This is an explanation of the datasets and the possible research questions which can be answered:

**Central Question – Can machine learning be used to predict/understand/analyse flash droughts?**

All the datasets are spatio-temporal in nature, which means these are pixelated data (3-D data) which has coordinates (X, Y) (spatial component) and value. Temporal means the time length of the data which is from 2001-2022.

Flash Droughts (FD) dataset is the variable which needs to be predicted/analysed. The data is in the folder with name “FD”. Each pixel will have one of the 3 distinct values, which are 0, 1 and 2. The pixels with value 2 means flash drought, while 0 and 1 show no flash drought. This is a daily dataset, which means for every day each pixel will have one value.

How can it be analysed – Just because a pixel has value 2 it does not mean flash drought. The same pixel should have that value for atleast 5 consecutive days. And if the surrounding pixels also show the same phenomenon, it means we have a cluster, which would be helpful for analysis purposes. It is always beneficial to examine a cluster rather than a single pixel.

Task 1 – Compute a repository of all these pixels which show FD based on the above criterion. This will be the training data, and the other pixels may be used for training data (you guys need to decide, if you need the non-FD data)

**Now, research problem 1 for Paper 1 – Can GNN’s accurately predict flash droughts?**

For this, there are many “climatic variables” which have been provided in the dataset, which are:

1. Rainfall (rain)
2. Mean Air Temperature (mean\_temp)
3. Soil Moisture (sm)
4. Runoff (runoff)
5. Vapour Pressure (vp)
6. Wind Speed (wind)
7. Evapotranspiration (ET)
8. Potential evapotranspiration (PET)

Now, one important question, FD is calculated using ET and PET. Will you use these datasets in the model to predict FD (ideally, I would not include it).

**Now, research problem 2 for Paper 2 – Quantifying the impact of vegetation on flash drought.**

This will use a ML model where we need to understand the relationship between FD and vegetation parameters. The challenge is the vegetation data has a different time scale but same pixel resolution when compared with FD data. There are 4 vegetation parameters, (1) soil moisture (daily) (2) EVI (16days) (3) NDVI (16 days) and (4) SIF (8 days).

The clusters which we identified earlier will be used to understand the relationship for different landcovers (forests, croplands, pastures). I will do this part once we have the idea on what outcomes can be expected from the model.

**Research problem 3 for Paper 3 – Can we develop a machine learning model to identify flash drought?**

The aim is to identify a flash drought event using all these variables (climatic and vegetation). A flash drought event can be any event which is different from the “normal”. This could be a difficult problem which we can discuss later.