Data Quality Analysis

Analyzing the accuracy of satellite precipitation measurements with Rain Gauges across different geographic and topographic regions in US.

Ankit Ghanghas and Pin-Ching Li

Lyles School of Civil Engineering, Purdue University

Datasets Description (includes variables and units)

GPM (Global Precipitation Measurement)

Our study involves precipitation datasets from on ground measurement and satellite project. GPM is known as NASA Global Precipitation Measurement. The satellite measurement GPM is calibrated with an algorithm called Integrated Multi-satellite Retrievals for GPM (IMERG). IMERG interpolate all satellite precipitation with microwave-calibrated infrared (IR) satellite estimates. There are three steps in IMERG system. Final run of IMERG is chosen to generate the research-level products.
The major variable of GPM dataset applied in our study is CalPrecipitation. The long name of CalPrecipitation is Daily accumulated precipitation (combined microwave-IR) estimate. The unit of CalPrecipitation is mm.

GHCN (Global Historical Climatology Network)

In the United States, rainfall data is collected by National Oceanic and Atmospheric Administration's (NOAA's) and is available for download from the National Centers for Environmental Information (NCEI). Our daily precipitation dataset is from a integrated database, which contains summaries across the globe: GHCN (Global Historical Climatology Network). NCEI is the database storing the GHCN and can be accessed by people with registered token.
GHCN dataset contains daily maximum and minimum temperature, precipitation, snowfall, and snow depth. Daily precipitation dataset is the variable we chose for analysis of satellite precipitation. The unit of GHCN Daily Precipitation is tenth of millimeter. There are hundreds of gage stations within a state, while some of them are poor in data coverage. For example, though there are 631 stations within Indiana or nearby Indiana, there are many stations without enough data coverage. As the histogram shown below, more than 300 stations suffer from insufficient data coverage (less than 95%).

Importance of Analysis : Rainfall Dataset is key to many hydrologic models and water management practices. Although the ground rain gage stations provide relatively accurate measure of the amount of rainfall, they are spatially sparse. The satellite data on the other hand can be used to measure the rainfall even in gage sparse locations and can be more useful but this has larger associated errors. Our study tries

to look at how the ground data is related to satellite data and do geographic and topographic location impact the relationship between the two.

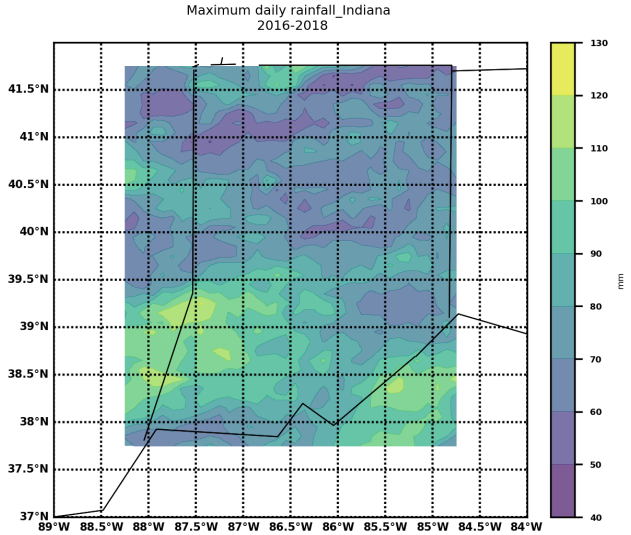
Availability of Dataset :

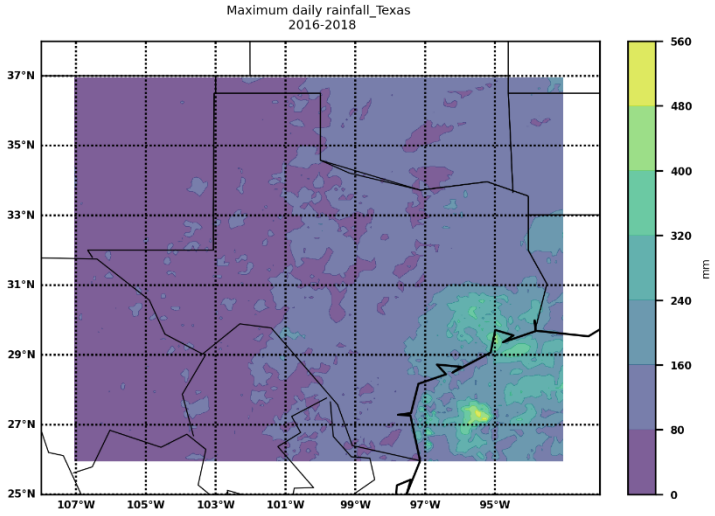
The GPM dataset covers the entire earth and is available at 30 min, hourly, daily and monthly time resolution. Here for our study we use daily time scale. The dataset is available from June 2000 to present but for the present study we focus on Jan2016 to Dec2018.

Based on the data coverage of station we select 196 sites in Indiana and 365 sites in Texas for our analysis.

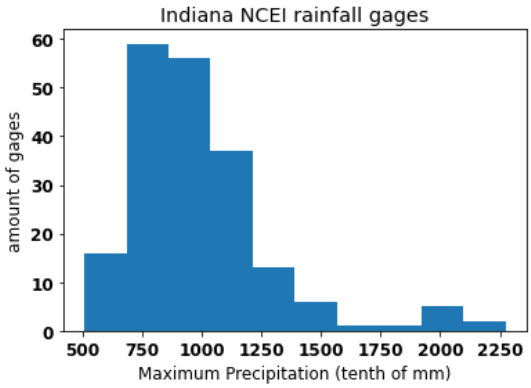
Graphical Analysis:

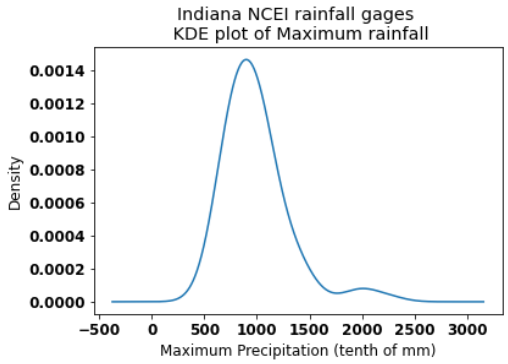
The graphical analysis of the data includes looking at maximum daily rainfall from GPM data for both the states. Histograms and KDE plots with gaussian kernel (width =0.5) are drawn to find out the precipitation influenced by spatial distribution within a state. Box Plots are used for checking the difference of max, mean and std of precipitation in different states. Finally the maximum precipitation are fit against normal distribution to look for any possible pattern across the states.

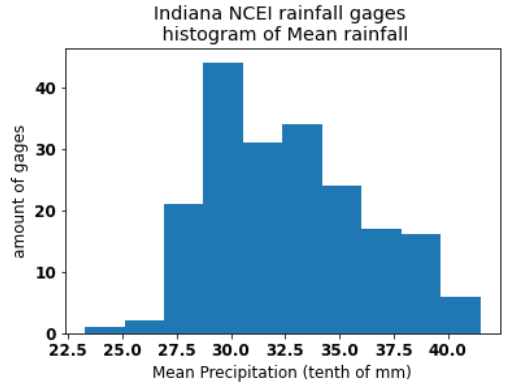


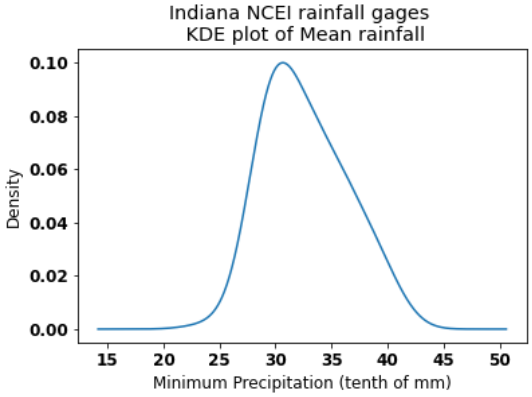


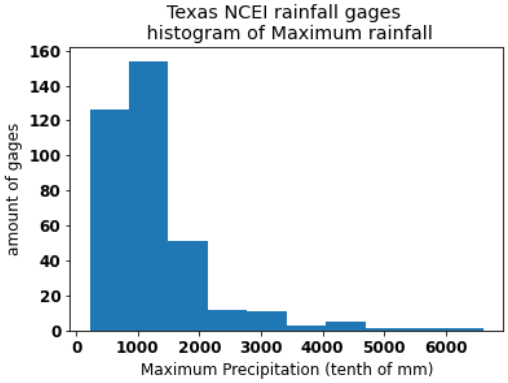
The plot of maximum daily rainfall gives us an idea of locations where it rains the most. As evident a large portion of Texas which is desert receives relatively low max rainfall, while the maximums are relatively similar in Indiana. This is also evident from the histograms and KDE Plots shown below. These plots tell us about the level of consistency in rainfall across the state.

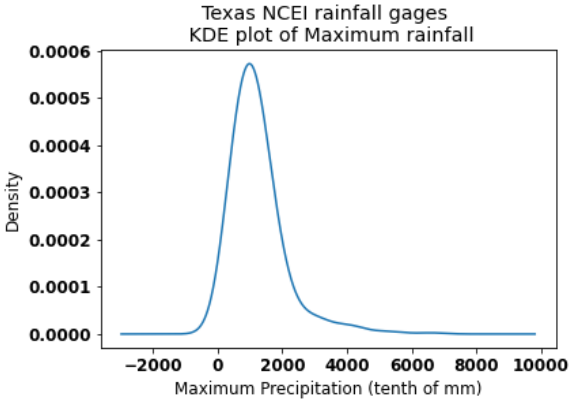

Indiana NCEI rainfall gages
amount of gages
Maximum Precipitation (tenth of mm)

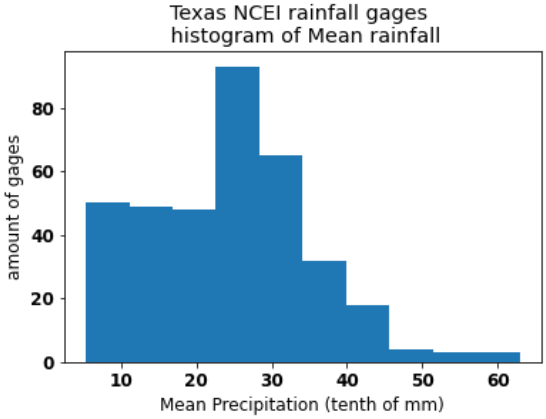

Indiana NCEI rainfall gages
KDE plot of Maximum rainfall
Density
Maximum Precipitation (tenth of mm)

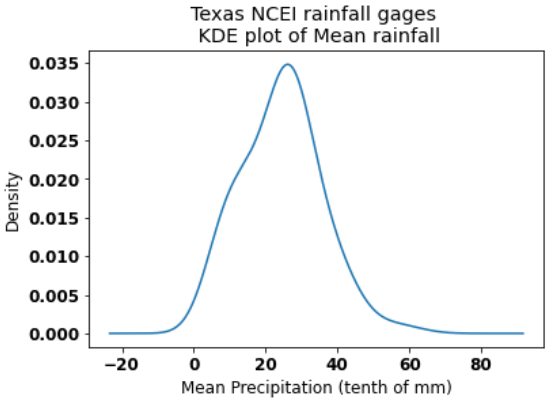

Indiana NCEI rainfall gages
histogram of Mean rainfall
amount of gages
Mean Precipitation (tenth of mm)

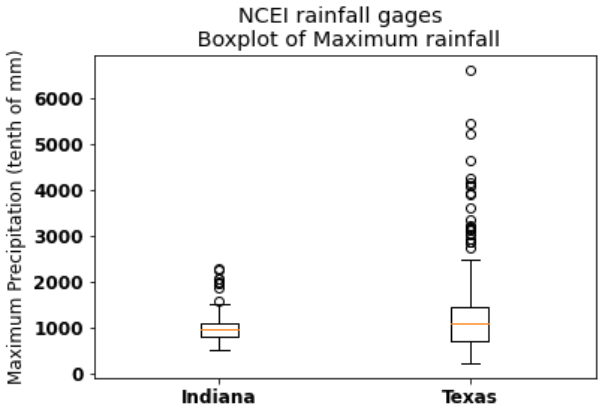

Indiana NCEI rainfall gages
KDE plot of Mean rainfall
Density
Minimum Precipitation (tenth of mm)

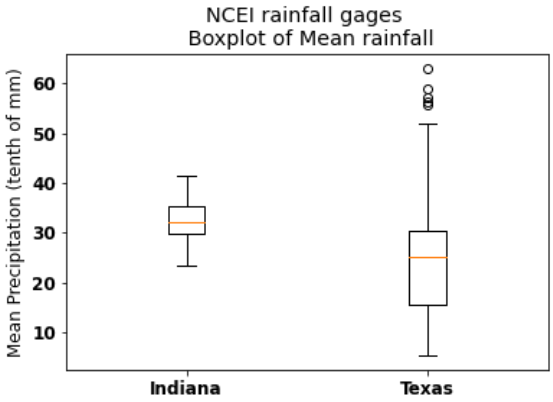

Texas NCEI rainfall gages
histogram of Maximum rainfall
amount of gages
Maximum Precipitation (tenth of mm)


Texas NCEI rainfall gages
KDE plot of Maximum rainfall
Density
Maximum Precipitation (tenth of mm)

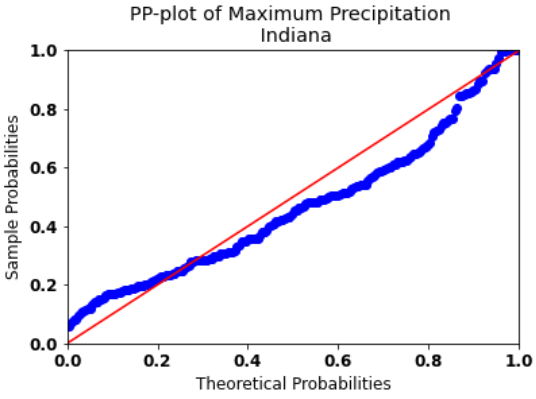

Texas NCEI rainfall gages
histogram of Mean rainfall
amount of gages
Mean Precipitation (tenth of mm)

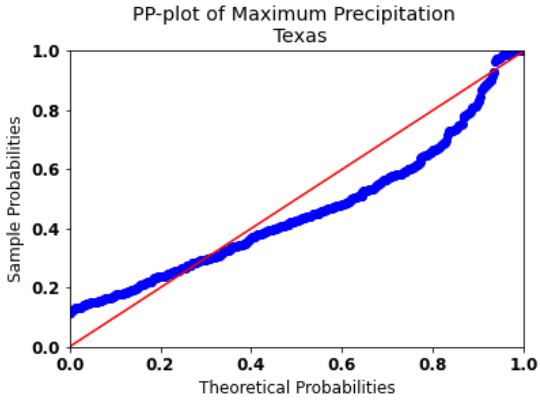

Texas NCEI rainfall gages
KDE plot of Mean rainfall
Density
Mean Precipitation (tenth of mm)





The box plots tells us about the amount of variability in rainfall amount across the state. From these plots we can infer that rainfall in Texas in more variable while it is relatively consistent in Indiana.





These plots indicate the level of agreement of The Maximum rainfall with normal distribution. We infer from the plots that maximum rainfall is not normally distributed in either of the states but it has relatively more deviation in Texas.

Data Quality Checking:

A lot of data quality checking is done while on the process of selecting the appropriate data to be used. The NCDC ground observations are segregated even during the downloading process to screen out stations that have less than 95% data coverage and stations that do not have no data collection in our time of interest (Jan 2016 to Dec 2018). The GPM grids are downloaded without any filters for the given time frame. The GPM data is available in raster form in grids while the NCDC data is point dataset, so we need to extract GPM data for the location of the ground NCDC stations in order to compare them. For extracting GPM dataset we map the geographic location of stations and take average and maximum (and form two kinds of datasets) of the values in the four nearest grids in order to deal with the error that

might be associated with lat, long projection coordinate systems of the two datasets. The maximum approach diminishes the possibility of no data points.

Once the datasets are retrieved they go through basic data quality checking. The entire dataset is looked for no data values that may be represented by (-9999.9) and all such values are set aside as NaN so that they do not interfere with the statistics.

Next we remove any Gross Error that may be found in the dataset. The values of precipitation less than 0 (rainfall cannot be negative) and more than 250 mm for Indiana and more than 750 mm for Texas are set to NaN. The value for Indiana is based on Maximum daily precipitation received in the state, while the value for Texas is much larger the normal yearly max values because of the fact that Hurricane Harvey hit the state in the time period of study.

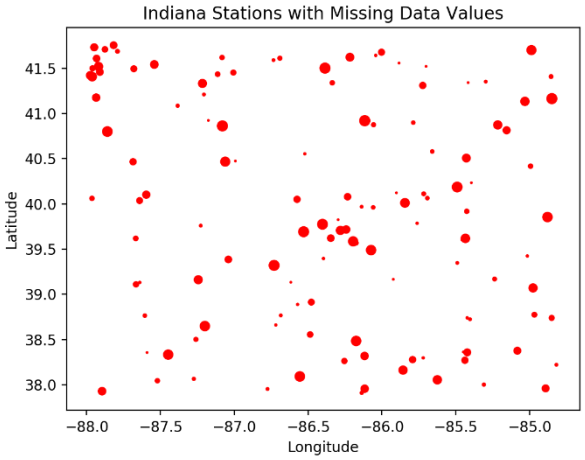
There are no other special data quality checks needed for our interest in the study.

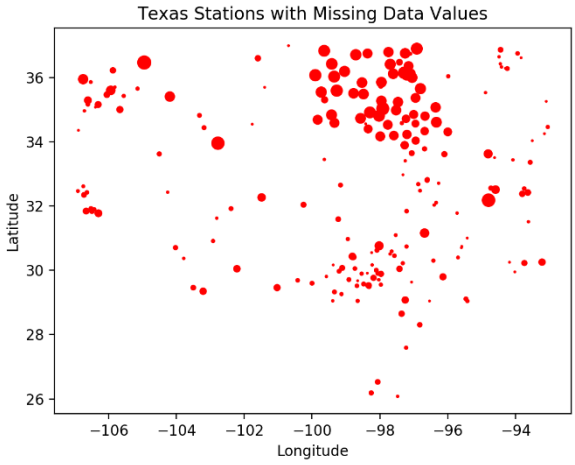
Results and Discussion.

Since the Number of stations is very large so it is not feasible to show the table for all the points but here we provide a summary for the results. (The csv files including corrected datasets and dictionary for data quality check, for the individual datasets can be found in the repository)

		Total No. Data Values	Gross Error
IN	GHCN	2075	0
	GPM _avg	0	0
	GPM_max	0	0
TX	GHCN	3330	0
	GPM_avg	0	0
	GPM_max	0	0

The plot below shows the variation of number of missing values across the states. The size of the dot indicates the number of missing values for that stations. The maximum number of values that fail the data quality check for a station in Texas is 77 while it is 47 for Indiana. (These are the number of days with missing values in the given time frame from 1st Jan 2016 to 31st Dec 2018).





The data quality check indicates that there are no erroneous and missing values in case of GPM dataset both average and max methods for both the states. There are a few missing values in the ground observations (GHCN dataset) but even this dataset does not contain any gross error like negative precipitation or excessively large precipitation. Texas has a cluster of points with greater number of missing data values in area near Dallas. Texas also has higher percentage of stations with missing values and has stations have relatively more missing values than stations in Indiana. A significant amount of missing data values are handled by carefully selecting the stations to be used for study.