Inverse Modeling of Surface Carbon Fluxes

Please read Peters et al (2007)

and

Explore the CarbonTracker website

Consider Linear Regression

- Given N measurements y_i for different values of the independent variable x_i, find a slope (m) and intercept (b) that describe the "best" line through the observations
  - Why a line?
  - What do we mean by "best"
  - How do we find m and b?
- Compare predicted values to observations, and find m and b that fit best
- Define a total error (difference between model and observations) and minimize it!

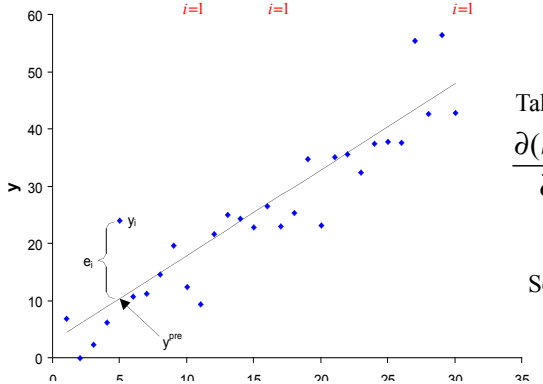
Defining the Total Error

- Our model is $y^{pre} = mx + b$
- Error at any point is just $e_i = y_i - mx_i - b$
- Could just add the error up: $\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} (y_i - mx_i - b)$
- Problem: positive and negative errors cancel … we need to penalize signs equally
- Define the total error as the **sum of the Euclidean distance** between the model and observations (sum of square of the errors):

$$SSE \equiv \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i^{pre})^2 = \sum_{i=1}^{N} (y_i - mx_i - b)^2$$

Minimizing Total Error (Least Squares)

$$SSE \equiv \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i^{pre})^2 = \sum_{i=1}^{N} (y_i - mx_i - b)^2$$



Take partial derivs

$$\frac{\partial(SSE)}{\partial m}, \frac{\partial(SSE)}{\partial b}$$

Set to zero

Solve for m and b

Minimizing the Error

(1) $\frac{\partial SSE}{\partial b} = 2Nb + 2m\sum x_i - 2\sum y_i = 0$ Solve for ***b***

(2) $b = \frac{2\sum y_i - 2m\sum x_i}{2N}$ Plug this result into other partial deriv, and solve for ***m***

(3) $\frac{\partial SSE}{\partial m} = 2b\sum x_i + 2m\sum x_i^2 - 2\sum x_i y_i = 0$

$$2\sum x_i\left\{\frac{1}{N}\sum y_i - \frac{m}{N}\sum x_i\right\} + 2m\sum x_i^2 - 2\sum x_i y_i$$

$$m\left\{\sum x_i^2 - \frac{\left(\Sigma x_i\right)^2}{N}\right\} = \sum x_i y_i - \frac{\sum x_i y_i}{N} \qquad m = \frac{N\sum x_i y_i - \Sigma x_i \Sigma y_i}{N\sum x_i^2 - \left(\Sigma x_i\right)^2} \quad (4)$$

Minimizing the Error (cont' d)

(4) $m = \frac{N\sum x_i y_i - \Sigma x_i \Sigma y_i}{N\sum x_i^2 - \left(\Sigma x_i\right)^2}$ Plug (4) into (2) and simplify:

$$b = \frac{\sum y_i}{N} - \frac{\sum x_i}{N}\left\{\frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\Sigma x_i\right)^2}\right\}$$

(5) $b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\Sigma x_i\right)^2}$

Now have simple "Least Squares" formulae for "best" slope and intercept given a set of observations

Geometric View of Linear Regression

- Any vector $\vec{a} = (x, y, z)$ can be written as a linear combination of the ***orthonormal basis*** set $(\hat{i}, \hat{j}, \hat{k})$
- This is accomplished by taking the dot product (or inner product) of the vector with each basis vector to determine the components in each basis direction
- Linear regression involves a 2D mapping of an observation vector into a different vector space
- More generally, this can involve an arbitrary number of basis vectors (dimensions)

Linear Regression Revisited

This notation can be rewritten in subscript notation: $d_i = \sum_j^M G_i m_i$

and applied to a familiar problem. Imagine that there are 2 data points (d_1, d_2) and 2 model parameters (m_1, m_2).

Then the system of equations could be explicitly written as:

$d_1 = G_{11}m_1 + G_{12}m_2$

$d_2 = G_{21}m_1 + G_{22}m_2$

Or in matrix form $\vec{d} = G\vec{m}$

With two points, this is just two slope-intercept form equations:

$y_1 = m\,x_1 + b$

$y_2 = m\,x_2 + b$

This is an "even-determined" problem - there is exactly enough information to determine the model parameters precisely, there is only one solution, and there is zero prediction error.

Generalized Least Squares

$$\begin{bmatrix} d_1 \\ d_2 \\ . \\ . \\ d_N \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & . & . & G_{1M} \\ G_{21} & G_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ G_{N1} & G_{N2} & . & . & G_{NM} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ . \\ . \\ m_M \end{bmatrix} \qquad \vec{m} = \begin{bmatrix} m_1 \\ m_2 \\ . \\ . \\ m_M \end{bmatrix}$$

$$d_1 = G_{11} m_1 + G_{12} m_2 + \ldots + G_{1M} m_M$$

$$\vec{m} = \hat{G}^{-1} \vec{d}$$

The *G's* can be thought of as partial derivatives and the whole matrix as a ***Jacobian***.

Linear Regression (again)

$$SSE = \sum_{i=1}^{N} \left(d_i - d_i^{pre} \right)^2 = \vec{e}^T \vec{e} = [e_1 e_2 \ldots e_N] \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_N \end{bmatrix} = e_1^2 + e_2^2 + \ldots e_N^2$$

$$= \left(d_1 - d_1^{pre} \right)^2 + \left(d_2 - d_2^{pre} \right)^2 + \ldots \left(d_N - d_N^{pre} \right)^2$$

$$SSE = \vec{e}^T \vec{e} = \left(\vec{d} - \hat{G}\vec{m} \right)^T \left(\vec{d} - \hat{G}\vec{m} \right)$$

Take partial derivatives w.r.t. m, set to zero, solve for m.

The solution is

$$\vec{m}^{est} = \left[\hat{G}^T \hat{G} \right]^{-1} \hat{G}^T \vec{d}$$

Matrix View of Linear Regression

$$\vec{d} = G\vec{m}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_N \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix}$$

Solution

$$\vec{m}^{est} = \left[\hat{G}^T \hat{G} \right]^{-1} \hat{G}^T \vec{d}$$

$$\hat{G}^T \hat{G} = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_1 & x_2 & . & . & x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\hat{G}^T \vec{d} = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_1 & x_2 & . & . & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\vec{m} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Matrix View of Regression (cont'd)

Perform the matrix inversion on $G^T G$:

$$\left(G^T G \right)^{-1} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}$$

Recall:

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\left(G^T G \right)^{-1} = \frac{1}{N \sum x_i^2 - \left(\sum x_i \right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sum x_i^2}{N \sum x_i^2 - \left(\sum x_i \right)^2} & \frac{-\sum x_i}{N \sum x_i^2 - \left(\sum x_i \right)^2} \\ \frac{-\sum x_i}{N \sum x_i^2 - \left(\sum x_i \right)^2} & \frac{N}{N \sum x_i^2 - \left(\sum x_i \right)^2} \end{bmatrix}$$

$$\vec{m} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Matrix View of Regression (cont'd)

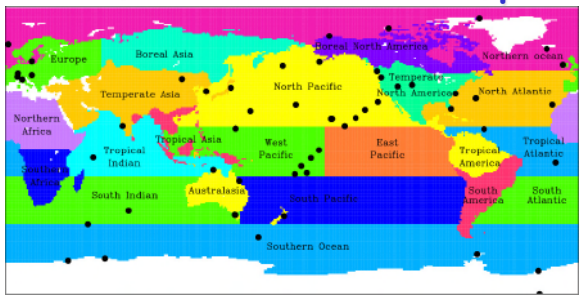
$$\vec{m} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\vec{m} = \begin{bmatrix} \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \\ \frac{-\sum x_i \sum y_i + N\sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \end{bmatrix}$$

$$m = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

TransCom Inversion Intercomparison



- Discretize world into 11 land regions (by vegetation type) and 11 ocean regions (by circulation features)
- Release tracer with unit flux from each region during each month into a set of 16 different transport models
- Produce timeseries of tracer concentrations at each observing station for 3 simulated years

Synthesis Inversion

- Decompose total emissions into ***M*** "basis functions"
- Use atmospheric transport model to generate ***G***
- Observe ***d*** and ***N*** locations
- Invert ***G*** to find ***m***

$$\begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_N \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & \cdot & \cdot & G_{1M} \\ G_{21} & G_{22} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ G_{N1} & G_{N2} & \cdot & \cdot & G_{NM} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \cdot \\ \cdot \\ m_M \end{bmatrix}$$

data *transport* *fluxes*

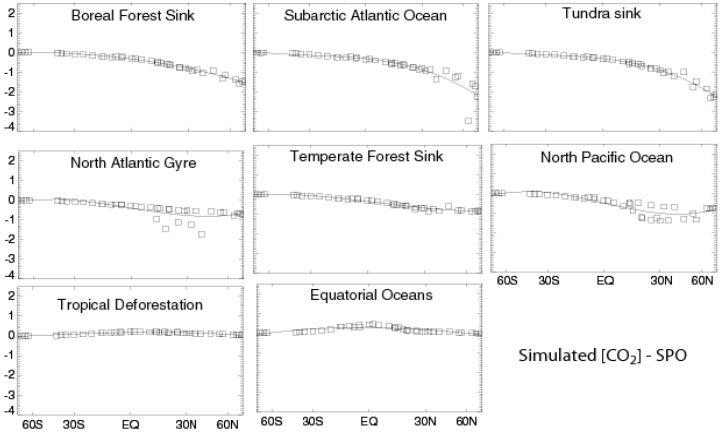
$$d_1 = G_{11}\, m_1 + G_{12}\, m_2 + \ldots + G_{1N}\, m_N$$

CO_2 sampled at **location 1**

partial derivative of CO_2 at **location 1** with respect to emissions of **type 2**

Strength of emissions of **type 2**

Near-Collinearity of Basis Vectors



"Dipoles"

Tracer	Description	Seasonal Terrestrial Biosphere		
		Fung	CASA	SiB2
T_3	Subarctic Atlantic	-0.40	-0.30	+0.05
T_4	North Atlantic Gyre	+0.69	+0.37	+0.57
T_5	North Pacific	+1.26	+0.64	+0.80
T_6	Equatorial Oceans	+1.62	+1.62	+1.62
T_7	Southern Gyres	-0.11	+0.15	+1.56
T_8	Antarctic Oceans	-0.50	-0.43	-0.30
Total Ocean Flux		+2.56	+2.06	+4.31
T_2	Tropical Deforestation	4.75	+6.04	**12.98**
T_{10}	NPP-based CO_2 Fertilization	-4.22	-4.41	**-9.56**
T_{11}	Water stress CO_2 Fertilization	-2.69	-5.62	**-14.41**
T_{12}	Temperate Forest Sink	+0.39	+0.14	+3.83
T_{13}	Boreal Forest Sink	-8.29	-3.68	-4.45
T_{14}	Tundra Sink	+4.59	+2.55	+4.19
Total Terrestrial Flux		-5.47	-4.97	-7.22
r.m.s. error (ppm)		0.39	0.35	0.41

a. Annual global fluxes in Gt C yr^{-1}. All three scenarios include 6.0 Gt C yr^{-1} due to fossil fuel emissions, and a net increase of 3.09 Gt C yr^{-1} in the atmosphere.

"Best Fit" Inverse Results

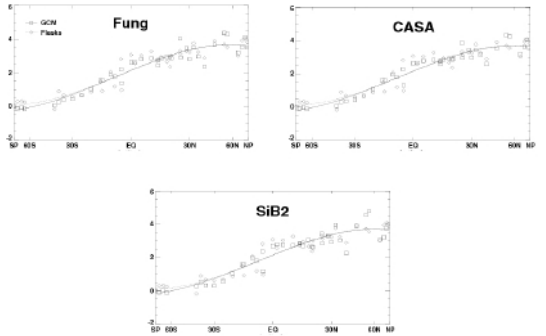
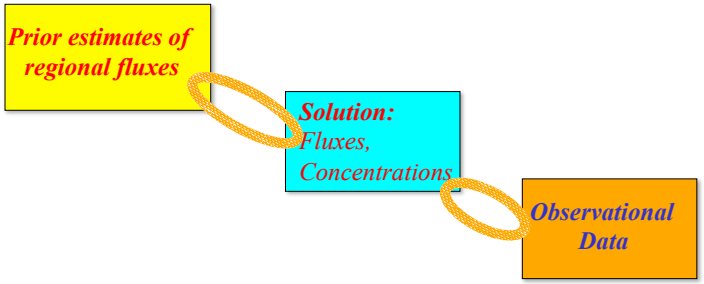


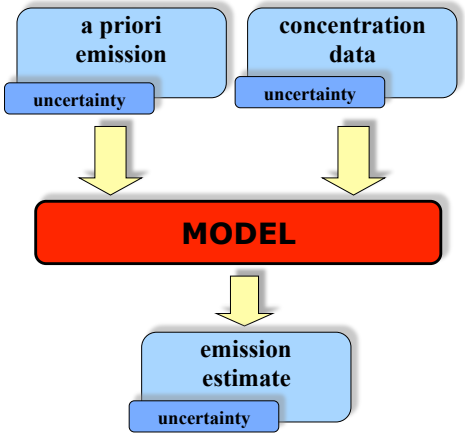
FIGURE 6.10: Simulated and observed CO_2 concentrations (ppm, difference from South Pole) at each of the NOAA flask stations using the "best fit" linear combinations of tracers (Table 6.2) derived from the data. Flask data are plotted as diamonds, and GCM data as squares. The solid curve (representing the GCM data) and the dotted curve (representing the observations) are cubic polynomials of the sine of latitude fitted to the station data.

Rubber Bands

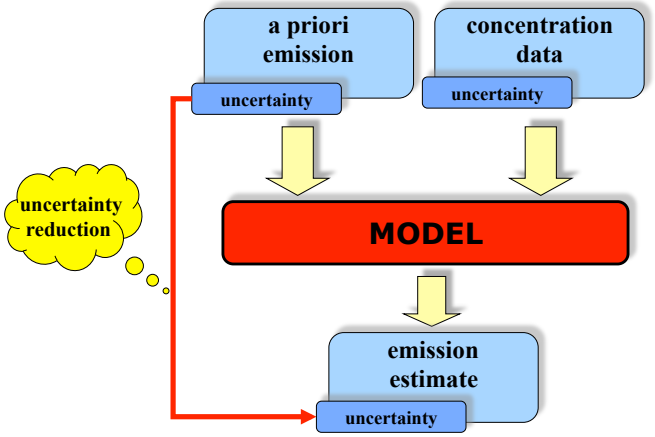


- Inversion seeks a compromise between detailed reproduction of the data and fidelity to what we think we know about fluxes
- The elasticity of these two "rubber bands" is adjustable

Bayesian Inversion Technique



Bayesian Inversion Technique



Bayesian Inversion Formalism

Our problem is ill-conditioned.
Apply prior constraints and minimize a more generalized cost function:

$$S(m) = \frac{1}{2}\left[\left(\hat{G}\vec{m} - \vec{d}_{obs}\right)^T \hat{C}_d^{-1} \left(\hat{G}\vec{m} - \vec{d}_{obs}\right) + \left(\vec{m} - \vec{m}_p\right)^T \hat{C}_m^{-1} \left(\vec{m} - \vec{m}_p\right) \right]$$

data constraint *prior constraint*

Solution is given by

$$\vec{m}_{est} = \vec{m}_p + \left(\hat{G}^T \hat{C}_d^{-1} \hat{G} + \hat{C}_m^{-1}\right)^{-1} \hat{G}^T \hat{C}_d^{-1} \left(\vec{d}_{obs} - \hat{G}\vec{m}_p\right)$$

Inferred flux; Prior Guess; Data Uncertainty; Flux Uncertainty; observations

Uncertainty in Flux Estimates

$$\vec{m}_{est} = \vec{m}_p + \left(\hat{G}^T \hat{C}_d^{-1} \hat{G} + \hat{C}_m^{-1}\right)^{-1} \hat{G}^T \hat{C}_d^{-1} \left(\vec{d}_{obs} - \hat{G}\vec{m}_p\right)$$

$$C_m^* = \left(\hat{G}^T C_d^{-1} \hat{G} + C_m^{-1}\right)^{-1}$$

- *A posteriori* estimate of uncertainty in the estimated fluxes
- Depends on transport (G) and *a priori* uncertainties in fluxes (C_m) and data (C_d)
- Does not depend on the observations per se!

Covariance Matrices

$$\mathbf{C}_d = \begin{bmatrix} \sigma_{d_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{d_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_{d_N}^2 \end{bmatrix}$$

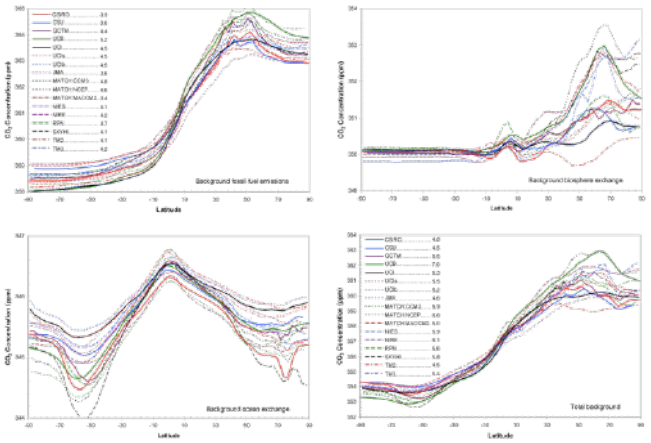
$$\mathbf{C}m = \begin{bmatrix} \sigma_{m_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{m_2}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_{m_M}^2 \end{bmatrix}$$

- Inverse of a diagonal matrix is obtained by taking reciprocal of diagonal elements
- How do we set these?
- (It's an art!)

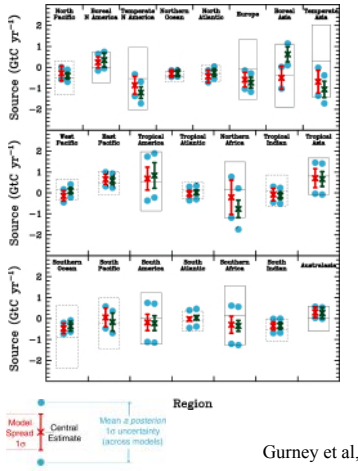
Accounting for "Error"

- Sampling, contamination, analytical accuracy (small)
- Representativeness error (large in some areas, small in others)
- Model-data mismatch (large and variable)
- Transport simulation error (large for specific cases, smaller for "climatological" transport)
- All of these require a "looser" fit to the data

Individual models: background fields



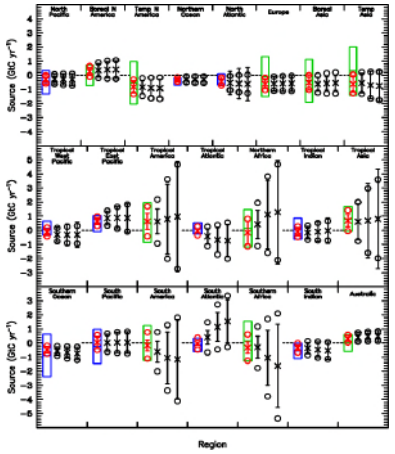
Annual Mean Results



- Substantial terrestrial sinks in all northern regions is driven by data (reduced uncertainty)
- Tropical regions are very poorly constrained (little uncertainty reduction)
- Southern ocean regions have reduced sink relative to prior, with strong data constraint
- Neglecting rectifier effect moves terrestrial sink S and W, with much reduced model spread

Gurney et al, *Nature*, 2002

Sensitivity to Priors



- Flux estimates and a posteriori uncertainties for data-constrained regions (N and S) are very insensitive to priors
- Uncertainties in poorly constrained regions (tropical land) very sensitive to prior uncertainties
- As priors are loosened, dipoles develop between poorly constrained regions