Talk Outline
• Science Motivation
• Philosophy of Modeling
• Model Construction
• Evaluation Against Data
Anthropogenic CO$_2$ uptake currently controlled by ocean circulation; but in future, what will be role of climate & biology?

For ocean acidification may want models to address many different aspects:
- patterns & trends in seawater chemistry
- population biology of individual species
- food-web & ecological interactions
- biogeochemical feedbacks
- socio-economic effects on fisheries & ecosystem services
"I am never content until I have constructed a mechanical model of the subject I am studying. If I succeed in making one, I understand; otherwise I do not."
- Lord Kelvin

"People don't understand the earth, but they want to, so they build a model, and then they have two things they don't understand,"
- Gerard Roe in “The Whale and the Supercomputer” by C. Wohlfarth
The Interdisciplinary Conundrum

Physics
• (relatively) simple set of (mostly) known governing equations (Navier-Stokes) => complex phenomenon
• unresolved scales

Chemistry
• mass balance equations
• chemical fields integrate over time/space variability
• don’t uniquely identify process/mechanism

Biology
• information as stories and conceptual pictures
• historical/evolutionary contingency
• no “biological” Navier-Stokes
• aggregate complicated dynamics across multiple-scales (genes, cells, populations, ecosystems)
Why do we build/use models?

- Quantitative dynamical framework
  • Are different data sets, rate estimates consistent?

- Design of experiments & observing systems
  • What, where and when do we sample?

- Hypothesis testing
  • If we add or change X, what happens?

- Forecasting
  • What will the ocean look like at some point in the future?

- Solving for unknown parameters and rates
  • Given things we can measure, can we estimate properties that are difficult to measure?
“Stocks” versus “Rates”

\[ C \quad \text{d}C/\text{dt} \]

Estimated from data
Unknowns
What is a model?

- **regression curves**
  variable Y is a function of other variables X and regression parameters p:
  \[ Y = f(\text{variables}, \text{parameters}) \]
  \[ \text{Chl} = p_0 + p_1 T \]

- **forward models (often time dependent)**
  Integrate forward in time to find Y:
  \[ \frac{dY}{dt} = \text{circulation} + f(\text{variables, parameters, forcing}) \]
  \[ \frac{d\text{Phyto}}{dt} = \text{circulation} + \mu \text{Phyto} (1-e^{-\frac{E}{\mu}}) - \lambda \text{Phyto} \]

- **inverse models**
  Invert the problem to find parameters from data:
  \[ \text{Parameters} = f(\text{circulation, data, forcing}) \]
What is a model? (continued)

-diagnostic versus prognostic models
  In diagnostic models, some variables may be prescribed based on observations:
  e.g., satellite chlorophyll => ecosystem model
  (no equation for dChl / dt)

-data assimilation
  Combine model equations and observations in a dynamically consistent fashion
  e.g., weather prediction
  analysis = f(model forecast, observations)

Prognostic, forward models needed to project into the future
From Word Problem to Equations

• Phytoplankton levels depend on nutrient inputs
• Perturbations relax back to some stable background level

\[
\frac{dP}{dt} = \mu_0 \left(1 - \frac{P}{C_p}\right)P
\]

• Functional form
  *Logistic model*

• State variable (concentrations)
  \(P\) (mmol C/m\(^3\)) phytoplankton

• Parameters
  \(\mu_0(1/d)\) and \(C_p\) (mmol C/m\(^3\))

Phytoplankton levels depend on nutrient inputs. Perturbations relax back to some stable background level. The functional form of the logistic model is given by:

\[
\frac{dP}{dt} = \mu_0 \left(1 - \frac{P}{C_p}\right)P
\]

- **\(\mu_0\)** (1/d) and **\(C_p\)** (mmol C/m\(^3\))

Phytoplankton levels depend on nutrient inputs. Perturbations relax back to some stable background level. The functional form of the logistic model is given by:

\[
\frac{dP}{dt} = \mu_0 \left(1 - \frac{P}{C_p}\right)P
\]

- **\(\mu_0\)** (1/d) and **\(C_p\)** (mmol C/m\(^3\))

**Diagram:**
- **\(dP/dt\)**
  - **\(\mu_0\)** >0 net growth
  - **\(0\)** <0 net loss

**Graph: P vs. time**
- **\(C_p\)**
- **\(P\)**
Simple NPZ Model

\[
\frac{dP}{dt} = \mu_0 \left( \frac{N}{k_N + N} \right) \left( 1 - e^{\alpha E / \mu_0} \right) P - g \left( \frac{P}{k_p + P} \right) Z - m_P P
\]

- Nutrient limitation
- Light limitation
- Grazing
- Mortality

\[
\frac{dZ}{dt} = ag \left( \frac{P}{k_p + P} \right) Z - m_Z Z
\]

\[
\frac{dN}{dt} = -\mu_0 \left( \frac{N}{k_N + N} \right) \left( 1 - e^{\alpha E / \mu_0} \right) P + (1 - a)g \left( \frac{P}{k_p + P} \right) Z + m_P P + m_Z Z
\]

- Three coupled ordinary differential equations
- Mass conservation
Discretization & Numerical Integration

$$\frac{dy}{dt} = f(t,y)$$

Discrete form
$$\Delta y = f(t,y)\Delta t$$

numerical method: Euler’s Method
$$y^{n+1} = y^n + f(t^n, y^n)\Delta t$$

- Subdivide time into discrete time-steps
- "Integrate" forward using $\Delta y/\Delta t$ approximation
- Numerical methods introduce errors
**Discretization & Numerical Integration**

**Runge-Kutta (2\textsuperscript{nd} order)**

\[ k_1 = f(t^n, y^n)\Delta t \]

\[ k_2 = f(t^n + \frac{1}{2}\Delta t, y^n + \frac{1}{2}k_1)\Delta t \]

\[ y^{n+1} = y^n + k_2 \]

- use a “trial step” to find gradient at mid-point
- even with 2x larger \( \Delta t \), more accurate integration
How do you estimate parameters and functional forms?

Laboratory & field incubations
  • P-E curves; nutrient uptake curves
  • elemental stochiometry

Comparative analysis
  • allometric relationships

Tuned or optimized against field data
  • mismatch between parameters and data
  • cross-site comparison

Previous models
Adding Circulation

\[ \frac{\partial P}{\partial t} + \vec{u} \cdot \nabla P - \kappa \nabla^2 P = \text{RHS} \]

- advection
- diffusion
- biological source/sink terms
Models have time/space scale limits

- Computational costs scale as \((\text{length})^3\) to \((\text{length})^4\) and (time)
- Typically get \(~2-3\) decades in space (more in time)
- Can not resolve all scales; parameterize sub-grid scale

Dickey (2003)
Coupled “Eco-biogeochemical” Elements

Physics (flow field; mixing)
- equations for resolved flow; level of approximation (e.g., primitive equations; quasi-geostrophy)
- forcing (winds, heat & freshwater fluxes, light, tides)
- parameterization of unresolved scales (mixing)
- model architecture (e.g., horizontal vs. isopycnal)

Chemistry ($CO_2$, $O_2$, nutrient fields)
- air-sea gas exchange
- elemental stoichiometries
- trace metal deposition and scavenging

Biology
- primary production, respiration, remineralization
- community structure and succession
- bio-optics
- etc.
- Aggregate into trophic levels/functional groups
- Rates/processes from limited culture/field studies
- Many aspects empirically based
- Data poor for validation (rates, grazing, loss terms)
Multi-scale models

- **cellular models**
  - gene expression, metabolism, energetics

- **population models**
  - individual based or continuous dist.
  - cell-cell interactions (LES & DNS)

- simulate ecological functions
  - "genotype" => "phenotype"
  - abandon "boxes"

- ecological/evolutionary rules for ecosystem assembly
  - maximize resilience or energy flow

- emergent behavior & selection
  - selection and niche adaptation
  - physiological plasticity & constraints
  - micro/macroevolution
Cell Physiology/Genomics

Diatom Genome
Armbrust et al., Science, 2004
• Initialize many potentially viable functional types of phytoplankton

• Assign attributes and parameter values from prescribed ranges with element of chance

• Explicit competition selects for fittest functional types
Ocean ecology and biogeochemistry are (still) data-driven sciences

How do we avoid the trap of:
“false models tested by inadequate data”

John Steele
Modern Air-Sea CO$_2$ Flux

Model

"Looks pretty good" test

χ (chi) by eye

Takahashi (2002)


J. Marine Systems Special Issue on Skill Assessment for Coupled Biological / Physical Models of Marine Systems
Vol. 76, Issue 1-2, 2009
4) $AE$ — the average error (bias)

$$AE = \frac{\sum_{i=1}^{n} (P_i - O_i)}{n} = \overline{P} - \overline{O},$$

1) $r$ — the correlation coefficient of the model predictions and observations:

$$r = \frac{\sum_{i=1}^{n} (O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n} (O_i - \overline{O})^2 \sum_{i=1}^{n} (P_i - \overline{P})^2}},$$

2) $RMSE$ — the root mean squared error (also referred to as root mean squared difference):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}},$$

Bias

Correlation

rms Error

Figure 2. Diagram for displaying pattern statistics. The radial distance from the origin is proportional to the standard deviation of a pattern. The centered RMS difference between the test and reference field is proportional to their distance apart (in the same units as the standard deviation). The correlation between the two fields is given by the azimuthal position of the test field.
Look at the magnitude & structure in model-data residuals

Log-Normal Variables (e.g., chlorophyll)

\[ X = \log(\chi) \]  

\[ \langle \chi \rangle_G = \sqrt[\frac{N}{i=1}^N] \chi_i = \exp(\langle X \rangle) \]  

geometric mean  

geometric bias
(no bias =>1)  

geometric rms error
(~normalized to typical data value)
3) **RI** — the reliability index

\[
RI = \exp \left( \frac{1}{n} \sum_{i=1}^{n} \left( \log \frac{O_i}{P_i} \right)^2 \right),
\]

6) **MEF** — the modeling efficiency

\[
MEF = \frac{\left( \sum_{i=1}^{n} (O_i - \bar{O})^2 - \sum_{i=1}^{n} (P_i - O_i)^2 \right)}{\sum_{i=1}^{n} (O_i - \bar{O})^2},
\]

\[
MEF = 1 - \frac{\text{RMSE}^2}{s^2}
\]

\[
\chi^2_i = \frac{1}{\nu} \sum_i (P_i - O_i)^2 / \varepsilon_i
\]

average factor model differs from data

predictions relative to observed mean

MEF = 1 great

MEF = 0 no better than obs. mean

MEF < 0 worse than obs. mean

Reduced Chi squared \(\Rightarrow \sim 1\)
Assessing Model Skill

Relationships between the truth, model and data (adapted from the ideas of Dan Lynch)

Prediction

Predictive Error

Truth

Observational Error

Prediction uncertainty (e.g. numerical error, parameter uncertainty)

Data assimilation is the art of reducing this distance

Observational accuracy, (e.g. measurement error, range of replicates etc.)

Poor Model Skill

Ideally model uncertainty lies within the range of observational uncertainty.

Good Model Skill

Ocean carbonate system determined by temperature, salinity & 2 of 4 parameters (pH, total carbon, alkalinity, pCO₂)
Add sensors to autonomous platforms (AUVs, gliders & floats)
Some Issues to Ponder

Representativeness of data $y_0$
- “footprint” of observation & mismatch with model grid
- local heterogeneity or point sources
- aliasing of unresolved frequencies/wavenumbers (e.g., diurnal cycle)
- data selection (i.e., exclude “unrepresentative” observations)

$R = R_{\text{instrument}} + R_{\text{representativeness}}$

$y_{\text{obs}} = y_{\text{true}} + \epsilon_{\text{obs}}$

$\epsilon_{\text{obs}} = \epsilon_{\text{random}} + \epsilon_{\text{systematic}}$

$E[\epsilon_{\text{obs}}^1, \epsilon_{\text{obs}}^2] \neq 0$
Modeling Methods for Marine Environments
David M. Glover, William J. Jenkins & Scott C. Doney

- data analysis
- modeling techniques
- ocean examples and applications
- MATLAB based demos and code
- detailed web notes (and perhaps some day a book)

(http://eos.whoi.edu/12.747/)
Matlab Primer

-can run from Matlab command window or “scripts” (m-files)
-use help & lookfor commands

-define variables (case sensitive) & standard functions:
  \[ a = 7.3 \times 10^{-7} \]
  \[ b = -\log_{10}(a) \]
  (follow with “;” if don’t want the answer echoed back)

-vector mathematics
  \[ C = [0:5:100] \Rightarrow C = [0 \ 5 \ 10 \ 15 \ldots \ 100] \]
  \[ C(4) \Rightarrow 15 \]
  \[ D = 10 \times C \]
  \[ E = C \times C \] (use “.*” for scalar math)

-plotting of 2-D and 3-D graphics
  \[ \text{plot}(C,D,’-’) \]
Matlab Primer

- for-loops to cycle over common set of commands
  for i=1:n
    F(i) = exp(-i*lambda)
  end

- call user-written functions or subroutines
  C = convert_to_centigrade(F)

- hands-on demonstration (m-files)
  - Euler vs. 2nd order Runge Kutta
  - "simple" phytoplankton model

- Ordinary Differential Equation (ODEs) integrators
  - find y(t) from y(t0) and equation for dydt=f(y,t,p)
  - define "function" to integrate e.g. "dydt" (m-file)
    [T,Y] = ode12s(‘dydt’,T,Y0)

CO₂ thermodynamics code [pH,pCO2, …]=f(DIC,ALK,T,S…)