EPS 236 Environmental Modeling and Data Analysis.
Fall Term 2021

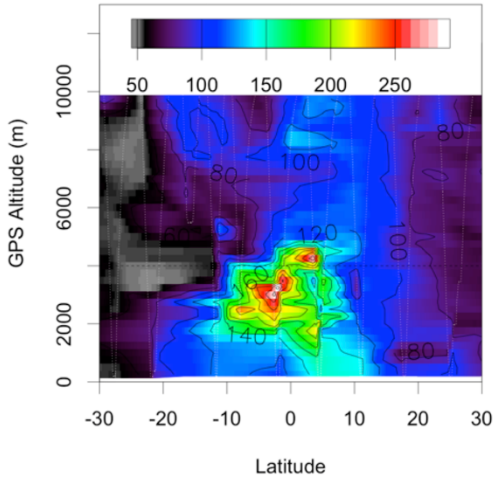
Prof. Steven C. Wofsy, Geo Museum Room 453. Telephone: 617 495-4566; email: swofsy@seas.harvard.edu
Prof. Daniel J. Jacob, Pierce Hall Room 110C. Telephone: 617 495-1794; email: djacob@seas.harvard.edu

Teaching Fellow:
Eleonora Maria Aiello emaiello@seas.harvard.edu

Location: Geo Museum 375 (TBD)
Time: Wed. and Fri., 1500 – 1615.
Office hours and section times TBD
*1st meeting **Friday September 3, 2020*.***

Course Website:
https://canvas.harvard.edu/courses/92071

CO pollutants in the Atlantic, Aug. 2016



Course overview

EPS 236 is a project-oriented, hands-on course that provides a graduate-level introduction to environmental modeling, data analysis, and data visualization. Topics include: data visualization, statistical inference, Bayes Theorem, optimal estimation, adjoint methods, Monte Carlo methods, time series analysis, denoising; principles and numerical methods for chemical transport and inverse models.

Prerequisites: Applied Mathematics 105b (ordinary differential equations) or preparation in atmospheric science are helpful, but not required; or permission of the instructors.

How to get help:

- For questions about content, email the instructors and expect a prompt reply.
- For technical difficulties with Canvas, 24 hour support is available. Click on the "Help" icon in the left navigation (located just below your Inbox) to access Canvas' 24/7 Live Chat, email support, and telephone hotline. Call or Live Chat for immediate assistance. Email response is typically 12-24 hours.

Time commitments expected for the course:

The time commitment for the course will be 8-10 hours per week. The partitioning between lectures and whole-class meetings ("synchronous") or work on your own or in small groups ("asynchronous"), will change from week to week as the course progresses.

Course content:

The course is divided into two parts:

1. Data analysis focusing on *gaining scientific knowledge* from complex environmental data sets and *quantifying sources of error.* Hands-on learning is emphasized. Data examples are drawn from atmospheric surface networks, aircraft observations, satellite sensors, COVID epidemiology, cell phones, etc.
Emphasis is on:

- concepts that underlie statistical inference
- applications of these concepts to data analysis and visualization.

Note: R will be used as a tool for <u>visualization</u>, <u>time series analysis</u>, <u>Monte Carlo methods</u>, and <u>statistical assessment</u>.

2(a). Models in environmental science: *linear models* (mathematical principles, time evolution, eigenvalues, eigenvectors); *stochastic models* (Markov chains).

2(b). Chemical transport models including basic principles, numerical methods, and inverse modeling (optimal estimation, Kalman filter, adjoint methods).

Course requirements

- **Tutorials:** Students must attend on-line tutorials offering instruction in coding R, principles and applications of various statistical methods, etc.
- **Credit:** Homework (project-oriented, bi-weekly, 30%), Projects (40%), class participation 10%, oral presentations and oral exam, 20%.
- **Exams:** There are no proctored exams.
- **Recommended**: Dalgaard, P. (2008) Introductory Statistics with R; *or similar*
- **Collaboration policy:**

For assignments in this course, you are encouraged to consult with classmates as you work on problem sets. However, after discussions with peers, make sure that you can work through the problem yourself and ensure that *answers you submit are the result of your own efforts*. You must cite any books, articles, websites, lectures, etc that have helped you with your work using appropriate citation practices. Access to solutions from previous years is strictly forbidden.

Use of laptop computers during class should be exclusively for in-class work; students should not access social media during class time.

Topics for Part 1 of EPS 236

Part 1a: Linear models, Markov chains, analysis via eigenvectors/eigenvalues. (Wofsy)

Linear models provide the conceptual framework for modeling many types of data. We study linear systems and explore their fundamental properties, e.g. in simulations of chemical species in the environment, including stochastic, inverse and adjoint models:

How do we structure a model and obtain estimates for the magnitudes of the parameters (the simplest “inverse modeling”)? We examine *model properties*: eigenvalues and eigenvectors, non-orthogonality, time-evolution operator, transient and steady-state behavior, tangent linear approximations for non-linear systems. *Applying the model*-how do we use these models as tools to improve our understanding?

Students will receive training to use R, to be utilized in applications to global chemical cycles, urban atmospheric structure and chemistry, etc. (Students already proficient in Matlab, Python, or similar applications may use one of those applications, but R provides course reference material). Note: *Excel and similar spreadsheet applications are not permitted for data analysis*.

Part 1b. Statistical inference; Time series, spatial data, and visualization (Wofsy)

Statistical inference is an essential part of the earth and environmental sciences. Probability distributions, variance, errors, and estimation of "uncertainty" of our models and data analysis are requirements for almost every paper in the field. We start this discussion with a very close look at how statistical inference is conceptualized and applied in a field where controlled experiments and repeated resampling of the same system are impossible *a priori*.

Part 1b topics: Statistical inference in the earth and environmental sciences. Probability distributions, variance, errors, and "uncertainty". Distributions and t tests; parametric and non-parametric regression. Analysis of data: linear regression, regressions with errors in dependent and independent variables, transformations of data; time series analysis, autocorrelated time series; error estimation: bootstrapping, correlated errors, bias, conditional sampling. Visualization of data: time series, scatter plots, missing data; smoothing and filling data using basic and advanced methods (interpolation, weighted least squares, the Savistky-Golay filter, Haar and Gaussian wavelets).

Part 1c. Hands-on Class Projects (Steve Wofsy and Teaching Fellow)

Environmental data often consist of a large number disparate observations directed towards understanding a particular phenomenon or set of phenomena. The data are often strictly incomparable in that they sample different spatial and/or temporal scales and different processes and attributes of the physical system. Examples include atmospheric trace gases measured from an aircraft, fluxes of these gases observed at points on the surface, long-term data acquired are remote stations on a weekly basis, and winds and temperatures obtained from radiosondes. We will use case studies to learn about data visualization and statistical inference in analysis of real data sets.

Class projects will be selected from various topics, using real data sets.

Lecture topics for Part 2.
Chemical Transport Models and Inverse Modeling (Daniel Jacob)

Lectures in Part 2 focus on the construction of chemical transport models (CTMs) in the atmosphere and oceans. Topics will start with the mass continuity equation, Eulerian and Lagrangian model frameworks, numerical solution of the advection equation and of chemical mechanisms, simulation of turbulence. The second set of lectures will focus on inverse modeling methods. Topics will include the general philosophy of inverse modeling, Bayes' theorem, optimal estimation, Kalman filters, adjoint methods.

Text: G.P. Brasseur and D.J. Jacob (2015), *Mathematical Modeling of Atmospheric Chemistry* (http://acmg.seas.harvard.edu/education/brasseur_jacob/index.html)

Lecture Schedule.

Part 1a. Linear modeling of environmental systems ("box models")

Note: There will be a training session in the use of the R programming language.

Sep 03. Introduction to EPS 236 – what we will be doing and why we will be doing it
Linear Modeling – Linear models are the common foundation for many data analysis frameworks, and for models of tracers and chemistry in the atmosphere and oceans. We start EPS 236 with an exploration of the analytical properties of linear models of complex systems, starting from a "Calculus II" level and proceeding to advanced concepts in ~ 2 lectures.

Part I: time evolution operator, solutions to the general problem.

Sep 08 and Sep 10. Linear Modeling; part II Analytical properties of linear systems; Markov chain equiv. Green's functions, Mean Age and Age spectrum; time evolution operators, tangent linear approximations for non-linear dynamical systems; application to estimating global fluxes of atmospheric tracers.

Sep 10. Introduction to the 1st Class Data Project:
A mulit-box model of greenhouse gases in the atmosphere (Workshop 1). Application of linear modeling methods, R coding skills development, stochastic modeling.

Part 1b. Statistical inference, regressions, curve fitting, confidence intervals, bootstrapping, MCMC —*focus on concepts and advanced applications*

Sep 15. Statistical Inference – a close look at the fundamentals from the point of view of atmospheric, marine, and environmental science: *maximum likelihood estimators and the chi square statistic.*

Sep 17. Linear regressions: Fitting a line (curve) to data;
Correlated parameters, degrees of freedom, overfitting.

Sep. 22. Type II regressions, York regressions, Fitexy (Chi-sq fitting)

Sep. 24. Confidence intervals, t-tests,
bootstrap error estimates, non-parametric assessment of data

1c. Time series and wavelet methods

Sep. 29. Data Filtering; Classifying data smoothing methods

Oct. 01. Modeling and analyzing atmospheric time series data (Workshop 2)

Oct. 06. Autoregressive data; systems with serial correlation

Oct. 08. Filtering and interpolation of data: wavelets and image processing

Oct. 13. Machine Learning Basics: Metropolis—Hastings optimization

Oct. 15. *Synthesis* – A look back at Maximum Likelihood and Chi-square frameworks that underlie many statistical models of data and associated statistical inference.

Oct. 20. Data sets for the final group projects
Data Workshop for the 2nd Class Project: students present to the class their data sets, and outline how they plan to apply their tools and models to their problems.

Oct. 22, 27, 29, and Nov. 3. Special Lectures and Discussions
The class can be divided into interested subgroups with common interests, which will deliver summaries to the whole class on Nov. 3.
Introduction to the raster framework in R
Remote sensing data: Efficient, memory-safe computation for large data sets in geographical coordinates, including links to NetCDF data sets, map projections, and more.
More image processing

Part 2a. Chemical transport models (Daniel Jacob)

Nov. 5. Numerical methods for advection

Nov 10. Parameterizations of turbulent transport

Nov. 12. Lecture and Workshop: Numerical solution of chemical mechanisms (Yang Li)

2b. Inverse modeling (Daniel Jacob)

Nov. 17. Applications of inverse modeling to atmospheric problems, Bayes' theorem

Nov. 19. Vector-matrix tools for inverse modeling
Analytical solution of the inverse problem

Dec. 1. Kalman filtering and 3-DVAR data assimilation
Adjoint methods and 4-DVAR data assimilation

Reading period 3 – 8 Dec: ***Individuals will work with the teaching staff on their final presentations.***

Dec 8. Final Presentations