Atmospheric Carbon and the Statistical Science of Measuring, Mapping, and Uncertainty Quantification

Noel Cressie
(ncressie@uow.edu.au)

Centre for Environmental Informatics (CEI)
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong (UOW), Australia

Acknowledgement: Collaborations with Andrew Zammit Mangion and Bohai Zhang in CEI
Middle Pliocene (approx. 3.6 million years ago)
Our planet (almost 50 years ago)

Credit: Astronaut Photography, NASA Johnson Space Center; 07 Dec 1972
Two monitoring stations – half a world apart

![Carbon Dioxide (CO₂) at Earth’s Surface](image)

Figure: Monthly mean atmospheric CO₂ (in ppm) at the **NOAA Mauna Loa Observatory (Hawaii, USA)** and the **CSIRO Cape Grim Baseline Air Pollution Station (Tasmania, Australia)**. The **smooth black line** represents seasonally corrected Mauna Loa data. (Credit: Paul Krummel, CSIRO; Pieter Tans, NOAA/ESRL; and Ralph Keeling, Scripps Institution of Oceanography)
**Figure:** Comparison of NASA global yearly data and NOAA/CSIRO point-level monthly data
Carbon is the fourth most abundant element in the universe. When bonded with oxygen it forms carbon dioxide (CO$_2$).

The Earth system is closed: Total carbon is constant.

A greenhouse gas (GHG) traps heat and increases Earth’s surface temperature. CO$_2$ is a GHG, and it is most responsible for GHG increases since the beginning of the industrial period.

Carbon cycles in the Earth system: The clubs are spinning, but one of the clubs (atmospheric carbon) is getting heavier and spinning differently.

Aerosols and clouds are spinning too, which, along with temperature increases, leads to changing precipitation patterns.
Figure: TCCON stations, produced by Reto Stöckli, NASA Earth Observatory (NASA Goddard Space Flight Center).
X is a four-dimensional spatio-temporal process of atmospheric CO$_2$:

$$X \equiv \{X(x, y, h; t) : (x, y) \in D_g, h > 0, t \in D_t \};$$

$(x, y) = (\text{lon}, \text{lat})$ on the geoid $D_g$; $h$ is geopotential height; $t$ is a time slice (e.g., a week) during a time period of interest $D_t$ (e.g., a given season).

Put $h \equiv S(x, y)$, the surface pressure. Then the CO$_2$ flux at Earth’s surface is a partial derivative w.r.t. time $t$,

$$X_F(x, y; t) \equiv \frac{\partial}{\partial t} X(x, y, S(x, y); t),$$

which defines a three-dimensional spatio-temporal process $X_F$ and the “sources" and "sinks" of CO$_2$.

**Goal**: Infer $X_F$ from data. Then manage and mitigate, that is, suppress the sources and enhance the sinks!
Figure: The OCO-2 launch (02 July 2014) and the OCO-2 spacecraft in orbit (artist concept); images are from https://www.nasa.gov/content/oco-2-launch/
Astronomy picture of the day: OCO-2 launch

Image credit: Rick Baldridge
From $X$, define the column average CO$_2$ using the Riemann integral w.r.t. height $h$:

$$X_C(x, y; t) \equiv \frac{1}{S(x, y)} \int_0^{S(x, y)} X(x, y, h; t) \, dh,$$

which defines a different (from $X_F$) three-dimensional spatio-temporal process $X_C$.

Remote sensing data measure $X_C$ imperfectly:

$$Y_C(x, y; t) = X_C(x, y; t) + \epsilon(x, y; t),$$

where $\epsilon$ is the measurement-error process. Henceforth, $Y_C$ denotes the remote sensing data.

Sometimes, $Y_C$ is written as “$X_{CO_2}$.”
Figure: One day of $Y_C$ data from the OCO-2 satellite
Figure: Sixteen days of $Y_C$ data from OCO-2 (boreal winter)
Figure: Sixteen days of $Y_C$ data from OCO-2 (boreal summer)
Spatio-temporal kriging (Descriptive)

Figure: S-T Fixed Rank Kriging (FRK) in a 16-day moving window (from YouTube video, https://www.youtube.com/watch?v=_bMHDJ6Y7Hc)
Data sources include:

- TCCON monitoring stations
- Towers
- Aircraft campaigns
- Remote sensing

The instruments are different; the data are spatio-temporal; the data have different supports; the space-time “cube" is still only sparsely filled; remote sensing data is a measure of $X_C$, an integral with respect to $h$, but it is not a (direct) measure of $X_F$, a partial derivative with respect to $t$.

**Goal:** Infer $X_F$ from data. Use statistical science!
Certainty and uncertainty

- All the clubs are spinning.

- We have observations that carry information (with more or less uncertainty). We have theories (more or less uncertain) that convert the information into knowledge (that is consequently more or less uncertain).

- The use of probabilities is a compelling way to quantify uncertainties. It has a coherent set of rules resulting in quantifications of uncertainty that can be expressed in terms of unitless numbers between 0 and 1, or can be expressed in terms of entropy, or can be expressed in terms of variances/covariances.

- Bayes’ Rule is the workhorse of uncertainty quantification when inferring $X_F$. 
Build a probability model for generic data \( Y \) given a generic process \( X \): \( \Pr(Y|X) \)

Build a probability model for the process \( X \): \( \Pr(X) \)

There are parameters \( \theta \) in the two models – ignore for the moment.

The **predictive distribution**, \( \Pr(X|Y) \), captures the uncertainty in \( X \) given the data \( Y \). Bayes’ Rule results in:

\[
\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}.
\]

Summarise the predictive distribution. For example, one might:

- use the first moment, \( E(X|Y) \), to predict \( X \); and
- use the second moment, \( \text{var}(X|Y) \), to quantify uncertainty.

**Statistical-science question**: Can we make this work for the flux process \( X_F \) and for \( Y = \text{“all data”} \) (or some well chosen subset of the data)?
The principal science objective of OCO-2 is to retrieve a global geographic distribution of CO$_2$ sources and sinks, $X_F$, at Earth’s surface.

- Currently, sources (e.g., fires and respiration, fossil fuels, fresh-water outgassing, volcanism) and sinks (e.g., oceans, photosynthesis, soils) and their changes over time are not known at high enough spatial resolution to develop mitigation strategies.

- Data $Y_C$ (plentiful observations on $X_C$) and in situ sources (plentiful-in-time and sparse-in-space observations on $X_F$) comprise $Y$.

- Data-assimilation schemes invert spatio-temporal data $Y$ to predict Earth’s CO$_2$ sources and sinks with $E(X_F|Y)$. Currently, data assimilation approximations and computational bottlenecks prevent calculation of $\text{var}(X_F|Y)$. 
Graphical model for predicting $X_F$ (Dynamic)

BIAS PARAMETERS AND ERROR SCALING

OCO-2 DATA

DISCREPANCY

$X_C$

$X_F$

FLUX MODEL PARAMETERS

Legend

Parameters ($\theta$)

Data ($Y_C$)

Processes ($X_C, X_F$)

$Y_{C,1}$ $Y_{C,2}$ $Y_{C,3}$

$\eta_1$ $\eta_2$ $\eta_3$

$X_{C,1}$ $X_{C,2}$ $X_{C,3}$

$X_{F,1}$ $X_{F,2}$ $X_{F,3}$

$a, q$

time (months)
Graphical model

\[ Y_{C,t} = X_{C,t} + A_t \beta + \Gamma(\gamma) \epsilon_t, \]

\[ X_{C,t} = \begin{bmatrix} X_{F,t} \\ X_{F,t-1} \\ \vdots \end{bmatrix} + S \eta_t, \]

\[ X_{F,t+1} = M X_{F,t} + v_t. \]

- \( \{Y_{C,t}\} \) and \( \{X_{C,t}\} \) are column-averaged CO\(_2\) variables; \( \{X_{F,t}\} \) are CO\(_2\) fluxes.
- \( A_t \) contains bias variables, such as observation mode (Land Nadir, etc.), co2\_grad\_del, and S31; \( \beta \) are the regression coefficients.
- \( \gamma = (\gamma_{LN}, \gamma_{LG}, \gamma_{OG})' \) contains observation-mode error-scaling factors.
- \( H_{t,t'} \) is the sensitivity of \( X_{C,t} \) to \( X_{F,t'} \).
- \( M = aI \), where \( a \) is an autocorrelation parameter between 0 and 1, and \( v_t \sim \text{Gau}(0, \text{diag}(q)) \) encodes spatial variability.
Unknown parameters: $\theta \equiv (\beta', \gamma', a, q')'$. 

Unknown process variables: $\{X_{C,t}\}, \{X_{F,t}\}, \{\eta_t\}$.

We have some prior knowledge about what these could be; for example, the autocorrelation parameter is between 0 and 1, and the mode-dependent bias is on the order of 0.1 ppm. Encode this prior knowledge using prior distributions.

Use Markov chain Monte Carlo (MCMC) to learn about both the processes and model parameters from OCO-2 data, and see their impact on flux uncertainty.

Build an Observing System Simulation Experiment (OSSE) first, to see if our methodology can recover the true fluxes.

Data used in the OSSE were simulated at the OCO-2 space-time locations and with the error characteristics of the 10-second averages compiled and used by the OCO-2 Flux Team.
Simulate the four-dimensional CO$_2$ field $X$, then simulate data $Y^{OSSE}$, and finally predict $X_F$, including quantification of its uncertainty, using Bayes’ Rule.

Statistical computations (MCMC) are highly complex.

We obtain an answer to the statistical-science question above: Yes, we can predict $X_F$ and quantify its uncertainty in the OSSE.

In the OSSE, the true flux process $X_F$ is known (derived from $X$, which is known).

Compare $E(X_F|Y^{OSSE}) \pm (\text{var}(X_F|Y^{OSSE}))^{1/2}$, and $E(X_F|Y^{OSSE}) \pm 2(\text{var}(X_F|Y^{OSSE}))^{1/2}$, to the true value $X_F$. 
Results show that we can learn about the true flux \((-\)\) at the TRANSCOM, one-month, space-time scale over most land regions, but not over ocean regions (TRANSCOM regions 12-22, denoted with a blue square).
In the OSSE, the S31 covariate (ratio of the signal in the strong CO₂ band to that in the O₂ A-band) has its coefficient = 7, but we fixed it to be 0 in the fitted model.

Getting the S31 bias parameter wrong leads to invalid flux inferences.
Replace $Y^{OSSE}$ with $Y^{OCO-2}$ and compute:

$$E(X_F|Y^{OCO-2}),$$
$$\text{var}(X_F|Y^{OCO-2}).$$

The “value-added” to current flux-inversion work is the addition of uncertainty quantification (UQ) through:

$$\left(\text{var}(X_F|Y^{OCO-2})\right)^{1/2}.$$  

UQ allows statistical comparisons to current estimates from other flux-inversion groups. It also allows sources and sinks of CO$_2$ to be identified. Then manage and mitigate; that is, suppress the sources and enhance the sinks.