EPS 236 Environmental Modeling and Data Analysis. Fall Term 2017

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: **Josh Benmergui**

Location: Geological Museum 375 (site of the former HUCE seminar room, near the Climate display in the Museum)
Time: Wednesday and Friday, 1000 – 1130.
Office hours and section times TBD

1st meeting ***Friday 01 September 2017*** *(no class on Wed. 30 August*

Course overview

EPS 236 is a project-oriented, hands-on course that provides a general introduction to modeling and data analysis suitable for students from many fields, with applications drawn from the atmospheric science, environmental engineering, and climate science. It is suitable for graduate students and advanced undergraduates in Earth and Planetary Sciences, Engineering Sciences/Environmental Science and Engineering, and allied natural science departments (e.g. Organismic and Evolutionary Biology, Chemistry and Chemical Biology, Physics; and at MIT, students from EAPS and Civil &. Environmental have often enrolled). *Prerequisite*: Applied Mathematics 105b or equivalent (may be taken concurrently); a course in atmospheric chemistry (EPS 133 or 200 or equivalent) is helpful, but not required; or permission of the instructors.

The course is divided into two parts:

> **Part 1. Models in environmental science** emphasizing *(a) linear models* (mathematical principles, time evolution operator, eigenvalues and eigenvectors; Markov chains), (b) *chemical transport models* including basic principles and numerical methods, and (c) *inverse modeling* (optimal estimation, Kalman filter, adjoint methods).
>
> **Part 2. Data analysis** focusing on assembling and cleaning data, understanding science content, and quantifying sources of error in analysis of complex data sets from environmental networks, satellite sensors, or individual instruments. Model concepts that underlie statistical inference and data analysis, and application of basic principles to real data, are emphasized. *R will be used as a tool for visualization, time series analysis, Monte Carlo methods, and statistical assessment.*

Course logistics and requirements

- **Credit:** Homework (50%), Projects (50%); oral exam replaces a written final.
- **Recommended**: Dalgaard, P. (2008) Introductory Statistics with R; *or similar*

- **Collaboration policy and electronic devices:**

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 any answers you submit are the result of your own efforts. In addition, 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. *Note: Use of laptop computers during class should be exclusively for in-class work with course material*.

Lecture topics for Part 1.
Chemical Transport Models and Inverse Modeling (Daniel Jacob)

This first set of lectures focus on the construction of chemical transport models (CTMs). Topics will include 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)

Requirements: bi weekly homework.

Lecture topics for Part 2.
Part 2a: Linear models, Markov chains, and analysis with eigenvectors/eigenvalues. (Steve Wofsy)

Linear models provide a widely used, basic conceptual framework for modeling many types of data. In the first part of the course, linear systems are examined to illustrate the fundamental properties of mass-conserving and non-mass-conserving systems simulating chemical species in the environment, including inverse and adjoint models, tangent linear approximations to non-linear systems, eigenmodes, non-orthogonal systems, tec..

Topics include the following: Setting up the conceptual model-how do we structure the model and obtain estimates for the magnitudes of the parameters (the simplest "inverse modeling")? *Solving the mode*l-eigenvalues and eigenvectors, the importance of non-orthogonality, the 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, which will be utilized in problems focused on 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 will provide the course reference material for Part Ib, data visualization). Excel and similar spreadsheet applications are not permitted.

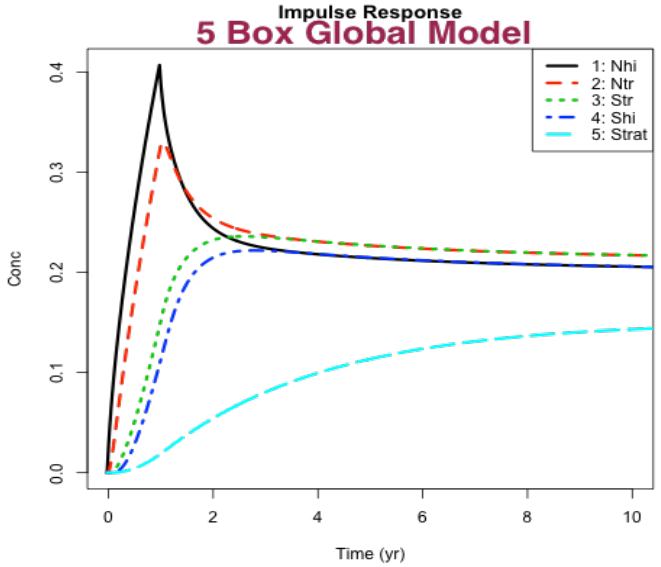
Part 2b. Time series and visualization (Steve Wofsy)

Part 1b topics: Foundations of statistical inference. Distributions and t tests; parametric and non-parametric regressions. Analysis of data: linear regression, regressions with errors in dependent and independent variables, transformations of data, correlated errors. 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 2c. Hands-on Class Projects (Steve Wofsy and Josh Benmergui)

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. Examples include*:

- Derive global emission rates from distributions of reactive and greenhouse gases as observed from ice cores, aircraft, and surface networks using a linear model framework.



- Model the instrument output and retrieval of geophysical quantities, end-to-end from the sensor to the final numbers, including a full bootstrap of confidence intervals, using real data.
- Use observed eddy fluxes in an Arctic environment to create a forward and adjoint model of CO_2 emissions from Arctic tundra. Assess climate sensitivity of the CO_2 emissions to the atmosphere.
- Assess the analytical and statistical properties of the NOAA station network for greenhouse gases.

Lecture Schedule.

Chemical transport models (Prof. Jacob)

Sep. 1: continuity equation, Eulerian and Lagrangian models
Sep. 6: numerical methods for advection
Sep. 8: numerical solution of chemical mechanisms

Inverse modeling

Sep. 13: applications of inverse modeling to atmospheric problems, Bayes' theorem
Sep. 15.: A simple example
Sep. 20.: Vector-matrix tools for inverse modeling
Sep. 22: Analytical solution of the inverse problem
Sep. 27: Kalman filtering and 3-DVAR data assimilation
Sep. 29: Adjoint methods and 4-DVAR data assimilation

Linear modeling of environmental systems ("box models")
Note: There will be a training session in the use of the R programming language.

Oct. 4 -- Linear Modeling; time evolution operator, solutions to the general problem.
Oct. 6 -- Linear Modeling; part II Analytical properties of linear systems; Markov chain equiv. Mean Age and Age spectrum; time evolution operators, tangent linear approximations for non-linear dynamical systems; application to estimating global fluxes of atmospheric tracers.

Introduction to regressions, curve fitting, confidence intervals, bootstrapping—focus on concepts and advanced applications

Oct, 11. -- Introduction to linear regressions: Fitting a line (curve) to data;
Oct. 13 -- Regressions; correlated parameters, degrees of freedom, overfitting.
Oct. 18. -- Introduction to the First Class Project: the 5-box model of greenhouse gases in the atmosphere.
Oct. 20. -- Type II regressions, Fitexy (Chi-sq fitting)
Confidence intervals, bootstrap error estimates, non-parametric assessment of data
Oct. 25 -- Confidence intervals; t-tests and bootstrap;

Time series methods

Oct. 27.— Data Filtering; Classifying data smoothing methods
Nov. 1. -- Workshop: Modeling and analyzing atmospheric time series data
Nov. 3. -- Autoregressive data; systems with serial correlation Oct. 8 -- Serial correlation
Nov. 8. -- Filtering and interpolation of data: wavelets and image processing
Nov. 10 -- Filtering and interpolation: Frequency domain, FFT, spectral decomposition
No classes in EPS 236 during Thanksgiving week.
Nov. 15, 17 -- hands-on class projects in groups of 2 or 3, interacting with instructors Wofsy and Benmergui; group presentations: oral, and annotated electronic media; poster preparation.
Nov. 29 Final lecture – the 5 deadly sins of data analysis, a.k.a. "Mistakes most commonly found during a thesis defense".
Dec. 1. – Student work groups present their work.