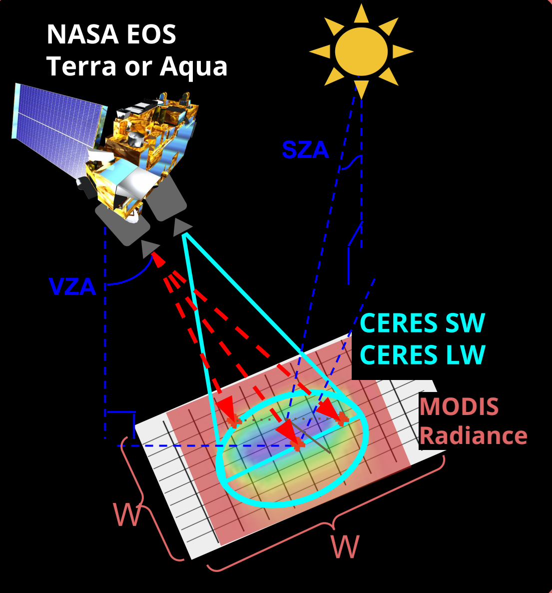
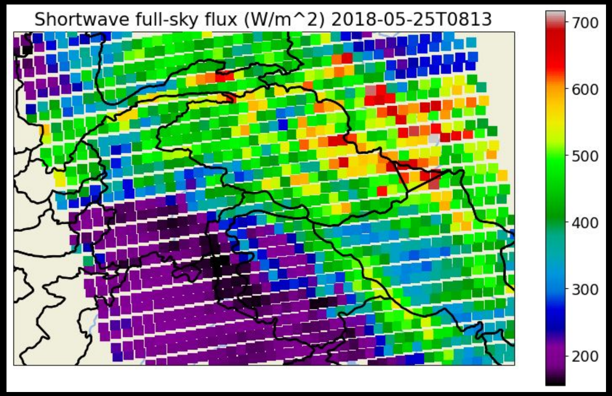
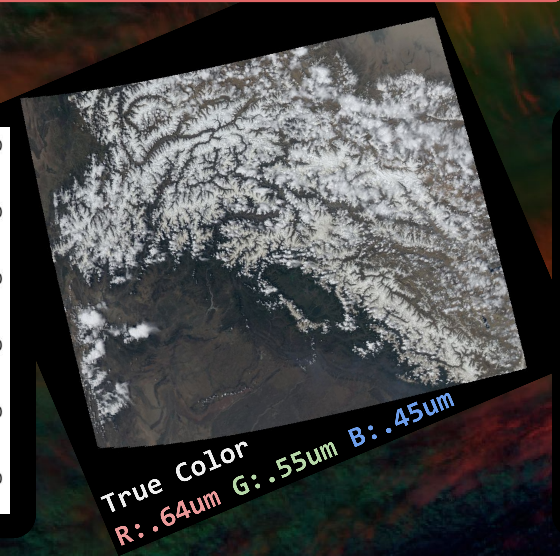
A Deep Learning approach to Estimate CERES Broadband Flux with 1km MODIS Pixels.

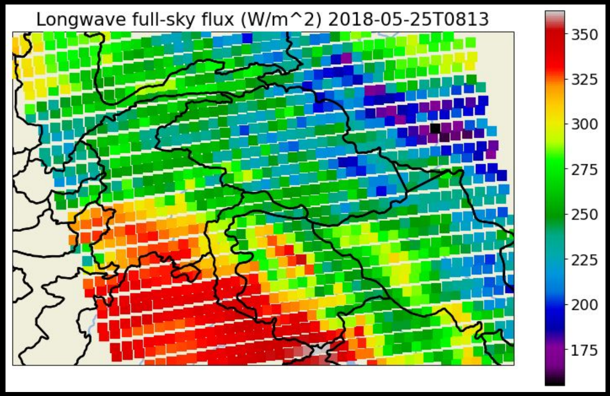
AES 690 - Project Update 2 - Mitchell Dodson

Goal: develop a **data-driven deep learning model** to characterize the relationship between EOS-MODIS' 36 spectral radiance bands, the sun/pixel/satellite viewing geometry, and the full-sky broadband radiative flux observed by the EOS-CERES instrument over a diverse set of regions.









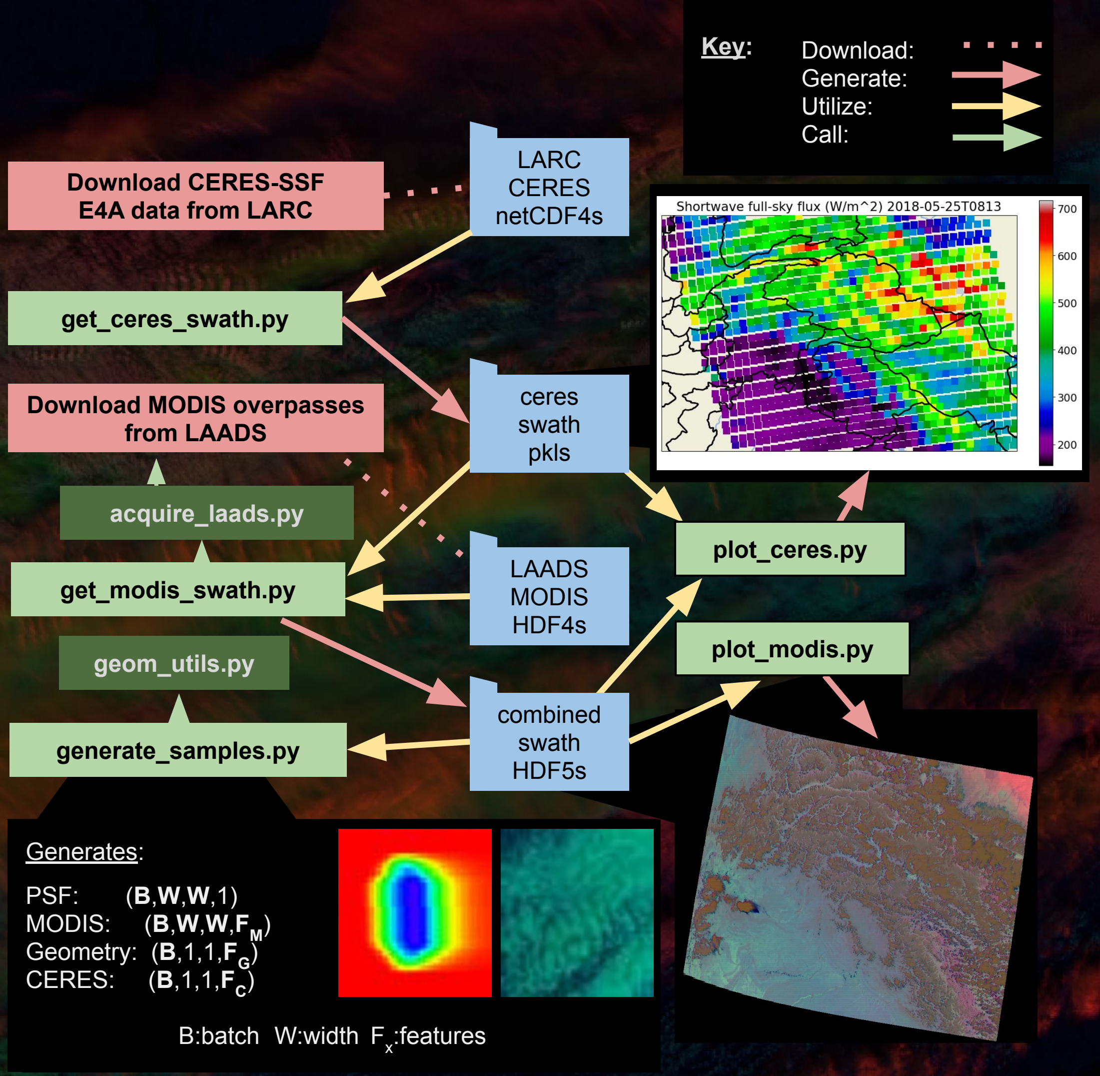
Data pipeline overview:

1 Use the LARC downloader to acquire CERES SSF data over a series of desired regions and time periods.

2 Select fields, restrict valid viewing geometry, and apply other constraints, then call **get_ceres_swath.py**, which extracts, subsets, and preprocesses CERES footprints, then groups them into individual overpasses alongside their meta-info in a new pickle file.

3 Select a series of MODIS bands to extract, and set file storage parameters in **get_modis_swath.py**, then run the script to download MxD021KM and MxD03 L1b granules from LAADS corresponding to each swath. Extract the desired bands in range of the CERES data, and combine the MODIS and CERES data for each overpass in a custom ‘swath’ HDF5.

4 **generate_samples.py** loads several swath HDF5s concurrently, selects CERES footprints and surrounding MODIS tiles, then calculates the CERES point spread function over each sub-domain. The PSF, MODIS/CERES data, and viewing geometry are interleaved and yielded as tensors.



Acquired regional data:

Using the described data pipeline, I acquired every third available EOS Terra and Aqua swath in the 3 years between Jan 1, 2018 and Dec 31, 2020, within the regional boundaries described below, for a total of about 2,600 overpasses, each of which contain about 300 individual samples on average.

(neus)

Label	Min	Max	Avg	Std
lat	34.263	42.930	38.566	2.414
lon	-77.322	-71.226	-74.367	1.493
vza	3.196	29.845	17.103	7.317
sza	37.378	46.063	41.750	2.078
swflux	90.989	611.449	282.029	120.906
lwflux	159.813	285.003	236.770	27.402
pct_clr	0.407	93.280	34.357	28.979
pct_l1	0.758	99.300	51.161	33.039
pct_l2	0.000	93.300	14.463	23.803
l1_cod	0.261	85.039	12.805	13.646
l2_cod	0.000	72.005	3.399	8.280
aer_land_pct	0.071	12.461	1.178	2.274
aod_land	0.000	0.380	0.037	0.063
aer_ocean_pct	0.006	65.302	7.309	14.265
aod_ocean	0.000	0.304	0.024	0.045
aod_ocean_small	0.000	0.180	0.015	0.028

Northeast United States
lat=[34,43], lon=[-69,-79]
Swath count: 532

(idn)

Label	Min	Max	Avg	Std
lat	-4.915	4.570	-0.170	2.646
lon	122.493	127.521	124.999	1.212
vza	3.631	29.836	17.166	7.275
sza	24.370	32.308	28.293	1.766
swflux	76.795	750.847	255.039	142.570
lwflux	129.861	291.400	237.621	33.949
pct_clr	0.370	96.402	33.312	28.465
pct_l1	0.139	99.266	42.657	31.592
pct_l2	0.000	97.317	24.029	28.747
l1_cod	0.119	88.985	7.904	11.022
l2_cod	0.000	75.253	3.334	7.609
aer_land_pct	0.000	66.529	2.391	7.737
aod_land	0.000	0.476	0.028	0.074
aer_ocean_pct	0.000	73.113	9.976	18.050
aod_ocean	0.000	0.317	0.035	0.056
aod_ocean_small	0.000	0.199	0.022	0.034

Indonesia
lat=[-5,5], lon=[120,130]
Swath count: 493

(alk)

Label	Min	Max	Avg	Std
lat	58.017	60.926	59.490	0.834
lon	-159.497	-154.426	-156.955	1.397
vza	6.192	27.520	16.523	5.521
sza	50.844	53.976	52.441	0.822
swflux	144.629	485.575	326.102	71.836
lwflux	183.409	252.280	219.779	15.006
pct_clr	1.557	76.016	25.339	19.179
pct_l1	1.941	97.623	60.437	28.288
pct_l2	0.000	81.810	14.198	20.830
l1_cod	1.684	71.662	17.545	13.079
l2_cod	0.000	54.584	4.973	8.809
aer_land_pct	0.008	35.277	5.602	7.379
aod_land	0.000	0.242	0.043	0.056
aer_ocean_pct	0.000	32.833	1.176	4.748
aod_ocean	0.000	0.161	0.008	0.027
aod_ocean_small	0.000	0.116	0.006	0.020

Alaska
lat=[58,61], lon=[-160,-154]
Swath count: 383

(seus)

Label	Min	Max	Avg	Std
lat	34.040	37.989	36.052	1.129
lon	-89.262	-85.598	-87.429	0.988
vza	6.326	29.145	17.635	6.152
sza	37.417	41.794	39.648	1.031
swflux	173.794	525.623	314.402	79.409
lwflux	188.285	277.697	242.395	20.083
pct_clr	6.995	80.235	38.927	19.716
pct_l1	1.329	90.458	47.711	25.542
pct_l2	0.003	74.999	13.343	19.167
l1_cod	1.436	58.629	13.910	10.614
l2_cod	0.000	48.194	4.084	7.625
aer_land_pct	0.313	14.549	4.637	3.751
aod_land	0.002	0.357	0.082	0.078
aer_ocean_pct	0.000	0.000	0.000	0.000
aod_ocean	0.000	0.000	0.000	0.000
aod_ocean_small	0.000	0.000	0.000	0.000

Southeast United States
lat=[34,38], lon=[-90,-85]
Swath count: 356

(hkh)

Label	Min	Max	Avg	Std
lat	32.033	37.975	35.050	1.691
lon	71.428	77.224	74.422	1.463
vza	3.056	29.836	16.732	7.363
sza	35.610	42.123	38.904	1.502
swflux	150.927	661.060	345.650	113.011
lwflux	168.895	312.938	244.834	30.726
pct_clr	1.460	97.151	41.651	27.731
pct_l1	0.154	97.405	41.977	29.464
pct_l2	0.000	89.662	16.353	22.979
l1_cod	0.254	74.625	9.265	9.618
l2_cod	0.000	61.636	3.812	7.255
aer_land_pct	0.000	81.474	13.372	22.328
aod_land	0.000	1.043	0.128	0.197
aer_ocean_pct	0.000	0.000	0.000	0.000
aod_ocean	0.000	0.000	0.000	0.000
aod_ocean_small	0.000	0.000	0.000	0.000

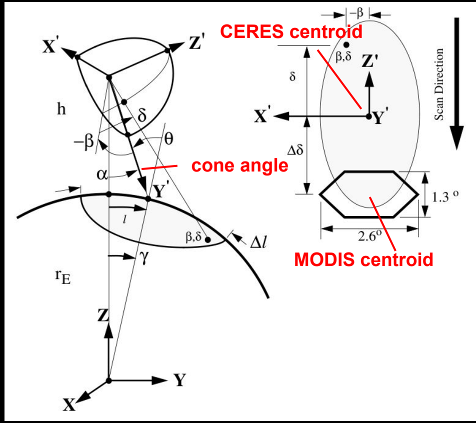
Hindu Kush Himalayas
lat=[32,38], lon=[69,79]
Swath count: 520

(azn)

Label	Min	Max	Avg	Std
lat	-7.978	-3.014	-5.503	1.419
lon	-69.103	-65.659	-67.383	0.910
vza	5.146	29.637	17.302	6.608
sza	26.745	31.586	29.156	1.095
swflux	171.321	693.881	342.504	110.928
lwflux	155.151	277.341	236.693	26.385
pct_clr	1.191	84.800	30.260	21.657
pct_l1	1.094	97.671	52.261	28.037
pct_l2	0.000	91.002	17.472	23.435
l1_cod	0.966	70.073	10.730	10.726
l2_cod	0.000	61.328	4.526	8.378
aer_land_pct	0.179	76.433	16.845	17.728
aod_land	0.002	0.525	0.135	0.122
aer_ocean_pct	0.000	0.000	0.000	0.000
aod_ocean	0.000	0.000	0.000	0.000
aod_ocean_small	0.000	0.000	0.000	0.000

Amazon Rainforest
lat=[-3,-8], lon=[-65,-70]
Swath count: 305

Challenge: Point Spread Function

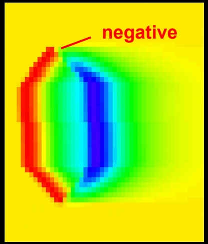


Equatorial geometry of the observation, and MODIS/CERES centroid offset schematic

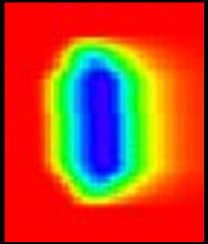
Implementing the CERES point spread function (**PSF**) over the co-located MODIS domain has presented a major development challenge. The algorithm to do so is specified in CERES ATBD subsystem 4.4, which I vectorized and implemented in **generate_samples.py** and **geom_utils.py**.

One issue I encountered is that the analytic equation approximating the PSF (specified to the right) was producing a result with very negative values near the aft border of the oblong CERES centroid.

I found that the cause of the error was that the ‘xi’ angle in the formula is assumed to be in radians rather than degrees, despite having been bounded and scaled previously by values in degrees.



Faulty PSF



Corrected PSF

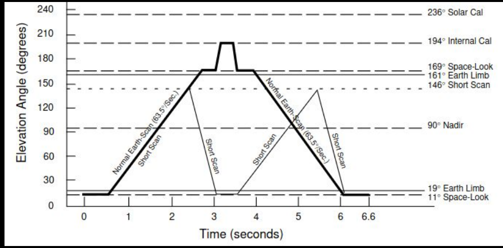
$$
\begin{aligned}
F(\xi) &= 1-(1+a_1+a_2)e^{-c_1\xi} \\
&\quad + e^{-6.35465\xi}[a_1\cos(1.90282\xi)+b_1\sin(1.90282\xi)] \\
&\quad + e^{-4.61598\xi}[a_2\cos(5.83072\xi)+b_2\sin(5.83072\xi)]
\end{aligned}
$$

$$
\begin{aligned}
a_1 &= 1.84205 & a_2 &= -0.22502 \\
b_1 &= 1.47034 & b_2 &= 0.45904 \\
c_1 &= 1.98412
\end{aligned}
$$

Empirical convolution equation (silently given in radians)

One related challenge that is still yet to be resolved is that the orientation of the MODIS grid centroid relative to the CERES footprint centroid is determined by the time derivative of CERES’ scan elevation (Earth cone angle) α, which alternates between cross-track sweeps, as diagrammed by the figure to the right.

Despite being specified in the ATBD, the derivative quantity isn’t included in the CERES SSF data product. Since the above geometry is entirely calculated from relative geodetic values, it also cannot be directly inferred from the data.



CERES’ alternating scan direction

The next steps:

The data tensors yielded by **generate_samples.gen_swaths_samples** are structured as tuples: **((M, G, P), C),** each containing **B** batch samples:

M := (**B**,**W**,**W**,**F**$_M$) MODIS reflectance or brightness temperature in **F**$_M$ bands
G := (**B**,1,1,**F**$_G$) Full-footprint viewing geometry **F**$_G$**:=(sol_zen, sat_zen, rel_azi)**
P := (**B**,**W**,**W**,1) 2D PSF probability distribution over the MODIS domain
C := (**B**,1,1,**F**$_C$) CERES footprint over the MODIS domain **F**$_C$**:=(lw_flux, sw_flux)**

The generator also provides options for linear normalization, and for constraining the parameters and sources of the data it returns.

The full dataset includes observations from 2 different sensors, which were collected over 6 distinct regions having widely varying characteristics.

These statistical sub-populations may make the flux prediction significantly more complex, so I will train separate models using stratified sampling techniques on subsets of the data.

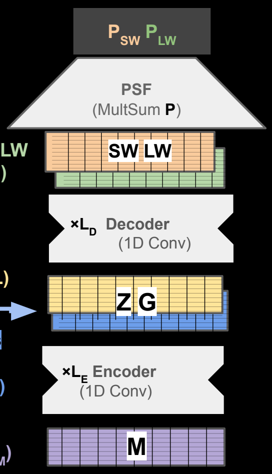
The basic model architecture consists of a sequence of 1D convolutions over the MODIS pixels to produce an independent latent representation of the radiance. The latent domain is concatenated with the viewing geometry (which is the same for all the pixels). Next, a second sequence of pixel-wise convolutional layers decodes the intermediate data to pixel-wise SW/LW predictions. Finally, the predictions are aggregated by a weighted sum over the PSF, and the loss is calculated with the result relative to the CERES footprints over the domain.

If time allows, I will experiment with applying the PSF aggregation to the latent grid before concatenating the resulting vector with the viewing geometry, and decoding with a simple multi-layer perceptron.

Since the entire model is pixel-wise, the full range of its predictions can be evaluated by modulating a single vector of input features. As such, it will also be interesting to see how predictions change with respect to the viewing geometry and MODIS spectral response.

Base Model Architecture
L_D and L_E are pixel-wise convolution layers.

Divergence:
$D = (P - \mathbf{F}_C)^2$



Model training plan:

First, do a broad random search of model hyperparameters on the simplest possible task (ie, a single region/season/satellite combo)

Then, using the most promising learning rate schedule and regularization method:

1. Train a large and small model on each individual region, using only one satellite.
2. Train a large and small model combining all swaths, seasons, and satellites.

Finally, evaluate each of the models on unseen data from all of the input sub-populations.