Laboratory for Atmospheric and Space Physics
University of Colorado **Boulder**

Data Systems / Science Data Processing
Scientific Programmer / Software Engineer
Interview Coding Challenge / Report from Alea Kootz
Revised Oct 2023

Thank you for this opportunity to demonstrate my skills and experience.

Table of Contents

Approach

I have chosen to do the entirety of the data analysis in a Jupyter notebook, as this is an unfamiliar dataset to me, and the interactivity of small re-runnable blocks is far more useful for failing quickly and learning, than a monolithic file. I intend to containerize the notebook in a Jupyter project provided image, specifically their jupyter/datascience-notebook image. With that image as the base, I'll use the Dockerfile to install a pair of packages that it lacks which I have found useful in data analysis contexts xarray, and scikit -image. The Jupyter notebook for this project will be hosted in a GitHub repository, which allows me to git clone the repository from the Dockerfile and simplifies the process for anyone else needing to recreate the container since they would not need to have any files other than the Dockerfile (and potentially the input data files) locally available.

Using the Docker Image

This assumes that the user has docker installed and a way run images with a bind mount.

1) from the directory containing the Dockerfile, on Mac/Linux run 'docker build --no-cache -t <your_chosen_container_name_here> -f alea_kootz_dockerfile .'
2) run the container and bind mount the local directory containing the data files to '/home/jovyan/work/lasp_dsse_hw/data_external' which is the directory path that the Jupyter notebook will find the input files, and where the output files will be written.
3) the container should run for about 20 seconds after which the container will automatically stop, and the output data files should be in the bind mounted directory.

Report of Results

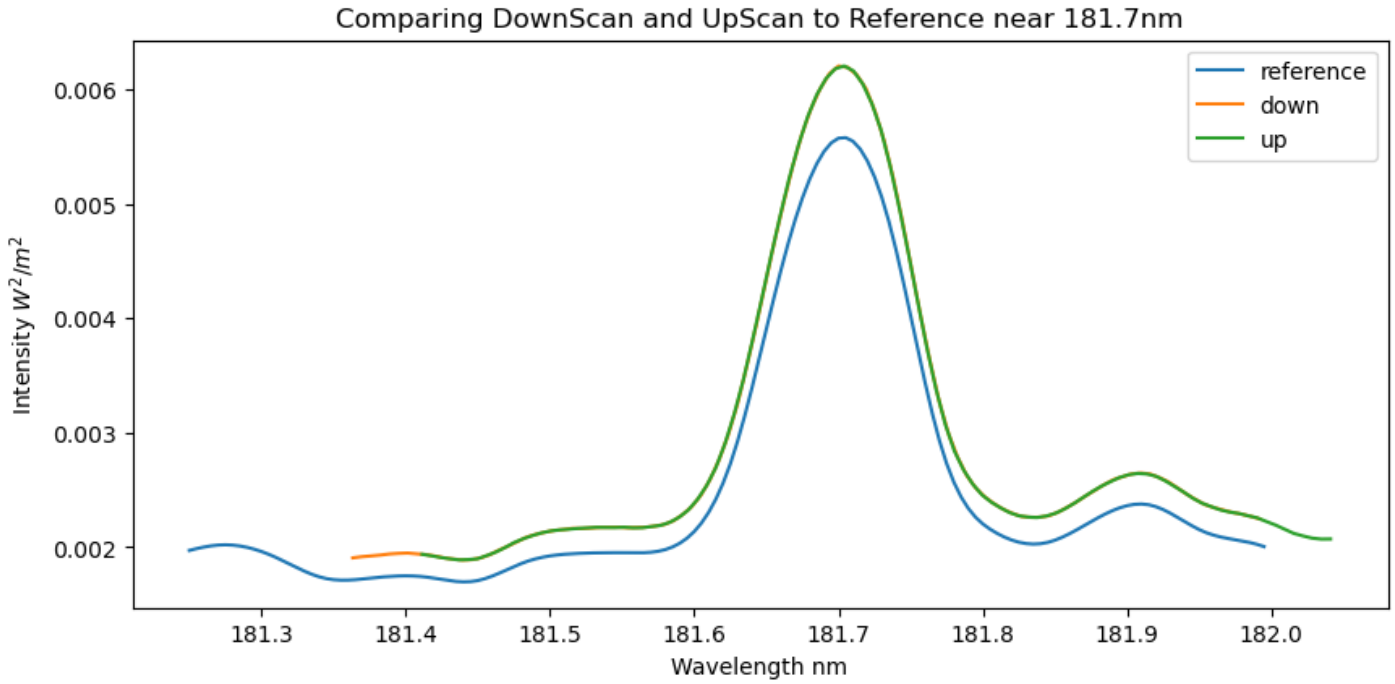
These results were obtained by following the instructions provided in the SDP Coding Challenge document with a a few changes to the suggested operations and assumptions. These are:

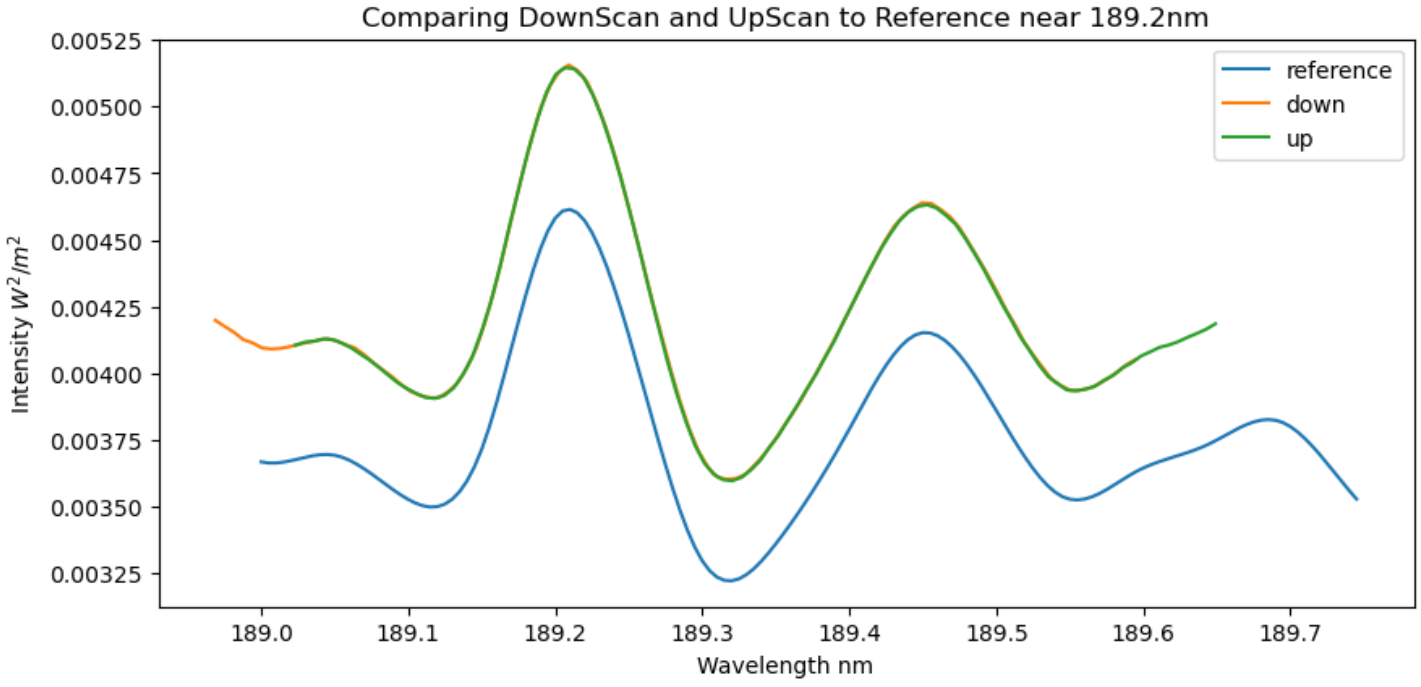
A) Dark subtraction, because the dark noise is thermally sourced for the single photo multiplier tube, instead of taking a median value from a full thermal cycle of the orbit, I applied a linear fit of the sensor temperature to the dark noise, and subtracted that fit, which improved the noise reduction of each measurement by a ratio of (1e-6 at the start and 1e-7 at the end) to the result data.
B) Order of operations for accounting for the change in sensor sensitivity due to temperature. Because I altered the way the darks were subtracted, I was confident that the resultant photon counts measured should be adjusted after the darks were removed, not before.
C) Doppler shifts in wavelength, because I had the data for it, and the math to implement it was straightforward enough that I felt it was definitely worth the effort to show improvements in the accuracy of the x axis of the result.
D) Doppler shifts in intensity, measured photons per second from the spacecraft's velocity toward or away from the sun, as the peak velocity of 1e-5 of the speed of light would increase and decrease the photon count rate by an equivalent amount. This actually decreased the agreement of the results, so I was tempted to omit it.
E) Planck constant, there were three values for the Planck constant from the SI standards during the operational life of the spacecraft. I chose to use the latest value, which has an 8.75e-8 difference from the previous value given in the documentation. I internally debated using the value given in order to show that my work is more aligned to the instructions I am given, but I ultimately decided that I would be happier being more accurate even if that was a deviation from the given value.

Those notes made, the agreement between the DownScan and UpScan spectral intensities is nearly prefect with the peak value comparison at 181.7nm being 3.30e-5. This is within the error of the spectral sampling bin width vs the width of the spectral peak. Further analysis of their agreement would require a polynomial fit of the spectral peak values to determine the true peak center value. There is no obvious peak to examine and compare for the low wavelength end of the spectrum so I omit any comparison on that side.

This intensity agreement extends to the high wavelength end of the data, where the intensity is quite high, and the DownScan had a very warm sensor, and the UpScan sensor is relatively cold, where the agreement at 189.2nm is 1.99e-3.

Agreement between the two scans established, let's cover their agreement with the provided reference spectrum. At the spectral peak near 181.7nm the sampled data is 8.24e-2 greater than reference. At 189.2nm the DownScan is 8.76e-2 and the UpScan is 8.55e-2 greater than reference. Again without a clear peak in the low wavelength range, I will omit a comparison there.





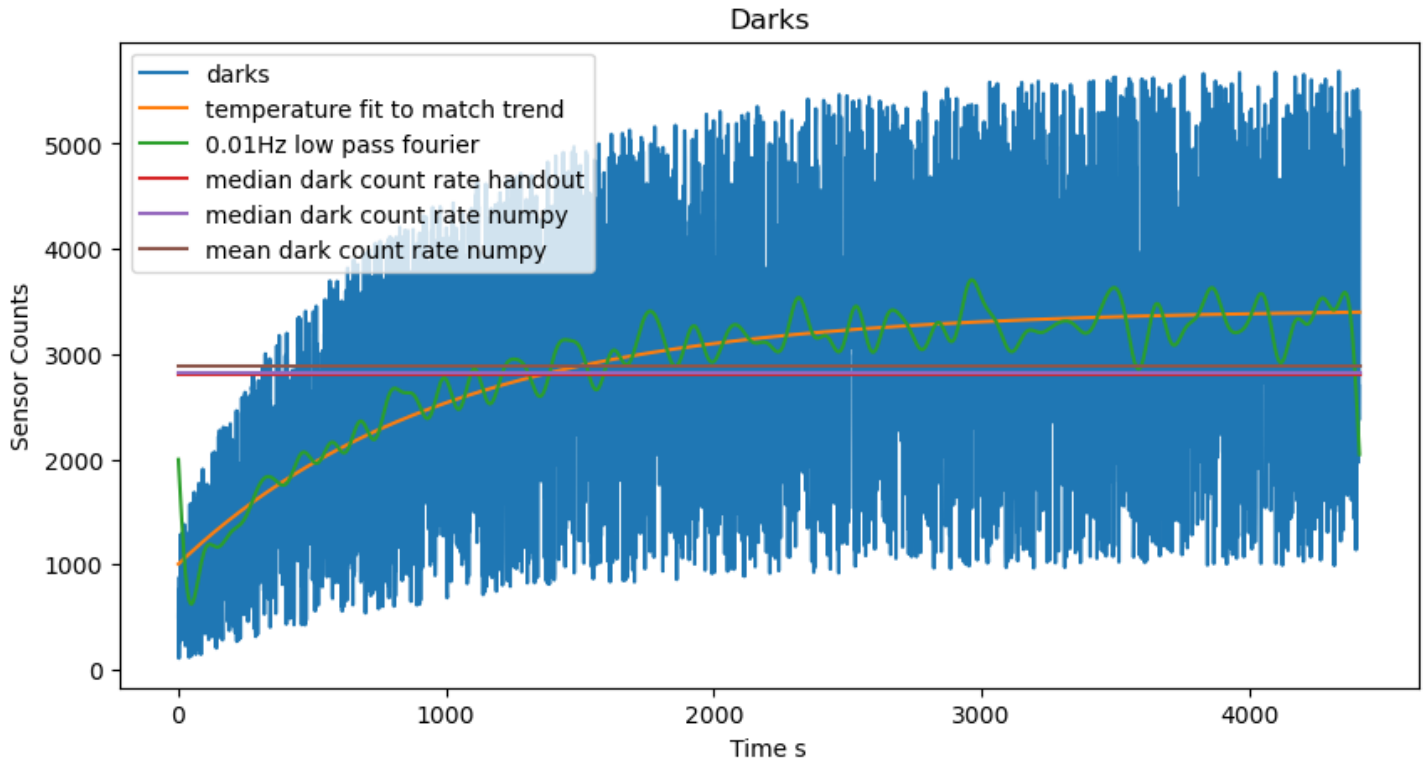
Technical Discussion

Understanding the Datasets

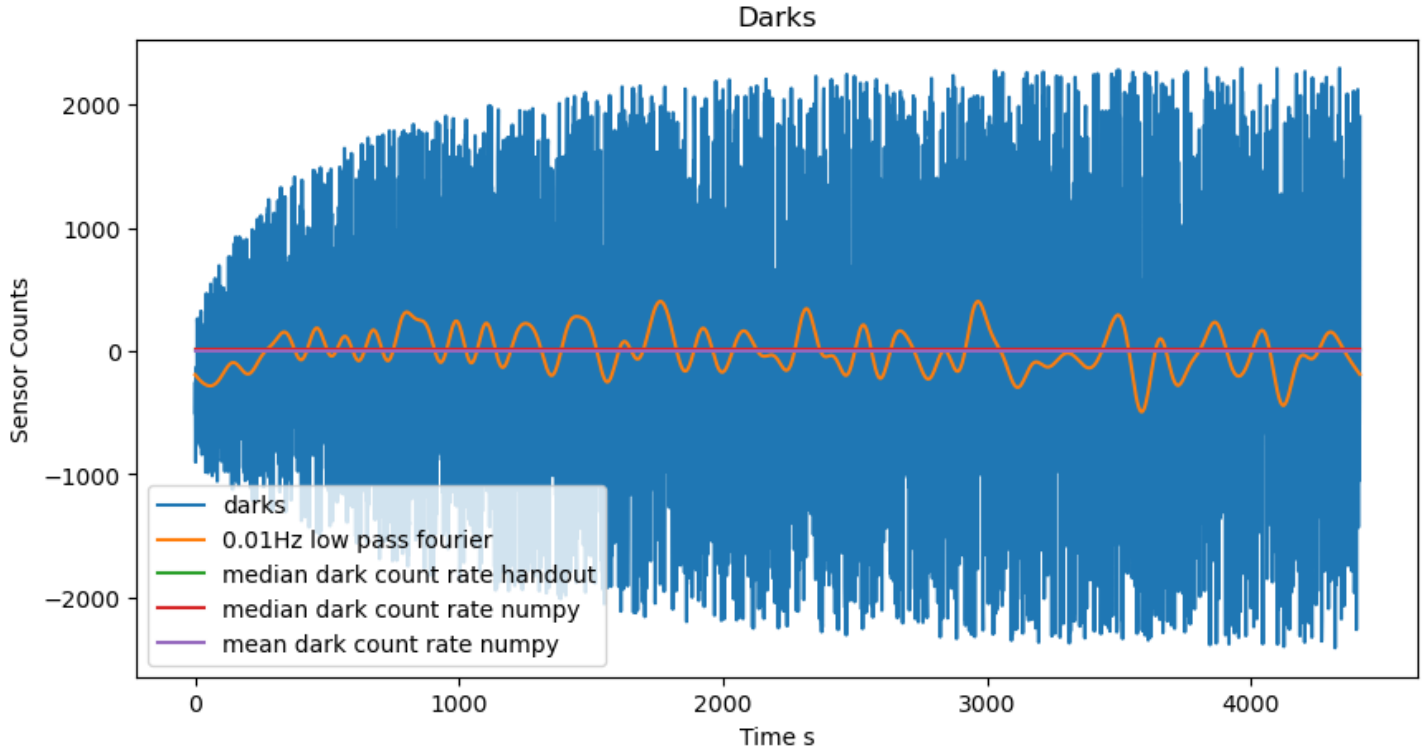
My first step was to import the all the data and convert the time units to seconds. Only barely followed by plotting the data from between the DownScan Dark and UpScan timestamps from the plans.txt file. This showed me what the DownScan and UpScan names meant, and the overall shape of the dark exposure I would need to remove from the data.

Dark Removal

With the goal of removing the noise from sensor heat as early on as possible, I chose that as the second step. I spent a few minutes plotting and comparing the heat of the sensor between passes, to understand what was driving the thermal noise. This lead me to make the decision to change the way the dark would be subtracted from the provided method of taking a median value for the observation period, to subtracting a fitted heat curve. I used SciPy's minimize() function to determine the best fitting linear weights of the heat curve for the dark exposure to the dark data, and then used those weights for the heat curves of the DownScan and UpScan to remove their darks. This included compensation for the different exposure times between the two scans and the dark. The following figure shows a few comparisons of the possible ways to fit value or curve to the darks.



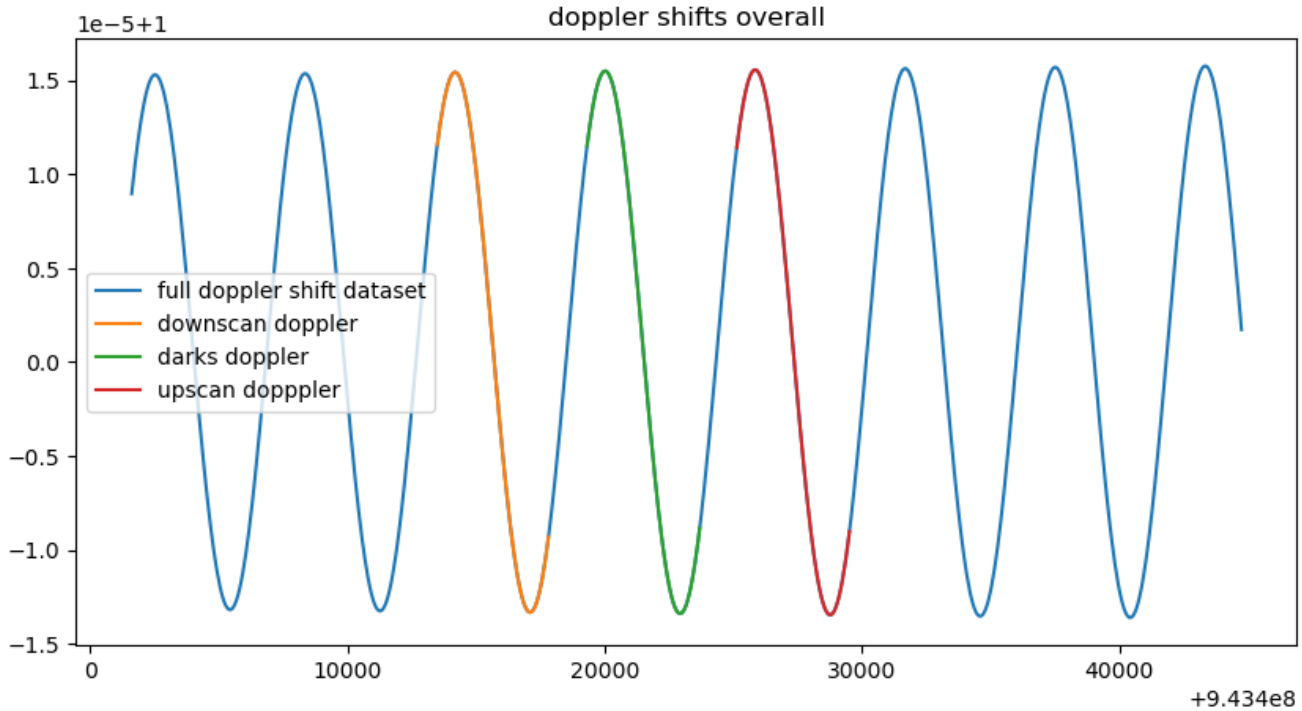
When the fitted temperature curve is removed from the Dark data, we can plot the results of running the same other methods of compensating for the darks on the corrected values. And we find that the median and mean methods agree that there is no compensation that they would provide against the corrected Dark data.



There is a slightly suboptimal behavior at the very cold side of the dark, this is likely due to a nonlinearity in the sensor's response there, but our fit is a simple linear scale of the hear curve so we cannot compensate for it.

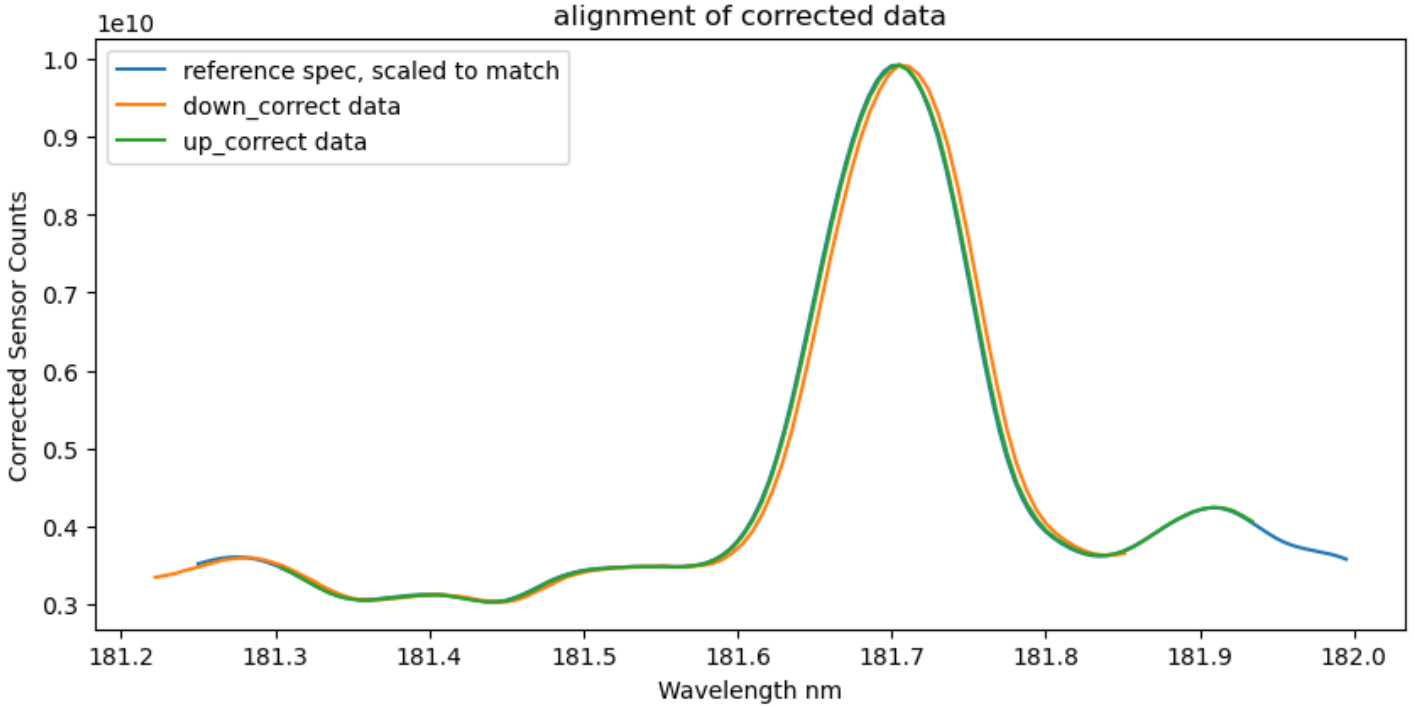
Doppler Shifts

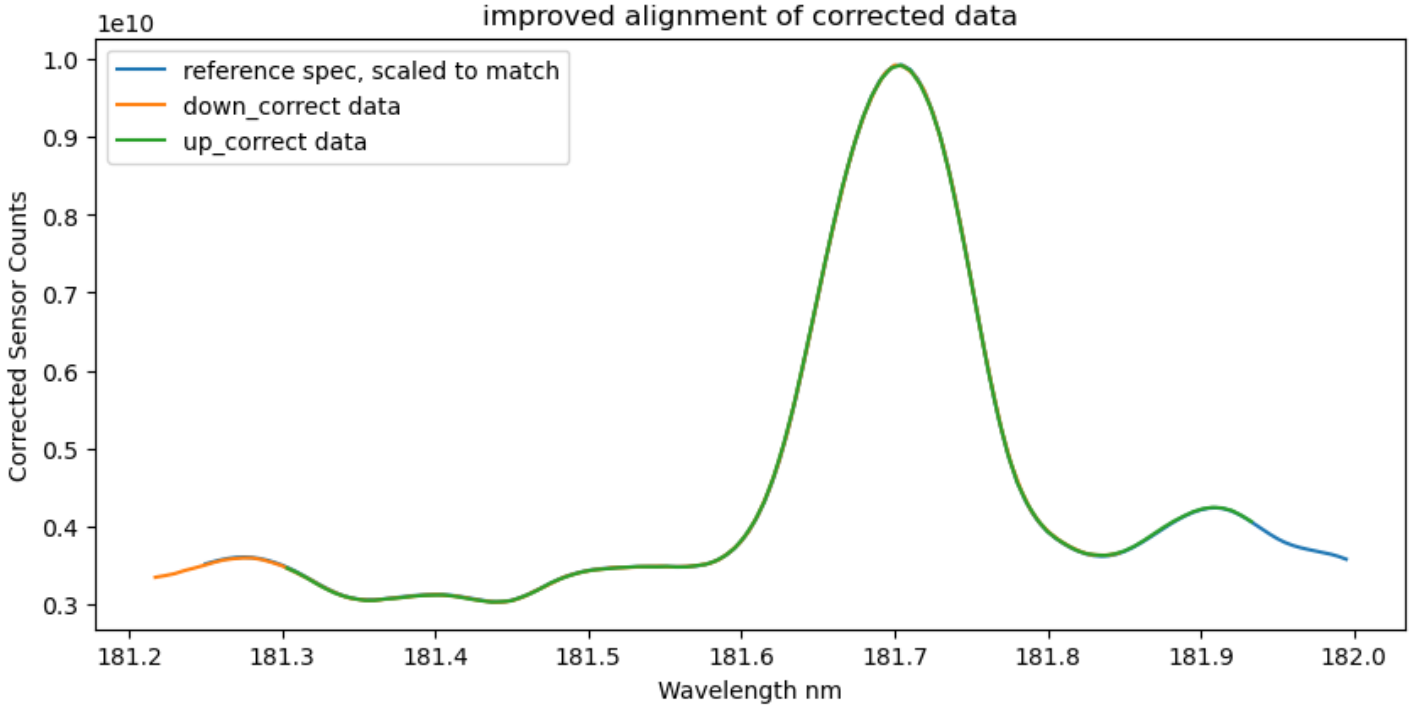
The provided problem prompt did say we could safely ignore it. And yet, the data to compensate for it was included, and the correction available was on the order of 1.5e-5, so I did my best to remove the effect of the spacecraft's velocity on both the wavelength and measured intensity.



Spectral Alignment

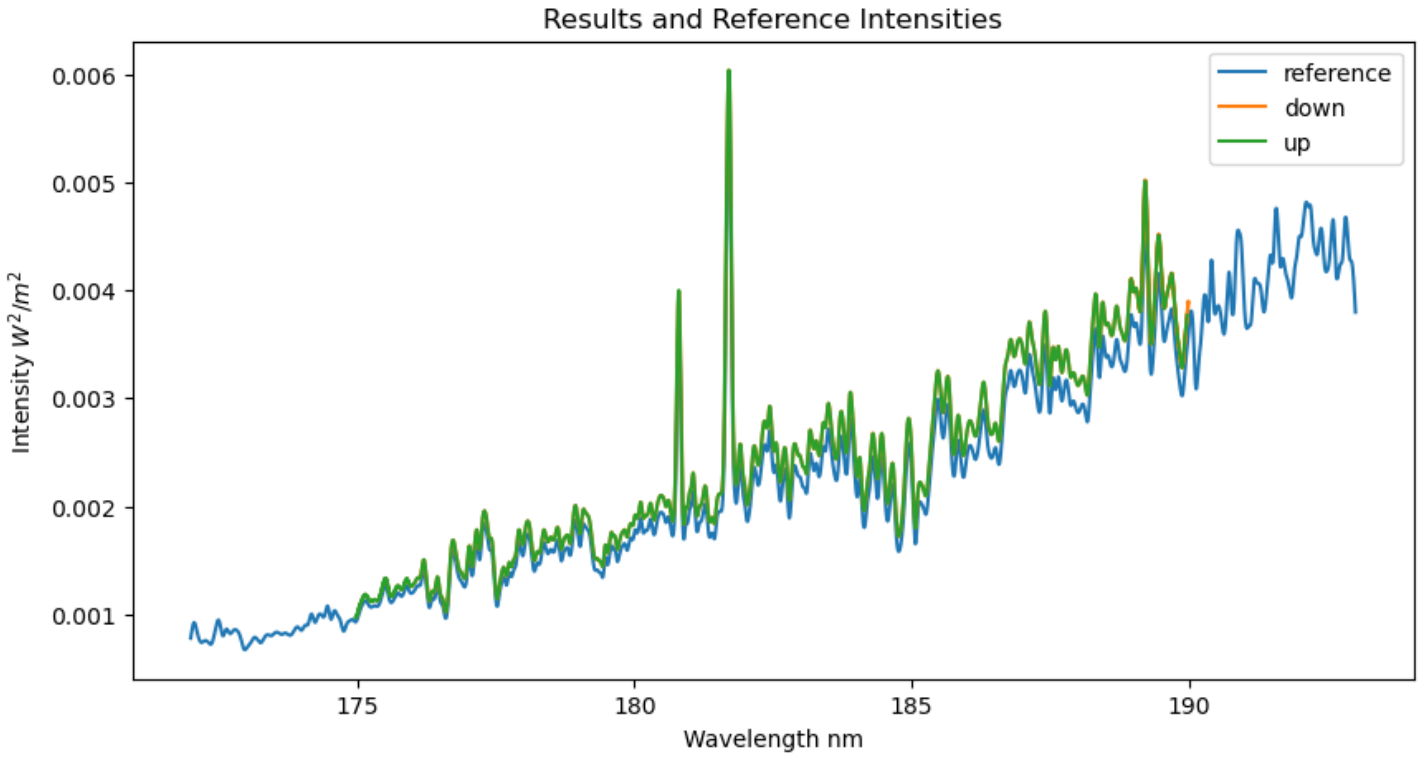
The naive alignment algorithm I implemented was to simply take the wavelength corresponding to the highest recorded values of each dataset near the spectral peak of 181.705nm and adjust the spectrums of both the scans to match that value. This capped the wavelength error at less than 6.33e-3nm. After plotting the resultant curves near the peak though, I chose to add small offsets by hand to slightly improve alignment.





Calculating Intensities

When calculating the intensity measured, something felt off. I usually memorize the more common derived physical constants that change over time as more precision is obtained in determining them, and the Planck constant didn't match what I thought I remembered. I had used the Planck constant quite a lot in the last few years working on DKIST, and so I double checked it. I used the newer value when I did the conversion.



The adjustment for distance to the sun was also done at this point, as suggested in the prompt.

Personal Note

Thank you for reading this, though I'm not sure that this fits within the bounds of a "brief report".

I leaned into discussing my thoughts and showing personality more than I would on a report intended for a presentation, as I intuit that this is one of the ways you'll learn about my work approach and the way that I think, so that you can determine if I'd be a good fit in the team.

I also left a lot of commented out code fragments, so you have the opportunity to see some of the things I tried but didn't use. This will be visible in the pdf of the completed jupyter notebook that will be output along with the results data files.