ATOC5860 - Homework 2 - due Thursday February 10, 2022

Please send your homework to Jen/Prof. Kay on Slack as a direct message.

Please Name Your Homework Files: "ATOC5860_HW2_LastName.pdf, .html, .ipynb"

Your submissions should include: 1) A .pdf document with responses to the questions below, 2) Your code in both .ipynb and .html format

Show all work including the equations used (e.g., by referring to the Barnes Notes).
Write in complete, clear, and concise sentences.
Eliminate spelling/grammar mistakes.
Label all graph axes. Include units.

PICK EITHER PROBLEM #1 or PROBLEM #2 TO COMPLETE. ***ONLY DO ONE.***

1) In this problem, you will assess the influence of autocorrelation on basic statistics. Specifically, you will use Monte Carlo techniques to explore how autocorrelation in an AR1 time series influences estimates of the sample mean and sample standard deviation. (50 points total)

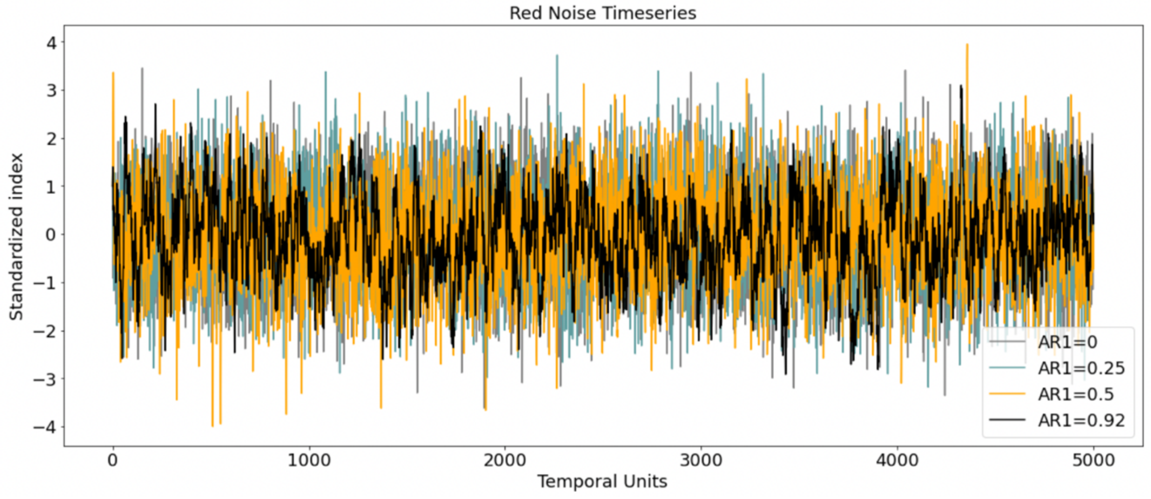
Generate four standardized red noise time series of length 5000. *Remember: "standardized" data have a mean of zero and a standard deviation of 1.* Time series one (T_1) should have a lag-1 autocorrelation ρ(1)=0; T_2 should have a lag-1 autocorrelation ρ(1)=0.25; T_3 should have a lag-1 autocorrelation ρ(1)=0.5, and T_4 should have a lag-1 autocorrelation ρ(1)=0.92. *For reference: The time series of yearly surface temperature anomalies over the southwest United States has ρ(1)=0.25. The monthly Cold Tongue index (a measure of ENSO) has ρ(1)=0.92.*

a) Plot time series for T_1, T_2, T_3, and T_4. (10 points)

Using Barnes Chpt. 2 Eq (69), we generate timeseries from:

$x(t) = ax(t-\Delta t) + b\epsilon(t)$

Here, a is the desired autocorrelation between timeseries values at $x(t)$ and $x(t-\Delta t)$. So, increasing a increases the amount of memory between two timestamp values and makes the timeseries more "red". $b\epsilon(t)$ is a random variable drawn from the standardized dataset to represent random noise.

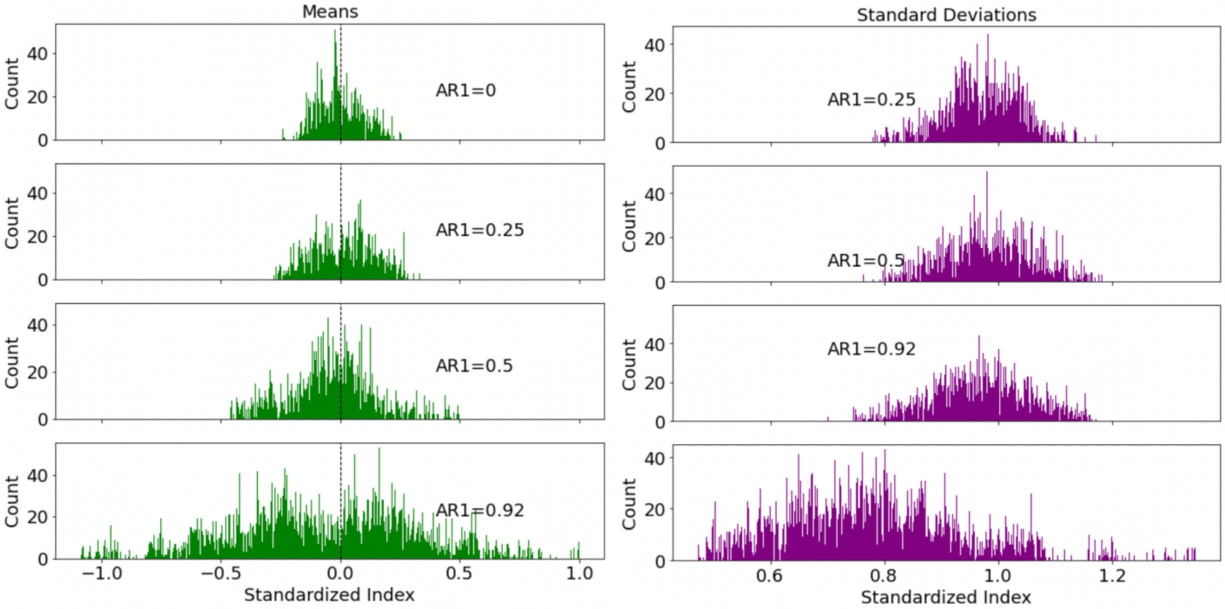


b) For each time series – Draw a random sample consisting of N=100 consecutive elements from the full time series. Note that you must use consecutive values since you are analyzing the persistence. Calculate the mean and standard deviation of the sample, and save your value. Repeat Monte Carlo style until you have a good distribution (i.e. do it 10,000 times). Look at the influence of persistence on your estimate of the mean and the standard deviation based on samples of size N=100. Plot histograms of your estimated sample means and standard deviations for T1, T2, T3, and T4. (10 points)

Here we calculate the mean and standard deviation from Barnes Chpt. 1 Eq. (5,6) as:

$$\bar{x} = \frac{1}{N}\sum x_i$$

$$x'^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2$$



c) Estimate how the number independent samples (N*) you have for T1, T2, T3, and T4 using Barnes Chapter 2 Equation (88). (15 points)

We use Barnes Chpt. 2 Eq. (88) to calculate the effective sample sizes where:

$$\boldsymbol{N^*} = \boldsymbol{N}\frac{\boldsymbol{1 - \rho(\Delta t)}}{\boldsymbol{1 + \rho(\Delta t)}}$$

Where $\boldsymbol{N^*}$ is the effective sample size, $\boldsymbol{N}$ is the population size, and $\boldsymbol{\rho(\Delta t)}$ is the autocorrelation at a lag of $\Delta \boldsymbol{t} = \boldsymbol{1}$. From this, we estimate that when the autocorrelation is 0, the effective number of samples remains at 5000. With an autocorrelation of 0.25, the effective sample size reduces to 3000. With an autocorrelation of 0.5, the effective sample size is 1667. Finally, with an autocorrelation of 0.92, the effective sample size reduces to 208.

d) Discuss your results in 1-2 paragraphs. Since you created red-noise time series, the population means are by definition equal to zero (μ=0) and the population standard deviations are by definition equal to one (σ=1). How does increasing the redness of a time series affect sample estimates of the mean and standard deviation? What are the broader implications for data analysis of time series? (15 points)

From the plot in part 1a, it is apparent that increasing the autocorrelation creates a “smoother” timeseries. For instance, values along the black curve with $\boldsymbol{\rho(1) = 0.92}$, remain large or small for a longer period than those on the gray curve with $\boldsymbol{\rho(1) = 0}$ that may oscillate between small and large values more frequently. This occurs because a higher autocorrelation increases the amount of memory, or red noise, between two timestamp values. So, it becomes less likely that a value can be significantly different from its preceding value as autocorrelation increases. Although the temporal variability becomes weaker as autocorrelation increases, the variance remains nearly the same.

Strong memory causes *individual* sample means to be higher or lower than 0 more often. This is why the histogram spread of means is larger in part 1b when the autocorrelation increases. For example, in the white noise sample, the spread of means is closer to 0 since many large and small, positive and negative means cancel each other out. When autocorrelation increases, larger (smaller) sample values will have larger (smaller) successive values, which increases (decreases) that individual sample’s mean. The standard deviations of samples with more red noise are typically lower than the standard deviations with more white noise. Because the samples with higher autocorrelations can’t change significantly as quickly, a greater range of values can be obtained in a given sample.

2) In this problem you will assess the relationship between two variables: Global mean temperature and ENSO. Use the two provided datasets that run from 1881 through 2017: 1) Global surface temperature anomalies from the GISTEMP dataset, 2) the previous December's ENSO Nino 3.4 Index. Data are in a comma-delimited text file available on the class google drive: homework2_problem2_data.csv. (50 points)

If at a later date you are looking at this assignment and you want to update this analysis to include more years – the original data are here:
GISTEMP data here: https://data.giss.nasa.gov/gistemp/
NINO data here: https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Nino34/

a) First, detrend your data. Next, standardize your data by subtracting the mean and dividing the standard deviation. *Note: Your data should both have a mean of 0 and a standard deviation of 1. After this, your data should both have a mean of 0 and a standard deviation of 1. Check that this is true before you proceed.* Then -- look at your data by making two plots. Plot #1: a time-series plot of standardized detrended global-mean temperature and standardized detrended ENSO index. Plot #2 scatter plot for the standardized detrended values of global-mean temperature (y-axis) versus the standardized detrended ENSO index (x-axis). (10 points)

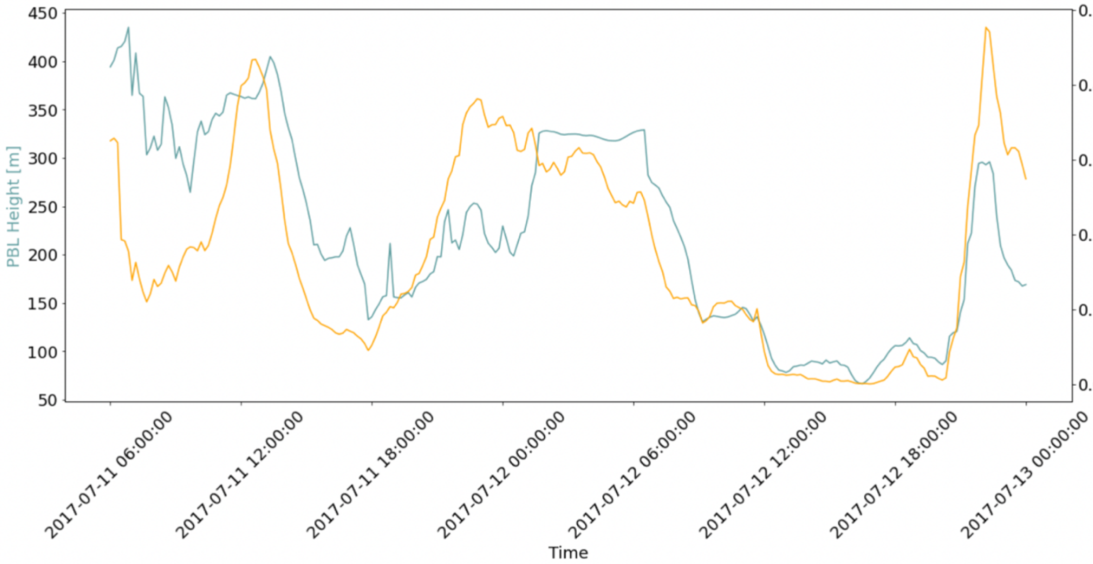
b) Calculate the correlation and regression coefficients between the standardized detrended global mean temperature and the standardized detrended ENSO index. The regression slope should be in units of standard deviation of global mean temperature per standard deviation of ENSO. Calculate the fraction of the variance in global mean temperature explained by ENSO variability. Is the correlation between these two variables statistically significant (i.e., different than 0) at the 95% confidence level? Assume 1 degree of freedom per year. (15 points)

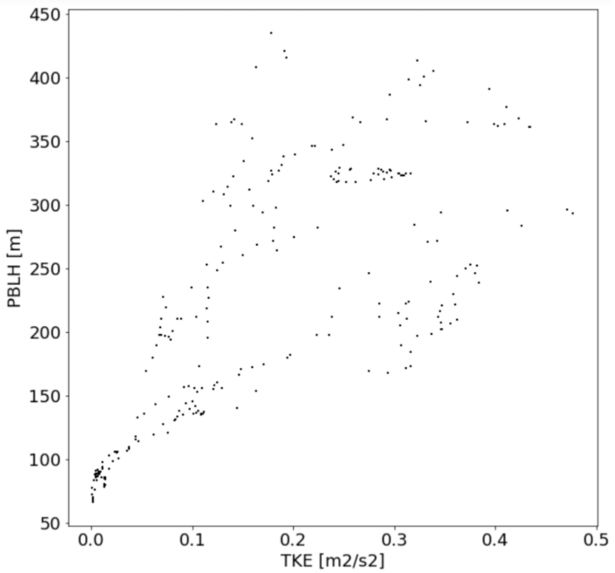
c) Using composite analysis, calculate the standardized detrended global mean temperature for years when the ENSO 3.4 anomaly exceeds +1 and -1 degree Celsius, respectively. Assess if the composite means are different at the 95% confidence level. *Hint: see example 1.4.2.2 comparison of two sample means problem in the Barnes lecture notes Chapter 1.* (10 points)

d) Discuss your results in 1-2 paragraphs. Discuss the similarities and differences between the results from the composite analysis and the regression/correlation analysis. Explain why they don't provide identical results. (15 points)

3) Select two variables (variable A, variable B) you are using in your research that you think might be related. Use what we have been learning in ATOC7500 to analyze them and assess if one can be used to predict the other using linear regression techniques. In other words – can you use variable X to predict variable Y using linear regression? (50 points)

a) **Plot your raw data as a line plot (i.e., as a function of the sampling direction, commonly time) and as a scatter plot. Explain why you decided which variable would be X and which would be Y. Explain what you see visually and if this gives you hope or not that you might be able to predict Y using X. (5 points)**



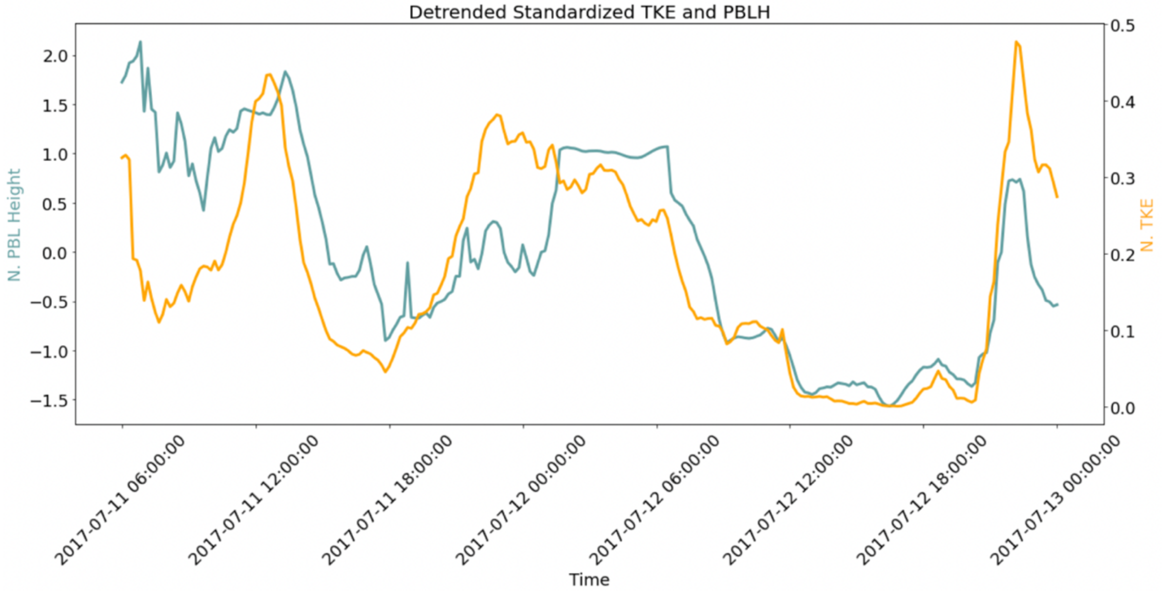


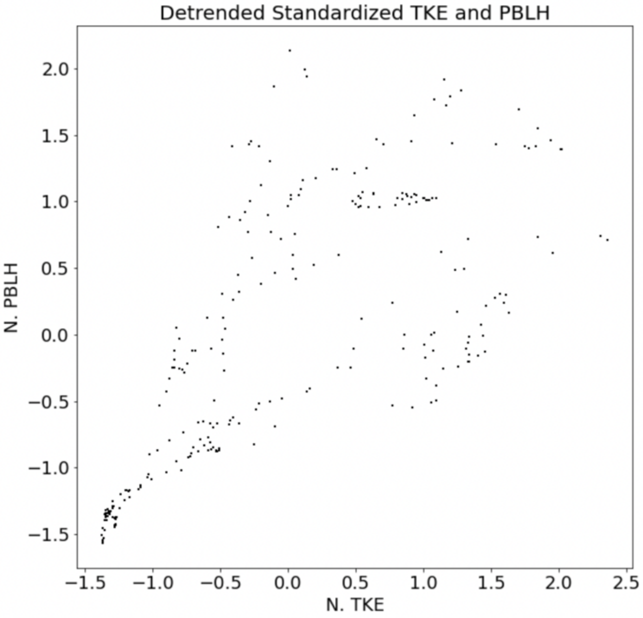
The plots below show planetary boundary layer height (PBLH) and turbulent kinetic energy (TKE) over time at a height of 138 meters. Each value is obtained from the same grid cell for the period 11-13 July, 2017 within a Weather Research and Forecasting (WRF) model run. The first 6 hours are removed as spinup since the model hasn't reached equilibrium yet.

Here, turbulent kinetic energy (TKE) is the independent variable X, and the planetary boundary layer height (PBLH) is the dependent variable Y. TKE is affected by the strength of buoyant eddies and by wind speed where mechanical shear creates turbulence. This turbulence induces mixing within the PBL. Greater amounts of mixing cause the PBL height to increase and vice versa. So, the TKE strength may directly affect PBLH.

Overall, I see that both timeseries tend to increase and decrease with each other. I also see that PBL height tends to temporally follow an increase or decrease of TKE. For instance, at 2017-07-11 12:00, TKE first increases and is followed by an increase in the PBL height a few timestamps after. This leads me to believe that TKE can be used to predict PBLH.

b) Standardize your data: subtract the mean and divide by the standard deviation. In other words, ensure that both variables have a mean of 0 and a standard deviation of 1. Also de-trend if needed. Plot your standardized data as an x-y plot and as a scatter plot. (5 points)





We standardize using Barnes Eq. 71: $z = \frac{x-\mu}{\sigma}$ where x is an independent value, μ is the population mean, and σ is the population standard deviation.

Standardized PBLH population mean: = 0

Standardized PBLH population standard deviation: 0.7

Standardized TKE population mean: $= 0$

Standardized TKE population standard deviation: 0.96

c) **Calculate the autocorrelation and estimate the number of independent samples for both of your variables. (10 points)**
We calculate the autocorrelation from Barnes Chpt. 2 Eq. 67:
$\gamma(\tau) = \frac{x'(t) \cdot x'(t+\tau)}{(N-\tau)\gamma(0)}$ where primes indicate perturbations from the mean, N is the timeseries length, τ is the lag, the dot product incorporates the sum, and the resulting value is normalized by $\gamma(0) = \overline{x'^2}$, the variance of the population without lag.

We calculate the effective sample size from Barnes Chpt. 2 Eq. 88:
$\boldsymbol{N}^* = \boldsymbol{N}\frac{\boldsymbol{1-\rho(\Delta t)}}{\boldsymbol{1+\rho(\Delta t)}}$ Where $\boldsymbol{N}^*$ is the effective sample size, $\boldsymbol{N}$ is the population size, and $\boldsymbol{\rho(\Delta t)}$ is the autocorrelation at a lag of $\Delta \boldsymbol{t} = \boldsymbol{1}$. The code uses three methods to determine the effective sample size but one is listed here for brevity.

The autocorrelation for the planetary boundary layer height timeseries is 0.985 with an effective sample size of 2.
The autocorrelation for the turbulent kinetic energy timeseries is 0.989 with an effective sample size of 1.

d) **Complete a regression between two variables. Assess the statistical significance taking into account the number of independent samples. Place confidence intervals on your regression coefficient. Provide any other statistical results that are helping you interpret your data. Provide a plot or two showing your results. (10 points)**

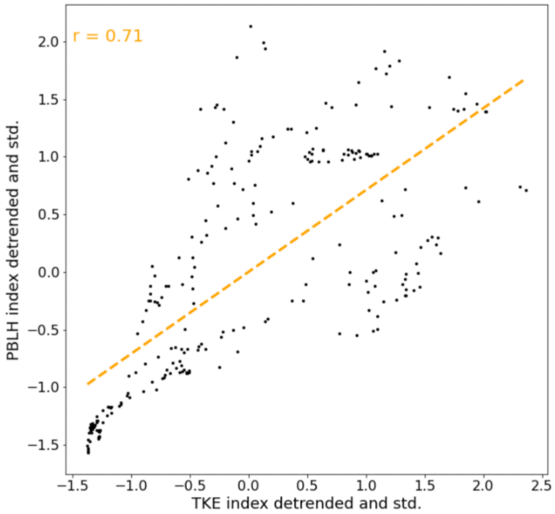
1) Calculate the regression coefficient using Barnes Chp. 2, Eq. 14
$$a1 = \text{slope} = \frac{\overline{x'y'}}{\overline{x'^2}} = 0.71$$
The y-intercept is 0. The correlation coefficient is also 0.71 when using the full timeseries. This follows that $a1 = r\frac{s_1}{s_2}$ where s_1 and s_2 equal 1 when datasets are standardized. Unfortunately, the correct correlation coefficient cannot be calculated using the effective sample size of 1. I understand that in other cases, I would use the effective sample size moving forward.
2) Calculate the confidence interval
We choose a confidence level of 95% so $\alpha = 0.05$. The null hypothesis is that the datasets do correlate. We evaluate using the Fisher-Z score with a two-sided approach. Both datasets have lengths of 253, and thus we may assume Normality from the central limit theorum. Using a 95% confidence interval, the critical t-score is 1.96. To reject the null hypothesis, the confidence interval must not contain $\rho = 0$.
Using Barnes Chp. 2 Eq. 52:
$$Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$
$$= 0.889$$
Using Barnes Chpt. 2 Eq. 53:
$\sigma_Z = \frac{1}{\sqrt{N-3}}$ We cannot use the effective sample size here since it would create an imaginary number, so, the full sample size is used. Again, I understand that I should use the effective sample size when possible to remove persistence.
From Barnes Chp. 2 Eq. 48

$Z - t_{crit}\sigma_z \leq \mu \leq Z + t_{crit}\sigma_Z$
$0.76 \leq \mu_z \leq 1.01$
From Barnes Chp. 2 Eq. 55
$\rho = \tanh(\mu_Z)$
So, $0.64 \leq \rho \leq 0.77$
Our calculated $\rho = r = 0.71$ lies within the 95% confidence interval. The 95% confidence interval does not contain $\rho = 0$ so we cannot reject the null hypothesis that the datasets correlate.

To check this, we evaluate again using the t-score with a two-sided approach. The t-score is used since it converges to the z-score at large N. We again choose a 95% confidence interval so $\alpha = 0.05$. The datasets are Normally distributed and the critical t-score is 1.96. Our null hypothesis is now that the datasets do not correlate. In order to reject the null hypothesis, the t-score much exceed the critical z-score. From Barnes Chpt. 2 Eq. 45:

$$t = \frac{r\sqrt{N-2}}{\sqrt{(1-r^2)}} = 16$$

Since $t > t_{crit}$, we can reject the null hypothesis that the datasets do not correlate. This result is consistent with the Fisher-Z approach.



e) Apply Granger Causality. Can X help you predict Y more than Y itself? (10 points)

Here, we use lags from 1 to 6 timesteps. This data is at a 10-min output frequency and turbulent eddies typically find maximum time scales of 1 hour. Thus, 6 timestamps correspond to the point at which new new eddies are generated.

In theory, we calculate the correlation of Y (PBL height) with itself at all lags from Barnes Chpt. 2 Eq. 119:
$y_t = a_0 + \sum a_\tau y_{t-\tau}$

In practice, we calculate the autocorrelation at each lag from Barnes Chpt. 2 Eq. 67:
$\gamma(\tau) = \frac{x'(t) \cdot x'(t+\tau)}{(N-\tau)\gamma(0)}$ where primes indicate perturbations from the mean, N is the timeseries length, and τ is the lag, the dot product incorporates the sum in the equation, and the value is normalized by $\gamma(0) = \overline{x'^2}$, the variance of the population without lag.

From this, the maximum autocorrelation of PBL height is 0.985. This corresponds to a variance explained, or Pearson's correlation coefficient r^2 of 97.14%.

Next, we use Barnes Chpt. 2 Eq. 121 to find the correlation between the two datasets at each lag value:
$y_t = b_0 + \sum b_\tau y_{t-\tau} + \sum c_\tau x_{t-\tau}$ where $c_\tau x_{t-\tau}$ is the addition of correlation using variable X (TKE) at each lag value.

From this the maximum variance explained has increased to 97.65%. Thus, we have obtained a small amount of additional unique information by adding TKE amount.

f) Discuss what have you learned. Can you use X to predict Y using linear regression? Are your results statistically significant? If Yes – can you explain the underlying physical reasons for a potential correlation between the two variables? If No – why? What are the next steps if you wanted to take this analysis further? Please explain in a paragraph or two. (10 points)

Finally, we perform an F-test to determine if the additional variance explained is statistically significant. We choose a 95% confidence interval so $\alpha = 0.05$. The null hypothesis is that the addition of TKE has not changed the amount of variance explained. We will reject the null hypothesis if the f-score is greater than the critical f-score of 1.23.
From Barnes Chpt1. Eq. 122:
$F = \frac{s_1^2}{s_2^2} = 1$. We may use this version of the equation since the standardized populations both have standard deviations of 1.
Because the f-score is less than the critical f-score, we cannot reject the null hypothesis that adding TKE provides significant added benefit.

From this study, we have learned that we can use turbulence strength to predict the planetary boundary layer height using linear regression. This is because added turbulence provides enhanced mixing in the atmosphere that raises the boundary layer height. For example, if the atmosphere is stably stratified, mixing creates isothermal temperatures which lift the nocturnal capping inversion and raise the boundary layer. However, while we receive added benefit from correlating lagged versions of PBLH with TKE, the effect is not significant. This is likely due to the PBL timeseries being extremely "red" with lots of memory. Thus, a lag of 1 timestamp has a high autocorrelation, making PBL height easy to predict using lagged versions of itself.

To take this analysis further, I would employ compositing. Namely, this region has land to the west and open ocean to the east. So, wind directions coming from land may introduce mechanical turbulence from a rough coast and buoyant turbulence produced by daytime surface heating. This turbulence may introduce enough mixing to affect the boundary layer height. However, wind directions coming from the east come from the open ocean. There is little mechanical turbulence since the ocean has a small roughness length and little buoyant turbulence since the sea surface can't heat up quickly from its high heat capacity. So, we may find that when winds come from the open ocean, turbulence is no longer a good predictor of the PBL height and there is no longer extra variance explained.

4) Future homework assignments will continue to require that you analyze your own data. Identify a dataset you wish to use and describe it here briefly. The dataset should have at least two dimensions (e.g., time&space, size distribution&mass). You will apply EOF/PCA analysis

to this dataset. Please discuss with the professor if you do not have a dataset in mind. (0 points)
In the future, I can use a simulation I ran offshore of our U.S. Outer Continental Shelf. This dataset has spatial dimensions [466,259,54] [west-east,south-north,heights] and has 289 times. Reading this in is a bit memory intensive and can cause crashes. So, I would want to subset it to a single vertical level. This dataset has turbines in it, so there is a reduction of wind speed and a source of turbulence in the grid cells that contain turbines.