EVALUATING THE PERFORMANCE OF GRIDDED RAINFALL DATASETS FOR THE DES MOINES RIVER BASIN

Revanth Mamidala
Ph.D. in Civil Engineering
mrevanth@iastate.edu

CE - 5900 Section 16 AI4CCEE
Final Project Report

Word Count: 1760 words + 1 table(s) $\times$ 250 = 2010 words

Submission Date: December 16, 2024

ABSTRACT
Precipitation is a crucial input for the Eco-hydrologic models in analyzing the dynamics of the nutrients in a watershed. The in-consistency and poor spatial representativeness of the ground-based observations and at the same time, the limitation of SWAT+ model to consider the raingauge station nearesr to the centroid of a sub-basin have highlighted the importance of usage of remotely sensed gridded rainfall datasets. However, the selection of suitable gridded dataset is sensitive to the region of interest and the timescale. In the current study, the performance of 7 gridded rainfall datasets that include CHIRPS, ERA5, GLDAS, GPM, GSMAP, PERSIANN, and PRISM was assessed by comparing them with ground-based observation data from 22 rain gauge stations in the Des Moines River basin using the statistical metrics such as Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative Bias (rBias %).

The PRISM dataset outperformed other gridded datasets displaying highest median R^2 value (0.63) and least median MAE (2.8 mm), wile the datasets GPM and GSMAP have exhibited an overestimating tendency thus suggesting their limited applicability in the watershed. The study directs the future research in extracting the PRISM rainfall values at the centroid of each sub-basin and use them as input for the SWAT+ model.

Keywords: Gridded Rainfall dataset, Googel Earth Engine, SWAT+, Eco-hydrologic modeling.

INTRODUCTION

Eco-hydrologic models such as SWAT+ are a state-of-the-art tools in understanding complex hydrologic systems and simulate the nitrate dynamics in an agricultural watershed to develop management strategies. Accurate precipitation data is a crucial input to the model (*1, 2*). Due to the sparse ground-based observation data, poor spatial distribution, and the limitation of SWAT+ in selection of rain gauge data of the nearest station to the centroid of the watershed has made the gridded precipitation datasets derived from remote sensing, reanalysis and interpolation techniques a better alternative (*2*). However, the reliability and applicability of these datasets vary with climatic zones, topography and the location of the watershed thus necessitating a detailed evaluation to identify the most suitable dataset for the Des Moines River basin(*3, 4*). In the current study, precipitation data was extracted from the gridded rainfall datasets available in Google Earth Engine and compared with the observed rain gauge data using statistical metrics.

DATA AND METHODS

Study Area

The analysis was conducted on the Des Moines River basin which is located between the coordinates 43.75° to 40.50° North and 92.90° to 95.70° West with a drainage area of 38,340 km^2 Stretching over 845 km in the states of Minnesota, Iowa, and Missouri. The sub-basins within the Des Moines River basin that were delineated using the USGS monitoring stations as outlets are as shown in figure 1.

Datasets

This section describes the gridded rainfall datasets used in the current study which were compared with the observed rainfall data.

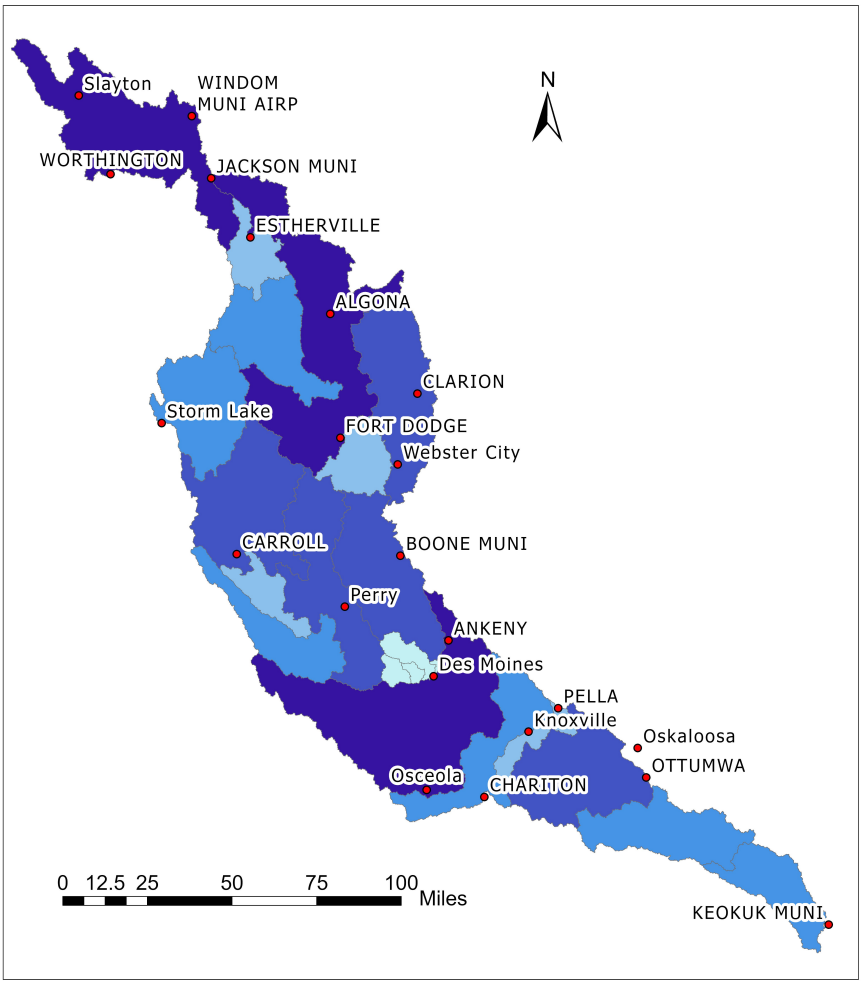

FIGURE 1 Illustration of sub-basins in Des Moines River Basin and the selected rain gauge stations in the basin.

Gridded Rainfall Datasets

The gridded rainfall datasets that were hosted in the Google Earth Engine (GEE) catalog that has a spatial scale covering the entire Des Moines River basin are used. The datasets were selected based on the literature review (*2, 5, 6*) and the details of the gridded datasets used in the study are shown in the Table 1. The temporal resolution of the datasets varied which were converted to daily temporal scale by using **.sum()** reducer in code editor of GEE. The snapshot of the GEE script is available at: `https://code.earthengine.google.com/362b351a7168f97c64511d015c847d7b?noload=true`. The GEE Java Script extracts the daily rainfall data at the coordinates of rain gauge stations shown in figure 1 in millimeters and is saved to CSV which was used in further analysis.

Observed Rainfall Data

The Iowa Environmental Mesonet (IEM) at Iowa State University contains a variety of meteorological data of which the automated airport weather observations, including ASOS(Automated Surface Observing System) data was used in the current study. 22 ASOS rain gauge locations in

TABLE 1 GEE catalog - Gridded datasets used in the study

Dataset	Temporal Resolution	Spatial Resolution	Availability Period
CHIRPS	Daily	~0.05 deg (5.5km)	1981 – present
ERA5	Daily	~0.25 deg (31km)	1979 – present
GSMAP	Hourly	~0.1 deg (11km)	2000 – present
GPM	Hourly	~0.1 deg (11km)	2014 – present
GLDAS	3-Hourly	~0.25 deg (31km)	2000 - present
PERSIANN	Daily	~0.25 deg (31km)	1983 – 2020
PRISM	Daily	~4km	1981 – present

the region of interest were chosen for the study which are shown in figure 1. The IEM data at the selected rain gauge locations was extracted using a data scraping python script (Supplementary document 2).

Data Analysis

The precipitation data extracted from the gridded and observed rainfall datasets were evaluated and compared using statistical metrics such as Coefficient of Determination R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Relative Bias (rBias).

Coefficient of Determination R^2

The values of R^2 range from $-\infty$ to 1 and are calculated as shown below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where:

- y_i is the actual value for observation i,
- $\hat{y}_i$ is the predicted value for observation i,
- $\bar{y}$ is the mean of the actual values, and
- n is the number of observations.

Root Mean Squared Error (RMSE)

The RMSE is always non-negative (RMSE ≥ 0) and is calculated as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

where:

- y_i is the actual value,
- $\hat{y}_i$ is the predicted value, and
- n is the number of observations.

Mean Absolute Error (MAE)

The values of MAE are always non-negative (MAE ≥ 0) and are calculated as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{3}$$

where:

- y_i is the actual value for observation i,
- $\hat{y}_i$ is the predicted value for observation i,
- n is the number of observations.

Relative Bias (rBias)

The Relative Bias (rBias) measures the systematic bias in the values extracted from a gridded dataset relative to the observed rainfall values at a particular gauge station. It helps identify whether the model tends to overestimate or underestimate the actual values. The values of rBias are calculated as shown in equation (4).

$$\text{rBias} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i} \times 100\% \tag{4}$$

where:

- y_i is the actual observed value for observation i,
- $\hat{y}_i$ is the predicted or extracted value for observation i,
- n is the number of observations.

Interpretation of rBias:

rBias > 0 Indicates an **overestimation** by the model or dataset.

rBias < 0 Indicates an **underestimation** by the model or dataset.

rBias $= 0$ Indicates no systematic bias; the predicted values match the observed values perfectly.

RESULTS AND DISCUSSIONS

A common problem that might arise in using the global dataset is their timezone mentioned in the timestamp. The PRISM dataset has a variation in timezone and contains local time zone of each pixel. At the same time, all the other gridded datasets are in Coordinated Universal Time (UTC) while the observed dataset is in Central Standard Time (CST). This variation might cause a difference in sub-daily comparisons, however from figure 2 it can be observed that there is no significant shift in rainfall values on a daily timescale. Therefore, in the current study no changes were applied to the timezones.

From the R^2 box plot in figure 3, the PRISM dataset has highest median R^2 (0.63) value followed by GPM (0.6), at the same time PRISM has higher variability in the R^2 values at different locations.GLDAS dataset has consistently underperformed with least median R^2 (0.42) values while CHIRPS, ERAS, PERSIANN, and GSMAP were observed to provide a moderate performance (R^2: 0.52-0.55) as far as R^2 value is considered.

The MAE values at all 22 stations were plotted in a box plot for each dataset (figure 4) indicates a varying levels of accuracy accross datasets with PRISM having the least median MAE (2.8 mm) while GPM displayed higest MAE of 6.5mm. The remaining datasets had a median MAE values ranging between 3-4 mm.

The RMSE values as illustrated in figure 5 reveal a distinct pattern of RMSE values across stations. At the stations MJQ and FOD, all the datasets have been underperforming while the datasets GPM and GSMAP displayed higher RMSE (12-18mm) values. This shows the sensitivity of rain gauge location and the dataset used. The R^2 values also show the similar trend with the rain gauges with poor performance of all datasets at the stations MJQ, FOD and EOK.

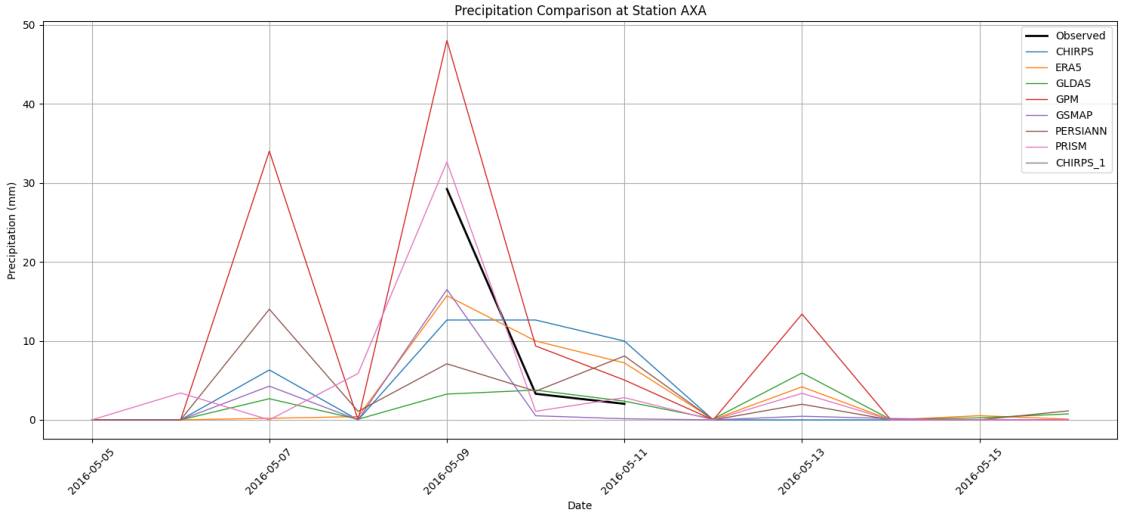

FIGURE 2 Comparison of gridded data with daily precipitation at station ALGONA (AXA) in Iowa between 2016-05-05 to 2016-05-16.

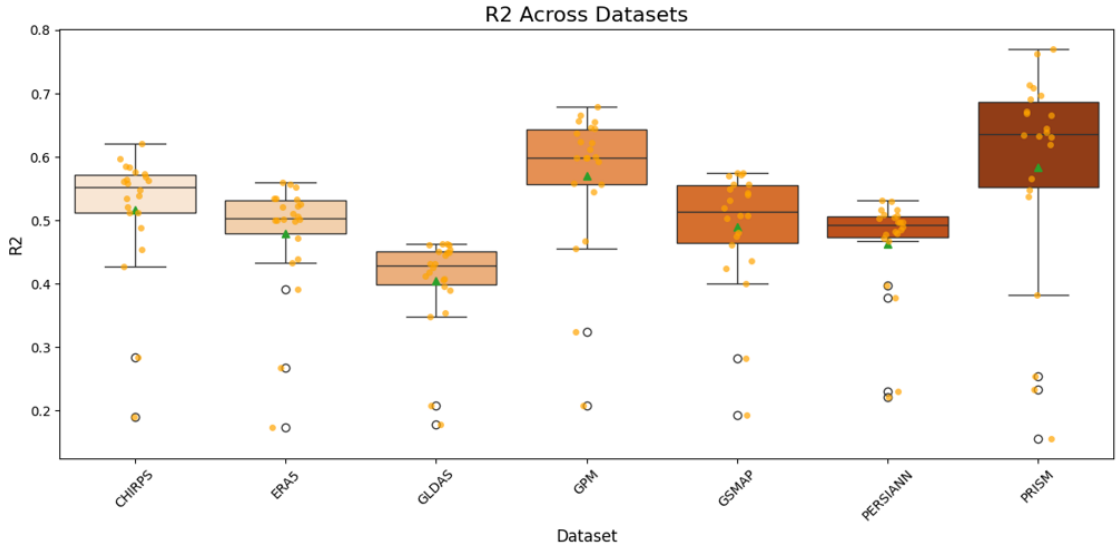

FIGURE 3 Box plot of R^2 value of all stations for each dataset.

The relative bias at all stations except MJQ show an overestimation of rainfall by all the gridded datasets. The major outliers were station EOK with a positive bias indicating overestimation while the staion MJQ has negative bias indicating underestimation of the rainfall. Within the datasets, GPM and GSMAP has shown a consistently higher overestimation thus showing that they are not applicable to the Des Moines River basin.

Finally, scatter plots were drawn to compare the performance of datasets at daily, monthly and annual timescales for each rain gauge station. From the comparison as shown in figure 8, the PRISM dataset has the least spread of 95th percentile region when compared to other gridded dataset in all timescales.

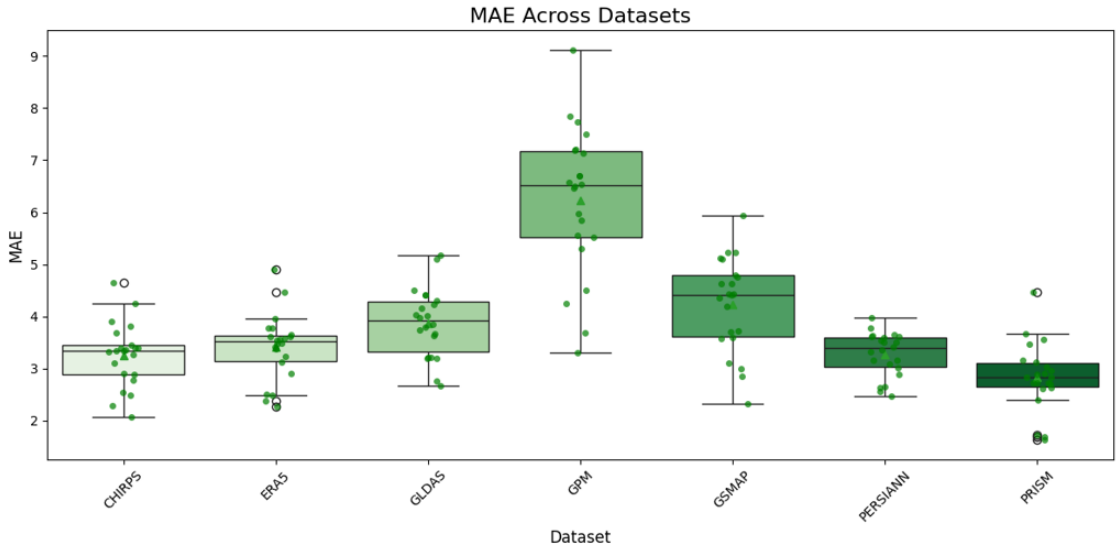

FIGURE 4 Box plot of MAE value of all stations for each dataset.

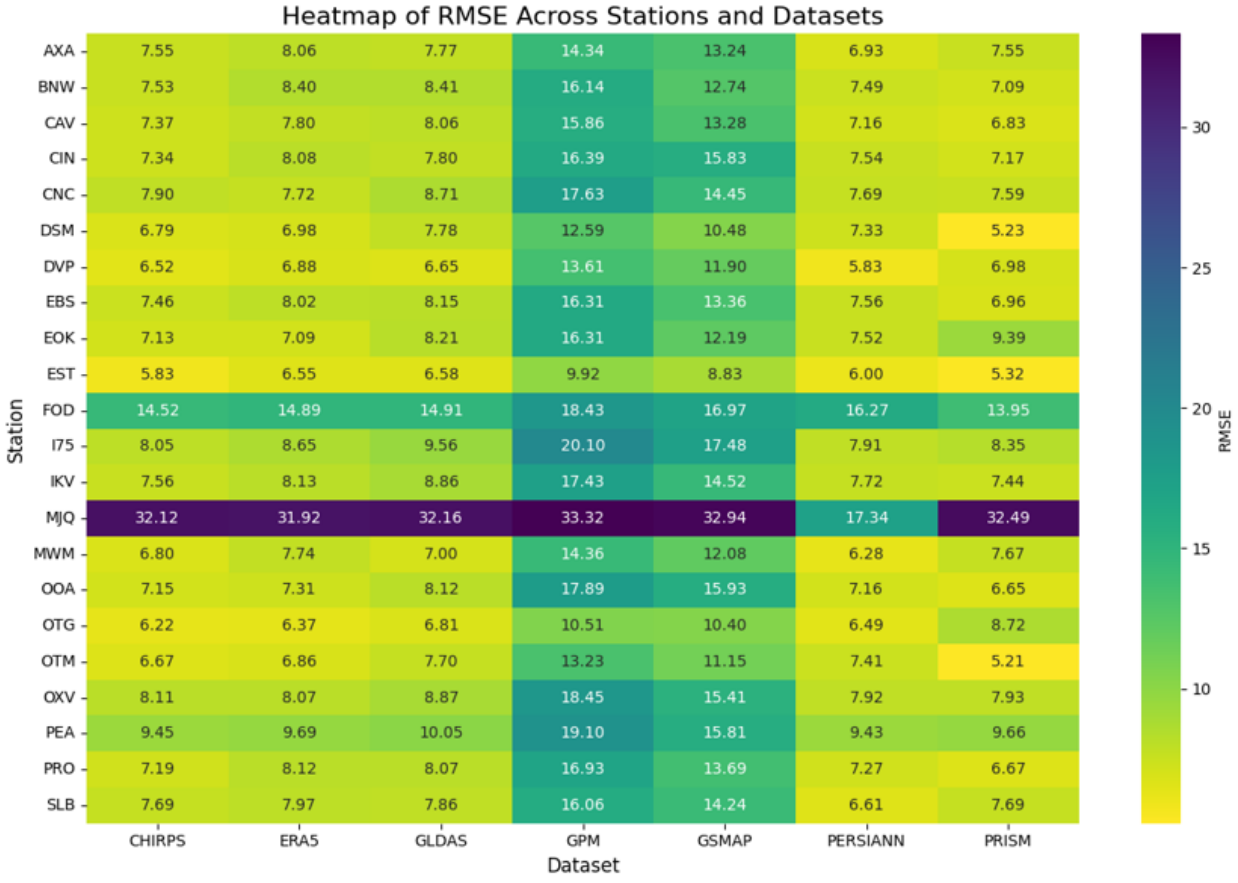

FIGURE 5 Heatmap of RMSE across stations and datasets for rainfall comparison.

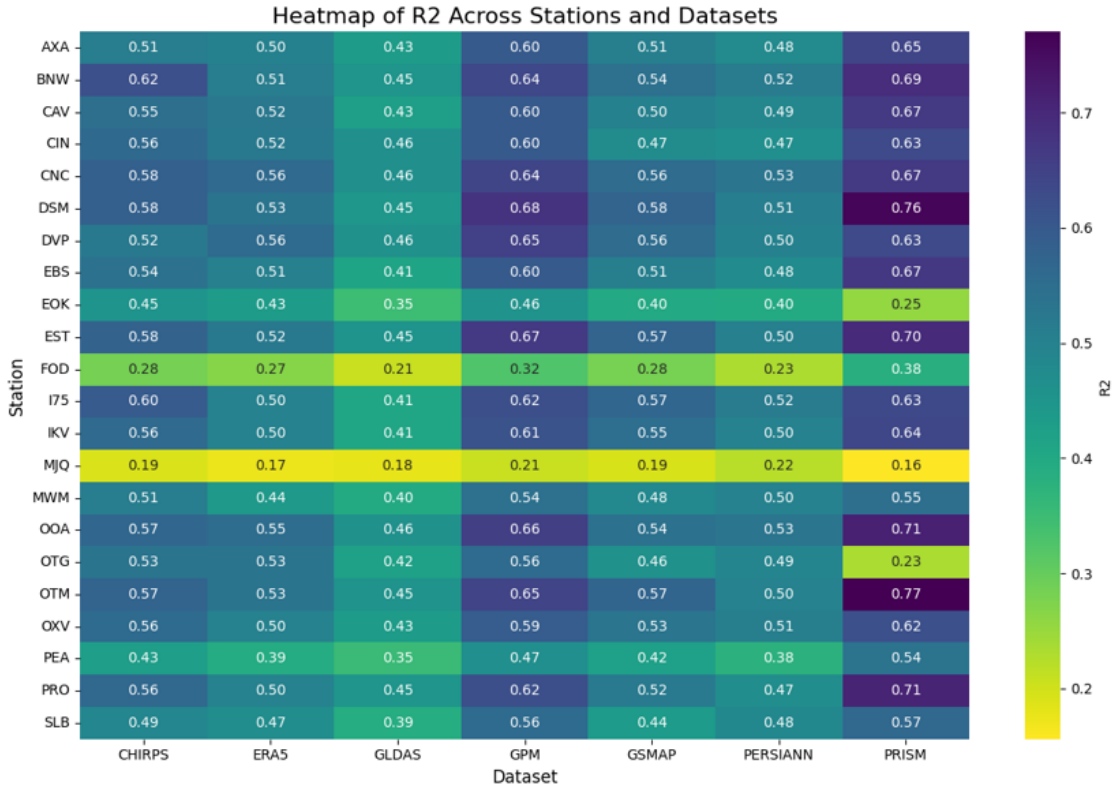

FIGURE 6 Heatmap of R^2 across stations and datasets for rainfall comparison.

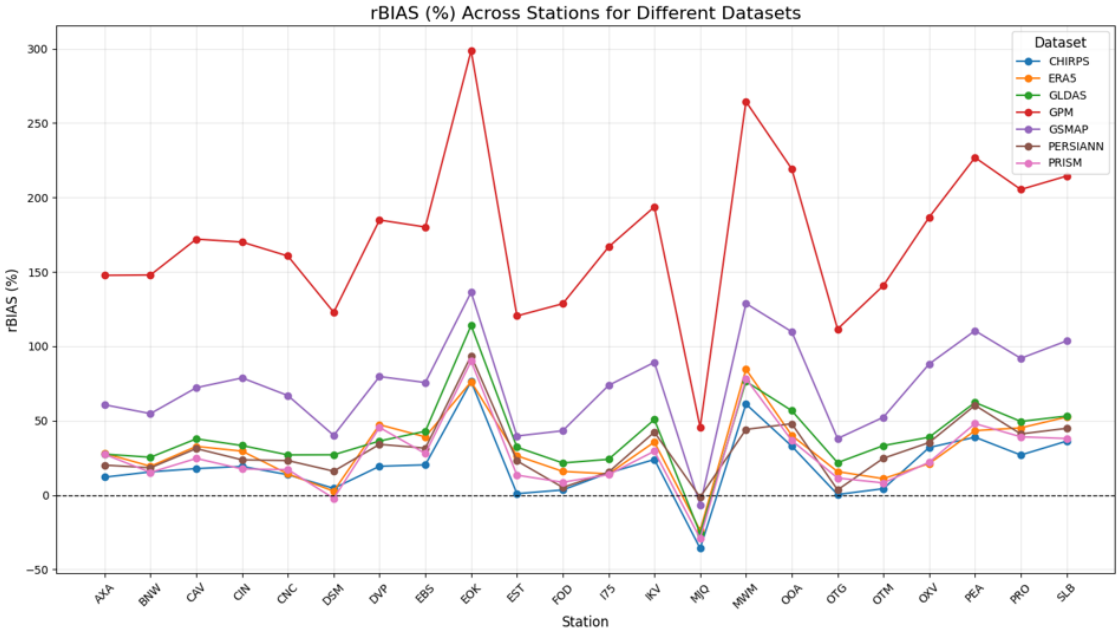

FIGURE 7 Line graph of rBias(%) across stations and datasets for rainfall comparison.

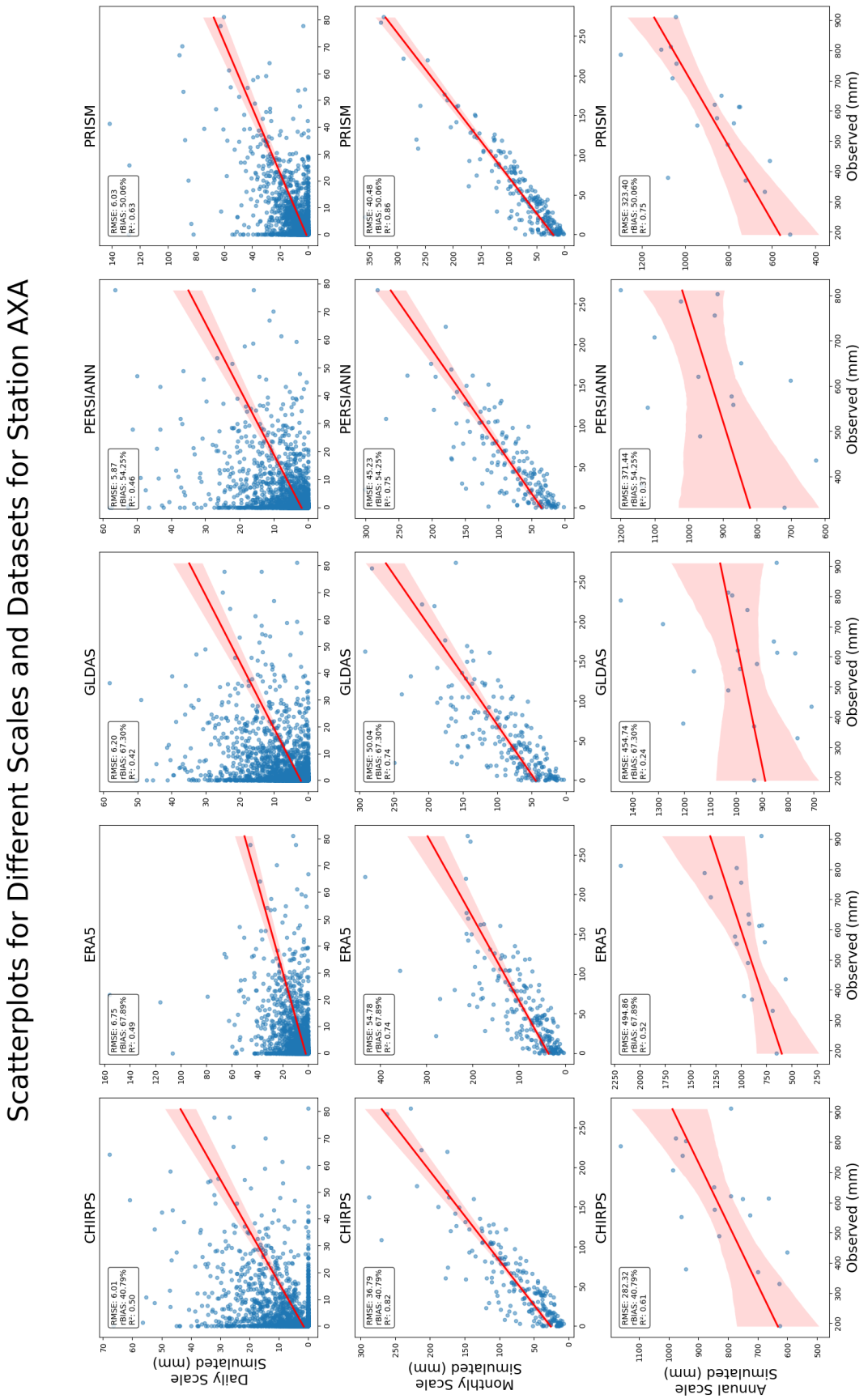

FIGURE 8 Scatter plots of R^2 at daily, monthly, and annual timescales.

CONCLUSIONS

This study highlights the performance of the open source gridded datasets (CHIRPS, ERA5, GLDAS, GPM, GSMAP, PERSIANN, PRISM) available in Google Earth Engine Catalog in the Des Moines River basin. The results has shown a vide variability of the datasets in different spatial and temporal scales indicating the importance of considering region of interest while choosing the gridded rainfall dataset for a Eco-hydrologic model. The study also emphasizes on using different statistical metrics in evaluating the overall performance of a dataset or a rain gauge station for better informed decisions.

Overall, PRISM dataset is found to be outperforming the other available gridded datasets considered in the study for the Des Moines River basin.

LIMITATIONS AND FUTURE DIRECTIONS

The different timezones of the datasets is a major limitation in the current study as the UTC timezone varies with 5-6 hours with CDT/CST in the Des Moines River basin. Therefore, converting the observed data and PRISM dataset into UTC timezone can give better comparison of the datasets in daily and sub-daily timescales.

The PRISM dataset can be used to extract the precipitation data at the centroids of each sub-basin shown in figure 1 and the values can be used as input to the SWAT+ model.

ACKNOWLEDGMENTS

The authors would like to thank **Dr. Anuj Sharma** for giving an opportunity to work on the project for the course **CE-5900 section 16 AI4CCEE** which has been a very helpful for our research.

SUPPLEMENTARY DOCUMENTS

- **Supplementary document 1:** Java Script to extract precipitation from Gridded Rainfall datasets using GEE.
- **Supplementary document 2:** Python Script to extract ground-based observed precipitation from IEM dataset using data scraping technique.
- **Supplementary document 3:** Python Script used to perform data analysis.

REFERENCES

1. Shokati, H., M. Mashal, A. A. Noroozi, and S. Mirzaei, Evaluating the Accuracy of Precipitation Products Over Utah, United States, Using the Google Earth Engine Platform. *Desert*, Vol. 28, No. 1, 2023, pp. 145–162.
2. Rincón-Avalos, P., A. Khouakhi, O. Mendoza-Cano, J. L.-D. l. Cruz, and K. M. Paredes-Bonilla, Evaluation of satellite precipitation products over Mexico using Google Earth Engine. *Journal of Hydroinformatics*, Vol. 24, No. 4, 2022, pp. 711–729.
3. Mankin, K. R., S. Mehan, T. R. Green, and D. M. Barnard, Review of Gridded Climate Products and Their Use in Hydrological Analyses Reveals Overlaps, Gaps, and Need for More Objective Approach to Model Forcings. *Hydrology and Earth System Sciences Discussions*, Vol. 2024, 2024, pp. 1–36.
4. Massmann, C., Evaluating the Suitability of Century-Long Gridded Meteorological Datasets for Hydrological Modeling. *Journal of Hydrometeorology*, Vol. 21, No. 11, 2020, pp. 2565–2580.
5. Banerjee, A., R. Chen, M. E. Meadows, R. Singh, S. Mal, and D. Sengupta, An analysis of long-term rainfall trends and variability in the uttarakhand himalaya using google earth engine. *Remote Sensing*, Vol. 12, No. 4, 2020, p. 709.
6. Paluba, D., V. Bližňák, M. Müller, and P. Štych, EVALUATION OF PRECIPITATION DATASETS AVAILABLE IN GOOGLE EARTH ENGINE ON A DAILY BASIS FOR CZECHIA, 2024.