

Random forests in water resources

Hristos Tyrallis, Georgia Papacharalampous, and Andreas Langousis

European Geosciences Union

EGU General Assembly 2021

Virtual Conference | 19-30 April 2021

Session NP4.1: Big Data and AI in the Earth Sciences



Abstract

Random forests is a supervised machine learning algorithm which has witnessed recently an exponential increase in its implementation in water resources. However, the existing implementations have been restricted in applications of Breiman's (2001) original algorithm to regression and classification models, while numerous developments could be also useful for solving diverse practical problems. Here we popularize random forests for the practicing hydrologist and present alternative random forests based algorithms and related concepts and techniques, which are underappreciated in hydrology. We review random forests applications in water resources and provide guidelines for the full exploitation of the potential of the algorithm and its variants. Relevant implementations of random forests related software in the R programming language are also presented.

History of random forests in hydrology

- Random forests appeared in [Breiman \(2001\)](#).
- Applications in several fields (agriculture, ecology, land cover classification, remote sensing and more).
- Data-driven models are important tools for hydrologists.
- Amongst data-driven models, random forests have a prominent position.
- [Parkhurst et al. \(2005\)](#) were the first to use random forests in hydrology.
- 203 papers using random forests in water science and technology during the period 2005-2018 (Scopus database).
- Several applications with diverse purposes.
- Following are based on the review of random forests in water science by [Tyralis et al. \(2019b\)](#).

Random forests and their variants

- Random forests are mainly used for prediction (in particular point predictions).
- Several variants exist:
 - Quantile regression forests for probabilistic predictions.
 - Conditional inference forests (unbiased variable importance measures).
 - Generalized random forests (generalization of random forests for adaptive, local estimation).
 - Distributional regression forests (version of generalized additive models for location, scale, and shape parameters (GAMLSS), using trees).
- Many more variants.

Reasons towards the use of random forests

- Increased predictive performance, as verified in competitions.
- They can capture non-linear dependencies between predictor and dependent variables.
- They are fast compared to other machine learning algorithms.
- They can be applied to large-scale problems.
- They are stable.
- They are robust to the inclusion of noisy predictor variables.
- They can handle highly correlated predictor variables.
- They can operate successfully when interactions are present.
- They are flexible.
- They are suitable for coping with high dimensional data.
- They can simultaneously incorporate continuous and categorical variables.
- Many more reasons.

Why to use random forests with caution

- Theoretical properties of random forests are not fully understood.
- Random forests cannot extrapolate outside the training range.
- Variable importance metrics are not always informative/reliable.
- Random forests are harder to interpret/understand compared to single trees.
- The automation of random forests may result in a slight decrease of their predictive performance compared to e.g. highly parameterized tree-based boosting.
- They cannot model adequately datasets with imbalanced data.
- Their original version is not suited for causal inference.

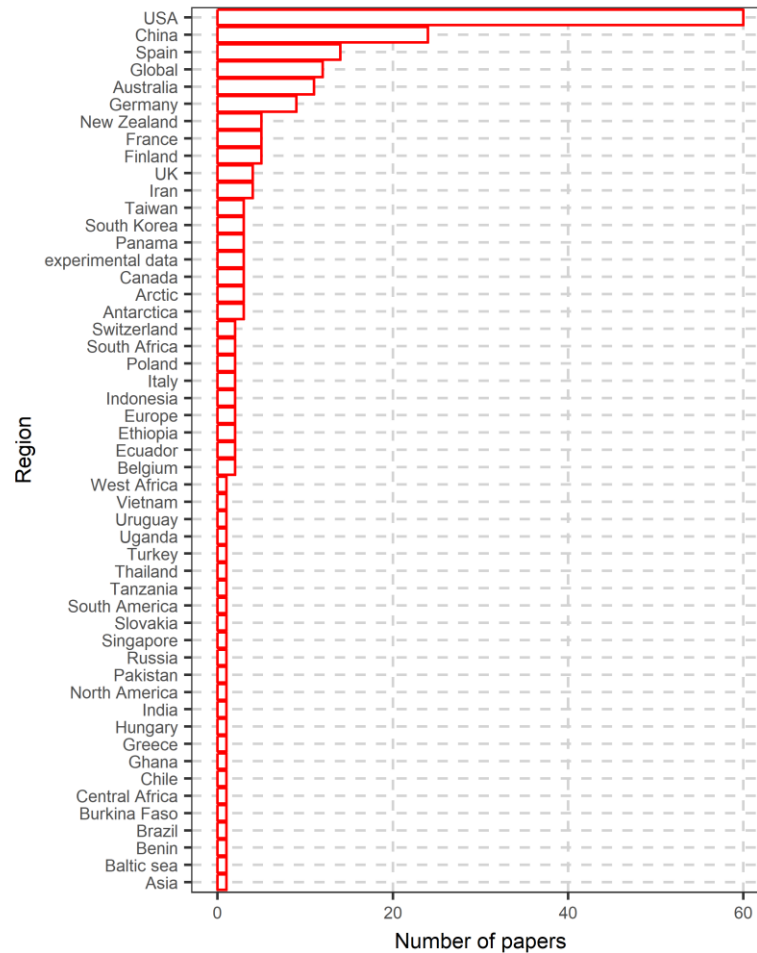
R software

- Some important R packages.
 - `randomForest` (main package).
 - `ranger` (fast implementation).
 - `quantregForest` (for quantile regression forests).
 - `party` (unbiased variable importance).
 - `grf` (generalized random forests).
 - `Boruta` (variable selection).
- Many more packages exist.

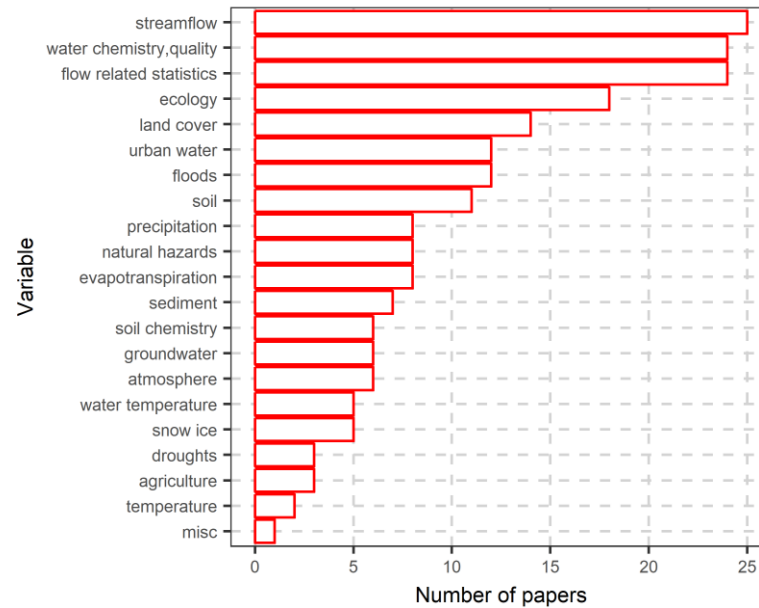
Application of random forests in water science



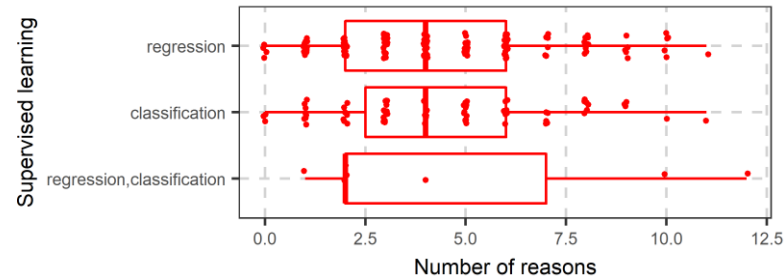
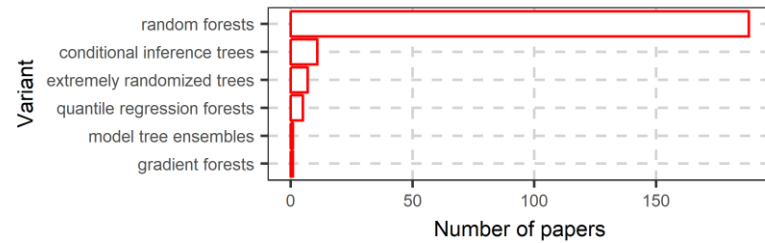
Application of random forests in water science



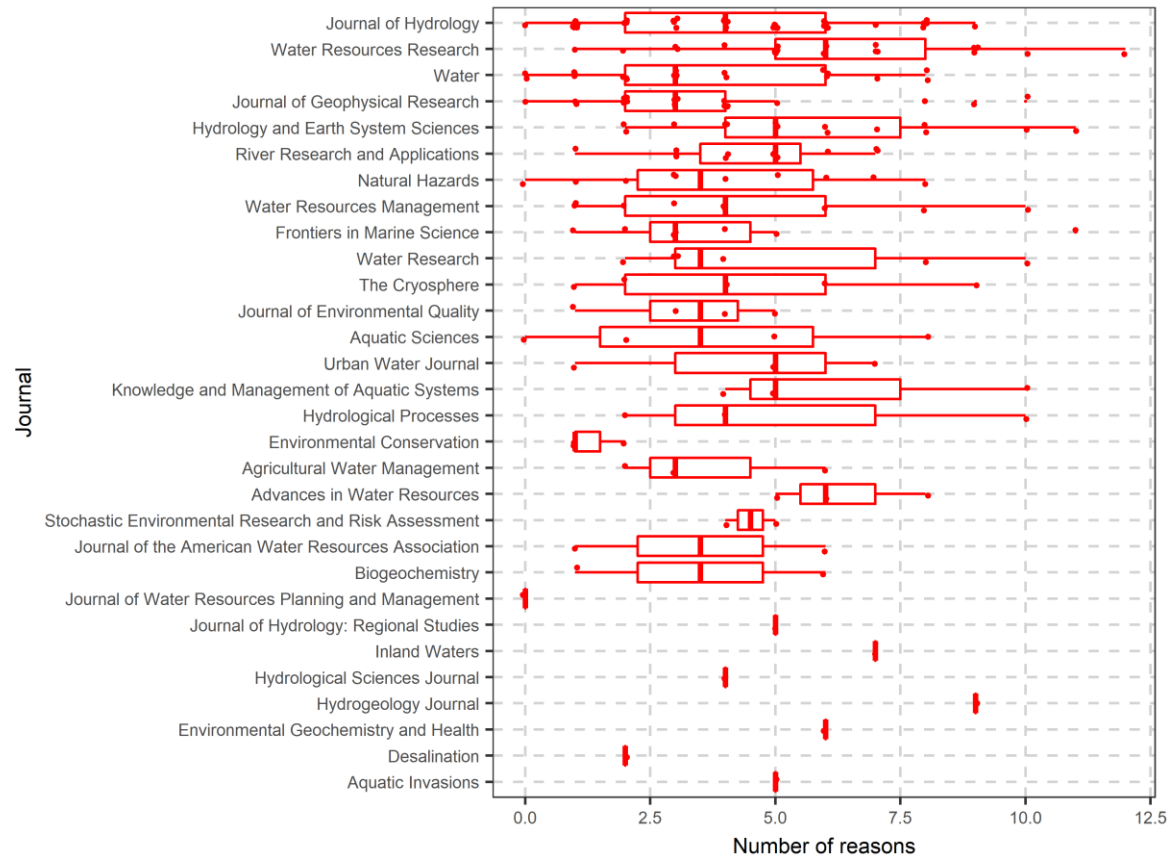
Application of random forests in water science



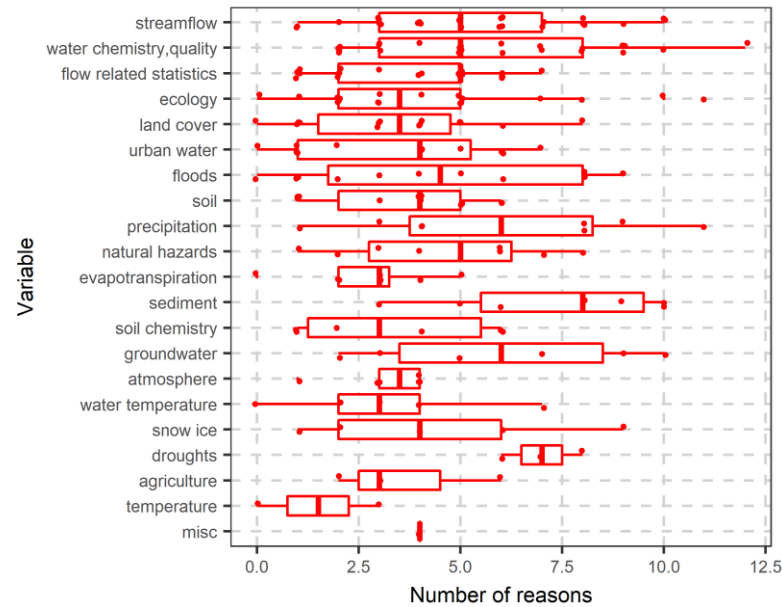
Application of random forests in water science



Application of random forests in water science



Application of random forests in water science



Some of our hydrology works with random forests

- Forecasting (Tyralis and Papacharalampous 2017, Papacharalampous et al. 2019a).
- Forecasting with predictor variables (Tyralis et al. 2021).
- Probabilistic predictions and statistical post-processing (Papacharalampous et al. 2019b, Tyralis et al. 2019a).
- Clustering application (Papacharalampous et al. 2021).
- Spatial modelling and interpolation (Tyralis et al. 2019c).

Concluding remarks and take-home considerations

- Random forests allow for explicit interpretation (contrary to the general class of data-driven models).
- In 33% of the reviewed water-related studies (67 out of 203), random forests were not the algorithm of focus, but they were used to complement other modelling approaches.
- The role of random forests as a useful complementary tool in water resources applications is related to their benchmarking nature.
- Large potential of random forests in water resources applications has been exploited only to a small degree.
- Most random forests variants have been implemented in the R programming language and are freely available.
- Only a few studies recognize possible shortcomings of random forests.

References

- [1] Breiman L (2001) Random forests. *Machine Learning* 45 (1), 5–32.
 - [2] Papacharalampous G, Tyralis H, Koutsoyiannis D (2019a) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment* 33 (2), 481–514.
 - [3] Papacharalampous G, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019b) Probabilistic hydrological post-processing at scale: Why and how to apply machine-learning quantile regression algorithms. *Water* 11 (10), 2126.
 - [4] Papacharalampous G, Tyralis H, Papalexiou SM, Langousis A, Khatami S, Volpi E, Grimaldi S (2021) Global-scale massive feature extraction from monthly hydroclimatic time series: Statistical characterizations, spatial patterns and hydrological similarity. *Science of the Total Environment* 767, 144612.
 - [5] Parkhurst DF, Brenner KP, Dufour AP, Wymer LJ (2005) Indicator bacteria at five swimming beaches—analysis using random forests. *Water Research* 39 (7), 1354–1360.
 - [6] Tyralis H, Papacharalampous G (2017) Variable selection in time series forecasting using random forests. *Algorithms* 10(4):114.
 - [7] Tyralis H, Papacharalampous G, Burnetas A, Langousis A (2019a) Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. *Journal of Hydrology* 577, 123957.
 - [8] Tyralis H, Papacharalampous G, Langousis A (2019b) A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* 11 (5), 910.
 - [9] Tyralis H, Papacharalampous G, Tantanee S (2019c) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. *Journal of Hydrology* 574, 628–645.
 - [10] Tyralis H, Papacharalampous G, Langousis A (2021) Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Computing and Applications* 33, 3053–3068
-