EGU General Assembly 2024

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

Predictive uncertainty estimation in satellite precipitation data correction using machine learning

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

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

Abstract

Predictive uncertainty estimates for precipitation data acquired through merging satellite and ground-based observations are usually not provided. Here, we present the first benchmark experiments on the use of machine learning algorithms for fulfilling the task of delivering such estimates. These experiments compared six machine learning algorithms (i.e., quantile regression, quantile regression forests, generalized random forests, gradient boosting machines, light gradient boosting machines and quantile regression neural networks) and relied on 15-year-long monthly data that originate from across the contiguous United States. The comparison referred to the ability of the machine learning algorithms in delivering predictive quantiles at various levels. The results allow the ordering from the best to the worst of the machine learning algorithms for the problem of interest.

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

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

1. Introduction

- ○ Gridded precipitation data are often formed by merging satellite and gauge-measured data.
- ○ However, uncertainty estimates for the precipitation data acquired in this manner are rarely provided.
- ○ A few studies focus on how to provide such estimates by using machine learning algorithms (Bhuiyan et al. 2018, Zhang et al. 2022, Glawion et al. 2023, Tyralis et al. 2023).
- ○ Still, the benefits that machine learning can bring to the task of interest have not been explored so far through benchmark tests.
- ○ The work summarized by this presentation has filled in this gap.

2. Summary of methods and comparative framework

Machine learning algorithms

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

Dependent variable

Gauge-measured precipitation at the location of interest

Predictor variables

- Precipitation at the four PERSIANN grid points that are closest to the location of interest
- Precipitation at the four IMERG grid points that are closest to the location of interest
- Distances between the location of interest and each of its closets PERSIANN grid points
- Distances between the location of interest and each of its closets IMERG grid points
- Elevation at the location of interest

Five-fold cross-validation

Quantile levels

{0.025, 0.050, 0.100, 0.250, 0.500, 0.750, 0.900, 0.950, 0.975}

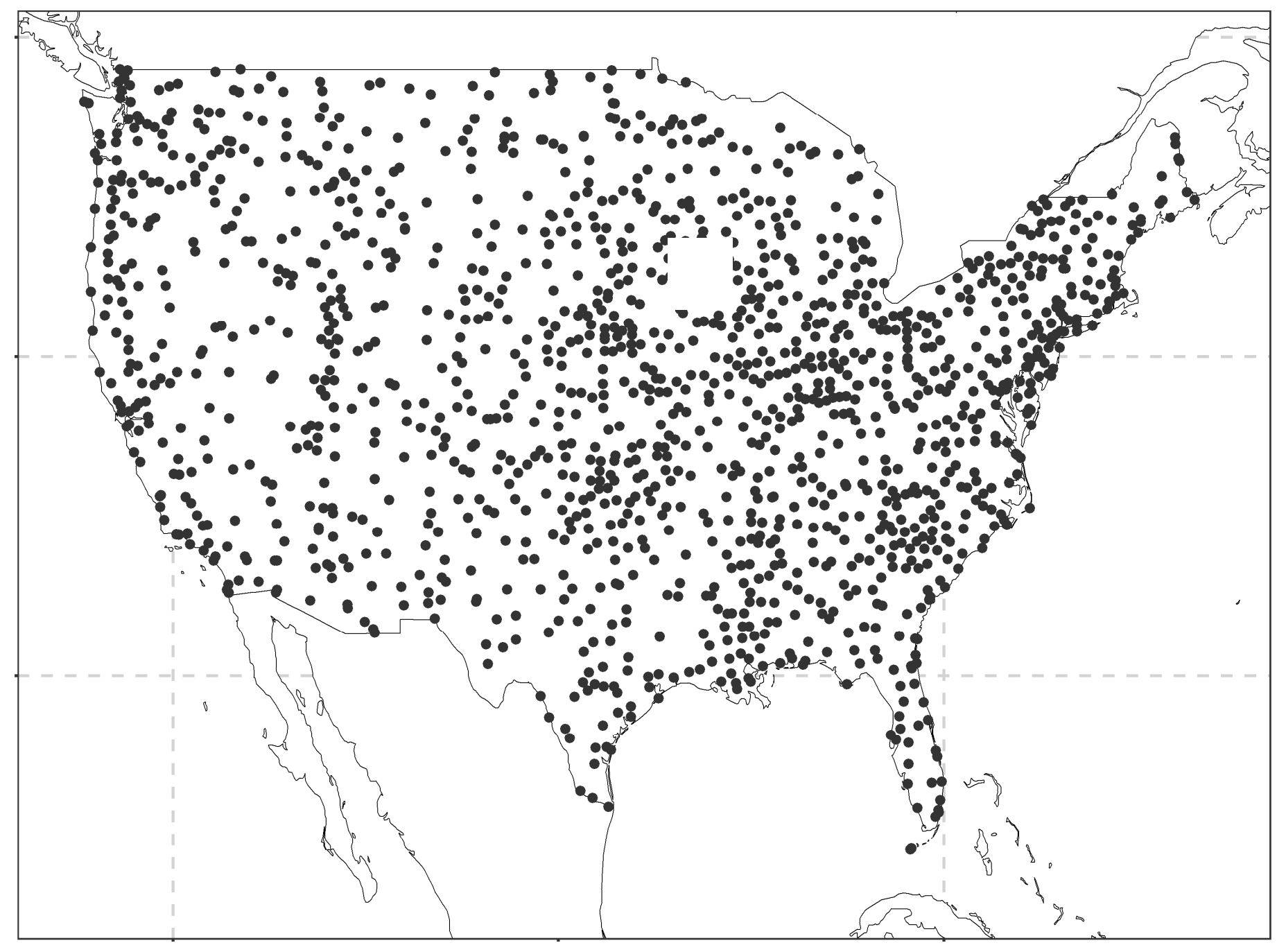
Metrics

- Sample coverage
- Quantile prediction skill
- Quantile scoring rule skill

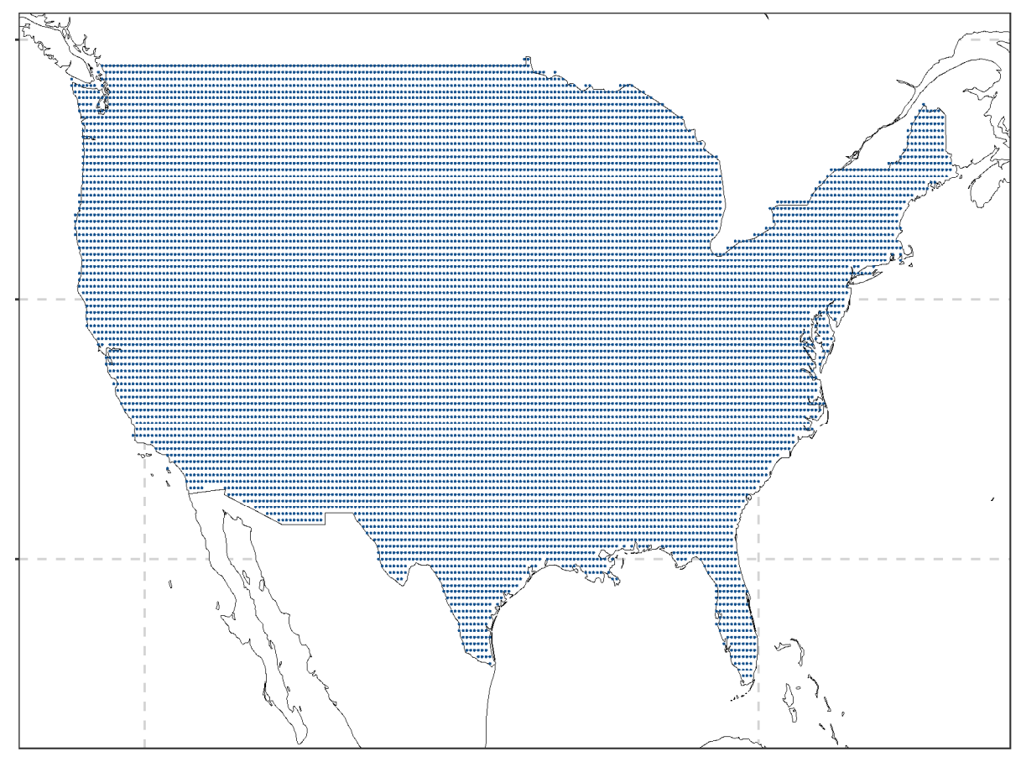
3. Summary of data

- ✓ **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.

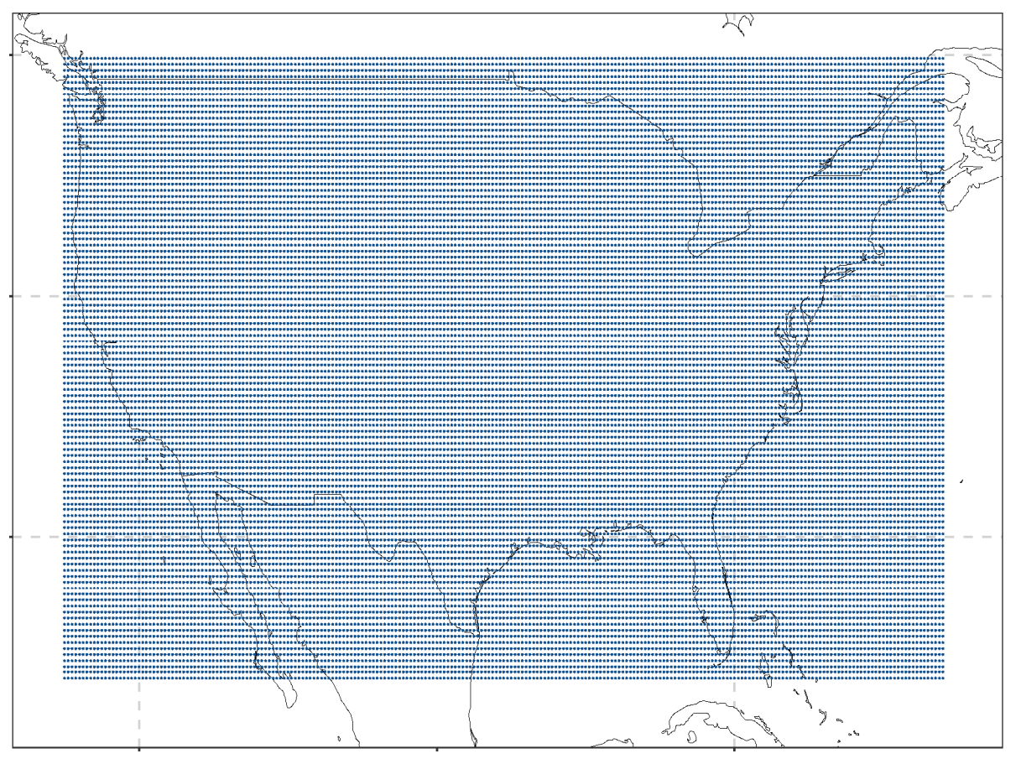
1 421 stations with data in the period 2001–2015



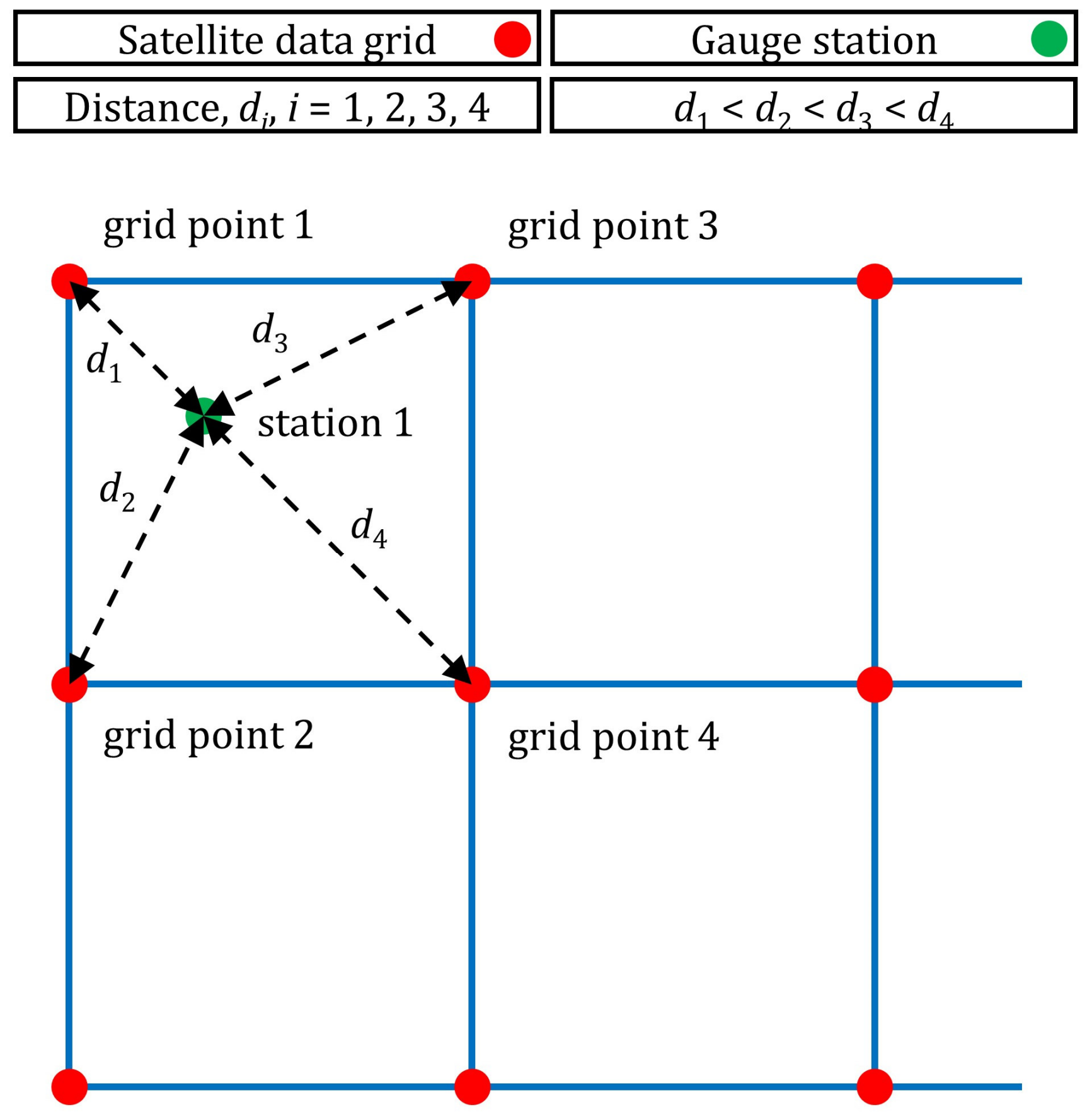
PERSIANN grid with data in the period 2001–2015



IMERG grid with data in the period 2001–2015

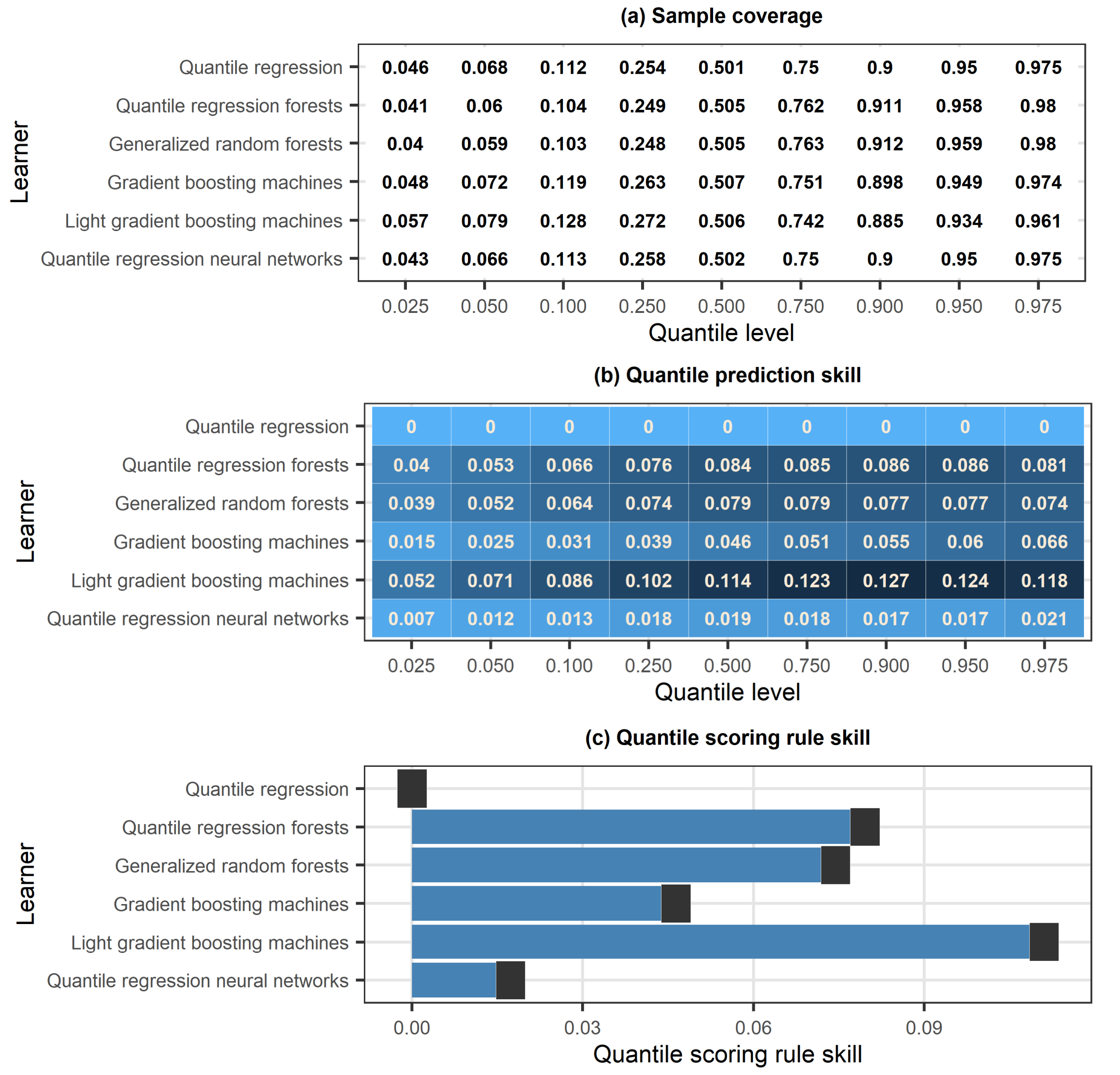


4. Spatial interpolation problem formulation



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

5. Comparison of machine learning algorithms



(a) Sample coverage

Learner	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975
Quantile regression	0.046	0.068	0.112	0.254	0.501	0.75	0.9	0.95	0.975
Quantile regression forests	0.041	0.06	0.104	0.249	0.505	0.762	0.911	0.958	0.98
Generalized random forests	0.04	0.059	0.103	0.248	0.505	0.763	0.912	0.959	0.98
Gradient boosting machines	0.048	0.072	0.119	0.263	0.507	0.751	0.898	0.949	0.974
Light gradient boosting machines	0.057	0.079	0.128	0.272	0.506	0.742	0.885	0.934	0.961
Quantile regression neural networks	0.043	0.066	0.113	0.258	0.502	0.75	0.9	0.95	0.975

Quantile level

(b) Quantile prediction skill

Learner	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975
Quantile regression	0	0	0	0	0	0	0	0	0
Quantile regression forests	0.04	0.053	0.066	0.076	0.084	0.085	0.086	0.086	0.081
Generalized random forests	0.039	0.052	0.064	0.074	0.079	0.079	0.077	0.077	0.074
Gradient boosting machines	0.015	0.025	0.031	0.039	0.046	0.051	0.055	0.06	0.066
Light gradient boosting machines	0.052	0.071	0.086	0.102	0.114	0.123	0.127	0.124	0.118
Quantile regression neural networks	0.007	0.012	0.013	0.018	0.019	0.018	0.017	0.017	0.021

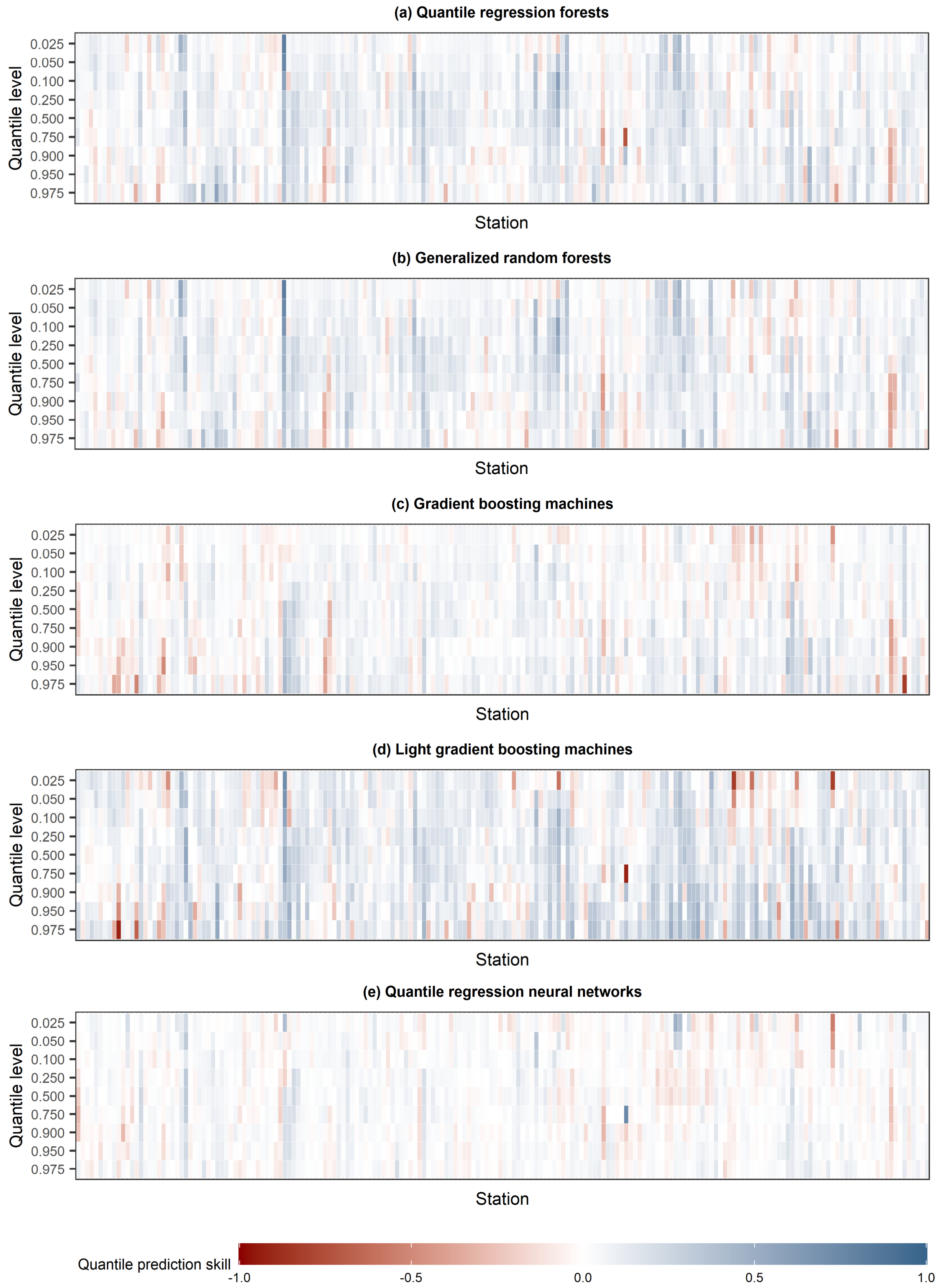
Quantile level

(c) Quantile scoring rule skill

Learner
Quantile regression
Quantile regression forests
Generalized random forests
Gradient boosting machines
Light gradient boosting machines
Quantile regression neural networks
0.00
0.03
0.06
0.09
Quantile scoring rule skill

6. Type-1 investigations across stations

Quantile prediction skill



(a) Quantile regression forests

(b) Generalized random forests

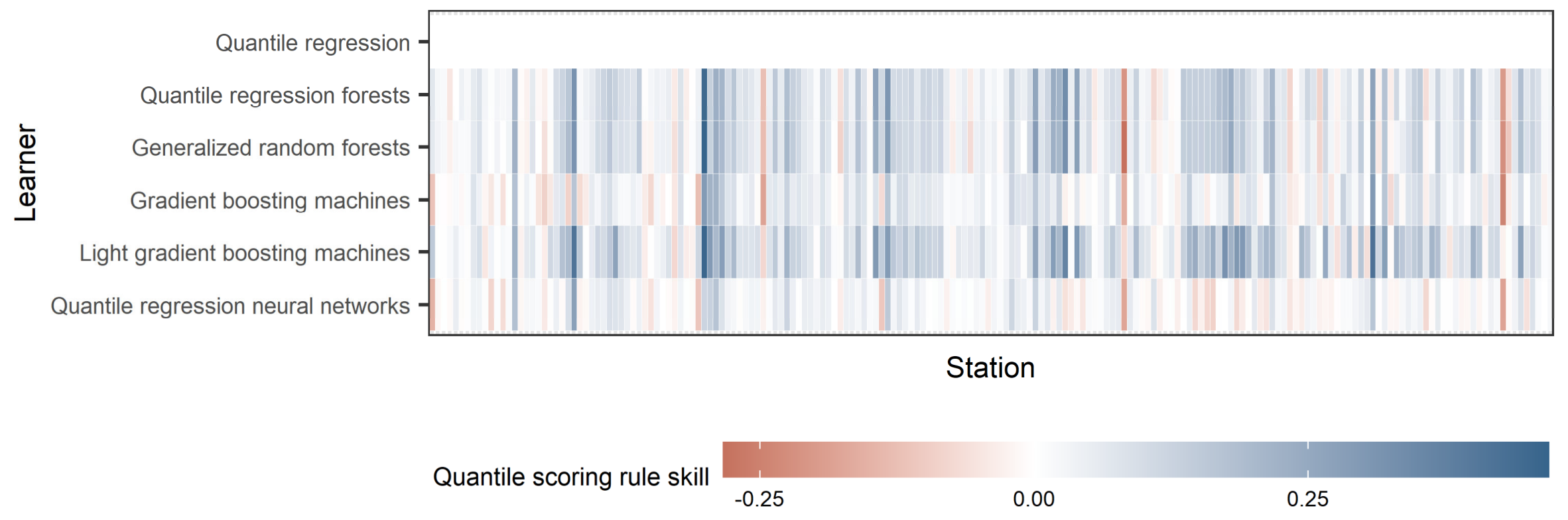
(c) Gradient boosting machines

(d) Light gradient boosting machines

(e) Quantile regression neural networks

7. Type-2 investigations across stations

Quantile scoring rule skill



8. Summary of findings

- ○ Light gradient boosting machines have the best performance.
- ○ The remaining algorithms can be ordered from the best to the worst as follows: quantile regression forests, generalized random forests, gradient boosting machines, quantile regression neural networks and quantile regression.

9. 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 (2024) Uncertainty estimation in satellite precipitation interpolation with machine learning. https://arxiv.org/abs/2311.07511 .

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.

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.

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.