,\mathcal{R}}||_2}{||\mathbf{f}_{V,k}||_2} \quad \text{and} \quad \eta_{\mathcal{R}'} = \frac{||\mathbf{f}_{k,\mathcal{R}'}||_2}{||\mathbf{f}_{V,k}||_2} \quad \text{for } k = 31.$$

Here $\mathbf{f}_{k,\mathcal{R}}$ and $\mathbf{f}_{k,\mathcal{R}'}$ are the emissions (from Step I) over $\mathcal{R}$ and $\mathcal{R}^{'}$. $\mathbf{f}_{V,k}$ is the true (Vulcan) emission field during the same period. These ratios are plotted as a function of the log-transformed number of compressive samples M_{cs} per 8-day period. We see that 10 projections per 8-day period is too few, leading to around 20% errors in $\mathbf{f}_{k,\mathcal{R}'}$ ($\eta_{\mathcal{R}'} \approx 0.2$). Beyond about 100 projections per 8-day period, $\eta_{\mathcal{R}'}$ oscillates around 0.1. The corresponding errors in $\mathbf{f}_{k,\mathcal{R}}$ are about 5% ($\eta_{\mathcal{R}} \approx 1.05$). In our study we used 300 random projections for each 8-day period. In contrast, had we imposed $\mathbf{F}_{\mathcal{R}'} = 0$ in all grid cells in $\mathcal{R}^{'}$, we would have generated over 3000 constraints per 8-day period (our 64×64 mesh has 4096 grid cells, of which 816 are in $\mathcal{R}$ and the rest, 3184 are in $\mathcal{R}^{'}$). This economy of computational effort in the imposition of the constraint is only partially due to the efficiency of random projections; a major simplification is achieved by the lower-dimensional model of $\mathbf{F}_{\mathcal{R}}$ using the wavelets identified by $\mathbf{X}$.

5.6 Impact of ε

In this section we study the impact of the measurement error ε. The nominal value used in this study is $\varepsilon = 10^{-2}$; we explore the impact of $\varepsilon = \{2.5, 5.0, 7.5, 10\} \times 10^{-2}$. The results, for the 31^{st} 8-day period are shown in Fig. 18. Top left, we plot the true emissions from the Vulcan inventory for reference. Top right, we plot results using $\varepsilon = 2.5 \times 10^{-2}$. We see enhanced "blockiness" as some of the finer wavelets cannot be estimated; whereas the true emissions are represented using 363 wavelets coefficients with a magnitude above 0.01, the reconstruction with $\varepsilon = 2.5 \times 10^{-2}$ recovers only 325 such coefficients. Bottom left, at a higher level of measurement errors ($\varepsilon = 10^{-1}$), the solution seems qualitatively different; low emission regions in the Western quadrants (Continental Divide and the deserts of the Southwest) show erroneous (and

higher) levels of emissions. This leads to the recovery for 377 wavelet coefficients with a magnitude above 0.01. Bottom right: we summarize the quality of the reconstruction via the reconstruction error (%) and the correlation between the true and estimated emissions. We see that the degradation of reconstruction quality is roughly linear in the measurement error ε.

5.7 Impact of the number of observation towers

In this section, we investigate the impact of reducing the number of observation towers. In Fig. 19 we plot the results from inversions performed with 35-, 25- and 15-tower configurations. In the left column, we plot reconstructed emissions during the 33^{rd} 8-day period, with the different tower configurations. The resolution at which the emissions can be estimated decreases with the number of towers; the differences between the 35- and 15-tower reconstructions are easy to detect. In the right column, we quantify the differences. In Fig. 19, top right, we plot the reconstruction error over 45 8-day periods. We see that while the difference in the reconstruction error between 35- and 25-tower inversions is not much, the 15-tower inversion is clearly inferior. This is also borne out in the plot of the correlation of the true and estimated emissions (Fig. 19, right middle), where the estimation performed with 15 towers provides the lower bound. Finally, in Fig. 19, bottom right, we plot the wavelet coefficients. The coefficients have been subjected to a hyperbolic tangent transformation for plotting clarity. The true wavelet coefficients with a magnitude above 0.01 are plotted with red crosses. The 35-tower reconstruction comes close to estimating the true coefficients; this is especially true for the wavelets with larger spatial support (i.e., low index). As the number of towers drop, the errors in the estimated wavelet values grow. Further, the number of wavelet coefficients with large magnitudes (above 0.01) steadily decreases. Thus while the true emissions are represented by 363 wavelet coefficients with magnitudes above 0.01, the estimated emissions, using 35, 25 and 15 towers have, respectively, 315, 292 and 287 coefficients. Therefore, in the absence of observations, the solution to Eq. 11 sets the wavelet coefficients to values near zero.

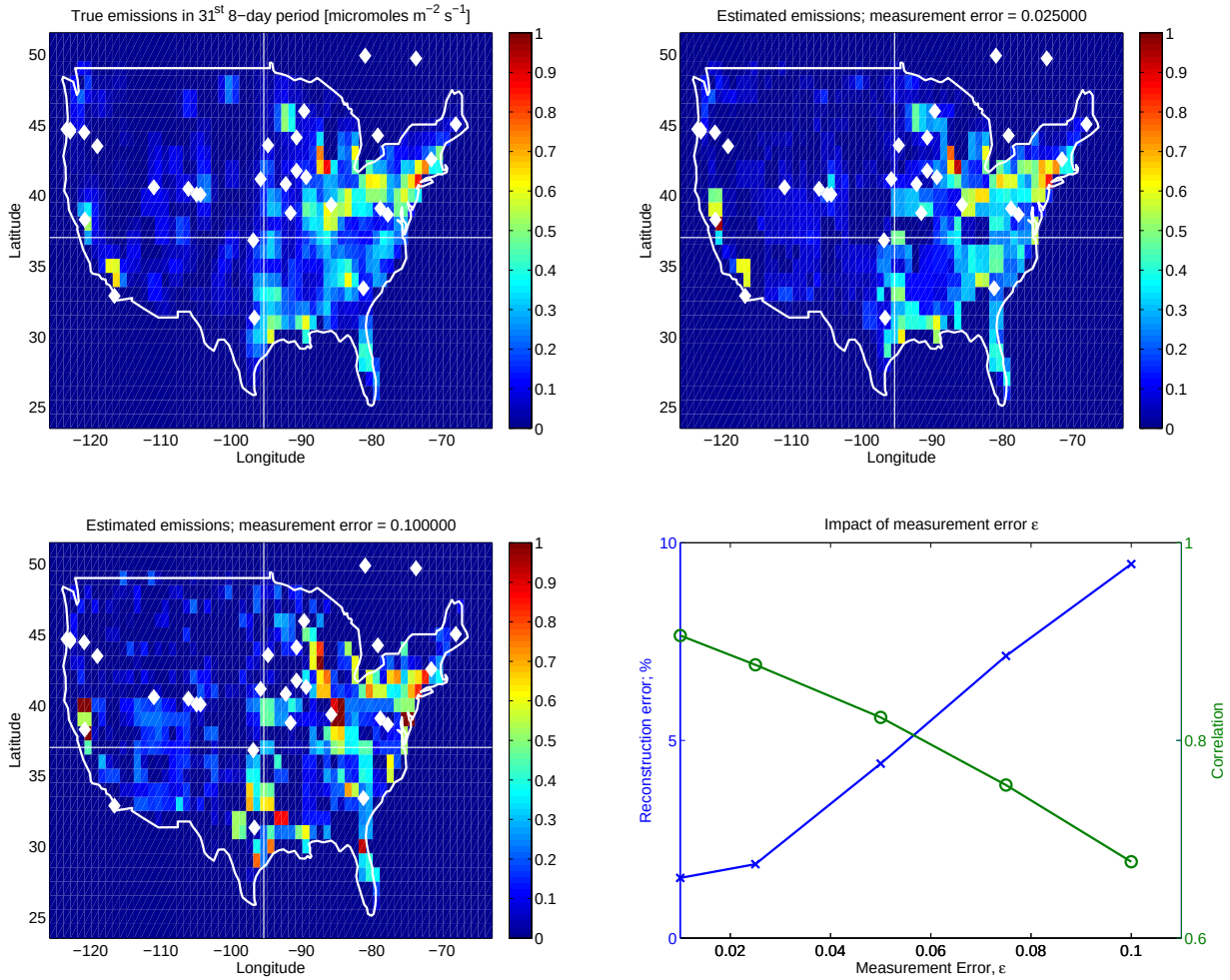


Figure 18: Impact of measurement error ε. Top left: We plot the true ffCO$_2$ emissions from the Vulcan inventory. Top right: We plot the estimates calculated using $\varepsilon = 2.5 \times 10^{-2}$. Bottom left, we plot the reconstructed emissions using $\varepsilon = 10^{-1}$; we see a clear degradation of the reconstruction. Bottom right: We plot the reconstruction error (%) and the correlation between the reconstructed and true emissions for various values of ε; a clear degradation is seen. Reconstruction errors and correlations are plotted on opposing Y-axes. All results are for the 31st 8-day period.

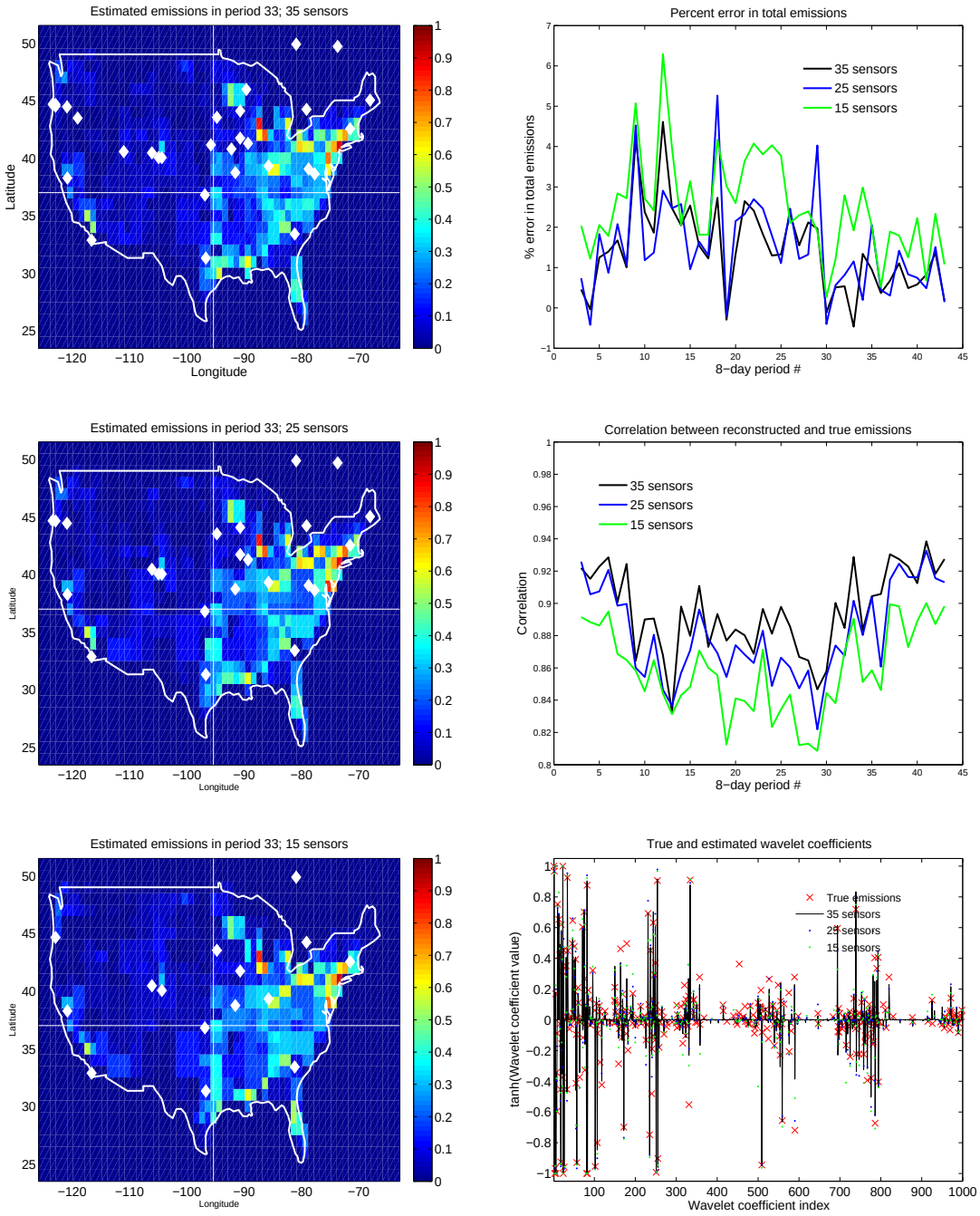


Figure 19: Left column: Estimated ffCO$_2$ emissions for the 33^{rd} 8-day period. Inversions were done with 35 (top), 25 (middle) and 15 (bottom) towers. As the number of towers decreases, we see that the resolution of estimated emissions decreases. Right column: At the top, we plot the error in the reconstructed emissions for the three tower sets. There is not much difference between reconstructions with 25 and 35 towers, but the 15-tower reconstruction is poor. Right (middle): The correlation between reconstructed and true emissions shows much the same trend as the reconstruction, with the 15-tower reconstruction having a far lower correlation over the entire 360-day duration. Right column, bottom: We plot the sparsity of the estimated wavelet coefficients. We see that as the number of towers decrease, the wavelet coefficient estimates deviate further from the true values (red crosses).

6 Conclusions

We have devised a multiresolution parametrization (also known as a multiscale random field or MsRF model) for modeling $ffCO_2$ emissions at 1° resolution. It is based on Haar wavelets, and is designed for use in atmospheric inversions. It uses easily observable proxies of human activity e.g., images of lights at night and maps of built-up areas to reduce its dimensionality as well as to build "prior" models of $ffCO_2$ emissions.

The MsRF model was tested in a set of synthetic-data inversions. Time-dependent $ffCO_2$ emission fields were estimated over the lower 48 states of the US, conditioned on 360 days of observations of time-varying $ffCO_2$ concentrations at a set of observation towers. In conjunction with an atmospheric transport model, a sparsity-enforcing optimization method, Stagewise Orthogonal Matching Pursuit (StOMP), was used to fit MsRF models and reconstruct a time-series of emission fields. It was found that less than half the parameters of the MsRF model could be estimated from the sparse data; the rest were identified by StOMP and set to zero. We also identified an efficient way of incorporating a prior model of emissions into the inversion.

The MsRF, being based on wavelets, models emissions on a dyadic square grid, whereas $ffCO_2$ emissions were restricted to an irregular region $\mathcal{R}$ (the lower 48 states of the US). Instead of specifying zero emissions in the grid-cells outside $\mathcal{R}$, we used concepts from compressive sensing (random projections) to achieve the same effect, but at a tenth of the computational effort and memory requirements (Sec. 5.5). The emission fields reconstructed using the MsRF were found to contain some regions (grid-cells) with negative emissions. Using these as a starting guess, we devised a simple method to iteratively impose non-negativity on the $ffCO_2$ emissions.

Primarily, our work demonstrated that observations of $ffCO_2$ concentrations could be used to update models of $ffCO_2$ emissions and improve their accuracy (see Fig. 10). Further, we assumed that these measurements could be obtained from existing towers, which were sited with a view of reconstructing biospheric, rather than $ffCO_2$, fluxes. This accounts for some of our inability to estimate fine-scale spatial structure in the emissions. Note that currently the network has expanded beyond the 35 towers included in this work, and could potentially furnish better estimates of $ffCO_2$ than reported here. Future work will examine how the inversion method described here could be used to design a monitoring network for $ffCO_2$, rather than biospheric CO_2 fluxes.

The MsRF models constructed using images of nightlights and built-up area maps provided reconstructions that differed in their fine-scale details (at the quadrant scale and below). This is a reflection of the paucity of observations and their inability to constrain the fine-scale features of the emission fields. The formulation also correctly showed that as the number of observation towers were decreased, larger numbers of MsRF parameters were set to zero and the reconstructed emission fields progressively lost their spatial fidelity, starting from the fine-scale details.

Our inversion formulation suffers from two drawbacks. It requires measurements of $ffCO_2$ concentrations at the measurement towers. While these concentrations can be back-calculated using $\Delta^{14}CO_2$ or CO measurements, they are less common and more expensive than CO_2 concentration measurements. The second drawback is the deterministic nature of the reconstruction - we do not provide error bounds on the estimates of the MsRF parameters (the wavelet coefficients). This can be rectified by adopting a Bayesian approach e.g., Kalman filters, but it is unclear how one would preserve the non-negative property of $ffCO_2$ emissions. This investigation is currently underway.

This page intentionally left blank

References

[1] Global Fire Emissions Database. http://www.globalfiredata.org/index.html.

[2] S. M. Gourdji, K. L. Mueller, V. Yadav, D. N. Huntzinger, A. E. Andrews, M. Trudeau, G. Petron, T. Nehrkorn, J. Eluszkiewicz, J. Henderson, D. Wen, J. Lin, M. Fischer, C. Sweeney, and A.M. Michalak. North American CO2 Exchange: Inter-comparison of modeled estimates with results from a fine-scale atmospheric inversion. *Biogeosciences Discussions*, 9:457–475, 2012. doi:10.5194/bg-9-457-2012.

[3] Project Vulcan webpage. http://www.purdue.edu/eas/carbon/vulcan/index.php.

[4] K. R. Gurney, D. L. Mendoza, Y. Zhou, M. L. Fischer, C. C. Miller, S. Geethakumar, and S. de la Rue de Can. High resolution fossil fuel combustion CO_2 emission fluxes in the United States. *Environmental Science and Technology*, 43:5535–5541, 2009.

[5] The United Nations Collaborative Program on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries. http://www.un-redd.org/.

[6] United Nations Framework Convention on Climate Change. http://unfccc.int/.

[7] 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.

[8] T. Oda and S. Maksyutov. A very high-resolution (1km $\times$ 1km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*, 11:543–556, 2011.

[9] P. J. Rayner, M. R. Raupach, M. Paget, P. Peylin, and E. Koffi. A new global gridded data set of CO_2 emissions from fossil fuel combustion: Methodology and evaluation. *Journal of Geophysical Research*, 115, 2010. D19306.

[10] J. G. J. Olivier, J. A. V. Aardenne, F. J. Dentener, V. Pagliari, L. N. Ganzeveld, and J. A. H. W. Peters. Recent trends in global greenhouse gas emissions: Regional trends 1970–2000 and spatial distribution of key sources in 2000. *Journal of Integrative Environmental Science*, 2:81–99, 2005.

[11] G. Marland. Uncertainties in accounting for CO_2 from fossil fuels. *Journal of Industrial Ecology*, 12(2):136–139, 2008.

[12] G. Marland, K. Hamal, and M. Jonas. How uncertain are estimates of CO_2 emissions? *Journal of Industrial Ecology*, 13(1):4–7, 2009.

[13] J. E. Sinton. Accuracy and reliability of China's energy statistics. *China Economic Review*, 12:373–383, 2001.

[14] Weather research and forecasting model. http://www.wrf-model.org/index.php.

[15] S. Houwelling, F.-M. Breon, I. Aben, C. Rodenbeck, M. Gloor, M. Heimann, and P. Ciais. Inverse modeling of CO_2 sources and sinks using satellite data: A synthetic inter-comparison of measurement techniques and their performance as a function of space and time. *Atmospheric Chemistry and Physics Discussions*, 3:5237–5274, 2003.

[16] K. L. Mueller, S. M. Gourdji, and A. M. Michalak. Global monthly averaged CO_2 fluxes recovered using a geostatistical inverse modeling approach: 1. Results using atmospheric measurements. *Journal of Geophysical Research*, 2008. D21114.

[17] L. Feng, P. I. Palmer, H. Bosch, and S. Dance. Estimating surface CO_2 fluxes from space-borne CO_2 dry air mole fraction observations using ensemble Kalman Filter. *Atmospheric Chemistry and Physics*, pages 2619–2633, 2009.

[18] F. Chevallier, R. J. Engelen, C. Carouge, T. J. Conway, P. Peylin, C. Pickett-Heaps, M. Ramonet, P. J. Rayner, and I. Xueref-Remy. AIRS-based versus flask-based estimation of carbon surface fluxes. *Journal of Geophysical Research*, 114, 2009. D20303.

[19] R. Nassar, D. B. A. Jones, S. S. Kulawik, J. R. Worden, K. W. Bowman, R. J. Andres, P. Suntharalingam, J. M. Chen, C. A. M. Brenninkmeijer, T. J. Schuck, T. J. Conway, and D. E. Worthy. Inverse modeling of CO_2 sources and sinks using satellite observations of CO2 from TES and surface flask measurements. *Atmospheric Chemistry and Physics*, 11:6029–6047, 2011.

[20] P. Ciais, P. Rayner, F. Chevallier, P. Bousquet, M. Logan, P. Peylin, and M. Ramonet. Atmospheric inversions for estimating CO_2 fluxes: methods and perspectives. *Climate Change*, 103:69–92, 2010.

[21] A. Chatterjee, A. M. Michalak, K. L. Mueller, J. L. Anderson, and V. Yadav. Towards reliable ensemble Kalman filter estimates. *Journal of Geophysical Research - Atmospheres*, 117, 2012. D22306.

[22] A. M. Michalak, L. Bruhwiler, and P. Tans. A geostatistical approach to surface flux estimation of atmospheric trace gases. *Journal of Geophysical Research*, 2004. D14109.

[23] G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley, MA, 1997.

[24] Tony F. Chan and Jackie (Jianhong) Shen. *Image Processing and Analysis – Variational, PDE, Wavelet, and Stochastic Methods*. Society of Applied Mathematics, 2005.

[25] J. C. Turnbull, A. Karion, M. L. Fisher, I. Faloona, T. Guilderson, S. J. Lehman, B. R. Miller, J. B. Miller, S. Montzka, T. Sherwood, S. Saripalli, C. Sweeny, and P. P. Tans. Assessment of fossil fuel carbon dioxide and other anthropogenic trace gas emissions from airborne measurements over Sacramento, California in spring 2009. *Atmospheric Chemistry and Physics*, 11:705–721, 2011.

[26] K. McKain, S. Wofsy, T. Nehrkorn, J. Eluszkiewicz, J. R. Ehrlinger, and B. B. Stephens. Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region. *Proceedings of the National Academy of Science*, 109(22):8423–8428, 2012.

[27] Energy Statistics Database. http://unstats.un.org/unsd/energy/edbase.htm.

[28] Carbon Monitoring and Action website. http://carma.org.

[29] T. D. Kelly and G. R. Matos. Historical statistics for mineral and material commodities in the United States. *U.S. Geological Survey Data*, 140, 2007.

[30] C. D. Elvidge, K. E. Baugh, E. A. Kihn, H. W. Kroehl, and E. R. Davis. Mapping city lights with nighttime data from DMSP operational linescan system. *Photogrammetric Engineering and Remote Sensing*, 63:727–734, 1997.

[31] C. N. H. Doll, J-P. Muller, and C. D. Elvidge. Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *Ambio*, 29:157–162, 2000.

[32] European Commission, Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL). Emission Database for Global Atmospheric Research (EDGAR), release version 4.1. http://edgar.jrc.ec.europa.eu/index.php. EDGAR/EUROPA Web-site.

[33] Carbon dioxide information analysis center webpage. http://cdiac.ornl.gov.

[34] R. J. Andres, J. S. Gregg, L. Losey, G. Marland, and T. A. Boden. Monthly, global emissions of carbon dioxide from fossil fuel consumption. *Tellus B*, 63(3):309–327, 2011.

[35] K. R. Gurney, I. Razlivanov, Y. Song, Y. Zhou, B. Benes, and M. Abdul-Massih. Quantification of fossil fuel CO_2 emissions on the building/street scale for a large city. *Environmental Science and Technology*, 46:12194–12202, 2012.

[36] Hestia webpage. http://www.purdue.edu/climate/hestia.

[37] J. S. Gregg, R. J. Andres, and G. Marland. China: Emission pattern of the world leader in CO_2 emissions from fossil fuel consumption and cement production. *Geophysical Research Letters*, 35, 2008.

[38] C. S. Potter, J. T. Randerson, C. B. Field, P. A. Matson, P. M. Virousek, H. A. Mooney, and S. A. Klooster. Terrestrial ecosystem production: A process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7(4):811–841, 1993.

[39] C. Rodenbeck, S. Houwelling, M. Gloor, and M. Helmann. CO_2 flux history 1982–2001 inferred from atmospheric transport data using a global inversion of atmospheric transport. *Atmospheric Chemistry and Physics*, 3:1919–1964, 2003.

[40] F. Chevalier, M. Fisher, P. Peylin, S. Serrar, P. Bousquet, F.-M. Breon, A. Chédin, and P. Ciais. Inferring CO_2 sources and sinks from satellite observations: Method and application to TOVS data. *Journal of Geophysical Research*, 110, 2005. D24309.

[41] W. Peters, J. B. Miller, J. Whitaker, A. S. Demming, A. Hirsch, M. C. Krol, D. Zupanski, L. Bruhwiler, and P. P. Tans. An ensemble data assimilation system to estimate CO_2 surface fluxes from atmospheric gas observations. *Journal of Geophysical Research*, 110, 2005. D24304.

[42] S. M. Gourdji, K. L. Mueller, K. Schaefer, and A. M. Michalak. Global monthly averaged CO_2 fluxes recovered using a geostatistical inverse modeling approach: 2. Results including auxiliary environmental data. *Journal of Geophysical Research*, 113, 2008. D21115.

[43] S. M. Gourdji, A. I. Hirsch, K. L. Mueller, A. E. Andrews, and A. M. Michalak. Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study. *Atmospheric Chemisty and Physics Discussions*, 9:22407–22458, 2009.

[44] S. Mallat. *A wavelet tour of signal processing*. Academic Press, San Diego, CA, USA, 1999.

[45] J. R. Williams and K. Amaratunga. Introduction to wavelet engineering. *International Journal for Numerical Methods in Engineering*, 37:2365–2388, 1994.

[46] J. Romberg. Imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):14–20, 2008. http://www-dsp.rice.edu/cs.

[47] Emmanuel Candes and Michael Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[48] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions Information Theory*, 52(12):5406–5425, 2006.

[49] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[50] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[51] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Applied and Computational Harmonic Analysis*, 10(1), 2001.

[52] Y. Tsaig and D. Donoho. Extensions of compressed sensing. *Signal Processing*, 86(3):533–548, 2006.

[53] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[54] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[55] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[56] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[57] J. Tropp and A. C. Gilbert. Signal recovery from partial information via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[58] D. L. Donoho, Y. Tsaig, I. Drori, and J-L Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. March 2006.

[59] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6), 2008.

[60] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Signal Processing*, 19(1), 2010.

[61] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk. Fast reconstruction of piecewise-smooth signals from random projections. In *Proceedings of Signal Processing with Adaptive Sparse Structured Representations*. Rennes, France, 2005.

[62] C. La and M. N. Do. Signal reconstruction using sparse tree representation. In *Proc. Wavelets XI at SPIE Optics and Photonics*. San Diego, CA, USA, 2005.

[63] Richard Baraniuk, Volkan Cevher, Marco Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982 – 2001, 2010.

[64] L. He and L. Carin. Exploiting structure in wavelet-based bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57(9):3488–3497, 2009.

[65] Global radiance calibrated nighttime lights. http://www.ngdc.noaa.gov/dmsp/download_radcal.html.

[66] P. Cinzano, F. Falchi, C. D. Elvidge, and K. E. Baugh. The artificial night sky brightness mapped from DMSP satellite Operational Linescan System measurements. *Monthly Notices of the Royal Astronomical Society*, 318:641–657, 2000.

[67] C. D. Elvidge, K. E. Baugh, J. B. Dietz, T. Bland, P. C. Sutton, and H. W. Kroehl. Radiance calibration of DMSP-OLS low-light imaging data of human settlements - a new device for portraying the Earth's surface entire. *Remote Sensing of Environment*, 68:77–88, 1999.

[68] M. R. Raupach, P. J. Rayner, and M. Page. Regional variations in spatial structure of nightlights, population density and fossil-fuel CO_2 emissions. *Energy Policy*, 38:4756–4764, 2009.

[69] Atlas of the biosphere. http://www.sage.wisc.edu/atlas/maps.php?datasetid=18&includerelatedlinks=1&dataset=18.

[70] Billie Miteva. billie_m@rocketmail.com. Personal communication.

[71] International Geosphere-Biosphere Programme. http://www.igbp.net/.

[72] J. C. Lin, C. Gerbig, S. C. Wofsy, A. E. Andrews, B. C. Daube, K. J. Davis, and C. A. Grainger. A near-field tool for simulating the upstream influence of atmospheric observations: the Stochastic Time-Inverted Lagrangian Transport (STILT) model. *Journal of Geophysical Research*, 108(D16), 2003. 4493, doi:10.1029/2002JD003161.

[73] S. M. Gourdji, A. I. Hirsch, K. L. Mueller, A. E. Andrews, and A. M. Michalak. Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study. *Atmospheric Chemisty and Physics*, 10:6156–6167, 2010. doi:10.5194/acp-10-6151-2010.

[74] W. C. Skamarock and J. B. Klemp. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational. Physics*, 227:3465–3485, 2008.

[75] T. Nehrkorn, J. Eluszkiewicz, S. C. Wofsy, J. C. Lin, C. Gerbig, M. Longo, and S. Freitas. Coupled Weather Research and Forecasting/Stochastic Time-Inverted Lagrangian Transport (WRF-STILT) model. *Meteorology and Atmospheric Physics*, 107:51–64, 2010. doi:10.1007/s00703-010-0068-x.

[76] Kimberly L. Mueller. *A data-driven multi-scale statistical investigation of regional sources and sinks to improve knowledge of terrestrial carbon cycling*. PhD thesis, University of Michigan, Ann Arbor, MI, 2011. Environmental Engineering.

[77] Emmanuel Cands and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969, 2007.

[78] L. Gan, T.T. Do, and T.D. Tran. Fast compressive imaging using scrambled block Hadamard ensemble. *16th European Signal Processing Conference*, 2008.

[79] Wotao Yin, Simon Morgan, Junfeng Yang, and Yin Zhang. Practical compressive sensing with Toeplitz and circulant matrices. *Signals*, 1:77440K–77440K–10, 2010.

[80] Tomas Tuma and Paul Hurley. On the incoherence of noiselet and Haar bases. In *International Conference on Sampling Theory and Applications*, 2009.

[81] C. Gerbig, A. J. Dolman, and M. Heimann. On observational and modeling strategies targeted at regional carbon exchange over continents. *Biogeosciences*, 6:1949–1059, 2009.

This page intentionally left blank

A Glossary

Table A.1: Definitions of Greek symbols

ε	Measurement error
ε_2	Convergence tolerance for StOMP
ε_3	Convergence tolerance for non-negativity imposition
ϕ	Basis vector
$\mathbf{\Phi}$	Matrix of basis vectors, $N \times L$
$\widetilde{\mathbf{\Phi}}$	Matrix of basis vectors for K periods, diag($\mathbf{\Phi}$)
$\widetilde{\mathbf{\Phi}}_{\mathcal{R}}$	Basis matrix for modeling $\mathbf{F}_{\mathcal{R}}$, $(N_{\mathcal{R}}K) \times (LK)$
$\widetilde{\mathbf{\Phi}}_{\mathcal{R}'}$	Basis matrix for modeling $\mathbf{F}_{\mathcal{R}'}$, $(N_{\mathcal{R}'}K) \times (LK)$
$\widetilde{\mathbf{\Phi}}'_{\mathcal{R}}$	$\widetilde{\mathbf{\Phi}}_{\mathcal{R}}$ with columns multiplied by weights from $\mathbf{f}_{pr}$
$\widetilde{\mathbf{\Phi}}'_{\mathcal{R}'}$	$\widetilde{\mathbf{\Phi}}_{\mathcal{R}'}$ with columns multiplied by weights from $\mathbf{f}_{pr}$
$\widetilde{\mathbf{\Phi}}'_{\mathcal{R}'}$	Basis matrix for modeling $\mathbf{F}_{\mathcal{R}'}$, $(N_{\mathcal{R}'}K) \times (LK)$
ψ	Random measurement vector for CS
$\mathbf{\Psi}$	Random projection matrix / Measurement matrix for CS
$\mathbf{\Upsilon}$	"Observations"; $\mathbf{Y}$ or $\Delta\mathbf{Y}$
$\mathbf{\Gamma}$	$\mathbf{G}$ or $\mathbf{G}'$

Table A.2: Definitions of symbols

CS	Compressive Sensing
D	Domain size, along one axis
$\mathbf{E}$	Non-negative ffCO$_2$ emissions, over K 8-day periods, in $\mathcal{R}$
$\mathbf{f}$	CO$_2$ emission, over a spatial domain. This is not a time-series.
$\mathbf{f}_k$	CO$_2$ emission averaged over a 8-day period k
$\mathbf{f}_{k,\mathcal{R}}$	ffCO$_2$ emissions averaged over a 8-day period k inside $\mathcal{R}$
$\mathbf{F}_{\mathcal{R}}$	CO$_2$ emissions time-series $\mathbf{F}_{\mathcal{R}} = \{\mathbf{f}_k\}, k = 1 \ldots K$, in the grid-cells covering the Lower 48 states of US, $\mathcal{R}$; $(N_{\mathcal{R}} K) \times 1$
$\mathbf{F}_{\mathcal{R}'}$	CO$_2$ emissions time-series in the grid-cells other than those covering the Lower 48 states of US; the complement of $\mathbf{F}_{\mathcal{R}}$; $(N_{\mathcal{R}'} K) \times 1$
$\mathbf{f}_{pr}$	Prior flux or a guess of what the $\mathbf{f}$ might be
$\overline{\mathbf{f}_V}$	Vulcan emissions at 1 degree resolution, averaged over 2002
$\overline{\mathbf{f}_V}^{(s)}$	Vulcan emissions at 1 degree resolution, averaged over 2002, as represented by a sparse basis set
$\mathbf{f}_{V,k}$	Vulcan emissions at 1 degree resolution, but averaged over the k^{th} 8-day period
$\mathbf{g}$	A 1D signal
$\mathbf{g}'$	A measured 1D signal
$\mathbf{G}$	The gain matrix; $(K_s N_s + M_{cs}) \times (LK)$
$\mathbf{G}'$	The "normalized" gain' matrix; $(K_s N_s + M_{cs}) \times (nzwK)$
$\mathbf{H}$	Transport model
K	Number of 8-day time-periods in a year, $K = 45$
K_s	The number of times a tower sensor senses in a year. Towers measure once every 3 hours
L	Number of non-zero weights in a wavelet decomposition; sparsity
M	The number of wavelet levels the grid can be decomposed to
M_{cs}	The number of random projections used to enforce zero flux outside $\mathcal{R}$
N	Number of grid cells; equal to $2^M \times 2^M$
$N_{\mathcal{R}}$	The number of grid-cells in $\mathcal{R}$
$N_{\mathcal{R}'}$	The number of grid-cells in $\mathcal{R}'$
N_s	Number of sensors
$\mathbf{R}$	Random projection matrix, used to enforce zero flux outside $\mathcal{R}$; $M_{cs} \times (N_{\mathcal{R}'} K)$
$\mathcal{R}$	Region of interest, lower 48 states of US
$\mathcal{R}'$	Region that is in the grid, but outside the ower 48 states of US
StOMP	Stagewise Orthogonal Matching Pursuit, [58]
US	United States
$\mathbf{w}$	Weights of a wavelet basis set
$\mathbf{w}'$	Weights of a wavelet basis set, normalized by weights of a guessed flux from a proxy
$W^{(s)}$	Set of weights corresponding to the sparsified set of wavelet bases
$\mathbf{X}$	A proxy for CO$_2$ emissions, e.g., nightlights
$\mathbf{X}^{(s)}$	Proxy, sparsified
$\mathbf{y}$	Modeled CO$_2$ concentrations
$\mathbf{y}^{obs}$	Measured CO$_2$ concentrations
$\mathbf{Y}$	LHS of optimization problem; $\mathbf{Y} = \{\mathbf{y}^{obs}, 0\}$; $(K_s N_s + M_{cs}) \times 1$

B Measurement locations

This is a list of CO_2 measurement towers that were used in this study. They are targeted at biospheric CO_2 fluxes and their locations are not optimized for $ffCO_2$ emissions.

Table B.3: List of observation towers and their locations

Tower symbol	Full name	Latitude & Longitude	Height above ground (meters)
AMT	Argyle	45.03 N, -68.68 W	107
ARM	Southern Great Plains	36.80 N, -97.50 W	60
BAO	Boulder Atmospheric Observatory	40.05 N, -105.01 W	300
BRW	Barrow	71.32 N, -156.61 W	17
CDL	Candle Lake	53.99 N, -105.12 W	30
CEN	Centerville	40.79 N, -92.88 W	110
CHI	Chibougamau	49.69 N, -74.34 W	30
CVA	Canaan Valley	39.06 N, -79.42 W	7
EGB	Egbert	44.23 N, -79.78 W	3
ETL	East Trout Lake	54.35 N, -104.99 W	105
FIR	Fir	44.65 N, -123.55 W	38
FRD	Fraserdale	49.88 N, -81.57 W	40
GAL	Galesville	44.09 N, -91.34 W	122
HDP	Hidden Peak, Snowbird	40.56 N, -111.65 W	18
HFM	Harvard Forest	42.54 N, -72.17 W	30
KEW	Kewanee	41.28 N, -89.97 W	140
LEF	Park Falls	45.95 N, -90.27 W	396
LJA	La Jolla	32.87 N, -117.26 W	5
LLB	Lac LaBiche	54.95 N, -112.45 W	10
MAP	Mary's Peak	44.50 N, -123.55 W	8
MEA	Mead	41.14 N, -96.46 W	122
MET	Metolius	44.45 N, -121.56 W	34
MOM	Morgan Monroe	39.32 N, -86.41 W	48
NGB	NGBER	43.47 N, -119.69 W	7
NWR	Niwot Ridge	40.05 N, -105.58 W	5
OZA	Ozark	38.74 N, -92.20 W	30
ROL	Round Lake	43.53 N, -95.41 W	110
SBL	Sable Island	43.93 N, -60.02 W	25
SCT	South Carolina Tower	33.41 N, -81.83 W	305
SNP	Shenandoah National Park	38.62 N, -78.35 W	17
SPL	Storm Peak Lab	40.45 N, -106.73 W	9
WBI	West Branch	41.73 N, -91.35 W	379
WGC	Walnut Grove	38.27 N, -121.49 W	483
WKT	Moody	31.32 N, -97.33 W	457
YAH	Yaquina Head	44.67 N, -124.07 W	13

DISTRIBUTION:

1	Jaideep Ray, 08954	MS 9159
1	B. van Bloemen Waanders, 01442	MS 1318
1	Technical Library, 08944 (electronic)	MS 0899

Sandia National Laboratories