Large-scale comparison of machine and statistical learning algorithms for blending gridded satellite and earth-observed precipitation data

Georgia Papacharalampous, Hristos Tyralis, Anastasios Doulamis, and Nikolaos Doulamis

National Technical University of Athens, School of Rural, Surveying and Geoinformatics Engineering

Abstract

An established way for improving the accuracy of gridded satellite precipitation products is to "correct" them by exploiting ground-based precipitation measurements, together with machine and statistical learning algorithms. Such corrections are made in regression settings, where the ground-based measurements are the dependent variable and the satellite data are predictor variables. Comparisons of machine and statistical learning algorithms in the direction of obtaining the most useful precipitation datasets by performing such corrections are regularly conducted in the literature. Nonetheless, in most of these comparisons, a small number of machine and statistical learning algorithms are considered. Also, small geographical regions and limited time periods are examined. Thus, the results provided tend to be of local importance and to not offer more general guidance. To provide results that are generalizable, we compared eight state-of-the-art machine and statistical learning algorithms in correcting satellite precipitation data for the entire contiguous United States and for a 15-year period. We used monthly data from the PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) gridded dataset and the Global Historical Climatology Network monthly database, version 2 (GHCNm). Our results suggest that extreme gradient boosting (XGBoost) and random forests are more accurate than the remaining algorithms, which can be ordered as follows from the best to the worst ones: Bayesian regularized feed-forward neural networks, multivariate adaptive polynomial splines (poly-MARS), gradient boosting machines (gbm), multivariate adaptive regression splines (MARS), feed-forward neural networks, linear regression.

This poster is based on Papacharalampous et al. (2023).

1. Previous studies on the topic

Study	Time scale	Spatial scale	Algorithms
He et al. (2016)	Hourly	South-western, central, north-eastern and south-eastern United States	Random forests
Meyer et al. (2016)	Daily	Germany	Random forests, artificial neural networks, support vector regression
Tao et al. (2016)	Daily	Central United States	Deep learning
Yang et al. (2016)	Daily	Chile	Quantile mapping
Baez-Villanueva et al. (2020)	Daily	Chile	Random forests
Chen et al. (2020a)	Daily	Dallas–Fort Worth in the United States	Deep learning
Chen et al. (2020b)	Daily	Xijiang basin in China	Geographically weighted ridge regression
Rata et al. (2020)	Annual	Chéliff watershed in Algeria	Kriging
Chen et al. (2021)	Monthly	Sichuan Province in China	Artificial neural networks, geographically weighted regression, kriging, random forests
Nguyen et al. (2021)	Daily	South Korea	Random forests
Shen and Yong (2021)	Annual	China	Gradient boosting decision trees, random forests, support vector regression
Zhang et al. (2021)	Daily	China	Artificial neural networks, extreme learning machines, random forests, support vector regression
Chen et al. (2022)	Daily	Coastal mountain region in the western United States	Deep learning
Fernandez-Palomino et al. (2022)	Daily	Ecuador and Peru	Random forests
Lin et al. (2022)	Daily	Three Gorges Reservoir area in China	Adaptive boosting decision trees, decision trees, random forests
Yang et al. (2022)	Daily	Kelantan river basin in Malaysia	Deep learning
Zandi et al. (2022)	Monthly	Alborz and Zagros mountain ranges in Iran	Artificial neural networks, locally weighted linear regression, random forests, stacked generalization, support vector regression
Militino et al. (2023)	Daily	Navarre in Spain	K-nearest neighbors, random forests, artificial neural networks

2. Summary of methods and metrics

Algorithms for spatial interpolation

- Linear regression (Hastie et al. 2009, pp 43–55)
- Multivariate adaptive regression splines (MARS; Friedman 1991, 1993)
- Multivariate adaptive polynomial splines (poly-MARS; Kooperberg et al. 1997, Stone et al. 1997)
- Random forests (Breiman 2001, Tyralis et al. 2019)
- Gradient boosting machines (gbm; Friedman 2001, Mayr et al. 2014, Tyralis and Papacharalampous 2021)
- Extreme gradient boosting (XGBoost; Chen and Guestrin 2016 , Tyralis and Papacharalampous 2021)
- Feed-forward neural networks (Ripley 1996, pp 143–180)
- Feed-forward neural networks with Bayesian regularization (MacKay 1992)

Variable importance metric

Random forests' permutation importance

Evaluation metrics

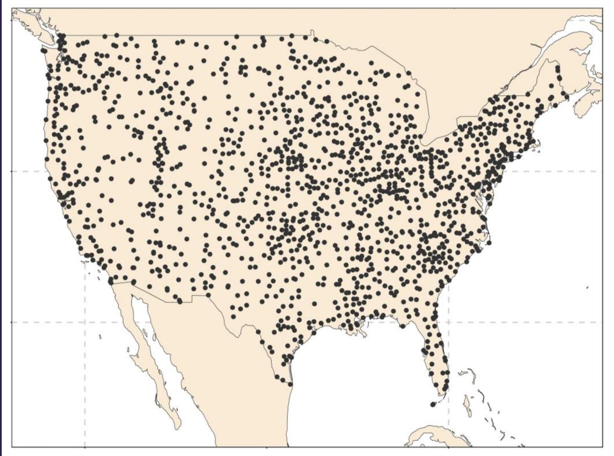
Median squared error → rankings, mean relative improvements, mean rankings

The comparison is made in a five-fold cross-validation setting.

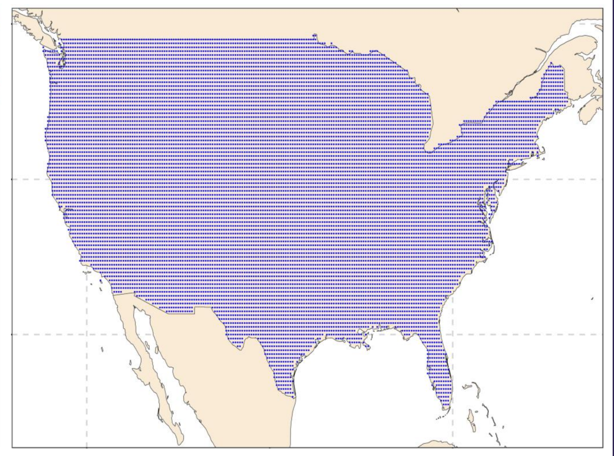
3. Summary of data

- **Total monthly precipitation data** from:
  - a) the Global Historical Climatology Network monthly database, version 2 (GHCNm; Peterson and Vose 1997); and
  - b) daily precipitation data of the current operational PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) system.
- **Elevation data** from the Amazon Web Services (AWS) Terrain Tiles application.

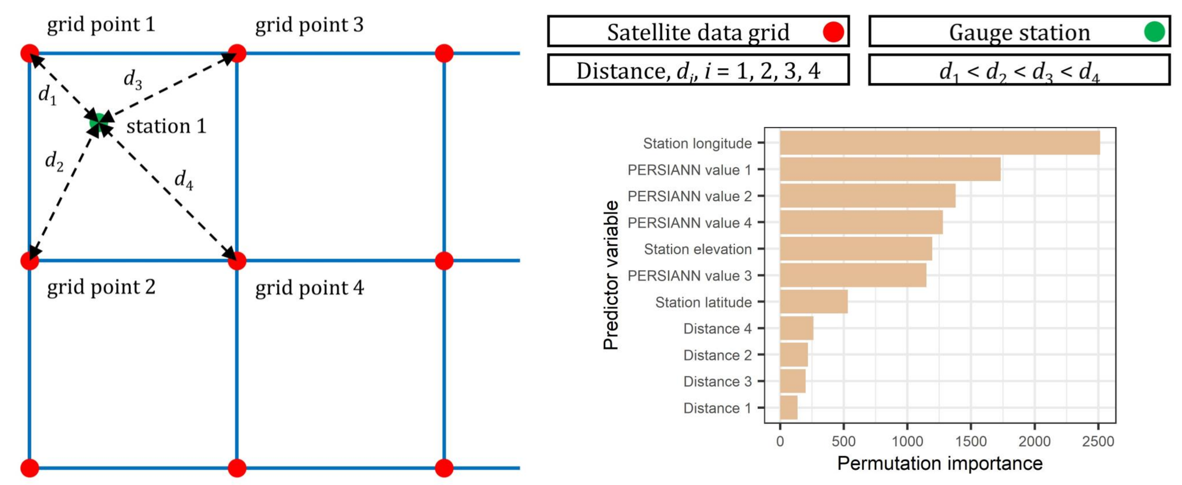
1 421 stations with data in the period 2001–2015



PERSIANN grid with data in the period 2001–2015



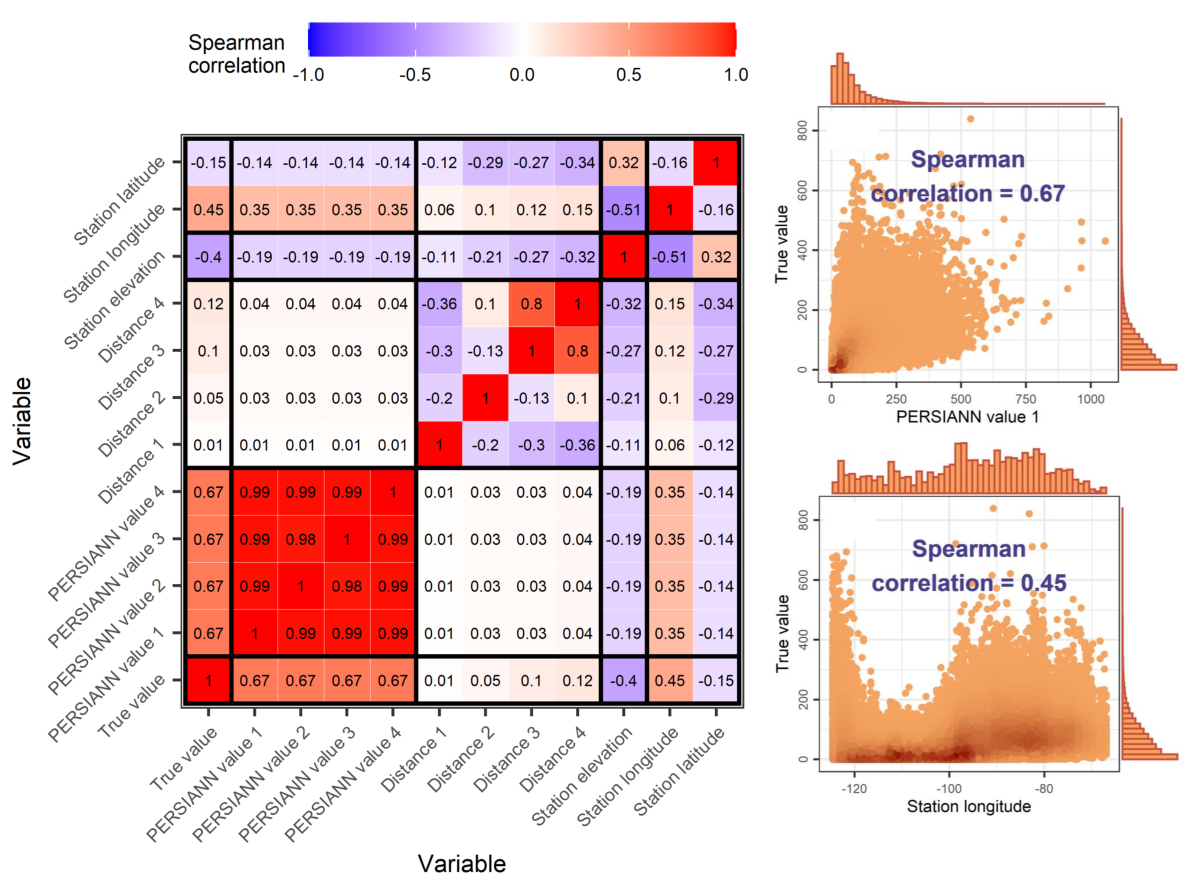
4. Regression setting and importance of predictors



- Grid points are the geographical locations with satellite data.
- Stations are the geographical locations with ground-based measurements.
- Distances 1−4 are the distances d_i, i = 1, 2, 3, 4.
- PERSIANN values 1−4 are the total monthly precipitation values at the grid points 1−4, respectively.

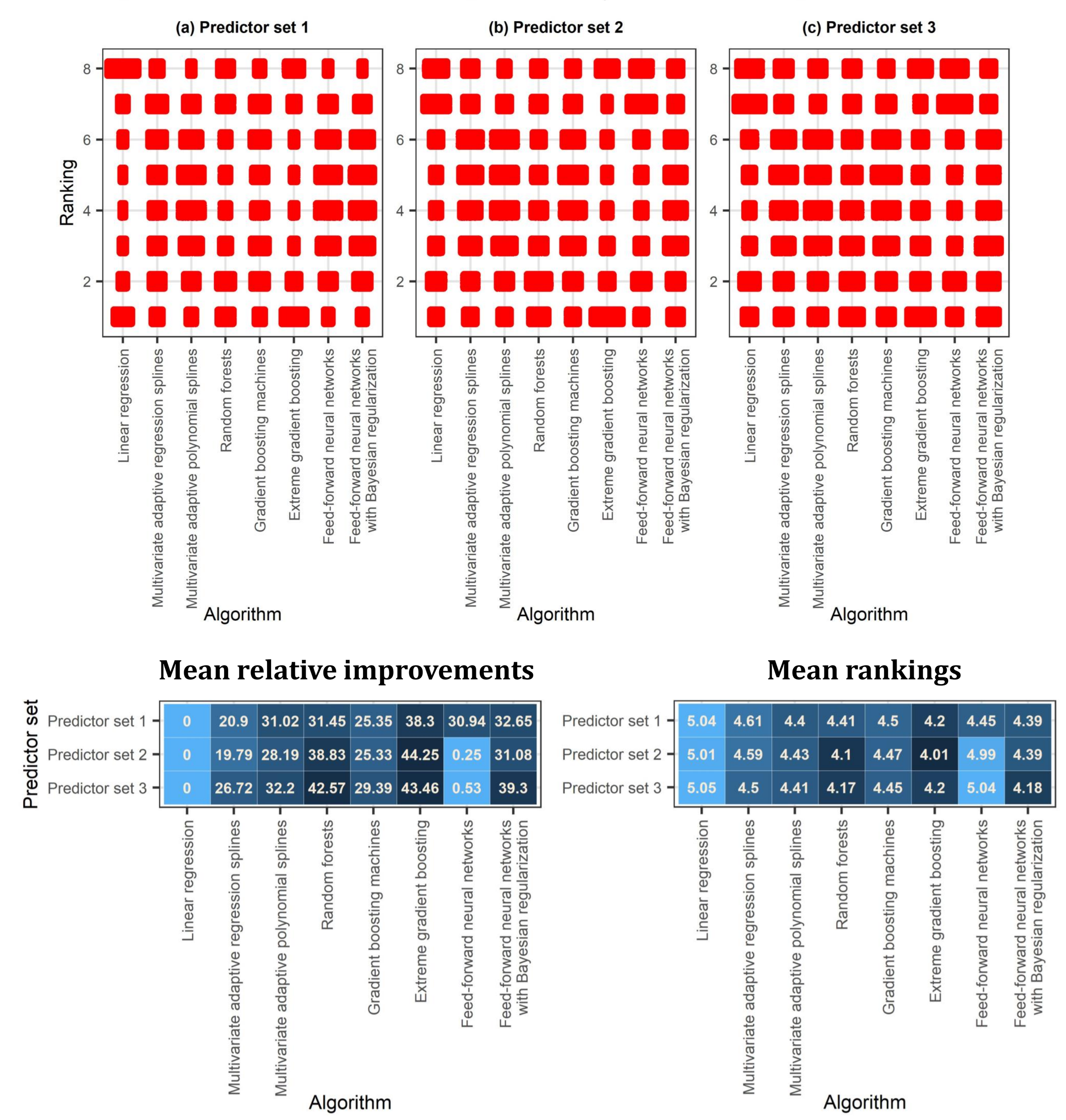
Predictor variable	Predictor set 1	Predictor set 2	Predictor set 3
PERSIANN value 1	✓	✓	✓
PERSIANN value 2	✓	✓	✓
PERSIANN value 3	✓	✓	✓
PERSIANN value 4	✓	✓	✓
Distance 1	×	✓	✓
Distance 2	×	✓	✓
Distance 3	×	✓	✓
Distance 4	×	✓	✓
Station elevation	✓	✓	✓
Station longitude	×	×	✓
Station latitude	×	×	✓

5. Relationships between variables

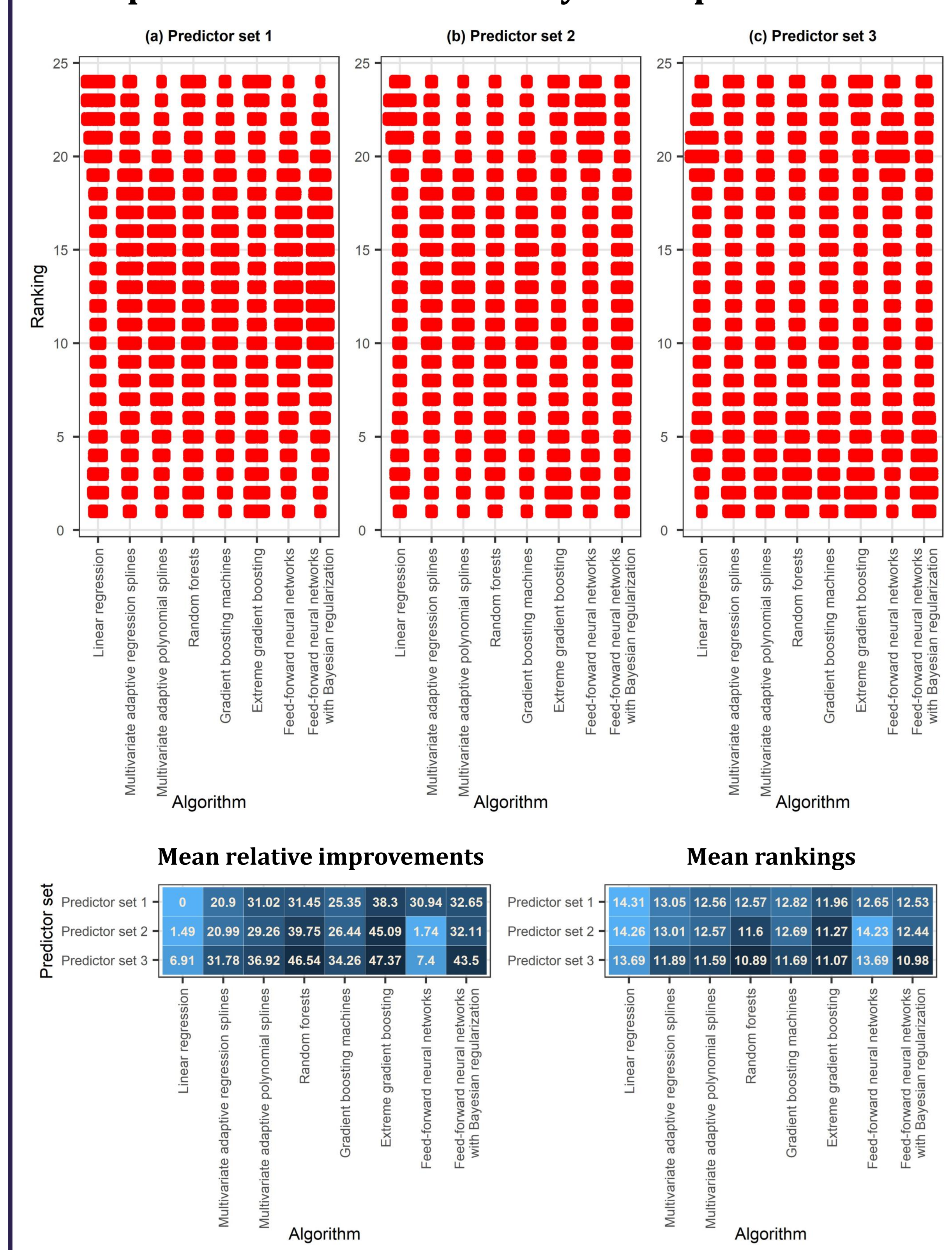


6. Comparison of algorithms

Computations made separately for each predictor set



Computations made collectively for all predictor sets



7. Summary of findings

- Extreme gradient boosting (XGBoost) and random forests are the most accurate algorithms.
- The former algorithm was found to be more accurate than the latter to a small extent based on the majority of the scores.
- The remaining algorithms can be ordered from the best- to the worst-performing as follows:
  - feed-forward neural networks with Bayesian regularization;
  - multivariate adaptive polynomial splines (poly-MARS);
  - gradient boosting machines (gbm);
  - multivariate adaptive regression splines (MARS);
  - feed-forward neural networks; and
  - linear regression.

8. Funding

This work was conducted in the context of the research project BETTER RAIN (BEnefiTTing from machine lEarning algoRithms and concepts for correcting satellite RAINfall products). This research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project Number: 7368).

References

Baez-Villanueva OM, Zambrano-Bigiarini M, Beck HE, McNamara I, Ribbe L, Nauditt A, Birkel C, Verbist K, Giraldo-Osorio JD, Xuan Thinh N (2020) RF-MEP: A novel random forest method for merging gridded precipitation products and ground-based measurements. Remote Sensing of Environment 239:111606. https://doi.org/10.1016/j.rse.2019.111606.
Breiman L (2001) Random forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324.
Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 785–794. https://doi.org/10.1145/2939672.2939785.
Chen H, Chandrasekar V, Cifelli R, Xie P (2020a) A machine learning system for precipitation estimation using satellite and ground radar network observations. IEEE Transactions on Geoscience and Remote Sensing 58(2):982–994. https://doi.org/10.1109/TGRS.2019.2942280.
Chen S, Xiong L, Ma Q, Kim J-S, Chen J, Xu C-Y (2020b) Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation products based on the geographically weighted ridge regression method. Journal of Hydrology 589:125156. https://doi.org/10.1016/j.jhydrol.2020.125156.
Chen C, Hu B, Li Y (2021) Easy-to-use spatial random-forest-based downscaling-calibration method for producing precipitation data with high resolution and high accuracy. Hydrology and Earth System Sciences 25(11):5667–5682. https://doi.org/10.5194/hess-25-5667-2021.
Chen H, Sun L, Cifelli R, Xie P (2022) Deep learning for bias correction of satellite retrievals of orographic precipitation. IEEE Transactions on Geoscience and Remote Sensing 60:4104611. https://doi.org/10.1109/TGRS.2021.3105438.
Fernandez-Palomino CA, Hattermann FF, Krysanova V, Lobanova A, Vega-Jácome F, Lavado W, Santini W, Aybar C, Bronstert A (2022) A novel high-resolution gridded precipitation dataset for Peruvian and Ecuadorian watersheds: Development and hydrological evaluation. Journal of Hydrometeorology 23(3):309–336. https://doi.org/10.1175/JHM-D-20-0285.1.
Friedman JH (1991) Multivariate adaptive regression splines. The Annals of Statistics 19(1):1–67. https://doi.org/10.1214/aos/1176347963.
Friedman JH (1993) Fast MARS. Stanford University, Department of Statistics. Technical Report 110. https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/LCS%20110.pdf.
Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451.
Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7.
He X, Chaney NW, Schleiss M, Sheffield J (2016) Spatial downscaling of precipitation using adaptable random forests. Water Resources Research 52(10):8217–8237. https://doi.org/10.1002/2016WR019034.
Kooperberg C, Bose S, Stone CJ (1997) Polychotomous regression. Journal of the American Statistical Association 92(437):117–127. https://doi.org/10.1080/01621459.1997.10473608.
Lin Q, Peng T, Wu Z, Guo J, Chang W, Xu Z (2022) Performance evaluation, error decomposition and tree-based machine learning error correction of GPM IMERG and TRMM 3B42 products in the Three Gorges reservoir area. Atmospheric Research 268:105988. https://doi.org/10.1016/j.atmosres.2021.105988.
MacKay DJC (1992) Bayesian interpolation. Neural computation 4(3):415–447. https://doi.org/10.1162/neco.1992.4.3.415.
Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms: From machine learning to statistical modelling. Methods of Information in Medicine 53(6):419–427. https://doi.org/10.3414/ME13-01-0122.
Meyer H, Kühnlein M, Appelhans T, Nauss T (2016) Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmospheric Research 169:424–433. https://doi.org/10.1016/j.atmosres.2015.09.021.
Militino AF, Ugarte MD, Pérez-Goya U (2023) Machine learning procedures for daily interpolation of rainfall in Navarre (Spain). Studies in Systems, Decision and Control 445:399–413. https://doi.org/10.1007/978-3-031-04137-2_34.
Nguyen GV, Le X-H, Van LN, Jung S, Yeon M, Lee G (2021) Application of random forest algorithm for merging multiple satellite precipitation products across South Korea. Remote Sensing 13(20):4033. https://doi.org/10.3390/rs13204033.
Papacharalampous GA, Tyralis H, Doulamis A, Doulamis N (2023) Comparison of machine learning algorithms for merging gridded satellite and earth-observed precipitation data. Water 15(4):634. https://doi.org/10.3390/w15040634.
Peterson TC, Vose RS (1997) An overview of the Global Historical Climatology Network temperature database. Bulletin of the American Meteorological Society 78(12):2837–2849. https://doi.org/10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.
Rata M, Douaoui A, Larid M, Douaik A (2020) Comparison of geostatistical interpolation methods to map annual rainfall in the Chéliff watershed, Algeria. Theoretical and Applied Climatology 141(3–4):1009–1024. https://doi.org/10.1007/s00704-020-03218-z.
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge. https://doi.org/10.1017/cbo9780511812651.
Shen Z, Yong B (2021) Downscaling the GPM-based satellite precipitation retrievals using gradient boosting decision tree approach over Mainland China. Journal of Hydrology 602:126803. https://doi.org/10.1016/j.jhydrol.2021.126803.
Stone CJ, Hansen MH, Kooperberg C, Truong YK (1997) Polynomial splines and their tensor products in extended linear modeling. Annals of Statistics 25(4):1371–1470. https://doi.org/10.1214/aos/1031594728.
Tao Y, Gao X, Hsu K, Sorooshian S, Ihler A (2016) A deep neural network modeling framework to reduce bias in satellite precipitation products. Journal of Hydrometeorology 17(3):931-945. https://doi.org/10.1175/JHM-D-15-0075.1.
Tyralis H, Papacharalampous G (2021) Boosting algorithms in energy research: A systematic review. Neural Computing and Applications 33(21):14101-14117. https://doi.org/10.1007/s00521-021-05995-8.
Tyralis H, Papacharalampous G, Langousis A (2019) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910.
Yang Z, Hsu K, Sorooshian S, Xu X, Braithwaite D, Verbist KMJ (2016) Bias adjustment of satellite-based precipitation estimation using gauge observations: A case study in Chile. Journal of Geophysical Research: Atmospheres 121(8):3790-3806. https://doi.org/10.1002/2015JD024540.
Yang X, Yang S, Tan ML, Pan H, Zhang H, Wang G, He R, Wang Z (2022) Correcting the bias of daily satellite precipitation estimates in tropical regions using deep neural network. Journal of Hydrology 608:127656. https://doi.org/10.1016/j.jhydrol.2022.127656.
Zandi O, Zahraie B, Nasseri M, Behrangi A (2022) Stacking machine learning models versus a locally weighted linear model to generate high-resolution monthly precipitation over a topographically complex area. Atmospheric Research 272:106159. https://doi.org/10.1016/j.atmosres.2022.106159.
Zhang L, Li X, Zheng D, Zhang K, Ma Q, Zhao Y, Ge Y (2021) Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. Journal of Hydrology 594:125969. https://doi.org/10.1016/j.jhydrol.2021.125969.