Vision and Questionnaire for MPO 624, Applied Data Analysis

Brian Mapes, mapes@miami.edu, RSMAS/MSC 366, Spring 2016

ADA is…

- …what makes living things alive
- …what distinguishes animals from plants
- …what distinguishes humans from animals
- …a key step in the scientific method of knowledge advancement
- …how one person convinces a rational other of a proposition

MPO is a quantitative science, often with known underlying governing equations and known drivers of phenomena through them), so let's narrow our scope about data down to *numbers*, and most often to continuous numbers rather than discrete or dichotomous (yes/no). I cannot teach you Twitter text mining, although we may admire some together at some point.

For such numerical data, our topics range from Applied Logic of puzzles and games (math operations, statistics and formal skepticism, algorithmic thinking), notated as math and pseudocode, to hands-on Computer Practices (systems, languages, strategies), to Communication Strategies (graphical information & evidence). That is still too much material, and not in a linear order, but we will see how far we can get.

Course goal:
To inspire and empower your lifelong professional journey of learning how to acquire, skeptically screen, repeatably and robustly analyze, synthesize into your worldview, and communicate new scientific information, gained from data about the world and science's models thereof.
Read it again.

Aside on graduate education and grades

In college you were exposed to a lot of material. Grading spurred you to attend, and run the material through your brain, and remember some. But in the undergraduate classroom it tends to feel like somebody else's knowledge, packaged for your shopping.

Graduate school is different. You have chosen a professional path, for your one and only life, in the science of Earth's atmosphere and ocean. Your opportunities will depend mainly on your portfolio of work and letters of recommendation. The credential (degree) is an important floor indicating demonstrated competence, but course grades are unimportant for anything besides admission to more graduate school. Falling below a B average is a mechanism for the School to terminate situations that are a waste of time and resources, so grades for good-faith efforts range from A to B. Grades (and peer scoring)

are low-stakes chits in a lighthearted but useful motivational game. More usefully, they show you that someone (me and your classmates) can notice the difference between mediocre *vs*. excellent, better *vs*. worse, doing your best *vs*. not. Hopefully that will inspire you to do things to a quality level you can stand up next to in public.

Since different students come to the course with different backgrounds and capabilities and goals (PhD vs. MPS for instance), I want us all to think of the class of 624'ers as a temporary tribe, a friendly family for a while, as we spend these 40 hours in a room together and 120 hours total as part of this course experience.

The best teaching is peer to peer, and although you have no obligations, who doesn't like to be helpful? *Collaboration is encouraged*. Some of you may already know more now than others will fully grasp at the end. That doesn't correspond to success or failure, or even to a better or worse grade necessarily. Read again the course goal.

Still, I will delineate a block of core academic material for Final Exam and Comps questions, so that you can be sure you mastered it securely.

The atmosphere and ocean are complicated systems, characterized by 4-dimensional variations of several variables like u(x,y,z,t), coupled together by universal laws of nature. Look at the sky or sea: What could it mean for a human brain to "understand" that? Often pictures are the intermediate objects of our knowledge. And then there are words: There are names for phenomena, which are like open "sub-systems" within our unbounded global fluids. There are names for abstract quantities (like variance or skill) that are used for bookkeeping. Sometimes the words are ambiguous or slippery. But an educated professional in the field needs to know them and understand them, weaknesses as well as strengths. So among other things, our course must simply teach and test a bunch of *notation* and *vocabulary* and *customs*.

And I get to offer my opinions on all that, which makes it fun for me. (You do too! Please care enough to form opinions.)

<u>Outline of topics and the scope of this course:</u>

Here is an outline of all the things I can think of that an educated student should recognize or "know" by the end. All of you know some of these things already. None of us know everything about all of them. But we can all benefit from revisiting familiar things, and it is nice to see how far we are from knowing nothing.

Items buried in the middle of this list may be big and important. Don't think of this as a priority list -- it is just my attempt to lay them out in an order from Basic to Applied:

1. Basic operations: notation & applied logic
    1.1. Sums: Σ and integral notation
    1.2. Multiplication of numbers

1.3. Differences and division: derivatives
1.4. Division to normalize (mean or average)
1.5. Sums of products of differences
    1.5.1. anomaly, variance, covariance, correlation
    1.5.2. matrix multiplication
        1.5.2.1. Projection
            1.5.2.1.1. Orthogonal decompositions (e.g. Fourier)
        1.5.2.2. Convolution
            1.5.2.2.1. Weighting and filtering, kernels
    1.5.3. Covariance matrices (w/ many applications below)
1.6. Units on those numbers: physicality

1.7. Probability theory and notation
    1.7.1. Likelihood and probability density
        1.7.1.1. p(A) – "PDF" (density, distribution)
            1.7.1.1.1. histograms and Frequentism
        1.7.1.2. p(A,B) joint distribution
        1.7.1.3. p(A|B) conditional "
        1.7.1.4. p(A) marginal "
    1.7.2. Limit theorems, tail shapes, extremes
    1.7.3. Independence and orthogonality
    1.7.4. Populations and samples
    1.7.5. Types of error (false positives vs. false negatives)
    1.7.6. Bayes' theorem and Bayesian Worldview

2. The equal sign and equations (relations)
    2.1. Laws of physics
        2.1.1. Fluid dynamics and thermodynamics (MPO)
    2.2. Interpretive equations and form fits
        2.2.1. Error as a term, least squares concept
        2.2.2. Signal and noise
        2.2.3. Response, sensitivity, susceptibility, Jacobians
        2.2.4. Time-domain dynamics
            2.2.4.1. "forcing", "feedback", "noise", "forecast", "error"

3. Statistical customs
    3.1. Broad cuts
        3.1.1. Exploratory, expository, explanatory
        3.1.2. Inductive vs. deductive tools
        3.1.3. Tools for too much data vs. too little
    3.2. Operations on a set of numbers
        3.2.1. Counting, summing, multiplying: mean, median, variance, covariance, correlation
        3.2.2. Lingo: anomaly, deviation, departure, quantile, skew, normal,…
        3.2.3. Relationship to continuous probability
    3.3. Operations on ordered sequences of numbers

        3.3.1. Lags, structure functions
        3.3.2. Fourier analysis concepts and language
    3.4. Hypothesis testing and “significance”
        3.4.1. A philosophy or state of mind
        3.4.2. Cognitive, screening, and incentive-based biases
        3.4.3. Null hypotheses and “p values”
        3.4.4. Common tests assuming Normal data: t test, F test
        3.4.5. Fishing trips and *a priori* vs. *a posteriori* significance
        3.4.6. Model selection
        3.4.7. Monte Carlo, bootstrap, cross validation, resampling mindset
    3.5. Estimation: the better way to think
        3.5.1. Model selection
        3.5.2. Curve fitting
    3.6. Synthetic data: think you understand your data? Make some.

4. Graphs and visual numbers
    4.1. Honest and distorted depictions: pitfalls
        4.1.1. frequency distributions and normalization
    4.2. Tables and line plots
    4.3. Scatter plots and curve fits
    4.4. 2D field depictions
        4.4.1. contours
        4.4.2. color -- style and perception

5. Evidence and results in MPO science
    5.1. Honesty, your preferences, and taking both sides
    5.2. Exploratory & expository: information-rich graphics
        5.2.1. Maps, charts, graphs, plots
        5.2.2. ‘Phase spaces’: histograms, scatterplots,…
        5.2.3. Breakdowns and decompositions
    5.3. Explanatory: an argument
        5.3.1. Bigger and smaller
        5.3.2. Similar and different
    5.4. Spectral analysis claims
    5.5. Covariance matrix decompositions
        5.5.1. Checkerboards and SVD
    5.6. Maps of coefficients from pointwise analysis
        5.6.1. “map and reduce” in software generality
    5.7. Interpretive equations and their estimators/fits

6. Data in computers
    6.1. Bit, byte, int, float, string; and arrays thereof
    6.2. Structures and objects
        6.2.1. Tuples and dictionaries in Python
    6.3. Labels and metadata
    6.4. Files and formats

6.5. Folders and datasets: YOUR choices matter

7. Computers
   7.1. Size: bytes and flops
   7.2. Operating systems
   7.3. Environments

8. Computing langages
   8.1. Low vs. high level
   8.2. Compiled vs. interpreted (scripting)
   8.3. Procedure vs. object oriented
   8.4. Particular languages
      8.4.1. Licensed: Matlab, IDL, …
      8.4.2. Free: Python, R, NCL, Octave, GDL, …
   8.5. Distributions, packages, environments

9. Programming and Processing
   9.1. Strategy and style: a state of mind
      9.1.1. "Working backward": from goal to activity
      9.1.2. The Resource: Your time spent to reliable job completion
         9.1.2.1. …but of future jobs too…and time of others you can help…
      9.1.3. Documenting, version control, replicability
      9.1.4. Trust, using other peoples stuff
   9.2. Algorithms and flow charts
   9.3. Libraries of capabilities
      9.3.1. Accessing data and I/O
      9.3.2. Processing steps
      9.3.3. Graphics as outputs/ results
   9.4. Literate programming, repeatability
      9.4.1. Notebooks
   9.5. Development environments

10. Practices and trends in data research
   10.1. Legacy – massive, but just 1-2 generations of *ad hoc*
   10.2. Betterment principles in the Software Age
      10.2.1. Repeatability of results
      10.2.2. Provenance of data
      10.2.3. Building on prior work; progress
   10.3. "Workflow" thinking

We might approach material in an odd order – perhaps 1,7,8,2,10,3,9,5,6,4. Also we will fall back on the book's framework at times, which will seem less ad hoc. This course is not about a bounded and organized block of material. Read again the course goal.

Meanwhile there will be a few assignments. Also you should be thinking about your course project all semester, not just at the end.

To better help me map out the order and level of class material this year, please answer the following survey (including pasting in figures where asked).

MPO 624 Intake Survey

Name, program, year:
Tiago Carrilho Bilo, MPO, First.

Advisor and research topic (if applicable):
Dr. William Johns (deep ocean circulation, MOC variability, boundary currents)

Career goals/ hopes:
Become a good and independent scientist able to ask relevant questions

Hopes for this course (please be as specific and detailed as you like, perhaps from the outline above):

- Learn the theoretical and operational aspects of data analysis methods that I will extensively use during my scientific life (e.g., EOFs);
- Improve my data visualization skills;
- Improve my coding skills.

Computer system you will work on (Windows, Mac, Linux; RAM if you know):
Mac

Describe your computer experience in narrative form. What are your thoughts, philosophy, worries, hopes about working with computers. Don't try to impress, just show me where we are all at here at the beginning. Read the course goal again.
In terms of computer systems, I have experience using Windows (for games), Linux and more recently Mac. Using Linux machines I learned how to use command lines and code in shell, C++, and create documents/presentations using LaTEX.

Since I got involved in physical oceanography some years ago I acquired some knowledge about data analysis and numerical modelling, therefore I had to learn programming in Matlab, Fortran 90, and GMT. Two years ago I migrated from Matlab to Python.

Have you used the command line in a terminal?
Favorite commands:
Yes. I do not have favorite commands, however what I really enjoy is data visualization in python (plt.plot, contour, contourf, and etc).

Have you edited code? If so, in what editor?
Favorite or proudest few lines of code (any language):
I edit codes all the time using the Sublime Text 2 or VI. Bellow, I present some code lines in python used to plot a very nice TS-diagram.

```
fig = plt.figure(fign,facecolor='w')
ax = plt.gca()
ax.set_xlim(33.,38.)
ax.set_ylim(0.,25.)


levels = [24.5,25.6, 26.9,27.38,27.53,27.88]

CR = ax.contourf(S1, T1, D1, levels, colors = ('r','gold','skyblue','mediumaquamarine','plum'), origin='lower')
cr = ax.contour(S1,T1,D1,levels=[25.6,26.9,27.38,27.53],colors='k',linewidths=2.5,linestyles='solid')
c = ax.contour(S1,T1,D1,levels=range(23,30,1),colors='grey',linewidths=1.5,linestyles='solid')



# CARBOM
ax.plot(carbv_ctd1.sal,carbv_ctd1.pot_t,'b.')
ax.plot(carbv_ctd2.sal,carbv_ctd2.pot_t,'b.')
ax.plot(carbv_ctd3.sal,carbv_ctd3.pot_t,'b.')
ax.plot(carbv_ctd4.sal,carbv_ctd4.pot_t,'b.')
ax.plot(carbv_ctd5.sal,carbv_ctd5.pot_t,'b.')
ax.plot(carbv_ctd6.sal,carbv_ctd6.pot_t,'b.')
ax.plot(carbv_ctd7.sal,carbv_ctd7.pot_t,'b.')

# CERES
ax.plot(ceres_ctd1.sal,ceres_ctd1.pot_t,'.',color='darkorange')
ax.plot(ceres_ctd2.sal,ceres_ctd2.pot_t,'.',color='darkorange')
ax.plot(ceres_ctd3.sal,ceres_ctd3.pot_t,'.',color='darkorange')
ax.plot(ceres_ctd4.sal,ceres_ctd4.pot_t,'.',color='darkorange')


ax.plot([33.,36.,36.,33.],[26.,26.,18.5,18.5],'k',lw=4.)

# Coastal Water
ax.text(33.1,19.,'Coastal Waters',fontsize=22)


# Water Masses Identification
ax.text(33.05,11.5,'TW',color='w',fontsize=30,rotation=30,alpha=0.8)
ax.text(33.1,5.,'SACW',color='w',fontsize=30,rotation=50,alpha=0.8)
ax.text(34.7,9.,'AAIW',color='w',fontsize=30,rotation=40,alpha=0.8)
ax.text(36.,13.2,'UCDW',color='w',fontsize=25,rotation=30,alpha=0.8)
ax.text(37.1,15.8,'NADW',color='w',fontsize=25,rotation=25,alpha=0.8)
```

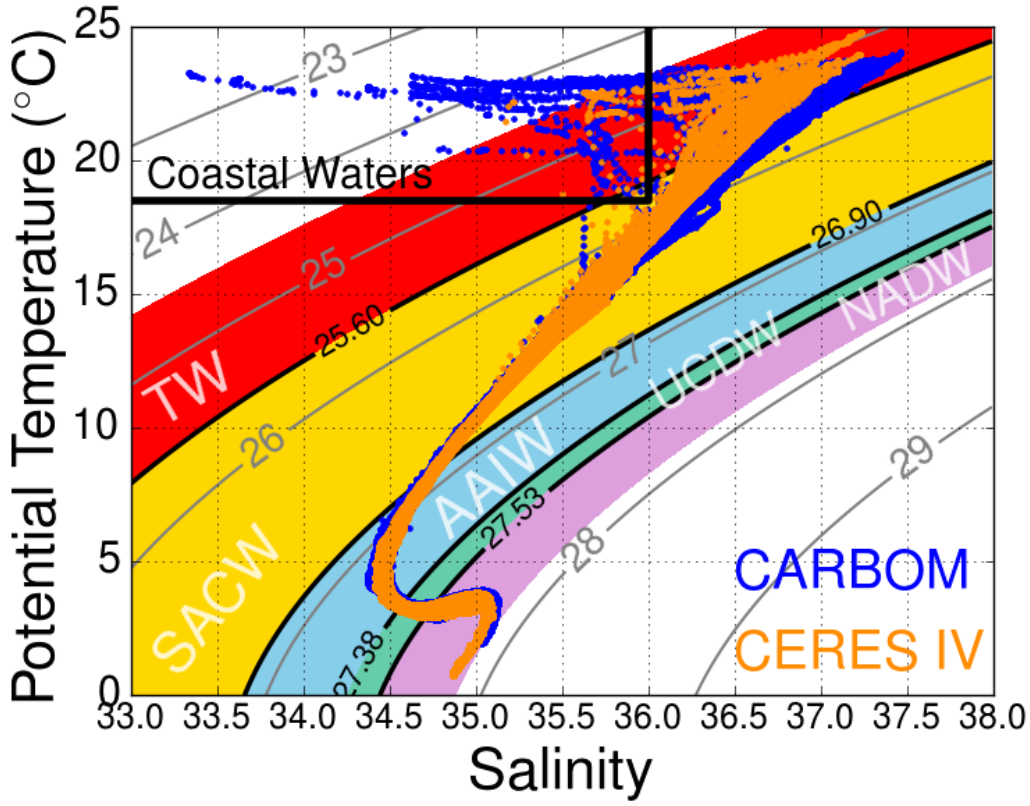
```
cr.levels = ['%1.2f'%(25.6),'%1.2f'%(26.9),'%1.2f'%(27.38),'%1.2f'%(27.53)]
ax.clabel(cr,cr.levels,colors='k',fontsize=16,fontweight='bold',inline=1,manual=True)
ax.clabel(c,fmt='%i',colors='grey',fontsize=21,inline=1,manual=True)

# Legends
ax.text(36.5, 4., 'CARBOM', fontsize=28,color='b')
ax.text(36.5, 1., 'CERES IV', fontsize=28,color='darkorange')


ax.set_xticks(np.arange(33.,38.5,0.5))
ax.set_xticklabels(np.arange(33.,38.5,0.5),fontsize=18)

ax.set_yticks(range(0,30,5))
ax.set_yticklabels(range(0,30,5),fontsize=22)

ax.set_ylabel(ur"Potential Temperature ($^{\circ}$C)",fontsize=30,fontweight='bold')
ax.set_xlabel(ur"Salinity",fontsize=30,fontweight='bold')


ax.grid()
plt.show()

return fig
```

Nicest figure you have made from data (paste, explain):

A figure you admire for graphical reasons (paste, explain):

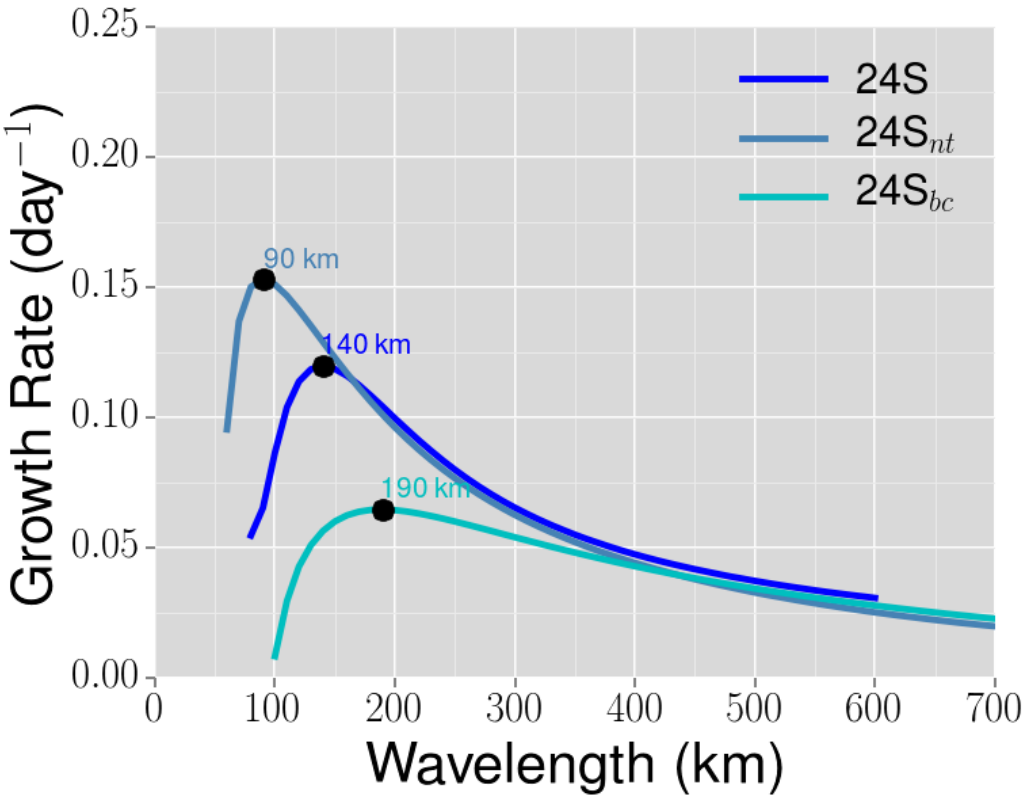
A figure you admire for content reasons (paste, explain):

The TS-Diagram bellow is the nicest figure and the one that I admire most (graphical and content reasons). This plot is beautiful, clear, efficient in showing the information and its content is rich. In the right contextualization it can be used to study a lot of aspects of the Brazil Current and its vicinities circulation.



A figure you are puzzled about or intrigued by (paste, explain why its puzzling):

During my masters, I was interested in the stability characteristics of the Brazil Current's vorticity waves at 24°S. One of the most puzzling result is: when baroclinic and barotropic instability are happening at the same time (curves 24S and $24S_{nt}$) the instability waves are shorter than waves triggered by "only" baroclinic instability ($24S_{bc}$). I was not able to explain why this happens based on stability theory, however the result is consistent with some observations of the Brazil Current meandering in the region.



Favorite ADA-related Web site or software not mentioned in course materials so far:
https://github.com

Initial brainstorming thoughts on a possible topic or ingredients for your term project.
You might as well choose your research or something related to your other coursework, so you can double-count the effort and do better work instead of just more work. Any or all of the following thought-provokers may be your springboard. Again, don't worry about impressing or grading or anything else, just help me (and us all, we will look these over on screen in class) understand where you're at and your interests.

(ONLY BRAINSTORMING)

Your application: what behavior of what system might you be trying to characterize, or compare to what other system or model or forecast?
The Meridional Overturning Circulation (MOC) and the North Atlantic western boundary current system.

Your data: what dataset(s) might contain the information you would like to explore or address? Feel free to use this as a springboard for a conversation with your advisor.
Currentmeters records of the Antilles and Deep Western Boundary Currents off the Bahamas (WOCE, MOCHA) and global ocean simulations (HYCON, ECCO2, GFDL models, etc) outputs.

Your analysis: what kinds of questions do you imagine you will ask about the system or the datasets? What would constitute an answer or an addressing of the issue? What kinds of figures would you like to create?

- Questions about the relationship between forcing mechanisms and the systems variability;

- The explanation of the phenomena that connect the forcing agents to the variability;

- Any kind simple line plots, contour maps, stickplots, boxplots, etc.