Fusion of satellite precipitation products and ground-based measurements using LightGBM with a focus on extreme quantiles

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

[1]National Technical University of Athens, School of Rural, Surveying and Geoinformatics Engineering
Session HS3.1 (Hydroinformatics: Data analytics, machine learning, optimisation)

Abstract

Satellite precipitation products are not accurate in representing the actual precipitation measured by gauges. To improve their accuracy, machine learning algorithms are applied in regression settings with ground-based measurements as dependent variables and satellite precipitation data as predictor variables. Here we examine the case of light gradient-boosting machine (LightGBM) for correcting daily IMERG (Integrated Multi-satellitE Retrievals for GPM) and PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) precipitation data using daily precipitation measurements in the contiguous US. Our demonstration especially focuses on the estimation of quantiles of the conditional probability distribution of daily precipitation at given points, with emphasis on extreme values.

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

2. Introduction

- Economic constraints limit the extent as well as the density of spatial coverage of areas with rain gauge stations.
- Gridded satellite datasets are used as a substitute of observed precipitation in hydrological applications.
- Gridded satellite datasets provide inaccurate estimates of actual precipitation.
- Merging gridded datasets with rainfall gauge-based measurements is a solution.
- Merging is done by applying machine learning algorithms in regression settings.
- The state-of-the-art algorithm in these regression settings is Breiman's (2001) random forests.
- In most studies merging satellite data and station observations, spatial point predictions are issued and assessed using the squared error scoring function, the absolute error scoring function or related skill scores (e.g. NSE and KGE).
- Prediction of quantiles of the conditional probability distribution at a dense grid of quantile levels can provide an approximation of the full probability distribution (Tyralis and Papacharalampous 2021, 2022).
- Our focus is on extreme quantiles of the conditional probability distribution (see e.g. Curceac et al. 2020, Tyralis and Papacharalampous 2023).
- The aim of the manuscript is to solve the problem of probabilistic prediction of precipitation with an emphasis on extreme quantiles in spatial interpolation settings.
- To this end, we propose to apply the Light Gradient Boosting Machine (LightGBM) algorithm (Ke 2017) trained with the quantile scoring function (Koenker and Bassett Jr 1978).
- LightGBM is compared with quantile regression forests (Meinshausen 2006).

3. Methods

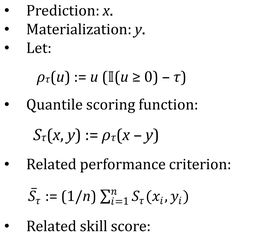
- LightGBM was trained using a quantile loss function (implementation: lightgbm R package, Shi et al. 2022).
- Qunatile regression forests (QRF) were used with default hyperparameter values (implementation: ranger R package, Wright 2022, Wright and Ziegler 2017).
- QRF is the reference algorithm for skill scores.
- LightGBM parameters optimized:

Parameter	Description	Values
`max_depth`	Max depth for tree model. `max_depth` can be used to limit the tree depth explicitly.	6, 8, 10
`min_data_in_leaf`	This is a very important parameter to prevent over-fitting in a leaf-wise tree. Its optimal value depends on the number of training samples and `num_leaves`. Setting it to a large value can avoid growing too deep a tree, but may cause under-fitting. In practice, setting it to hundreds or thousands is enough for a large dataset	20, 100, 200, 500, 1 000
`learning_rate`	Shrinkage rate. As a general rule, if one reduces `num_iterations`, then he should increase `learning_rate`	0.02, 0.05, 0.1
`num_iterations`	Number of iterations. The `num_iterations` parameter controls the number of boosting rounds that will be performed. Since LightGBM uses decision trees as the learners, this can also be thought of as "number of trees"	400
`num_leaves`	Max number of leaves in one tree. This is the main parameter to control the complexity of the tree model	20, 40, 60, 80, 100, 200, 500

4. Data

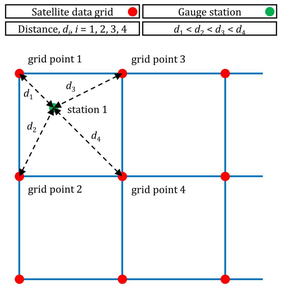
- Daily earth-observed precipitation retrieved from the Global Historical Climatology Network daily (GHCNd).
- Gridded satellite precipitation from the current operational PERSIANN system as well as the GPM IMERG late Precipitation dataset.
- Elevation data retrieved from the Amazon Web Services (AWS) Terrain Tiles application.
- Station data: 7 261 stations with **daily** precipitation in the period 2014-2015.

Stations



PERSIANN grid

5. Regression setting



Predictor variable	Predictor set
PERSIANN value 1	✓
PERSIANN value 2	✓
PERSIANN value 3	✓
PERSIANN value 4	✓
IMERG value 1	✓
IMERG value 2	✓
IMERG value 3	✓
IMERG value 4	✓
PERSIANN distance 1	✓
PERSIANN distance 2	✓
PERSIANN distance 3	✓
PERSIANN distance 4	✓
IMERG distance 1	✓
IMERG distance 2	✓
IMERG distance 3	✓
IMERG distance 4	✓
Longitude	✓
Latitude	✓
Station elevation	✓

6. Scoring functions and skill scores

- Quantile level of interest: τ.
- Prediction: x.
- Materialization: y.
- Let:

$$\rho_\tau(u) := u\,(\mathbb{I}(u \geq 0) - \tau)$$

- Quantile scoring function:

$$S_\tau(x, y) := \rho_\tau(x - y)$$

- Related performance criterion:

$$\bar{S}_\tau := (1/n)\sum_{i=1}^{n} S_\tau(x_i, y_i)$$

- Related skill score:

$$S_{\tau,\text{skill}} := 1 - \bar{S}_{\tau,\text{LightGBM}}/\bar{S}_{\tau,\text{QRF}}$$

- Frequency performance criterion:

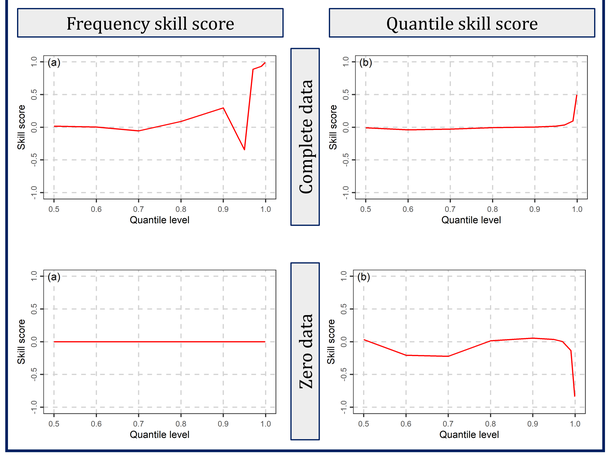
$$\overline{\text{FR}}_\tau := |(1/n)\sum_{i=1}^{n} \mathbb{I}(y_i \leq x_i) - \tau|$$

- Related skill score:

$$\text{FR}_{\tau,\text{skill}} := 1 - \overline{\text{FR}}_{\tau,\text{LightGBM}}/\overline{\text{FR}}_{\tau,\text{QRF}}$$

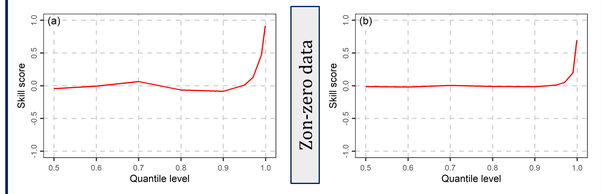
7. Skill scores

Frequency skill score | Quantile skill score



8. Skill scores

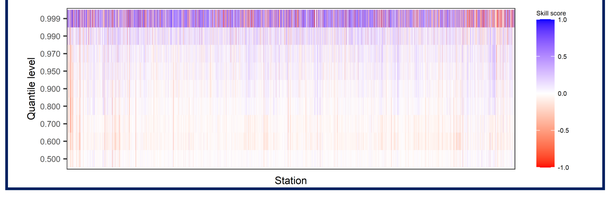
Frequency skill score | Quantile skill score



- For complete data and FR, algorithms are equal for $\tau \leq 0.8$, and LightGBM outperforms QRF for higher quantile levels.
- For complete data and quantile skill score, LightGBM outperforms QRF for $\tau \geq 0.97$.
- For zero data and FR, algorithms are equal.
- For zero data and quantile skill score, QRF outperforms LightGBM for $\tau \geq 0.97$.
- For non-zero data and FR, algorithms are equal for $\tau \leq 0.95$, and LightGBM outperforms QRF for higher quantile levels.
- For non-zero data and quantile skill score, LightGBM outperforms QRF for $\tau \geq 0.97$.

9. Skill scores at each station

- It is also of interest to understand how the algorithms perform at each station separately.
- Here, we examine the case of the quantile scoring function based skill score.
- Stations with skill scores lower than −1 were removed. The reason is that, some skill score values were as low as −10 or less, which would create some artefacts in the representation of the results. The conclusions are not affected by the removal,.
- Furthermore, we removed stations were both algorithms had a mean score equal to 0 (in which case the skill score is not defined).
- The skill score increases as the quantile level $\tau \to 1$.
- The skill score varies between stations at the same quantile level, although the variation is relatively small. A notable departure of the skill scores from 0 is observed for quantile levels $\tau \geq 0.97$.



10. Discussion

- LightGBM in general performs better compared to QRF when assessed with the quantile scoring function.
- LightGBM does not uniformly outperform QRF at all quantile levels.
  - At lower quantile levels, the two algorithms seem to behave similarly.
  - At higher quantile levels LightGBM clearly outperforms QRF.
- A possible explanation for the behaviour at lower quantile levels is based on the high proportion of zeros in the dataset. In particular, QRF is an algorithm based on bootstrapping therefore, it is possible to resample zero values. On the other hand, LightGBM is based on the minimization of the quantile scoring function that may be suboptimal when the dataset is highly intermittent.
- At extreme quantile levels, LightGBM clearly outperforms QRF with regards to all skill scores while the difference increases with increasing τ, while the skill score tends to 1 as $\tau \to 1$. A possible explanation is that QRF cannot predict values that are not in the range of the training set.
- While QRF outperforms LightGBM with regards to the quantile skill score at higher quantile levels when observed precipitation is zero, the inverse happens when observed precipitation is higher than zero. The performance of both algorithms in the complete test set favours LightGBM, since absolute values of quantile scores are lower in general when observed precipitation is zero compared to non-zero observed precipitation, consequently the largest part in the average score belongs to non-zero values.

11. Conclusions

- We proposed issuing probabilistic predictions of daily precipitation in spatial settings of merging gauge-based measurements and satellite precipitation products using LightGBM.
- LightGBM outperforms the state-of-the-art in such settings quantile regression forests when predicting extreme quantiles of the conditional probability distribution of the response variable, while both algorithms show similar performance when predicting quantiles at the centre of the probability distribution.
- The difference in the performance of the methods increases in favour of LightGBM as the quantile level (at which the methods are compared) increases and tends to 1.
- Confidence on the results is built through the comparison of the algorithms in a large dataset that includes observed precipitation in the contiguous US.
- An intuitive explanation of the results has also been provided.
- On the other hand, quantile regression forests have equalized when predicting quantiles at the centre of the conditional probability distribution, due to the highly intermittent nature of precipitation, combined with their bootstrap-based structure, that seems to be more suitable in this case compared to algorithm structures that are based on the quantile scoring function.

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

Breiman L (2001) Random forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Curceac S, Atkinson PM, Milne A, Wu L, Harris P (2020) Adjusting for conditional bias in process model simulations of hydrological extremes: An experiment using the North Wyke Farm Platform. Frontiers in Artificial Intelligence 3:82. https://doi.org/10.3389/frai.2020.565859.

Ke G, Meng Q, Finkey 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, Bassett Jr G (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643.

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

Shi Y, Ke G, Soukhavong D, Lamb J, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y, Titov N (2022) lightgbm: Light Gradient Boosting Machine. R package version 3.3.4. https://CRAN.R-project.org/package=lightgbm.

Tyralis H, Papacharalampous G (2021) Quantile-based hydrological modelling. Water 13(23):3420. https://doi.org/10.3390/w13233420.

Tyralis H, Papacharalampous G (2022) A review of probabilistic forecasting and prediction with machine learning. https://arxiv.org/abs/2209.08307.

Tyralis H, Papacharalampous G (2023) Hydrological post-processing for predicting extreme quantiles. Journal of Hydrology 617(Part C):129082. https://doi.org/10.1016/j.jhydrol.2023.129082.

Tyralis H, Papacharalampous GA, Doulamis N, Doulamis A (2023) Merging satellite and gauge-measured precipitation using LightGBM with an emphasis on extreme quantiles. https://arxiv.org/abs/2302.03606.

Wright MN (2022) ranger: A fast implementation of random forests. R package version 0.14.1. https://CRAN.R-project.org/package=ranger.

Wright MN, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software 77(1):1–17. https://doi.org/10.18637/jss.v077.i01.