Final: Solutions

1. (10 points) Consider a small data set that can be expressed as a matrix:

$$A = \begin{bmatrix} 1 & 2 & 0 & 3 \\ 2 & 1 & 2 & 1 \\ 3 & 3 & 2 & 4 \\ -1 & 1 & -2 & 2 \end{bmatrix}$$

What is the rank of this matrix? How much of the variance in the data can be expressed by the first EOF? How much by the first two EOFs?

Solution: Although this is a 4×4 matrix, the SVD of the matrix has 2 non-zero eigenvalues, indicating that the matrix is rank 2. The first mode EOF explains $\lambda_1^2/\sum_{i=1}^2 \lambda_i^2 = 80.33\%$ of the variance. The second mode EOF explains the remaining 19.67% of the variance, so that 100% of the variance is explained by the first two modes.

2. (20 points) You would like to estimate the meridional eddy temperature flux $\langle v'T'\rangle$ using a current meter and a thermistor. How long should you deploy your instruments in order to be able to distinguish the mean flux from zero (using 5% confidence limits)? At present you have made some preliminary measurements that have yielded the following information (some of which may prove irrelevant):

a. The lagged covariance of temperature can be fit to a Gaussian with an e-folding scale of 35 days.

b. The lagged covariance of velocity can be fit to a Gaussian with an e-folding scale of 10 days.

c. The lagged covariance of $\{v'T'\}$ can be fit to a Gaussian with an e-folding scale of 30 days.

d. The standard deviations of v, T, and $\{v'T'\}$ are all 2 in the appropriate units.

e. The means of v, T, and $\{v'T'\}$ appear to be about 1.

Solution: This problem involves two steps. First we need to determine how many effective degrees of freedom are required, and then we need to decide how long a current meter deployment that represents. The standard deviation of an averaged estimate $\overline{v'T'}$ is $\sigma_{v'T'}/\sqrt{N_E}$. For data with Gaussian statistics (which we'll assume here) roughly 95% of observations are within two standard deviations of the mean. So if we want to have 95% certainty of finding a mean that's distinct from zero, we'll need $2\sigma_{v'T'}/\sqrt{N_E} < \overline{v'T'}$. This means that $\sqrt{N_E} > 4$, and $N_E > 16$. (More formally, we'd require that $N_E > (1.96 \cdot 2)^2 = 15.37$.

We know that

$$N_E = \frac{T}{\int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{T_e^2}\right)\, dt}$$

so

$$T = N_E \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{T_e^2}\right)\, dt$$

where N_E is 16 and T_e is the e-folding scale, 30 days. Thus, $T = 16\sqrt{\pi}30$ days $= 851$ days.

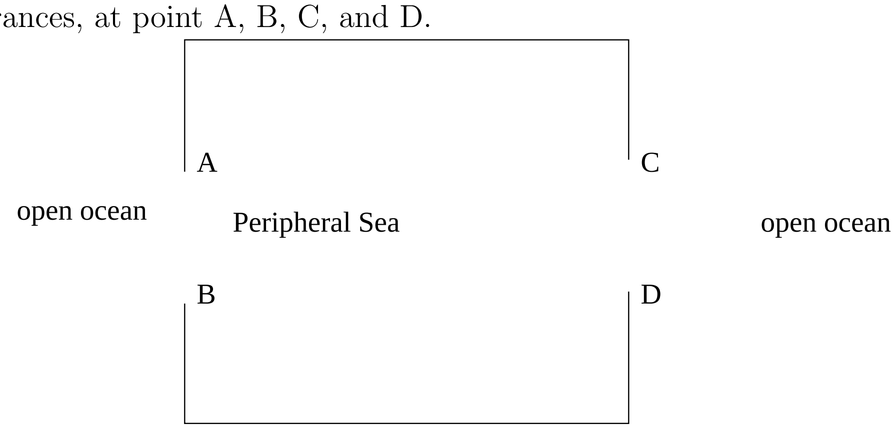
3. (20 points) You have collected the following data, which are irregularly spaced in time. In this part of the world, the covariance of temperature is assumed to be Gaussian with an

t (yrs)	T (°C)	σ_T (°C)
1920.	17.1	4.
1922	15.	4.
1923	14.8	2.
1923.5	14.4	2.
1927	19.	1.

e-folding timescale of 2 years. The variance of the signal is assumed to be 4 at zero time lag, and the variance of the measurement noise varies depending on the method used to collect each observation, as indicated in the table. Use the formalism of objective mapping to estimate the temperature at this location in 1925 and the uncertainty of your estimate. (Hint: be sure to remove the mean before doing the calculation.)

Solution: This is a relatively straight-forward objective mapping problem. First define a data-data covariance matrix: $C_{T(t)T(t+\Delta t)} = 4\exp(-(\Delta t)^2/2^2) + \sigma_n^2\delta(\Delta t)$. This produces a 5 by 5 matrix. Now, compute the data-grid covariance: $C_{T(t)\hat{T}(1925)} = 4\exp(-(1925-t)^2/2^2)$. This will be a 5 by 1 vector. Finally, compute your solution $\hat{T} = \overline{T} + C_{T(t)T(\hat{1925})} C^{-1}_{T(t)T(t+\Delta t)}(T - \overline{T}) = 16.41°$. The uncertainty is $\sigma^2 = C(0) - C_{T(t)\hat{T}(1925)} C^{-1}_{T(t)T(t+\Delta t)} C^T_{T(t)\hat{T}(1925)} = 2.92(°)^2$ so $\sigma = 1.71°$.

4. (20 points) On planet X, the Peripheral Sea has two entrances, both of which are deep, flat-bottomed channels that are substantially wider than the local Rossby radius, as shown in the figure. You have collected four hydrographic stations, one on either side of the the channel entrances, at point A, B, C, and D.



How can you estimate the total transport through each of the channels? Set up the matrix equations that you would use to solve for this transport. What additional information would you need to determine your solution?

Model answer: This problem is asking you to set up an inverse problem using 4 stations. To begin setting up the problem, we'll assume that mass, heat and salt are all conserved within the Peripheral Sea, in a broad range of isopycnal layers. For each density class we can define a balance: $(u_{\text{shear}_{AB}} + u_{\text{bot}_{AB}})s_{AB} = (u_{\text{shear}_{CD}} + u_{\text{bot}_{CD}})s_{CD}$ where $u_{\text{shear}_{AB}}$ is the geostrophic shear between stations A and B relative to the reference depth, and $u_{\text{bot}_{AB}}$ is the reference depth velocity, and s_{AB} is the mean concentration of density, potential temperature, or salinity at the station locations in the isopycnal layer. This leads to a matrix equation with two unknowns, representing the inflow and outflow reference velocities, and as many equations as we define refernce levels. Thus:

$$
\begin{bmatrix}
z_{1_{AB}} L_{AB} \rho_{1_{AB}} & z_{1_{CD}} L_{CD} \rho_{1_{CD}} \\
z_{2_{AB}} L_{AB} \rho_{2_{AB}} & z_{2_{CD}} L_{CD} \rho_{2_{CD}} \\
\vdots & \vdots \\
z_{n_{AB}} L_{AB} \rho_{n_{AB}} & z_{n_{CD}} L_{CD} \rho_{n_{CD}} \\
z_{1_{AB}} L_{AB} \theta_{1_{AB}} & z_{1_{CD}} L_{CD} \theta_{1_{CD}} \\
z_{2_{AB}} L_{AB} \theta_{2_{AB}} & z_{2_{CD}} L_{CD} \theta_{2_{CD}} \\
\vdots & \vdots \\
z_{n_{AB}} L_{AB} \theta_{n_{AB}} & z_{n_{CD}} L_{CD} \theta_{n_{CD}} \\
z_{1_{AB}} L_{AB} S_{1_{AB}} & z_{1_{CD}} L_{CD} S_{1_{CD}} \\
z_{2_{AB}} L_{AB} S_{2_{AB}} & z_{2_{CD}} L_{CD} S_{2_{CD}} \\
\vdots & \vdots \\
z_{n_{AB}} L_{AB} S_{n_{AB}} & z_{n_{CD}} L_{CD} S_{n_{CD}}
\end{bmatrix}
\begin{bmatrix}
u_{\text{ref}_{AB}} \\
u_{\text{ref}_{BC}}
\end{bmatrix}
=
\begin{bmatrix}
-u_{1_{AB}} \rho_{1_{AB}} z_{1_{AB}} L_{AB} - u_{1_{CD}} \rho_{1_{CD}} z_{1_{CD}} L_{CD} \\
-u_{2_{AB}} \rho_{2_{AB}} z_{2_{AB}} L_{AB} - u_{2_{CD}} \rho_{2_{CD}} z_{2_{CD}} L_{CD} \\
\vdots \\
-u_{n_{AB}} \rho_{n_{AB}} z_{n_{AB}} L_{AB} - u_{n_{CD}} \rho_{n_{CD}} z_{n_{CD}} L_{CD} \\
-u_{1_{AB}} \theta_{1_{AB}} z_{1_{AB}} L_{AB} - u_{1_{CD}} \theta_{1_{CD}} z_{1_{CD}} L_{CD} \\
-u_{2_{AB}} \theta_{2_{AB}} z_{2_{AB}} L_{AB} - u_{2_{CD}} \theta_{2_{CD}} z_{2_{CD}} L_{CD} \\
\vdots \\
-u_{n_{AB}} \theta_{n_{AB}} z_{n_{AB}} L_{AB} - u_{n_{CD}} \theta_{n_{CD}} z_{n_{CD}} L_{CD} \\
-u_{1_{AB}} S_{1_{AB}} z_{1_{AB}} L_{AB} - u_{1_{CD}} S_{1_{CD}} z_{1_{CD}} L_{CD} \\
-u_{2_{AB}} S_{2_{AB}} z_{2_{AB}} L_{AB} - u_{2_{CD}} S_{2_{CD}} z_{2_{CD}} L_{CD} \\
\vdots \\
-u_{n_{AB}} S_{n_{AB}} z_{n_{AB}} L_{AB} - u_{n_{CD}} S_{n_{CD}} z_{n_{CD}} L_{CD}
\end{bmatrix}
$$

where L_{AB} is the distance between points A and B, and $z_{n_{AB}}$ is the vertical separation between layers. What's missing from this? There's no formal error covariance built into the matrix equation here, but some adjustment should probably be made based on the a priori data uncertainties to help constrain the solution. Error bars should be estimated for the reference velocities. In addition, we haven't included air-sea heat and freshwater fluxes or Ekman transport. With a little a priori information about the flux sizes, it would be nice to build them into the right-hand side. Diapycnal fluxes between layers have also been neglected, but could be built into the left-hand side equations. In this case, it would be nice to have a little a priori notion about the flux sizes that we expect.

Finally, there's a lot of basic information about planet X that we'll need. What's its rotation rate? What's the latitude of the sea? How big is gravity? What are all the dimensions of the system?

5. (30 points) After carefully analyzing numerical model output for the North Pacific, oceanographer X has published a paper stating that fluctuations at 60°N are inversely correlated with fluctuations at 40°N. Moreover, in the model output, these fluctuations appear to be linked to the Pacific Decadal Oscillation (PDO).

You'd like to examine whether the behavior of the ocean resembles the model. Unfortunately, you are stranded on a desert island, and the only data that you have available are irregularly spaced XBT profiles and a time series of monthly PDO variations. The friend who sent you the PDO index rather cryptically remarked that she hoped you had a good singular value decomposition package, since she was sure you'd need either EOFs or a least-squares fitting technique to make sense of this data.

Write a clear discussion, explaining how you could use (a) EOFs and (b) least-squares fitting to analyze your data. Note that there may be more than one approach to this problem (and EOFs and least-squares fitting are not the only ways to treat this data).

For the EOF calculation, how would you actually go about sorting the data into a usable form? What would the elements of your matrix be, how would you handle this matrix, and how would you interpret the results of your analysis? What are the limitations of this analysis?

For the least-squares fitting, how would you organize your data and carry out your calculations? What quantities might you fit to what other quantities? What could you hope to learn from this analysis? What problems might you encounter?

Model answer: This question is asking you to outline a methodology for two separate calculation comparing irregularly spaced XBT data with the Pacific Decadal Oscillation index.

EOFs: Let's start by looking at EOFs. EOFs normally describe the modes of variability of one data field. Here we have two data types, so we might imagine using canonical correlations to describe the joint variability of the XBT observations and the PDO index. However, canonical correlations require that we can construct a covariance matrix relating the two data sets. The PDO is just a vector, so that would limit us to a rather uninteresting covariance vector. (We'd do better if we could go back to the original data used to devise the PDO index, but in this case we're stuck.) So let's compute EOFs for the XBT data and then correlate the EOF modes with the PDO. To do this, first construct a data-data covariance matrix for the XBTs. The PDO is a temporal mode, so we'll want to derive a temporal mode from our EOFs. Since the data are irregularly spaced, we'll construct the covariance matrix based on the available data, and will not try to grid the data. Here's one approach. First, extract temperature data from a single representative depth: perhaps an average temperature between 50 and 100 m. Remove the mean. (You should probably also remove an annual cycle, using a least-squares fit.) Then bin the data in time and space. Now, take time period one and compute the mean covariance of all of the observations with other observations from the same spatial bin (or with themselves). That will be C_{11}. Next compute the covariance of observations at time 2 with observations at time 1 that happen to be collocated. That will be C_{12} and C_{21}. Continue in this way to fill out the covariance matrix, $\mathbf{C}$. Now compute its SVD (or simply find the eigenvectors of $\mathbf{C}$) to obtain the temporal mode EOFs. This will produce $\langle \mathbf{Y}\mathbf{Y}^T \rangle \mathbf{V} = \mathbf{C}\mathbf{V} == \lambda^2 \mathbf{V}$. (The $\mathbf{U}$ and $\mathbf{V}$ matrices are the same, because the covariance matrix is symmetric.) The spatial modes are then of the form $\alpha = v^T Y$, where Y is the data matrix and v the temporal modes (the columns of $\mathbf{V}$. Since the data matrix is gappy, we might not be too happy with this version of the spatial modes. Alternatively, we could compute a separate covariance matrix for the spatial modes.

Here are a few additional considerations. If the XBT data are quite uniform in extent, then it may be realistic to compute a full data matrix in space and time, in which case the time and space modes of the EOFs can be computed directly from the SVD. It's also worth commenting a little on the spatial bins: you might choose to lump all data from each latitude into a single bin. However, this strategy could suppress aspects of the variability that may matter for, so it's probably preferable to sort your data into latitude and longitude bins.

Finally, to answer the question about whether the ocean matches the model we need to look at two things. First, is there a spatial mode in which variability at 40°N is inversely correlated with variability at 60°N? Second, does this mode correlate with the PDO? This analysis will pose several limitations: the spatial coverage of the XBT data may not be adequate to observe the signal that we'd like to see. If there are few observations in a bin, then our covariances may be noisy, which could lead to erroneous or confusing results. Moreover, if the signal were really a propagating mode, the EOF would not be tuned to capture it.

Least-Squares: Least-squares fitting is a very different tool. When we least-squares fit, we impose a functional structure. In this case, for lack of any better idea, let's least-squares fit the XBT data to the PDO. There are a number of ways we could do this, but here's one simple strategy: Sort the XBT data into geographic bins. Each bin will then have a time series of XBT data, and we can again extract a representative value (e.g. mean temperature in the upper 100 m depth). For each bin, least-squares fit a constant, an annual cycle (why not?) and the best fit coefficient to match the PDO. Evaluate locations where the "gain" linking the PDO and XBT data is high. Is there a reversal in the fitted coefficients for the PDO between 60° N and 40°N? Are the correlation coefficients statistically significant? How much of the data variance is explained?

Alternatively, fit all of the data with the PDO times a spatial structure (a plane, perhaps, or a functional fit to the first mode spatial EOF from your previous analysis). This has the advantage of retaining the exact spatial and temporal information in the observations, although there is nothing particularly optimal about the function that we're fitting to the data. Results of such a fit may prove difficult to interpret, however, since we've imposed so much information on the data using the fit.

Finally, Jessica asked, "Where is the desert island?" Here are a couple of possibilities: Monuriki (from the film *Castaway*), Palmyra (not actually a desert, but it belongs to the Nature Conservancy, and has been considered for possible management by Scripps).