**NOTES:**

Started working with V5. Though I have scripts in GEE to sanitize the shape files, I chose to do it in QGIS this time.

Remember that the actual version of the data is outside final. Everything has been copied to final.

Filtered by year. Josef’s ‘preprocess\_by\_shape’ hasn’t been done – for now I am not caring about multipolygons etc.

**Imp!** Buffering hasn’t been done as well. Do it in the future.

**Imp!** Getting the satellite data from getDownloadURL() - I am ignoring the metadata in the tiff files. I would like to take care of this in the future. Right now I am using txt files to store the band names (timesteps)

**UPDATE:** No longer using getdownloadurl. Simply a img.sample in a multiprocessing environment.

**Im1 -**  The proper way to run these scripts would be to change the run option in spyder to run in CWD instead of renaming CWD to the directory of the file, which is the default behaviour. Make sure you realize this, lest you get stuck on errors when executing some other project codes. (Actually, changing the seetings for the codes in this project shouldn’t change other codes – you have to manually define the run options for every file in spyder. So technically, you shouldn’t get any errors as long as you make sure every file in this project has CWD as FINAl for imports to work properly)

**IMP!** Hm doesn’t have nulls in unique\_id but optical.csv does (parallel text to csv 2 ). It shouldn’t – this looks like its coming from sjoin, where 2 lat long pairs don’t belong to any polygon. Check kthis later.

Be careful about global sowing dates, base temps etc with remapping target. I think remapping should come after, since original names might be important for base temps and everything.

Why should you even loop inputs? Create multiple instances of the run and do them parallel. Merge outputs when necessary. **NO,**  It would be useless to run all preprocessing steps just to run another ML model .

At some point MASK\_CLDPRB was changed to CLDPRB

CLDPRB itself has nan values. Check it.

**IMP** Implementing pipeline\_executables in every file. I think this imposes that all the optional parameters be in config, and that it never saves files. Save has to be done in the main file. Currently this either has to read or execute, but there should be 3rd option, to ignore.

Because of the change of YYYYMMDD\_\_ , you will get some bugs in the existing version of the code. Its okay, though.

-> Be careful about whether you are stating timesteps from 0 or 1. Especially in feature addition this is important.

The functionality for aggregating points of the same field has been removed from gdd script.

Cloud probability can exist even if all other bands are null. This is coded into the gdd function by just checking for len(bands)-1 along with len(bands)

Even in 20 day interval, GDD gets nulls if cldprb>75 is masked.

Currently having to copy everytime data is being sent into the pipeline. This creates a lot of redundancy.

Removing last columns that had null values.

Could the gdd error be because of passing dataframes as arguments?

From now, every downloaded version of data has its own parallel text to csv.

++++++++++++++

#################################################################3333333

The GDD error is quite weird. It is also related to the cloud error. Check and solve both.