What is your research question?

What are the spatial and temporal distributions of groundwater levels in California?

What is the data that you will use?

I am going to use the State of California’s Department of Water Resources (DWR) Periodic Groundwater Levels dataset, which contains seasonal and long-term groundwater level measurements. There are groundwater elevation measurements at various well sites across California. When the groundwater measurements were taken do not seem to be standardized across the well sites, so I need to find a group of sites that have the same temporal dimensions to use. For this reason, I am not sure exactly how many observations my dataset will have yet. But I will be using variables including site lat/long, as well as a variable describing the Ground Surface Elevation (GSE) at the well site in feet referenced to the North American Vertical Datum of 1988. If I have time, I may include data on temperature or precipitation.

What methods will you use?

I would like to use PCA on the data. I could apply PCA on a single year to understand the potential spatial pattern of groundwater levels across California. Then use the timeseries data to explore the temporal pattern. If there are missing days in the data (i.e. measurements were not collected on some days) I may need to aggregate up to monthly averages. If there are months missing, I may need to interpolate the data. I might decide to interpolate daily data or monthly average data.