Problems Week 3

Due Monday, October 18, 2021

We've talked about least-squares fitting, and we've talked about the χ^2 distribution, so it's time to explore both more closely. Here we'll use wave height, wind speed, and air temperature data from National Data Buoy Center buoy 46047. The data are available here:

https://www.ndbc.noaa.gov/historical_data.shtml

Download the "standard meteorological" data files for buoy 46047 for 2015 and 2016. (If you're curious you can grab the subsequent 4 years as well: 2017, 2018, 2019, and 2020.)

1. **Visual evaluation.** Plot the time series of wind speed, wave height, water temperature, and air temperature from 46047. Since the data are text files with a header, in Matlab you can read with:

   ```
   importdata
   ```

   Read the help page. You'll want to specify a blank space delimiter between files and 2 header lines.

   How are missing data identified in the records?

2. **Monthly means.** Since there's quite a bit of variability, average the data to produce monthly means for 2015 and 2016. Plot the means for each month and standard error of the mean. Data are provided at varying frequencies, but consecutive data are not independent. For the purposes of this problem set, let's assume that the data provide one independent sample every 7 days. Hint: In Matlab, use:

   ```
   errorbar
   ```

   to plot the data with uncertainty ranges.

3. **Least-squares fit.** Treating the two years separately, least-squares fit a mean and an annual cycle to the four data records. What is the mean, and what is the amplitude of the annual cycle? (Total amplitude should be determined from the square root of the sum of the squares of the sine and cosine amplitudes.) Are the fitted coefficients similar for the two years?

4. **Least-squares fit a semi-annual cycle.** Augment your annual cycle least-squares fit with a semi-annual cycle. What is the amplitude of the semi-annual cycle? Does the augmented fit give you a different annual cycle?

5. **χ^2 and the misfit.** What is the squared misfit of your least-squares fits? You can compute this as

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \sum_{j=1}^{M} a_{ij} x_j)^2}{\sigma_i^2} \sim N - M. \qquad (1)$$

In other words, the misfit for each point should be roughly equal to the uncertainty. We lose a degree of freedom for each function that we use to fit. How much does the misfit change if you fit with an annual cycle only or with an annual plus semi-annnual cycle? Use the χ^2 distribution to evaluate whether your fits are improved by adding the semi-annual fit? On the basis of the χ^2 distribution, are you overfitting the data, or choosing the wrong model for your data? How would your results change if you assumed that you had one independent sample (i.e. one degree of freedom) per day, instead of one per week?