Using generalized cross-validation to select parameters in inversions for regional carbon fluxes

Nir Y. Krakauer,1 Tapio Schneider,2 James T. Randerson,3 and Seth C. Olsen1

Received 21 April 2004; revised 6 July 2004; accepted 26 August 2004; published 14 October 2004.

[1] Estimating CO2 fluxes from the pattern of atmospheric CO2 concentrations with atmospheric transport models is an ill-posed inverse problem, whose solution is stabilized using prior information. Weights assigned to prior information and to CO2 concentrations at different locations are quantified by parameters that are not well known, and differences in the choice of these parameters contribute to differences among published estimates of the regional partitioning of CO2 fluxes. Following the TransCom 3 protocol to estimate CO2 fluxes for 1992–1996, we find that the partitioning of the CO2 sink between land and oceans and between North America and Eurasia depends on parameters that quantify the relative weight given to prior flux estimates and the extent to which CO2 concentrations at different stations are differentially weighted. Parameter values that minimize an estimated prediction error can be chosen by generalized cross-validation (GCV). The GCV parameter values yield fluxes in northern regions similar to those obtained with the TransCom parameter values, but the GCV fluxes are smaller in the poorly constrained equatorial and southern regions. INDEX TERMS: 0322 Atmospheric Composition and Structure: Constituent sources and sinks; 1610 Global Change: Atmosphere (0315, 0325); 3260 Mathematical Geophysics: Inverse theory; 3337 Meteorology and Atmospheric Dynamics: Numerical modeling and data assimilation.

Citation: Krakauer, N. Y., T. Schneider, J. T. Randerson, and S. C. Olsen (2004), Using generalized cross-validation to select parameters in inversions for regional carbon fluxes, Geophys. Res. Lett., 31, L19108, doi:10.1029/2004GL020323.

1. Introduction

[2] To complement direct measurements of, for example, land or ocean carbon uptake, inverse modeling is widely used to estimate CO2 fluxes from the observed spatial and temporal variations of atmospheric CO2 concentrations. Different inverse modeling studies, however, have reached apparently contradictory conclusions on the longitudinal and land/ocean partitioning of CO2 fluxes. For example, studies using similar data sets have variously placed a sink of order 1 Pg C yr−1 in temperate North America [Fan et al., 1998], in north Asia [Bousquet et al., 1999a], or in the north Atlantic and Pacific [Rayner et al., 1999].

[3] Estimating CO2 fluxes involves the solution of a linear system

\[ Ax = b + \varepsilon, \]  

where \( b \) is a vector of observed variations in CO2 concentrations, \( \varepsilon \) is a vector of random errors with zero mean and with covariance matrix \( \text{cov}(\varepsilon) = C_\varepsilon \), \( x \) is an unknown vector of CO2 fluxes, and \( A \) is a transport operator that relates CO2 fluxes to CO2 concentrations.

[4] The solution of this inverse problem is generally not well constrained by the CO2 concentrations (i.e., the problem is ill-posed) and must be stabilized through the use of prior information or regularity constraints [e.g., Hansen, 1998]. Estimates of CO2 fluxes are usually constrained to be close to CO2 fluxes specified a priori [e.g., Enting, 2002] by minimizing an object function of the form

\[ J = (Ax - b)^T C_b^{-1} (Ax - b) + \lambda^2 (x - x_0)^T C_x^{-1} (x - x_0), \]  

consisting of the sum of the least squares object function (first term) and a penalty term (second term) that penalizes deviations of the solution \( x \) from a given prior estimate \( x_0 \). The covariance matrix \( C_b \) represents uncertainty about the prior estimate \( x_0 \), and the regularization parameter \( \lambda \) sets the weight of the term involving the prior information relative to the least squares term. (See also the online supplement1.)

[5] In CO2 inversions, the covariance matrices \( C_b \) and \( C_x \) are usually taken to be diagonal, with diagonal entries \( c_b \) and \( c_x \) equal to assumed variances of the local CO2 concentration errors and of the regional prior flux distributions. The prior fluxes \( x_0 \) and their assumed variances \( c_x \) are typically chosen ad hoc from a range of reasonable values, as are the error variances \( c_b \) of the CO2 concentrations. It is known that inversion results can sensibly depend on the choice of such inversion parameters [Bousquet et al., 1999b; Rayner et al., 1999; Law et al., 2003], but a unified approach to quantifying this source of uncertainty and to choosing inversion parameters systematically has not been pursued.

[6] Several methods are available to choose inversion parameters systematically [e.g., Hansen, 1998, chapter 7]. One method applied in inverse problems in such fields as meteorological data assimilation [Wahba et al., 1995] and geodesy [Ditmar et al., 2003] is generalized cross-validation (GCV) [Golub et al., 1979]. A form of leave-one-out cross-validation, GCV chooses parameters by minimizing an estimated mean-squared error of predictions with the model specified by the parameters. For the CO2 problem, this

1Auxiliary material is available at ftp://ftp.agu.org/apend/gl/2004GL020323.
means that GCV chooses inversion parameters by minimizing an estimated mean-squared error of predictions of CO₂ concentrations given estimated CO₂ fluxes and the model of atmospheric transport. Although such predictions are generally not the objective of inverse modeling of CO₂ fluxes, GCV provides a useful heuristic for choosing inversion parameters.

[7] We applied GCV to estimate two parameters in the TransCom 3 framework for inverse modeling of CO₂ fluxes [Gurney et al., 2002]. The two inversion parameters considered control the weighting of the prior flux estimates and the relative magnitudes of error variances of CO₂ concentrations at different locations. We examine how the CO₂ flux estimates depend on the values assigned to these parameters, choose parameter values by GCV, and discuss how the flux estimates thus obtained differ from those obtained with the original TransCom parameter values.

2. Methods

2.1. Inverse Modeling Framework

[8] The TransCom 3 annual-mean inversion [Gurney et al., 2002] uses measured CO₂ concentrations to estimate CO₂ fluxes from 11 land and 11 ocean regions after subtracting the estimated effects of fossil fuel emissions, of ocean carbon uptake, and of covariances between the seasonality of the biosphere and that of atmospheric transport (the seasonal “rectifier effect”). The data used are the mean 1992–1996 CO₂ concentrations at 76 stations and the mean growth rate of atmospheric CO₂ concentrations during that period, the latter constraining the sum of the regional fluxes. We followed the TransCom 3 protocol except for excluding one station, Darwin, whose mean concentration is anomalously high and apparently reflects local sources [Law et al., 2003], leaving 75 stations. (The TransCom 3 data and flux patterns are available at http://transcom. colostate.edu/TransCom_3/T3 Input Data/. The prior fluxes and variances used in the TransCom 3 inversion are given in the online supplement to Gurney et al. [2002].)

[9] The transport operator A for this inverse problem specifies the modeled impact of the magnitude of each of the TransCom 3 regional flux patterns on each data point. We modeled atmospheric CO₂ transport with the Model of Atmospheric Transport and Chemistry (MATCH) [Rasch et al., 1994; Mahowald et al., 1997], run at T21 resolution (about 5.5° * 5.5°) with 26 vertical levels. The model was driven with one year of winds from a simulation with the National Center for Atmospheric Research’s Community Climate Model version 3 [Kiehl et al., 1998]. In this configuration, MATCH has been used to study column CO₂ variability [Olsen and Randerson, 2004]. Other configurations of MATCH participated in the TransCom model intercomparisons [Gurney et al., 2003].

2.2. Parameters Varied

[10] Like other recent inverse modeling studies, TransCom 3 uses prior estimates of regional CO₂ fluxes to stabilize the flux estimates. The weighting assigned to the prior estimates of regional CO₂ fluxes determines the relative influence of the data and of the prior estimates on the flux estimates. Assigning a very large weight to the prior estimates leads to a solution that is close to the prior estimates and fits the data poorly; a very small weight leads to an unstable solution that typically exhibits spurious large-amplitude fluxes.

[11] Here we scale the weighting given to the prior estimates by a parameter λ that we vary from 0.5 to 3 (equation (2)). The parameter λ can be interpreted as the regularization parameter in what is known as Tikhonov regularization or ridge regression [e.g., Hansen, 1998, chapter 5]. Alternatively, in a Bayesian inversion framework, the parameter λ can be thought of as a scaling factor of the prior standard deviations, so that instead of the error variances $c_{x,0}$ assigned by TransCom to the prior estimates, the error variances are of the form $c_{x} = \lambda^{-2}c_{x,0}$. (See also the online supplement.)

[12] The relative magnitudes of the error variances $c_{b}$ assigned to the CO₂ concentration data control the relative weights of CO₂ concentrations at different locations in the inverse model. The data error variances should reflect uncertainty due to error in the transport model and error in the regional source patterns assumed in the inversion as well as measurement and calibration error [Kaminski et al., 2001]. (Estimated errors in the forward model can be taken into account explicitly in the solution of inverse problems [Van Huffel and Vandewalle, 1989; Golub et al., 1999]. Using such methods to estimate CO₂ fluxes appears to be worth pursuing, but we will not do so here.)

[13] Investigators have typically either assumed the same error variance for all stations [Fan et al., 1998; Rayner et al., 1999] or have, as in TransCom, taken the error variance at each station to be proportional to the high-frequency variance of CO₂ concentrations at the station. In fact, the error variance may well be the sum of components that scale in each of the two ways [cf. Rödenbeck et al., 2003]. Therefore, we allow a gradation from uniform station error variances $c_{b,e}$ to the TransCom error variances $c_{b,0}$ by varying a parameter τ between 0 (uniform error variances) and 1 (TransCom error variances). The i-th station variance is given by $c_{b,i} = (c_{b,e})^{\tau}(c_{b,0})^{1-\tau}$, and the magnitude of the uniform error variance is chosen such that the diagonal covariance matrices $C_{b,e}$ and $C_{b,0}$ have the same determinant.

3. Results

3.1. Effects of Parameter Choices on Flux Estimates

[14] The large northern land sink inferred in the TransCom 3 inversions is stable with respect to changing the parameter values, decreasing by ~25% (0.7 Pg C yr⁻¹) with increasing λ (Figure 1a). The longitudinal distribution of the sink changes as τ is varied (an effect noted by Gurney et al. [2002] as explaining part of the difference between the TransCom flux distribution and that inferred by Fan et al. [1998]). Most of this shift is due to two stations at the ITN tower in North Carolina [Bakwin et al., 1995], whose low CO₂ concentrations imply a large North American sink but which are assigned large errors, hence little weight, in TransCom. At low τ, the North American sink is inferred to be higher by 0.3 Pg C yr⁻¹ (Figure 1b) and the estimated sinks in Europe (Figure 1c) and Asia correspondingly decrease.

[15] By contrast, the estimated flux distribution in equatorial and southern regions depends strongly on the parameter values chosen. Changing λ can shift fluxes of 2 Pg C yr⁻¹ between equatorial and southern land regions (Figures 1d and 1e), with particularly large flux magnitudes at low λ, where the prior constraints are weaker.
The sensitivity of the equatorial and southern regions’ flux estimates to parameter choices affects the global partitioning of fluxes between land and ocean. The total ocean flux estimate varies by up to 0.8 Pg C yr\(^{-1}\) as \(t\) is changed (Figure 1f), with compensating changes in the land flux estimate. (Because of the global growth rate constraint, the sum of land and ocean fluxes stays approximately constant at \(2.8\) Pg C yr\(^{-1}\) as inversion parameters are varied.)

### 3.2. Choosing Parameters by GCV

The GCV function is minimized when \(t = 0.7\) and \(\lambda = 2.1\) (Figure 2). TransCom’s error model (\(\tau = 1\)) or an intermediate error model (\(0 < \tau < 1\)) thus seems more appropriate than an error model with equal error variances for all stations (\(\tau = 0\)). The GCV choice of \(\lambda\) implies that the TransCom standard deviations of the prior fluxes may be too large by about a factor of 2, intimating that TransCom’s estimated CO\(_2\) fluxes overfit the data.

For northern regions, using the GCV parameter choices in place of the TransCom parameter values changes the estimated fluxes by less than 20% (Figures 1a–1c). For many equatorial and southern regions (Figures 1d and 1e), however, the GCV flux estimates are much closer to the prior values than the TransCom estimates, suggesting that the station network provides little information about CO\(_2\) fluxes in these regions.

### 4. Implications for Future Inverse Modeling Studies

Parameter-choice methods such as GCV offer some scope for reducing uncertainty in and variations between inverse modeling studies. Although selecting what parameters to estimate will remain arbitrary, the fact that even our restricted variation of parameters resulted in differences in estimated fluxes that are comparable to the TransCom inter-model variation and to the a posteriori uncertainties estimated from the data and prior flux uncertainties [Gurney et al., 2002] suggests that choosing parameters systematically is necessary.

GCV could be particularly useful for choosing relative weights for blocks of data from different sources [Gao, 1994]. As inversions for carbon fluxes add additional types of data, such as CO\(_2\) isotopic composition [Ciais et al., 1995; Rayner et al., 1999], spectroscopically determined column CO\(_2\) [Yang et al., 2002], and measurements of other gases such as CO and O\(_2\), GCV can be used to select a weighting for each type that reflects not only measurement accuracies but also model capability to simulate each type of data.

### 5. Conclusion

The choice of parameters can significantly contribute to variations in CO\(_2\) flux estimates obtained in inverse modeling studies. Methods such as GCV that choose parameters systematically by optimizing a given objective
function can improve inversion results. We recommend that uncertainty in inversion parameters be considered in future inverse modeling protocols and that formal parameter choice methods be used where appropriate.

[22] Acknowledgments. We thank K. R. Gurney for help with the TransCom 3 protocol, Z. H. Yang for reviewing drafts of this paper, and P. J. Rayner and an anonymous referee for helpful comments. NYK was supported by a graduate fellowship from the Betty and Gordon Moore Foundation. T. S. gratefully acknowledges support from NASA (NAGS-11200) and NOAA (NA05OAR4310059).

References

Bakwin, P. S., P. P. Tans, C. L. Zhao, W. Ussler, and E. Quesnell (1995), Measurements of carbon-dioxide on a very tall tower, Tellus, Ser. B, 47(5), 535–549.

Bousquet, P., P. Ciais, P. Peylin, M. Ramonet, and P. Monfray (1999a), Inverse modeling of annual atmospheric CO2 sources and sinks: 1. Method and control inversion, J. Geophys. Res., 104(D21), 26,161–26,178.

Bousquet, P., P. Peylin, P. Ciais, M. Ramonet, and P. Monfray (1999b), Inverse modeling of annual atmospheric CO2 sources and sinks: 2. Sensitivity study, J. Geophys. Res., 104(D21), 26,179–26,193.

Ciais, P., P. P. Tans, M. Trolier, J. W. C. White, and R. J. Francey (1995), A large Northern-Hemisphere terrestrial CO2 sink indicated by the 13C/12C ratio of atmospheric CO2, Science, 269(5257), 1098–1102.

Ditmar, P., J. Kusche, and R. Klees (2003), Computation of spherical harmonic coefficients from gravity gradiometry data to be acquired by the GOCE satellite: Regularization issues, J. Geod., 77(7–8), 465–477.

Enting, I. G. (2002), Inverse Problems in Atmospheric Constituent Transport, 392 pp., Cambridge Univ. Press, New York.

Fan, S., M. Gloor, J. Mahilman, S. Pacala, J. Sarmiento, T. Takahashi, and P. Tans (1998), A large terrestrial carbon sink in North America implied by atmospheric and oceanic carbon dioxide data and models, Science, 282(5388), 442–446.

Gao, F. (1994), Fitting smoothing splines to data from multiple sources, Commun. Stat. Theory Methods, 23(6), 1665–1698.

Golub, G. H., M. Heath, and G. Wahba (1979), Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics, 21(2), 215–223.

Golub, G. H., P. C. Hansen, and D. P. O’Leary (1999), Tikhonov regularization and total least squares, SIAM J. Matrix Anal. Appl., 21(1), 185–194.

Gurney, K. R., et al. (2002), Towards robust regional estimates of CO2 sources and sinks using atmospheric transport models, Nature, 415(6872), 626–630.

Gurney, K. R., et al. (2003), TransCom 3 CO2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, Tellus, Ser. B, 55(2), 555–579.

Hansen, P. C. (1998), Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion, 247 pp., Soc. for Ind. and Appl. Math., Philadelphia, Pa.

Kaminski, T., P. J. Rayner, M. Heimann, and I. G. Enting (2001), On aggregation errors in atmospheric transport inversions, J. Geophys. Res., 106(D5), 4703–4715.

Kiel, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch (1998), The National Center for Atmospheric Research Community Climate Model: CCM3, J. Clim., 11(6), 1131–1149.

Law, R. M., Y. H. Chen, and K. R. Gurney (2003), TransCom 3 CO2 inversion intercomparison: 2. Sensitivity of annual mean results to data choices, Tellus, Ser. B, 55(2), 580–595.

Mahowald, N. M., P. J. Rasch, B. E. Eaton, S. Whittlestone, and R. G. Prinn (1997), Transport of 222radon to the remote troposphere using the Model of Atmospheric Transport and Chemistry and assimilated winds from ECMWF and the National Center for Environmental Prediction/NCAR, J. Geophys. Res., 102(D23), 28,139–28,151.

Olsen, S. C., and J. T. Randerson (2004), Differences between surface and column atmospheric CO2 and implications for carbon cycle research, J. Geophys. Res., 109, D02301, doi:10.1029/2003JD003968.

Rasch, P. J., X. X. Tie, B. A. Boville, and D. L. Williamson (1994), A 3-dimensional transport model for the middle atmosphere, J. Geophys. Res., 99(D1), 999–1017.

Rayner, P. J., I. G. Enting, R. J. Francey, and R. L. Langenfelds (1999), Reconstructing the recent carbon cycle from atmospheric CO2, 813C and O2/N2 observations, Tellus, Ser. B, 51(2), 213–232.

Rödenbeck, C., S. Houweling, M. Gloor, and M. Heimann (2003), CO2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919–1964.

Takahashi, T., et al. (2002), Global sea-air CO2 flux based on climatological 13C/12C and O2/N2 observations, Tellus, Ser. B, 54(1), 197–211.

Van Huffel, S., and J. Vandevelle (1989), Analysis and properties of the generalized total least squares problem, J. Geophys. Res., Part II, 94(9–10), 1601–1622.

Vanuffelen, G., and J. Vandevelle (1989), Analysis and properties of the generalized total least squares problem Ax ≈ B when some or all columns in A are subject to error, SIAM J. Matrix Anal. Appl., 10(3), 294–315.

Wahba, G., D. R. Johnson, F. Guo, and J. J. Gong (1995), Adaptive tuning of numerical weather prediction models: Randomized GCV in 3-dimensional and 4-dimensional data assimilation, Mon. Weather Rev., 123(11), 3358–3369.

Yang, Z. H., G. C. Toon, J. S. Margolis, and P. O. Wennberg (2002), Atmospheric CO2 retrieved from ground-based near IR solar spectra, Geophys. Res. Lett., 29(9), 1339, doi:10.1029/2001GL004537.

N. Y. Krakauer and S. C. Olsen, Division of Geological and Planetary Sciences, California Institute of Technology, MC 100–23, Pasadena, CA 91125, USA. (niryk@caltech.edu)

J. T. Randerson, Earth System Science, University of California, Irvine, CA, USA.

T. Schneider, Division of Geological and Planetary Sciences and Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA, USA.