STATS 650: Midterm Data Mining Proposal on Hydro-Climatic Changes in the Indian Subcontinent

Group Number 10: Deep Shah (931002178)[1], Anshul Yadav (630000108)[1], Akhil Rajput (630000107)[1], Apurva Shinde (932002312)[2], Pranav Mahajan (730007851)[3], and Sidharth T (427009745)[3]

[1]Department of Civil Engineering, Texas A&M University
[2]Department of Computer Science and Engineering, Texas A&M University
[3]Department of Industrial Engineering, Texas A&M University

October 10, 2022

Abstract

Climate change with an increasing population poses tremendous challenges to freshwater availability and food security for the Indian subcontinent. Climate projections show that a majority of the Indian subcontinent will undergo severe water stress by the end of this century, hence, it is essential to understand the current long-term changes (January 2000 to February 2021) in hydro-climatic variables for resilient planning and water management in south Asia (SA). Our objective is to identify how water availability has changed in the current world, what drives these changes, how much is due to natural changes and how much is due to human alterations, and lastly how it is linked to the global climate variability and circulation patterns. This study has implications for drought and flood monitoring, weather forecasting, disaster management, and freshwater planning, which can provide crucial insights to stakeholders, policy-makers, water managers, and hydrologists for efficient decision-making.

1 Motivation and Background

Weak summer monsoons and anthropogenic warming (climate change) put significant pressure on water resources in the Indian subcontinent (Mishra et al., 2012). Studies have reported a decline in the monsoon rainfall over the last few decades, which is alarming and can influence the agriculture, financial, and economic facets of the Indian subcontinent (Koll et al., 2015). Human activities significantly influence freshwater availability in South Asia. For instance, due to excessive groundwater pumping, the groundwater is declining at the highest rate in northern India (Rodell et al., 2009).

Climate change along with increasing population can exacerbate water stress. Hence, it is crucial to understand the current challenges, drivers, and possible reasons behind the hydro-climatic changes. This can help predict future freshwater availability and associated hydro-climatic extremes such as megadroughts and floods and possible solutions that are required to be applied now. The long-term assessment of the hydro-climatic variables and their roles in hydro-climatic extremes was limitedly focused on. Therefore, our study fills this crucial research gap and provides an assessment of hydroclimatic changes in the Indian subcontinent, their potential drivers, human contributions, and the influence of large-scale climate variability on it.

1.1 Importance of data mining in our study

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. NetCDF file format (that we used here) supports array oriented scientific data used for climatology. Different python libraries support this type of data extending its functionality to multi-dimensional arrays. Using different data mining and visualization techniques, here we provide quality maps to comprehend the hydro-climatic changes and trends over Indian subcontinent.

2 Science Questions

Our work assists in addressing the following science questions:

1. How have the hydro-climatic variables such as precipitation, air temperature, soil moisture, runoff, and evapotranspiration changed across the Indian subcontinent?

2. What are the key drivers of those changes in different seasons (i.e., Monsoon, Pre-monsoon, or Post-monsoon)?

3. How are these changes in variables related to large-scale teleconnection patterns of climate change?

4. How do these changes influence the hydroclimatic extremes (droughts and floods)?

3 Objectives

These are the major objectives of the paper; however we only present the preliminary analysis that supports achieving the following:

1. To analyze the long-term changes in each hydro-climatic variable.

2. To analyze the trend and seasonality of each variable.

3. To identify the drivers behind those changes.

4. To link the regional changes to the global circulation patterns and evaluate the influence of global climate variability on those changes.

5. To provide an overall picture to the stakeholders, policymakers, water managers, and decision-makers to revise/update policies accordingly.

4 Study Area

We use the south asian region (Indian subcontinent) as our study area captures the strong seasonality and spatial climate variability. Moreover, it is one of the most water-stressed areas in the world. Figure 1 shows the geographical location of our study area.

5 Dataset

We obtained the hydro-climate data from the Global Land Data Assimilation Systems (GLDAS) which uses sophisticated numerical models of physical processes to integrate data from multiple ground and space-based observing systems in order to produce fields of water and energy states and fluxes (Rodell et al., 2004). The data was downloaded from the National Aeronautics and Space Administration (NASA) earthdata website (https://www.earthdata.nasa.gov/). We obtained the gridded data (0.25 × 0.25 degrees) from January 2000 to February 2021. The data is at a monthly temporal scale and in NetCDF4 file format. This dataset has about 5 million rows and 8 columns in raw format.

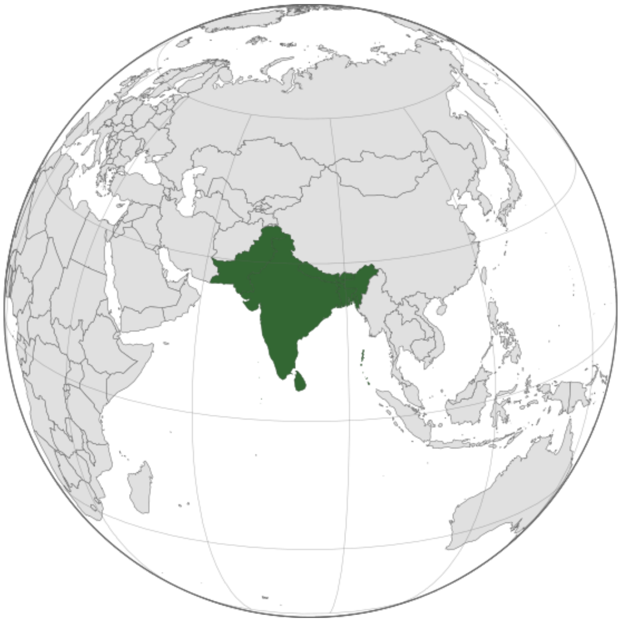
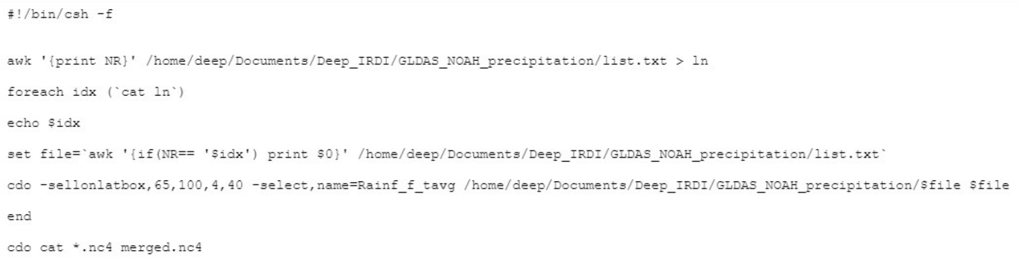


Figure 1: Geopolitical map of the Indian subcontinent

6 Data preparation

The data was available on a global scale. However, since our area of interest is the Indian subcontinental region, we clipped the data for south Asia using the Climate Data Operators (CDO). CDO is well known to handle climate data and very fast as compared to other programming options. Hence, we first clipped the global data to Indian subcontinent/south Asia and then merged all monthly (.nc4) files to make it a single (.nc4) file for each variable using CDO. Then all the exploratory data analysis was performed using python. The following figure (Fig. 2) shows the code written to clip and merge for rainfall only. The same was repeated to get each variable's merged file.



```
#!/bin/csh -f

awk '{print NR}' /home/deep/Documents/Deep_IRDI/GLDAS_NOAH_precipitation/list.txt > ln

foreach idx (`cat ln`)

echo $idx

set file=`awk '{if(NR== '$idx') print $0}' /home/deep/Documents/Deep_IRDI/GLDAS_NOAH_precipitation/list.txt`

cdo -sellonlatbox,65,100,4,40 -select,name=Rainf_f_tavg /home/deep/Documents/Deep_IRDI/GLDAS_NOAH_precipitation/$file $file

end

cdo cat *.nc4 merged.nc4
```

Figure 2: Code written to clip and merge data to get a single file for the entire temporal duration

6.1 Missing Data

We observe some NA values in the dataset which represents the data collected over the ocean. For the purpose of this analysis, these are not required as our study only focuses only on landmass.

6.2 Data Cleaning and Preprocessing

1. The collected climate data is in nc4 format for which each variable is converted into a numpy array and then collectively into a dataframe for statistical analysis.

2. Evaporation and Precipitation data variables are converted from kgm-2s-1 to mm/day for better understanding of the data variables and improved visual insights.

3. For temperature data variables we have considered the first 254 values (Jan 2000- Feb 2021) since the dataset contains repeated values.

6.3 Data ethics

We would like to acknowledge the National Aeronautics and Space Administration (NASA) for freely providing data for the research purpose. Our analysis using this data will help the water community to better understand the hydroclimatic changes over the Indian subcontinent region. This will help humanity to better understand the water problems and associated solutions.

7 Methods overview

The steps adopted in the study are summarized here. These steps are just a preliminary analysis. To answer all of our research questions, further detailed analysis is required, which is out of scope for this project.

1. Download and prepare the data for our study region using CDO.

2. Load the dataset into an ipython notebook to examine the data.

3. Using the NetCDF package, analyze metadata information to get a better understanding of the units, dimensions, and variable names and check if they are the correct format.

4. Identify the missing values and descriptive statistics to see if we need to calculate any missing parameters.

5. Merge multiple files to create a new pandas data frame for easier analysis.
6. Create a time series of each variable to visualize the temporal changes.
7. Plot the trend in air temperature to confirm that the air temperature is increasing and our data supports the existing findings.
8. Construct the standardized anomalies to remove the seasonality since each variable has strong seasonality. This is done to identify years where the variable has more/less than the expected (mean) value. This is helpful to visualize years which are relatively dry (droughts) or wet (floods).
9. Plot the histograms to visualize the distribution of each hydro-climatic variable.
10. Prepare the Quantile-Quantile (Q-Q) plots to check if any/all variables follow a normal distribution.
11. Create seasonal cycle plots along with the ± one standard deviation envelope to better understand how each variable responds during different months.
12. Create spatial plots of each variable for one drought and one flooding event. This is done to visualize the condition of each variable during extreme events and the possible driver or climatic condition behind it.
13. Calculate the correlation among the variables.
14. Lastly, we show spatio-temporal changes in soil moisture to visualize periodic changes and associated spatial variability using a GIF.

8 Preliminary Results

In this section we present the preliminary data analysis and early results using basic data exploration techniques.

8.1 Temporal trends of each variable

We plot the temporal trend of hydro-climate variables, as realized by taking the mean of all spatial areas for each month (Figure 3). Based on physical relationships, parameters such as evaporation, precipitation, stormwater runoff and soil moisture do not show significant monotonic trends as these depend on a variety of factors. Air temperature on the other hand,

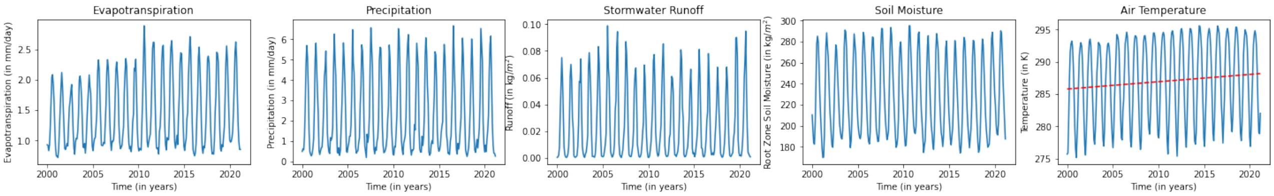


Figure 3: Time series plot of the variables in the dataset with a trend line for air temperature.

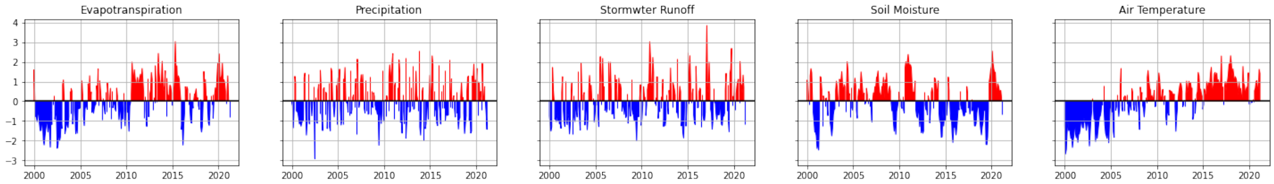


Figure 4: Anomalies detected in the variables

shows a clear, increasing trend due to climate change and hence, is shown by a linear trend (red line) in Figure 3.

8.2 Anomalies Detection

To remove the influence of seasonality, we constructed the standardized anomalies of each variable (Figure 4). The positive anomalies show higher than expected (mean) values and negative anomalies show lower than expected values. For example, months having negative precipitation anomalies (blue) show dry periods (Figure 4). Similarly, months having positive temperature anomalies show more than expected temperature (indicating climate change).

8.3 Data characteristics

Histograms of hydro-climatic variables are shown in Figure 5. Evapotranspiration, precipitation, and stormwater runoff are highly positively skewed. Soil moisture and Air Temperature have a slight positive and negative skewed distribution, respectively.

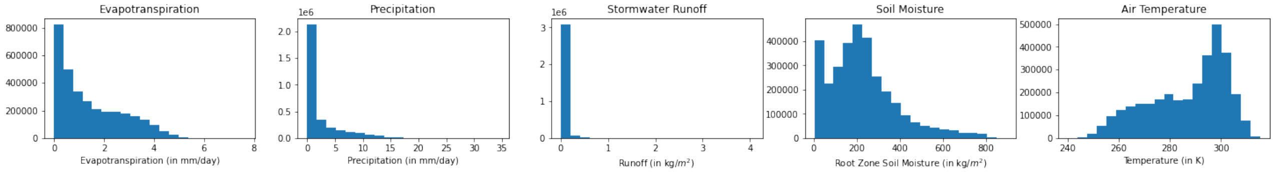


Figure 5: Distribution of the hydro-climatic variables

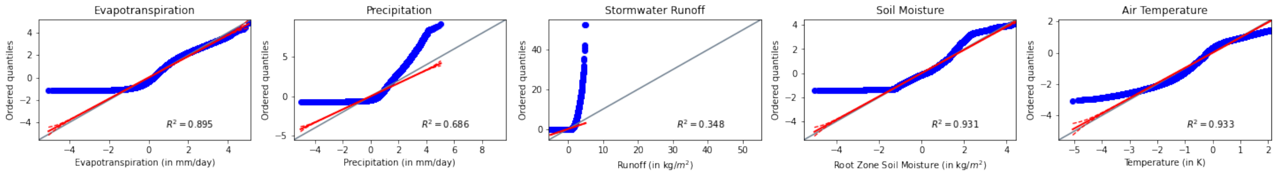


Figure 6: Normality test using q-q plot for the hydro-climate variables

8.4 Normality Test

Q-Q Plots, as shown in Figure 6, indicate that none of the hydro-climate variables have a perfect Normal Distribution.

8.5 Seasonality

To understand the seasonality in each variable, we created seasonal based on mean value for each month for all years. The plots (Figure 7) represent expected seasonal behaviour in a typical year. The shaded envelop shows uncertainty as in ± one standard deviation.

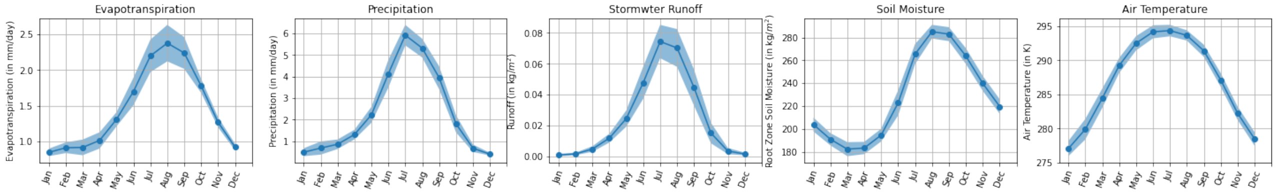
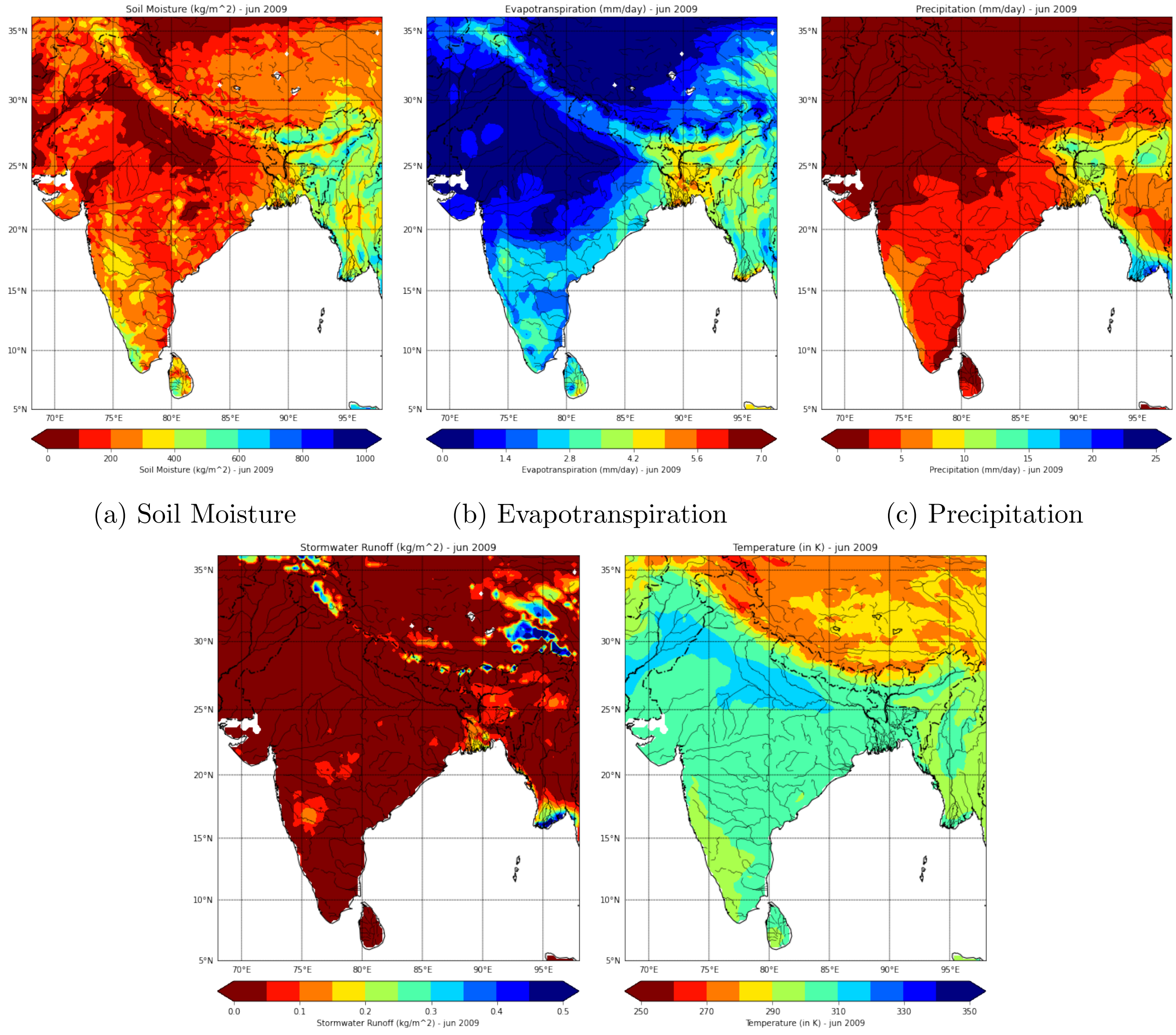


Figure 7: Evapotranspiration, Precipitation, Stormwater Runoff, Soil Moisture, Air Temperature during different months. The shaded regions show the values within one standard deviation.

8.6 Spatial plot of variables at extreme events

To visualize the different climatic conditions during dry (June 2009) and wet years (April 2020), we create the following spatial maps (Figure 8 and 9).


Soil Moisture (kg/m^2) - jun 2009
Evapotranspiration (mm/day) - jun 2009
Precipitation (mm/day) - jun 2009
35°N
30°N
25°N
20°N
15°N
10°N
5°N
70°E
75°E
80°E
85°E
90°E
95°E
0
200
400
600
800
1000
Soil Moisture (kg/m^2) - jun 2009
0.0
1.4
2.8
4.2
5.6
7.0
Evapotranspiration (mm/day) - jun 2009
0
5
10
15
20
25
Precipitation (mm/day) - jun 2009

(a) Soil Moisture

(b) Evapotranspiration

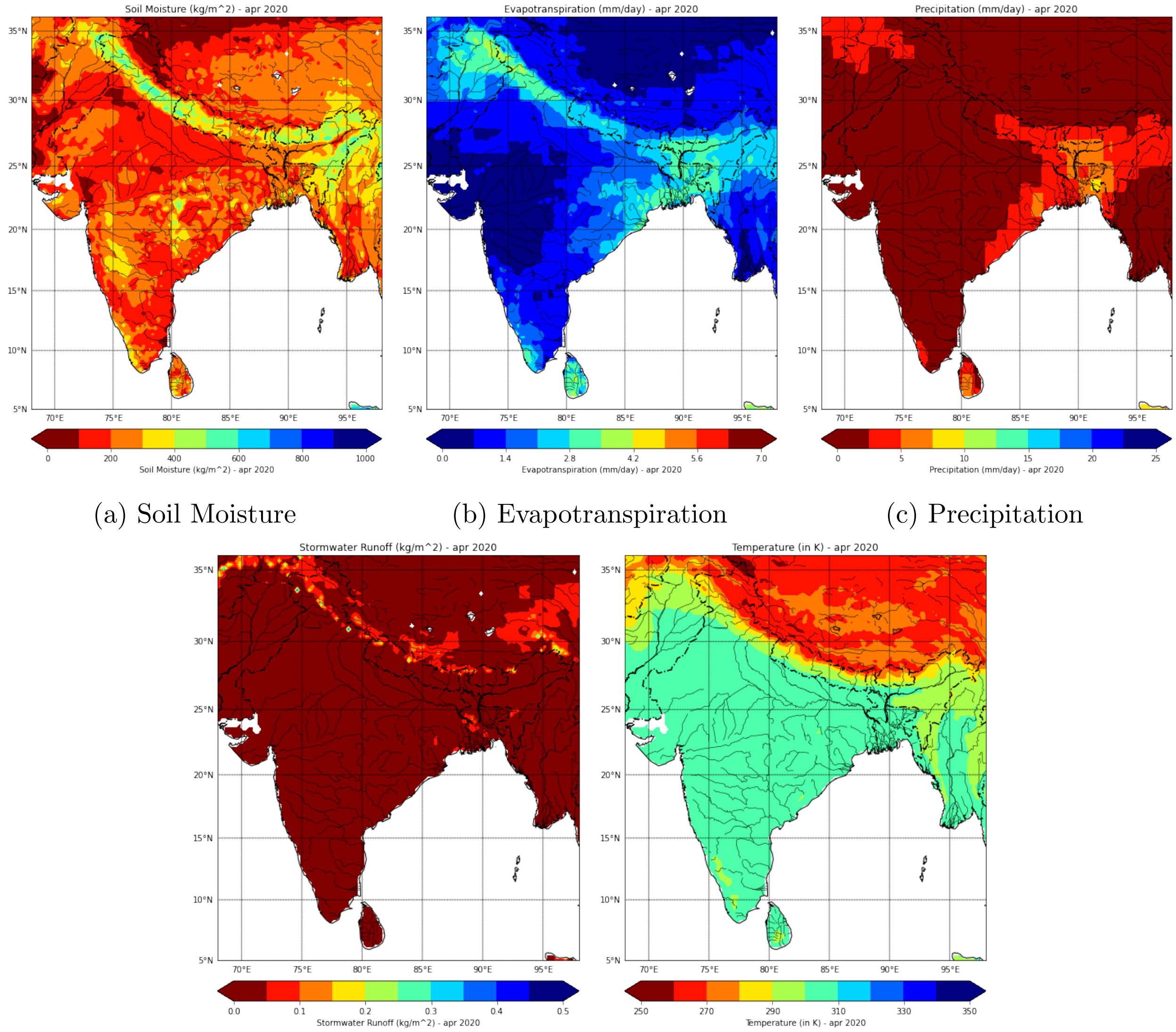
(c) Precipitation

Stormwater Runoff (kg/m^2) - jun 2009
Temperature (in K) - jun 2009
0.0
0.1
0.2
0.3
0.4
0.5
Stormwater Runoff (kg/m^2) - jun 2009
250
270
290
310
330
350
Temperature (in K) - jun 2009

(d) Stormwater Runoff

(e) Air Temperature

Figure 8: Observed extreme values in hydro-climatic variables for the month of June in drought year (2009)


Soil Moisture (kg/m^2) - apr 2020
Evapotranspiration (mm/day) - apr 2020
Precipitation (mm/day) - apr 2020
35°N
30°N
25°N
20°N
15°N
10°N
5°N
70°E
75°E
80°E
85°E
90°E
95°E
0
200
400
600
800
1000
Soil Moisture (kg/m^2) - apr 2020
0.0
1.4
2.8
4.2
5.6
7.0
Evapotranspiration (mm/day) - apr 2020
0
5
10
15
20
25
Precipitation (mm/day) - apr 2020
Stormwater Runoff (kg/m^2) - apr 2020
Temperature (in K) - apr 2020
0.0
0.1
0.2
0.3
0.4
0.5
Stormwater Runoff (kg/m^2) - apr 2020
250
270
290
310
330
350
Temperature (in K) - apr 2020

(a) Soil Moisture

(b) Evapotranspiration

(c) Precipitation

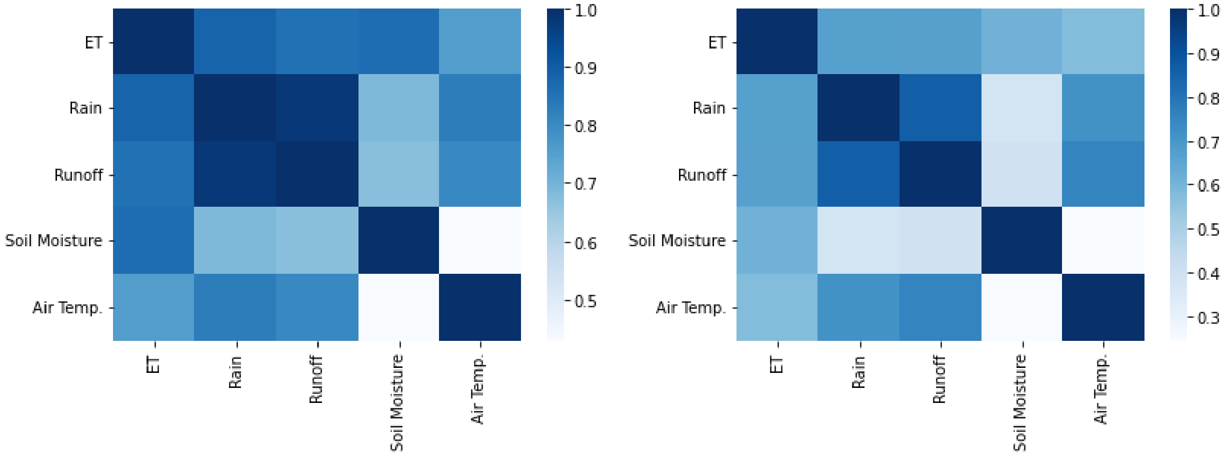
(d) Stormwater Runoff

(e) Air Temperature

Figure 9: Observed extreme values in hydro-climatic variables for the month of April in flood year (2020)

8.7 Correlation between the data variables

We calculate the correlation between the hydro-climatic variables based on Person's and Kendall's tau method (Figure 10a and 10b, respectively). There is a strong correlation between rainfall and stormwater runoff, which is expected. There is a weak correlation between air temperature and soil moisture as we considered only root zone soil moisture.



(a) Pearson's method

(b) Kendall method

Figure 10: Heatmap plot of correlation between data variables using a. Pearson's method and b. Kendall method

8.8 Animated plot of soil moisture change

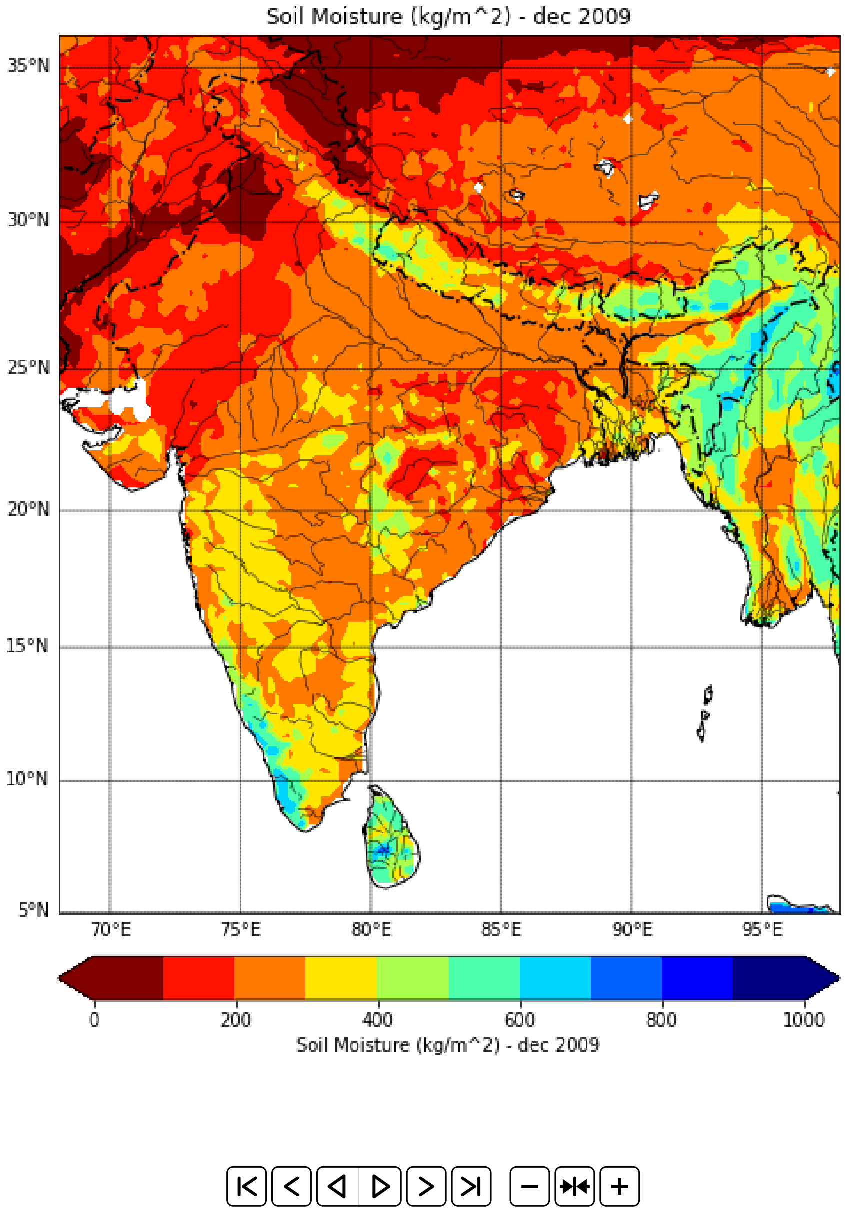


Figure 11: Evolution of soil moisture over the Indian subcontinent during the drought of 2009

9 Future Work

Work presented in this project only evaluates the relationship between the variables and exploratory data analysis. To understand the drivers of changes in these hydro-climatic variables due to human intervention, a detailed analysis is required. Since it is difficult to build a detailed model which incorporates all of the physical processes, models such as spatio-temporal GNN (Graph Neural Network) framework can help. Data variables such as LULC (Land use land cover), Reservoir operations, soil types, irrigation, hydrological catchment parameters, etc.

References

Mishra, V., Smoliak, B. V., Lettenmaier, D. P., & Wallace, J. M. (2012). A prominent pattern of year-to-year variability in indian summer monsoon rainfall. *Proceedings of the National Academy of Sciences*, *109*(19), 7213–7217.

Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al. (2004). The global land data assimilation system. *Bulletin of the American Meteorological society*, *85*(3), 381–394.

Rodell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in india. *Nature*, *460*(7258), 999–1002.

Roxy, M. K., Ritika, K., Terray, P., Murtugudde, R., Ashok, K., & Goswami, B. (2015). Drying of indian subcontinent by rapid indian ocean warming and a weakening land-sea thermal gradient. *Nature communications*, *6*(1), 1–10.