Quantifying predictive uncertainty in satellite precipitation data correction using ensemble learning

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

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

EGU General Assembly 2024

H.F.R.I. Hellenic Foundation for Research & Innovation

Abstract

We present the first ensemble learning methods for quantifying predictive uncertainty in satellite precipitation data correction, as well as the large-scale comparison of these methods. Ensemble learning was performed by combining in multiple ways a variety of machine learning algorithms that are particularly suited for the task of interest. Monthly precipitation data from across the contiguous United States supported the comparison, which predominantly relied on skill scores and referred to the ability of the ensemble learning methods in delivering predictive quantiles at many levels. The results allow the ordering from the best to the worst of the ensemble learning methods.

This poster is based on Papacharalampous et al. (2024b).

A review on predictive uncertainty estimation with machine learning can be found in Tyralis and Papacharalampous (2024).

1. Introduction

- Satellite data are not accurate but available at a dense spatial grid.
- Gauge-measured data are accurate but available in gauged locations.
- Thus, satellite and gauge-measured data are often merged for forming gridded precipitation data that are more accurate than the satellite ones.
- Still, uncertainty estimates for the data obtained in this way are sparsely provided.
- A few studies focus on the use of machine learning algorithms for providing such estimates (Bhuiyan et al. 2018, Zhang et al. 2022, Glawion et al. 2023, Tyralis et al. 2023, Papacharalampous et al. 2024a).
- This presentation outlines the first ensemble learning methods (Sagi and Rokach 2018; Wang et al. 2022) formulated for the task.
- Additionally, it presents the large-scale comparison of these methods.

2. Summary of methods and comparative framework

Ensemble learners (see 3)

Individual machine learning algorithms

- Quantile regression – QR (Koenker and Bassett 1978, Koenker 2005)
- Quantile regression forests – QRF (Meinshausen and Ridgeway 2006)
- Generalized random forests – GRF (Athey et al. 2019)
- Gradient boosting machines – GBM (Friedman 2001)
- Light gradient boosting machines – LightGBM (Ke et al. 2017)
- Quantile regression neural networks – QRNN (Taylor 2000, Cannon 2011)

Dependent variable

Gauge-measured precipitation at the location of interest

Predictor variables (see also 4 and 5)

- Distance-based weighted precipitation at the four PERSIANN grid points that are closest to the location of interest
- Distance-based weighted precipitation at the four IMERG grid points that are closest to the location of interest
- Elevation at the location of interest

Random division into 3 datasets of equal length

Quantile levels

{0.025, 0.050, 0.075, 0.100, 0.200, 0.300, 0.400, 0.500, 0.600, 0.700, 0.800, 0.900, 0.925, 0.950, 0.975}

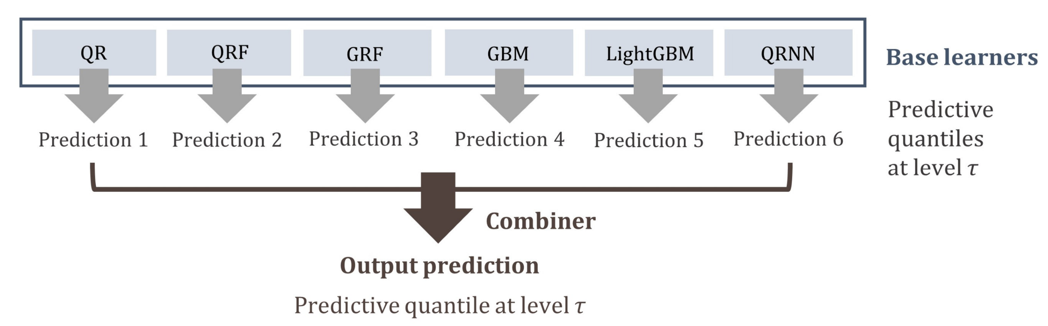
Metrics

- Quantile skill score
- Sample coverage

3. Ensemble learners

- Mean combiner
- Median combiner
- Best learner
- Stacking (Wolpert 1992) with QR as the combiner
- Stacking with QRF as the combiner
- Stacking with GRF as the combiner
- Stacking with GBM as the combiner
- Stacking with LightGBM as the combiner
- Stacking with QRNN as the combiner

How does each ensemble learner work?



What is the difference between the ensemble learners?

They utilize different combiners.

Which are the combiners introduced in this study?

Stacking with one of the following:

QR | QRF | GRF | GBM | LightGBM | QRNN

These combiners predict by optimizing the quantile loss and by using the predictions of the base learners as predictor variables.

Which are the benchmark combiners?

Mean combiner

This combiner delivers the mean of the predictions of the base learners.

Median combiner

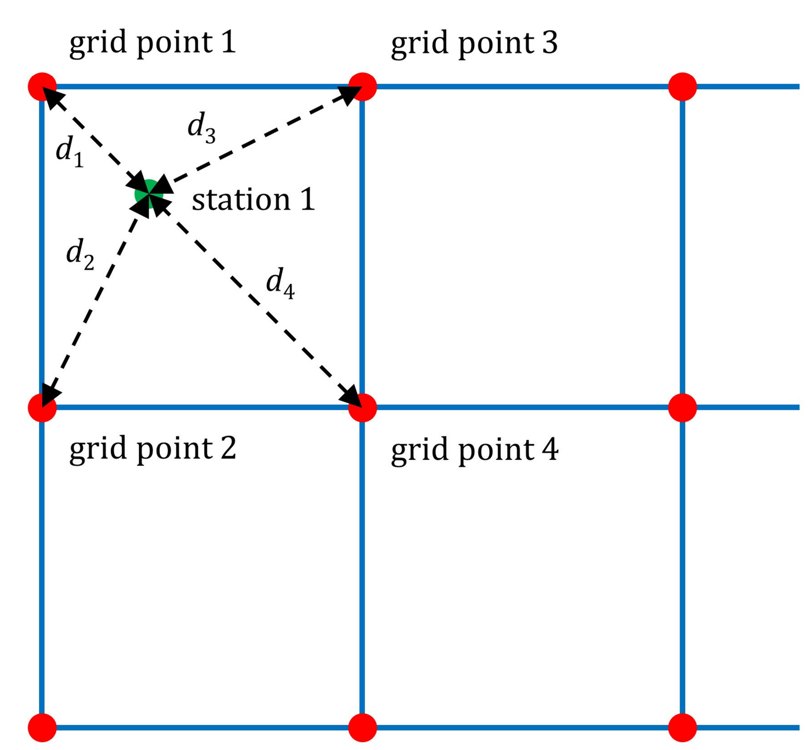
This combiner delivers the median of the predictions of the base learners.

Best learner

This combiner selects one base learner based on the quantile loss over train data and delivers its prediction.

4. Spatial interpolation problem formulation

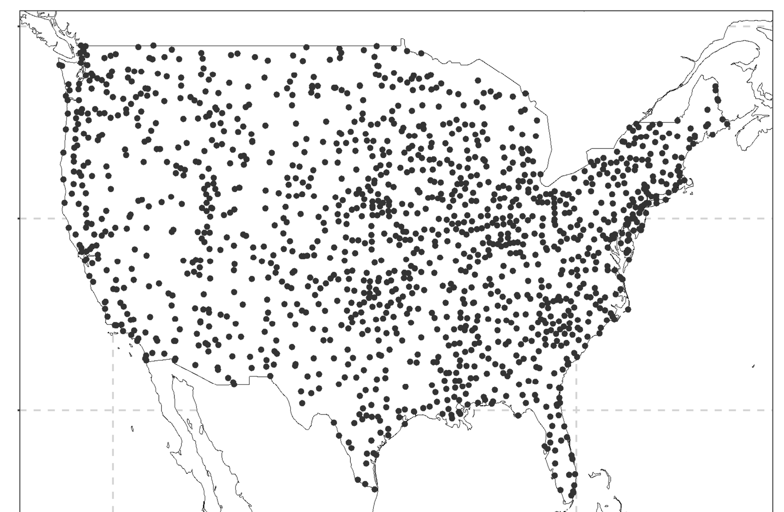
Satellite data grid	Gauge station
Distance, d_i, $i = 1, 2, 3, 4$	$d_1 < d_2 < d_3 < d_4$



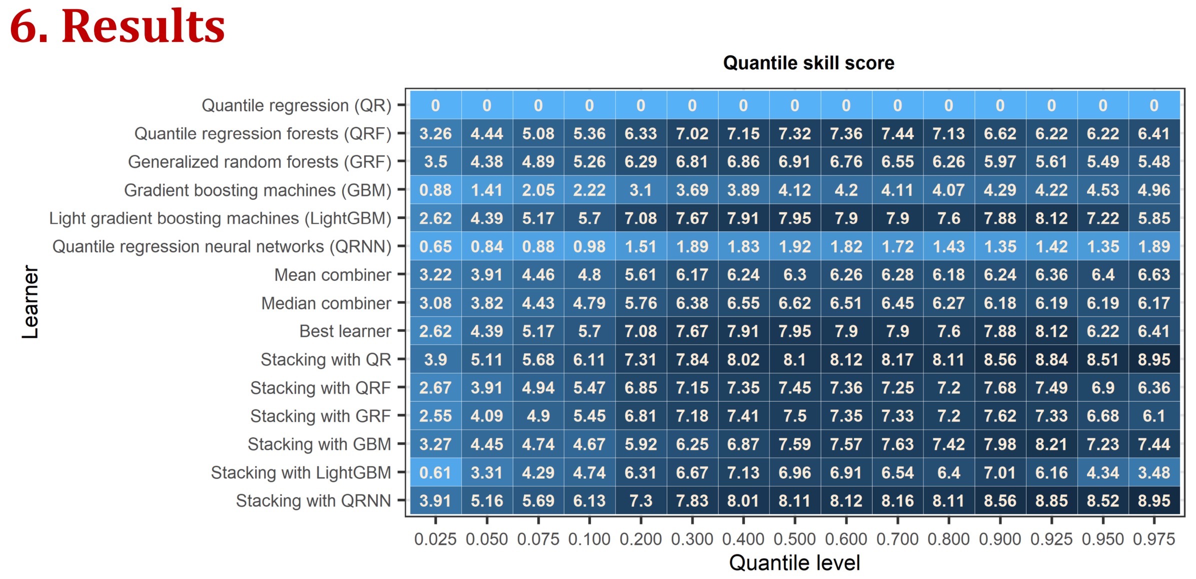
5. Summary of data

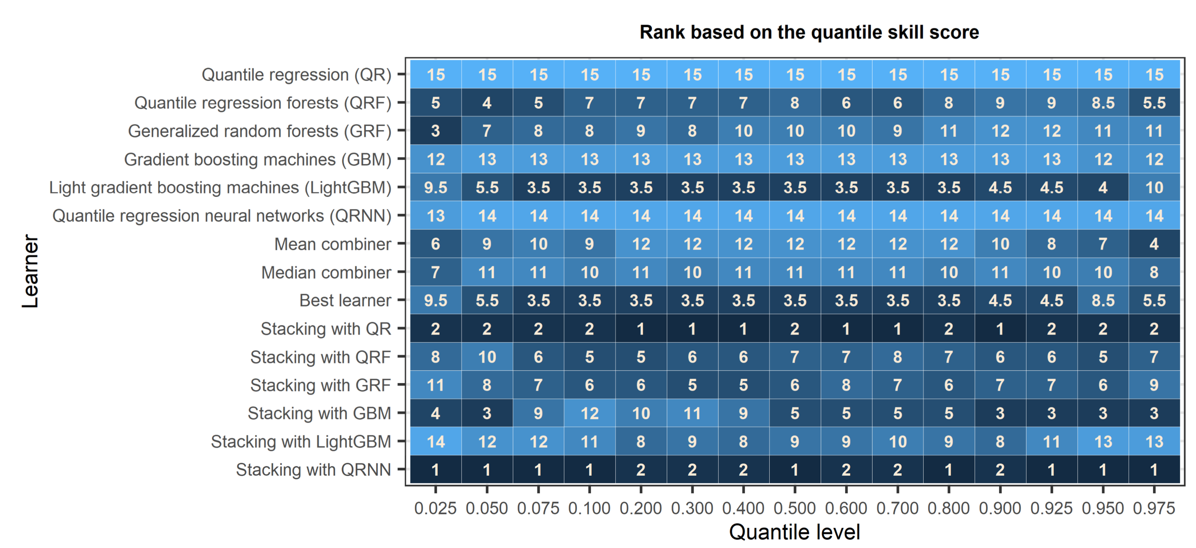
- ✓ **Total monthly precipitation data** from:
  - The Global Historical Climatology Network monthly database, version 2 (GHCNm; Peterson and Vose 1997).
  - Daily precipitation data of the current operational PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) system (Hsu et al. 1997, Nguyen et al. 2018, 2019).
  - Daily precipitation data of the GPM IMERG (Integrated Multi-satellitE Retrievals) late Precipitation L3 1 day 0.1 degree x 0.1 degree V06 dataset (Huffman et al. 2019).
- ✓ **Elevation data** from the Amazon Web Services (AWS) Terrain Tiles application.

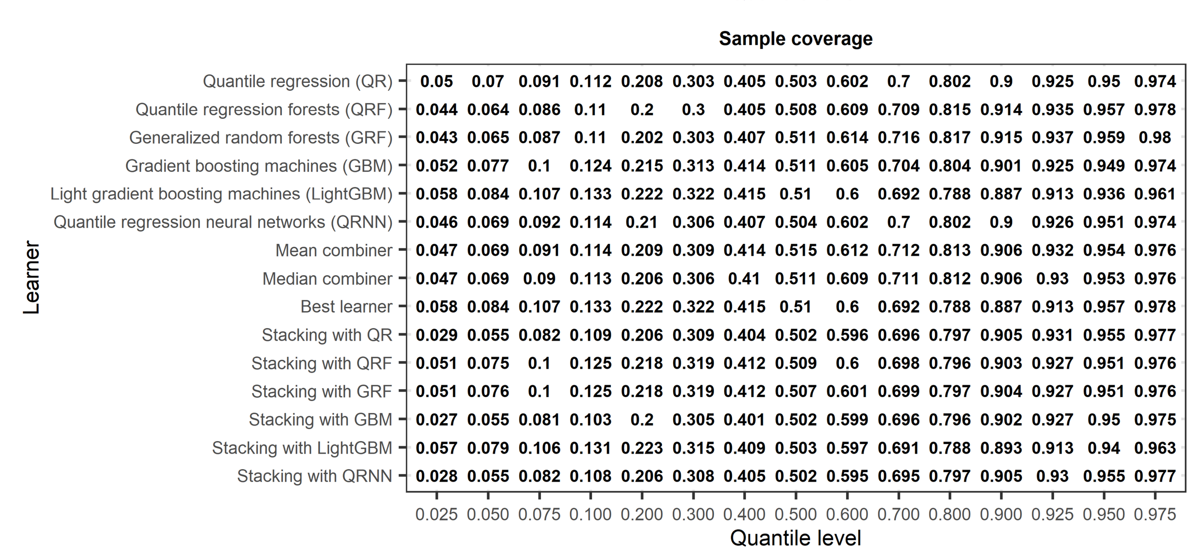
1 421 stations with data in the period 2001–2015



6. Results







7. Summary of findings and conclusions

- Overall, stacking with quantile regression and stacking with quantile regression neural networks are the best algorithms for the problem of interest.
- Still, the relative performance of the algorithms (ensemble learners and individual machine learning algorithms) should be expected to depend on the technical problem.
- Therefore, large-scale comparisons of the same algorithms in other technical problems would also be useful.

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

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. Annals of Statistics 47(2):1148–1178. https://doi.org/10.1214/18-AOS1709.

Bhuiyan MAE, Nikolopoulos EI, Anagnostou EN, Quintana-Seguí P, Barella-Ortiz A (2018) A nonparametric statistical technique for combining global precipitation datasets: Development and hydrological evaluation over the Iberian Peninsula. Hydrology and Earth System Sciences 22(2):1371–1389. https://doi.org/10.5194/hess-22-1371-2018.

Cannon AJ (2011) Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Computers and Geosciences 37(9):1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005.

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.

Glawion L, Polz J, Kunstmann HG, Fersch B, Chwala C (2023) spateGAN: Spatio-temporal downscaling of rainfall fields using a cGAN approach. https://doi.org/10.22541/essoar.167690003.33629126/v1.

Hsu K-L, Gao X, Sorooshian S, Gupta HV (1997) Precipitation estimation from remotely sensed information using artificial neural networks. Journal of Applied Meteorology 36(9):1176–1190. https://doi.org/10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2.

Huffman GJ, Stocker EF, Bolvin DT, Nelkin EJ, Tan J (2019) GPM IMERG Late Precipitation L3 1 day 0.1 degree x 0.1 degree V06, Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: [2022-10-12], https://doi.org/10.5067/GPM/IMERGDL/DAY/06.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30:3146–3154.

Koenker RW (2005) Quantile regression. Cambridge University Press, Cambridge, UK.

Koenker RW, Bassett Jr G (1978). Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643.

Meinshausen N, Ridgeway G (2006) Quantile regression forests. Journal of Machine Learning Research 7:983–999.

Nguyen P, Ombadi M, Sorooshian S, Hsu K, AghaKouchak A, Braithwaite D, Ashouri H, Rose Thorstensen A (2018) The PERSIANN family of global satellite precipitation data: A review and evaluation of products. Hydrology and Earth System Sciences 22(11):5801–5816. https://doi.org/10.5194/hess-22-5801-2018.

Nguyen P, Shearer EJ, Tran H, Ombadi M, Hayatbini N, Palacios T, Huynh P, Braithwaite D, Updegraff G, Hsu K, Kuligowski B, Logan WS, Sorooshian S (2019) The CHRS data portal, an easily accessible public repository for PERSIANN global satellite precipitation data. Scientific Data 6:180296. https://doi.org/10.1038/sdata.2018.296.

Papacharalampous GA, Tyralis H, Doulamis N, Doulamis A (2024a) Uncertainty estimation in satellite precipitation interpolation with machine learning. https://arxiv.org/abs/2311.07511 .

Papacharalampous GA, Tyralis H, Doulamis N, Doulamis A (2024b) Uncertainty estimation in spatial interpolation of satellite precipitation with ensemble learning. https://arxiv.org/abs/2403.10567.

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.

Sagi O, Rokach L (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249. https://doi.org/10.1002/widm.1249

Taylor JW (2000) A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of Forecasting 19(4):299–311. https://doi.org/10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V.

Tyralis H, Papacharalampous G (2024) Artificial Intelligence Review 57:94. https://doi.org/10.1007/s10462-023-10698-8.

Tyralis H, Papacharalampous GA, Doulamis N, Doulamis A (2023) Merging satellite and gauge-measured precipitation using LightGBM with an emphasis on extreme quantiles. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16:6969–6979. https://doi.org/10.1109/JSTARS.2023.3297013.

Wang X, Hyndman RJ, Li F, Kang Y (2022) Forecast combinations: An over 50-year review. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2022.11.005.

Wolpert DH (1992) Stacked generalization. Neural Networks 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Zhang Y, Ye A, Nguyen P, Analui B, Sorooshian S, Hsu K (2022) QRF4P-NRT: Probabilistic post-processing of near-real-time satellite precipitation estimates using quantile regression forests. Water Resources Research 58(5):e2022WR032117. https://doi.org/10.1029/2022WR032117.