European Geosciences Union
General Assembly 2022
Vienna | Austria | 23–27 May 2022

Session HS3.1: Hydroinformatics: data analytics, machine learning, systems analysis, optimization

Feature-based clustering of hydroclimatic time series

Georgia Papacharalampous[1], and Hristos Tyralis[2]

[1] Department of Water Resources and Environmental Modeling, Faculty of Environmental Sciences, Czech University of Life Sciences, Kamýcá 129, Praha-Suchdol 16500, Prague, Czech Republic

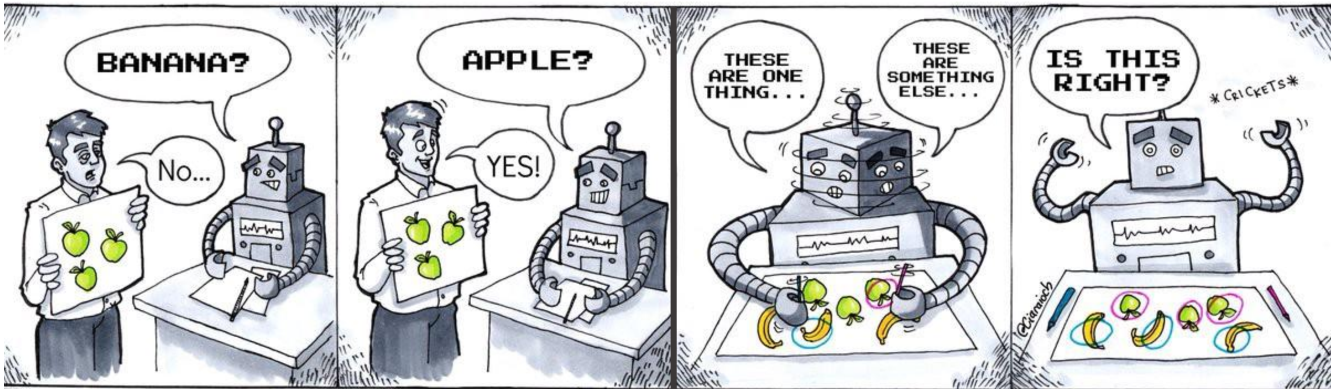
[2] Air Force Projects Authority, Hellenic Air Force, Mesogion Avenue 227–231, 15561 Cholargos, Greece

papacharalampous.georgia@gmail.com @GeorgiaPapachar

Machine learning basics with emphasis on clustering

Figure source: https://medium.com/data-solstice/wait-machines-can-learn-part-2-25e5d642652f

Definition



Supervised Learning

Unsupervised Learning

Task example

Probabilistic forecasting

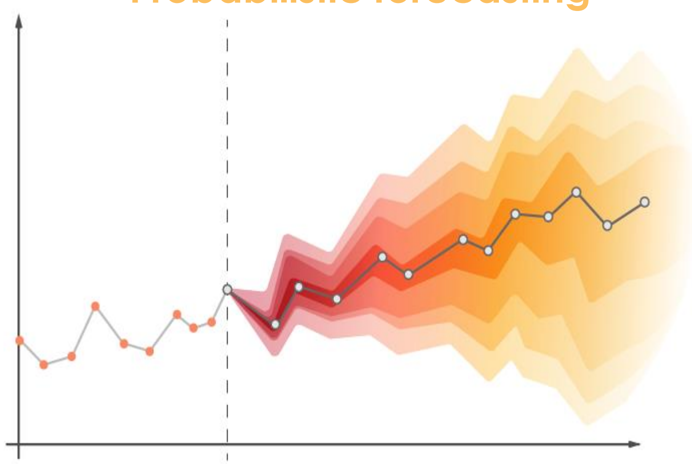


Figure source: https://medium.com/analytics-vidhya/time-series-forecasting-c73dec0b7533

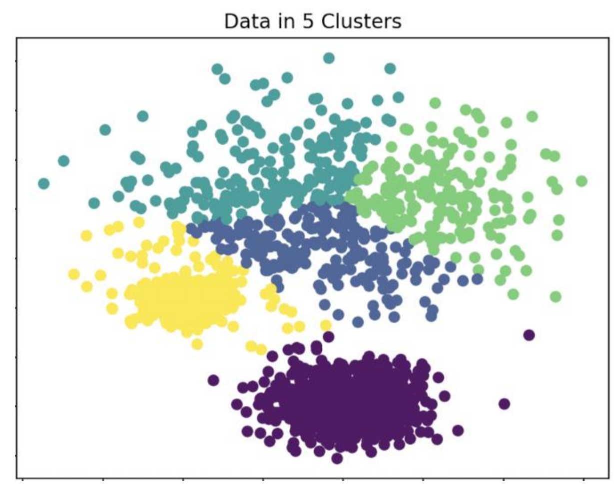


Figure source: https://www.ml-science.com/k-means-clustering

Cluster analysis

The importance and usefulness of cluster analyses

- A **clustering method** formulates and automates similarity guided groupings.
- Such groupings can support **technical and operational applications**.
- They can also support explorations by revealing **structures and patterns** in the data.
- Among others, they usually reveal interesting **spatial patterns**.
- The right above holds because real-world features are correlated in space.

Feature autocorrelation in space

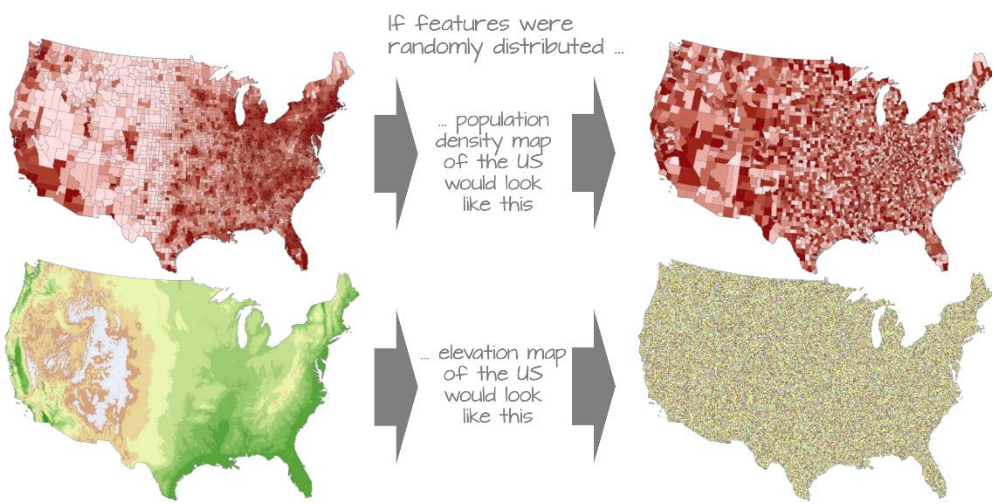


Figure source: https://mgimond.github.io/Spatial

Spatial patterns revealed through clustering

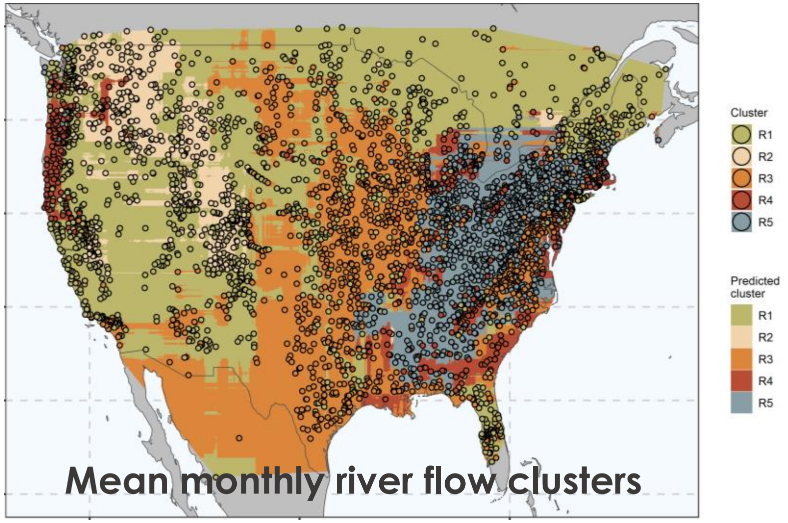
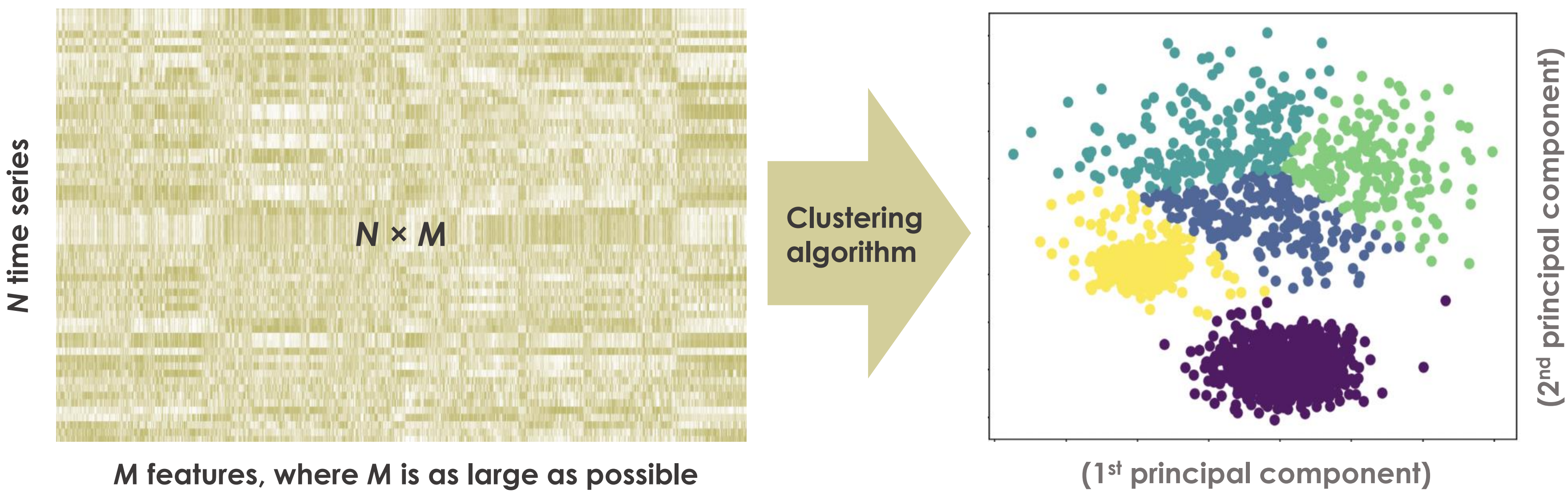


Figure source: Papacharalampous et al. (2021)

Massive feature extraction for hydroclimatic clustering

- There are two ways for improving clustering performance:
  - ✓ Improving the **clustering algorithm**;
  - ✓ Finding new informative **features** to cluster upon.
- Therefore, **massive feature extraction** (Fulcher and Jones 2014; Fulcher 2018) was proposed for clustering hydroclimatic time series by Papacharalampous et al. (2021).
- Indeed, this concept can lead to performance improvements, as it considerably increases the amount of information exploited by the clustering algorithm.



Examples of time series features and their compilations

- The existing general-purpose time series features and time series feature categories include **autocorrelation**, **partial autocorrelation**, **long-range dependence**, **entropy**, **temporal variation**, **seasonality**, **trend**, **lumpiness**, **stability**, **nonlinearity**, **linearity**, **spikiness**, **curvature** and **many more features**.
- Time series features are of fundamental interest in **stochastic (statistical) hydrology**.
- Examples of compilations of time series features for data science applications can be found in Wang et al. (2006), Fulcher et al. (2013), Hyndman et al. (2015), Kang et al. (2017, 2020) and Hyndman et al. (2020).
- Massive (or extensive) time series feature compilations introduced and successfully computed in stochastic hydrology can be found in Papacharalampous et al. (2021, 2022a,b).

Time series decomposition for feature extraction

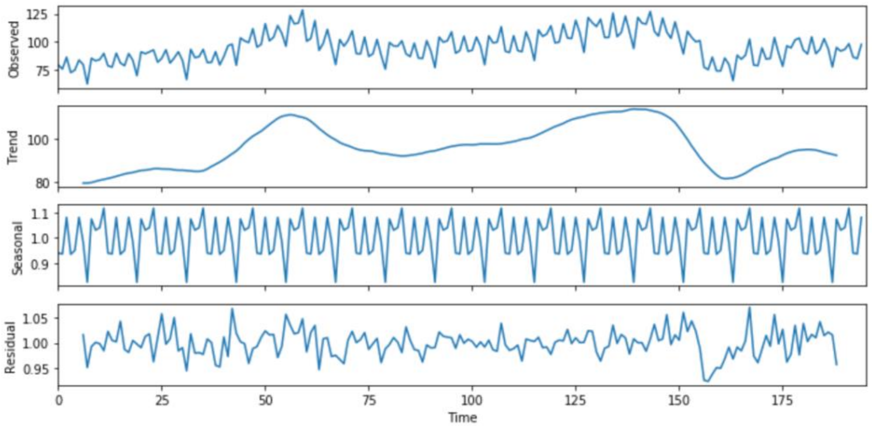


Figure source: https://datasciencebeginners.com/2020/11/25/time-series-forecast-and-decomposition-101-guide-python

Random forests for clustering upon numerous features

- **Random forests** (Breiman 2001) can be applied in unsupervised mode for **clustering** as detailed in Liaw and Wiener (2002).
- Their following **properties** (Tyralis et al., 2019, Section 2.8.1) make them appealing for clustering in general, and clustering upon numerous time series features in particular:
  - ✓ They demonstrate high performance compared to other algorithms.
  - ✓ They can handle highly correlated features.
  - ✓ They can operate successfully when interactions are present.
  - ✓ They are invariant to monotone transformations of the features.
- Lastly, they support the application of **explainable machine learning** through feature importance metrics.

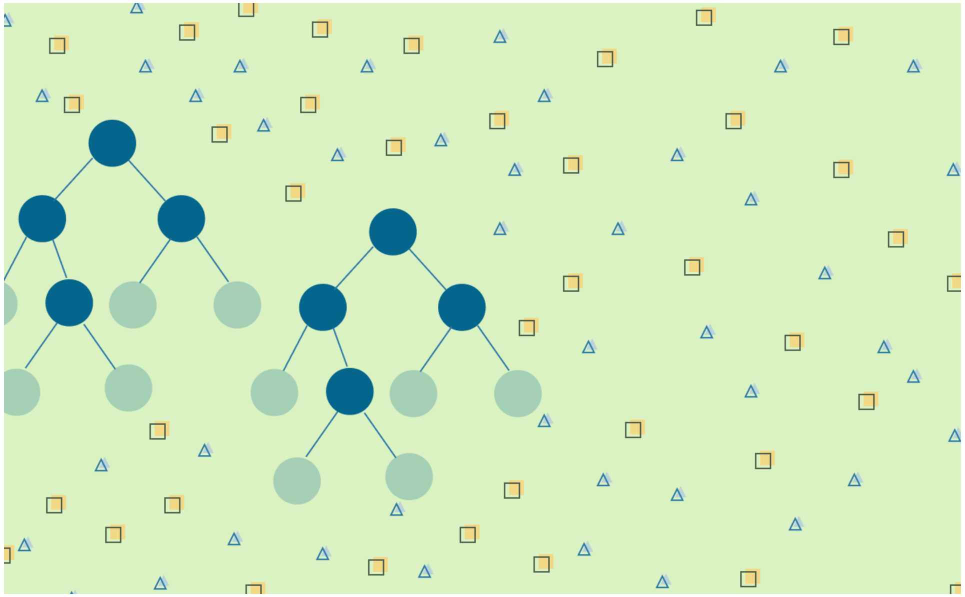


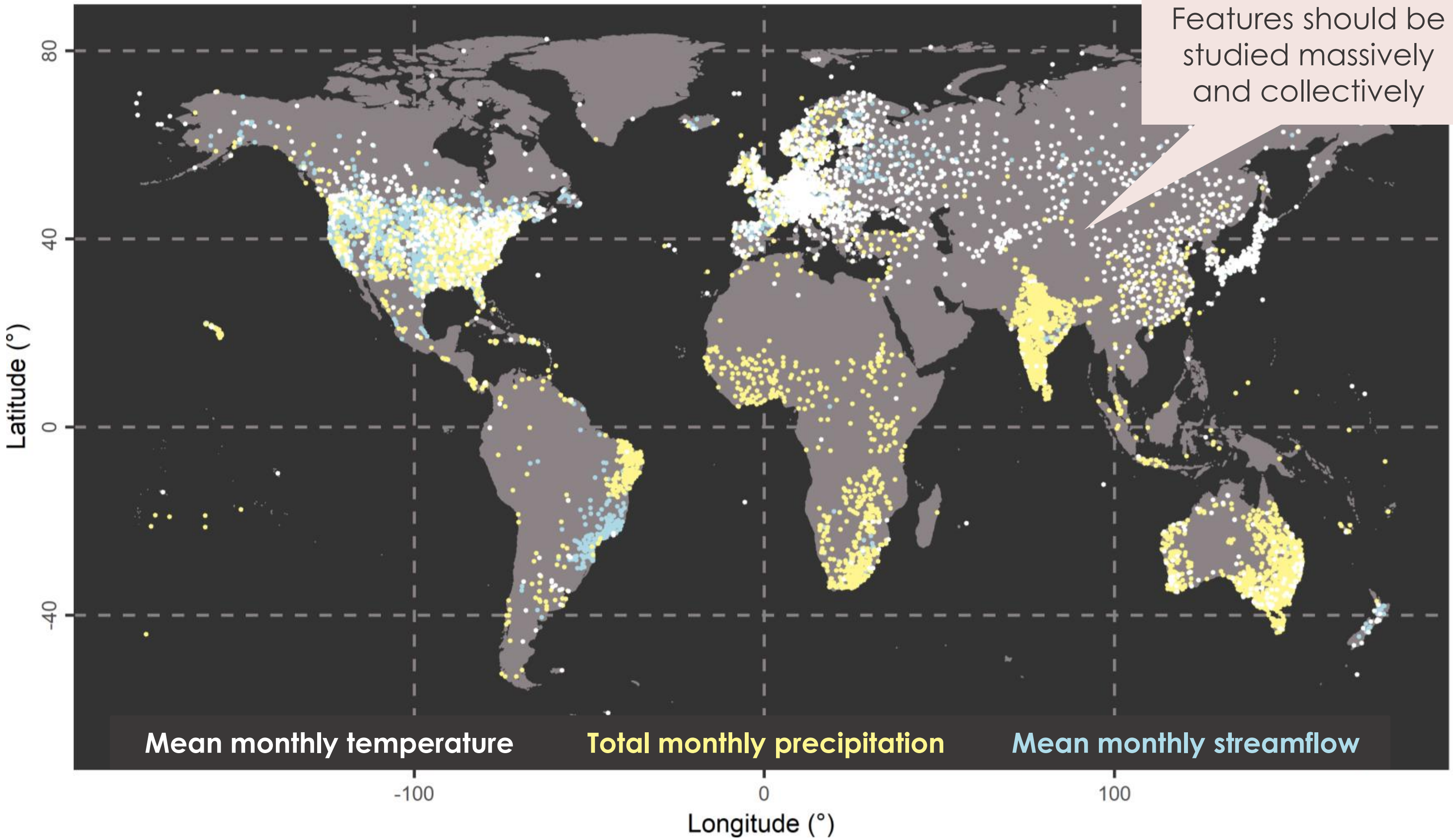
Figure source: https://towardsdatascience.com/random-forests-algorithm-explained-with-a-real-life-example-and-some-python-code-affbfa5a942c

Massive feature extraction for hydroclimatic clustering

Benefitting from 59 largely diverse time series features

Further reading: Papacharalampous et al. (2021)

Features should be studied massively and collectively

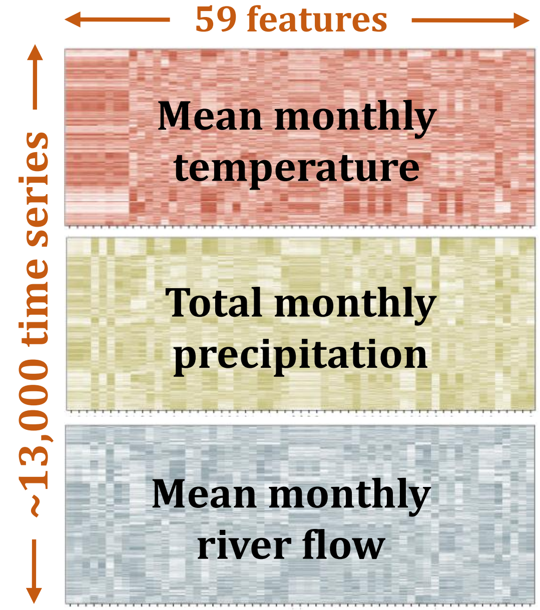


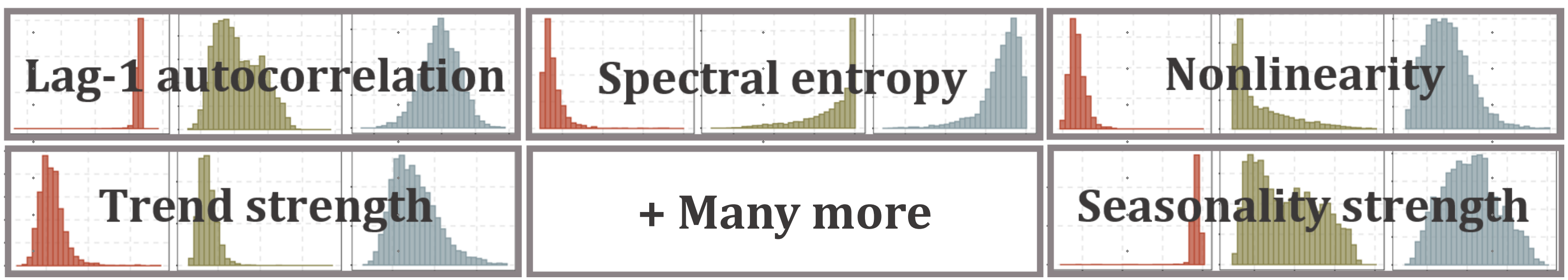
Massive feature extraction for hydroclimatic clustering

A compilation of 59 largely diverse features

x_acf1, ac_9, x_acf10, diff1_acf1, diff1_acf10, diff2_acf1, diff2_acf10, seas_acf1, firstzero_ac, firstmin_ac, embed2_incircle_1, embed2_incircle_2, trev_num, motiftwo_entro3, walker_propcross, x_pacf5, diff1x_pacf5, diff2x_pacf5, seas_pacf, localsimple_mean1, localsimple_lfitac, sampen_first, std1st_der, spreadrandomlocal_meantaul_50, spreadrandomlocal_meantaul_ac2, histogram_mode_10, outlierinclude_mdrmd, fluctanal_prop_r1, crossing_points, entropy, flat_spots, arch_acf, garch_acf, arch_r2, garch_r2, alpha, beta, gamma, lumpiness, stability, max_level_shift, time_level_shift, max_var_shift, time_var_shift, max_kl_shift, time_kl_shift, ARCH.LM, nonlinearity, unitroot_kpss, hurst, trend, spike, linearity, curvature, e_acf1, e_acf10, seasonal_strength, peak, trough

Massive feature extraction

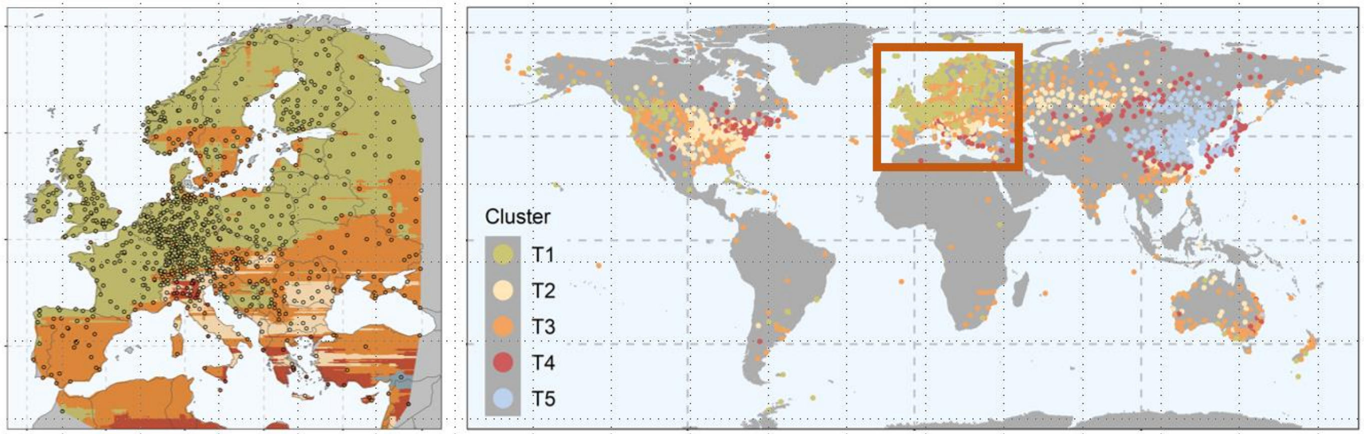




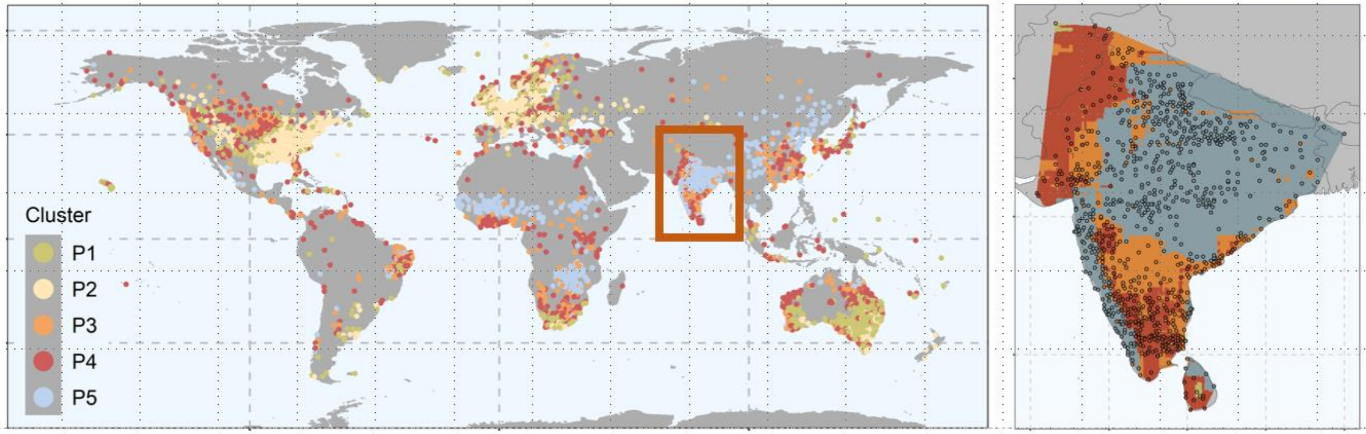
Further reading: Papacharalampous et al. (2021)

Hydroclimatic clusters based on 59 time series features

Mean monthly temperature

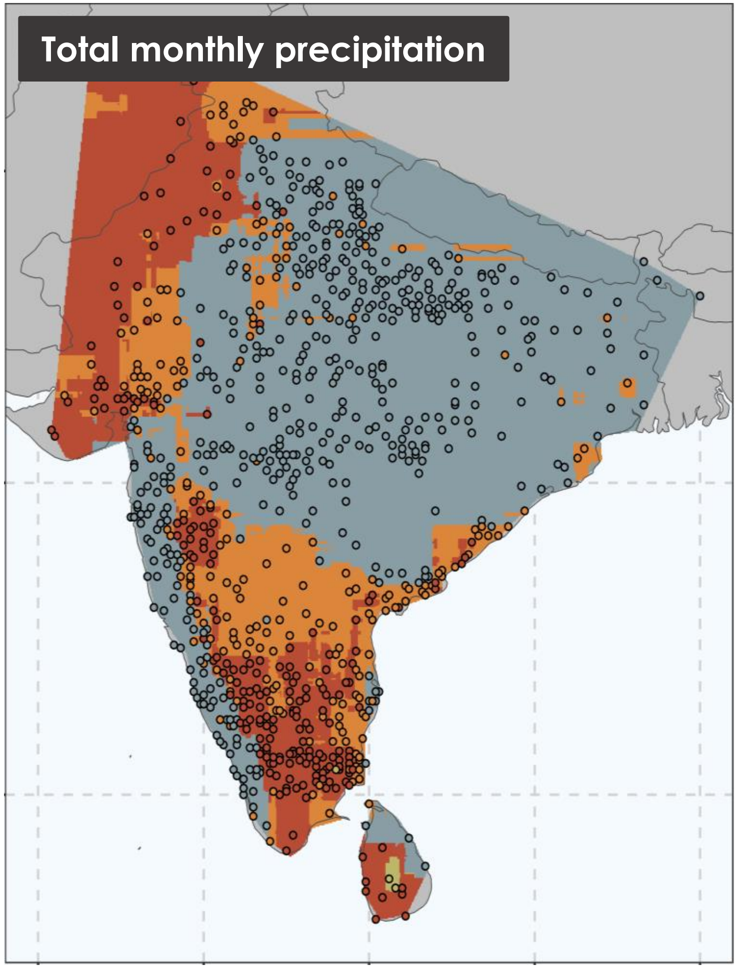


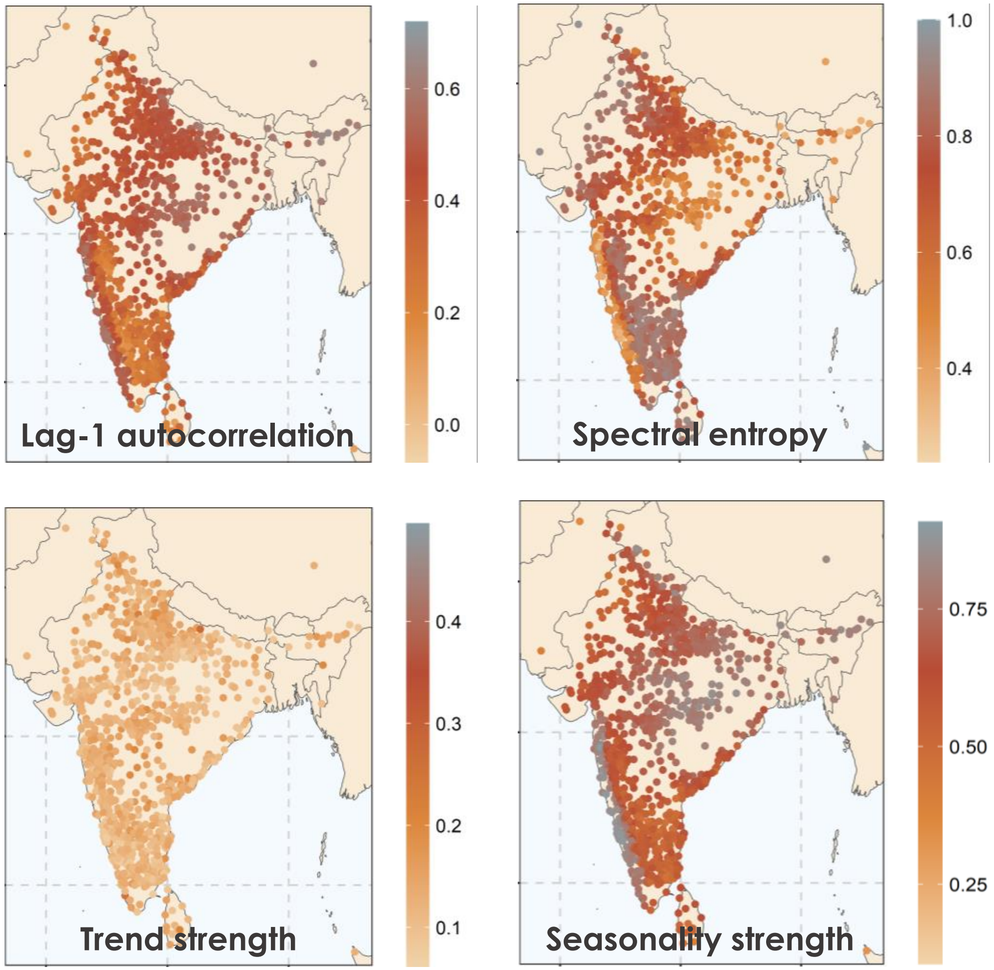
Total monthly precipitation



Further reading: Papacharalampous et al. (2021)

Hydroclimatic clusters based on 59 time series features



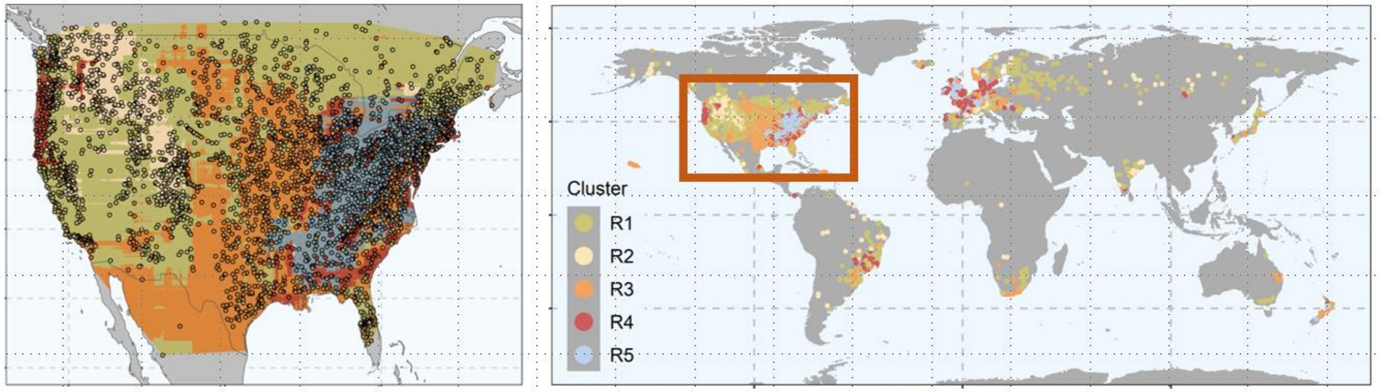


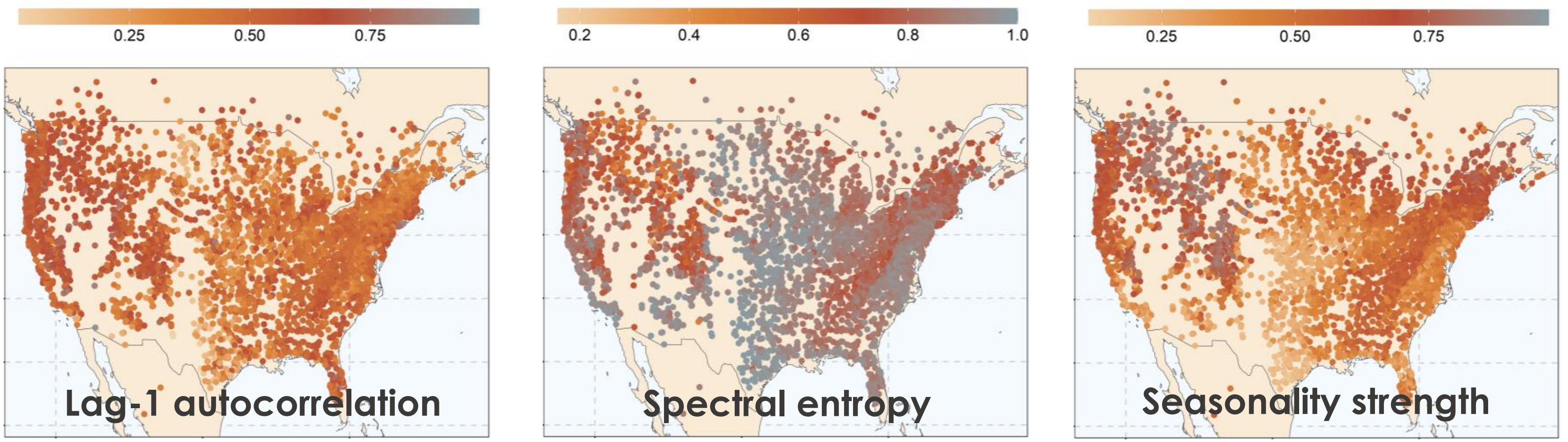
The legends present the global ranges of the feature values.

Further reading: Papacharalampous et al. (2021)

Hydroclimatic clusters based on 59 time series features

Mean monthly river flow



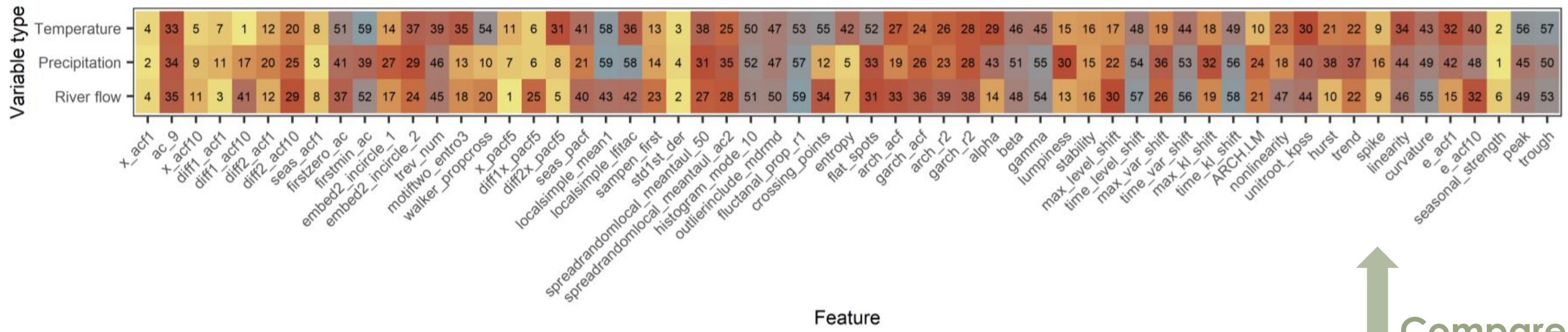


The legends present the global ranges of the feature values.

Further reading: Papacharalampous et al. (2021)

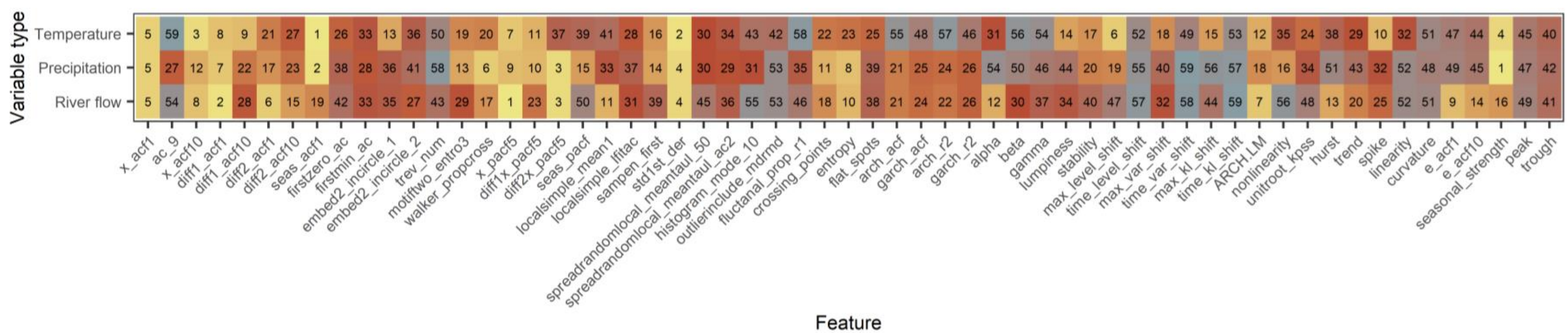
Feature importance in clustering different variable types

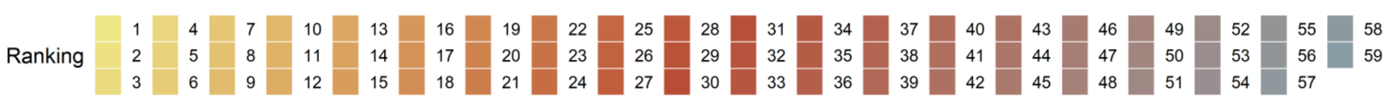
Rankings of the features from the most (1st) to the least (59th) important ones in clustering



Compare

Rankings of the features from the most (1st) to the least (59th) contributing ones to the 1st principal component

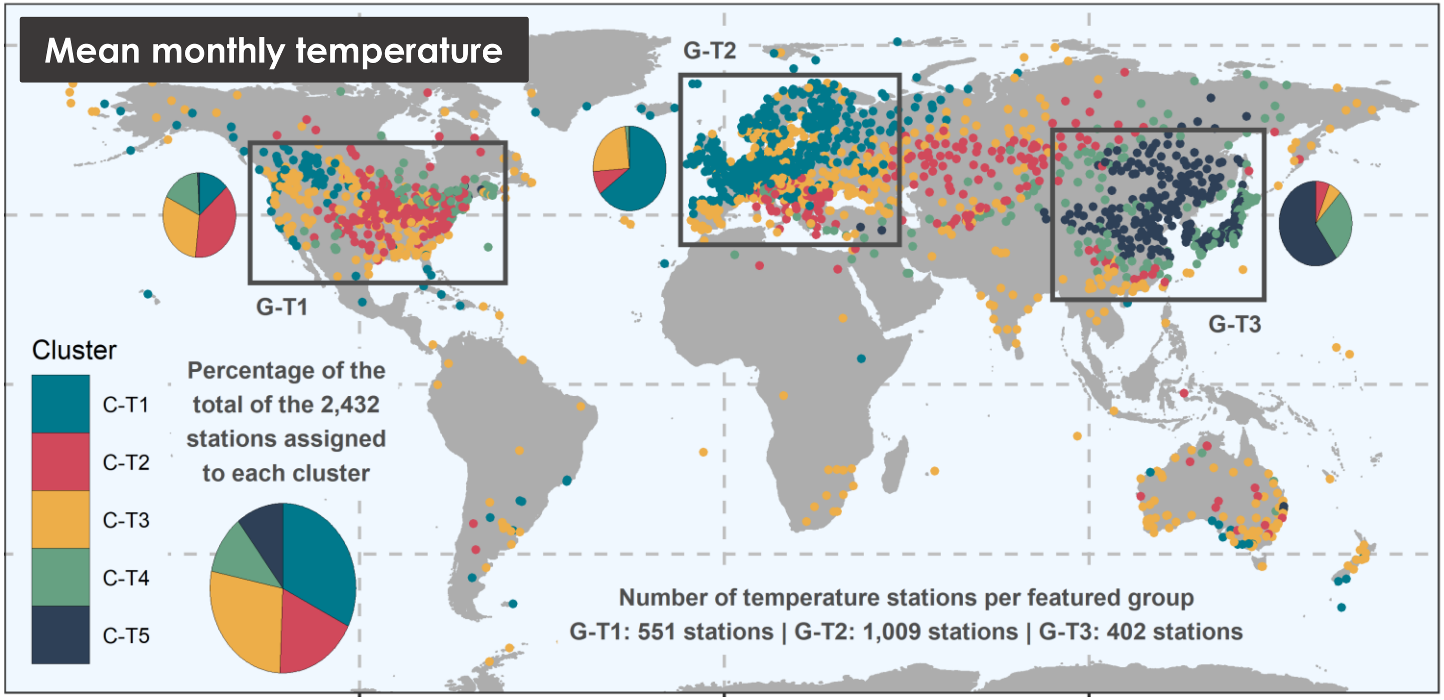




Further reading: Papacharalampous et al. (2021)

Forecastability comparisons across hydroclimatic clusters

Temperature time series forecastability in terms of Nash-Sutcliffe efficiency in the different clusters

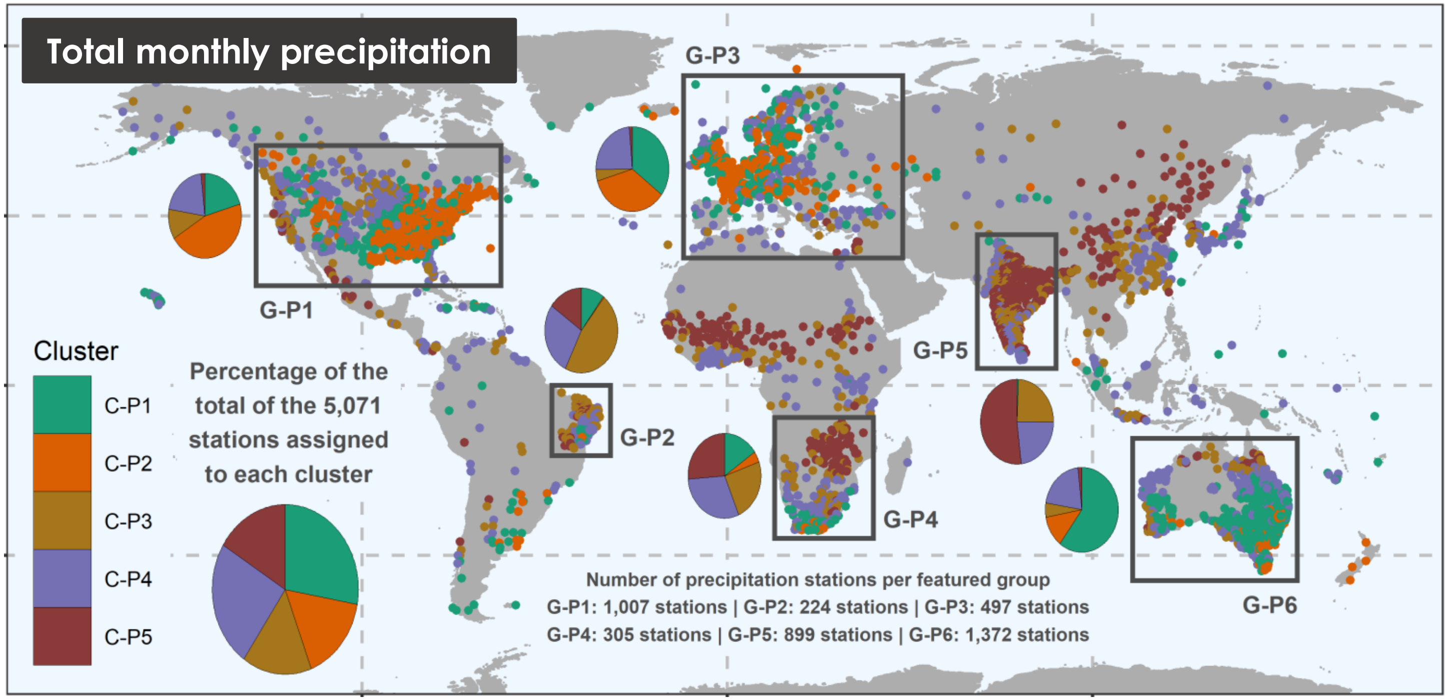


The clusters have been obtained by implementing a close variant of the method by Papacharalampous et al. (2021).

Further reading: Papacharalampous et al. (2022b)

Forecastability comparisons across hydroclimatic clusters

Precipitation time series forecastability in terms of Nash-Sutcliffe efficiency in the different clusters

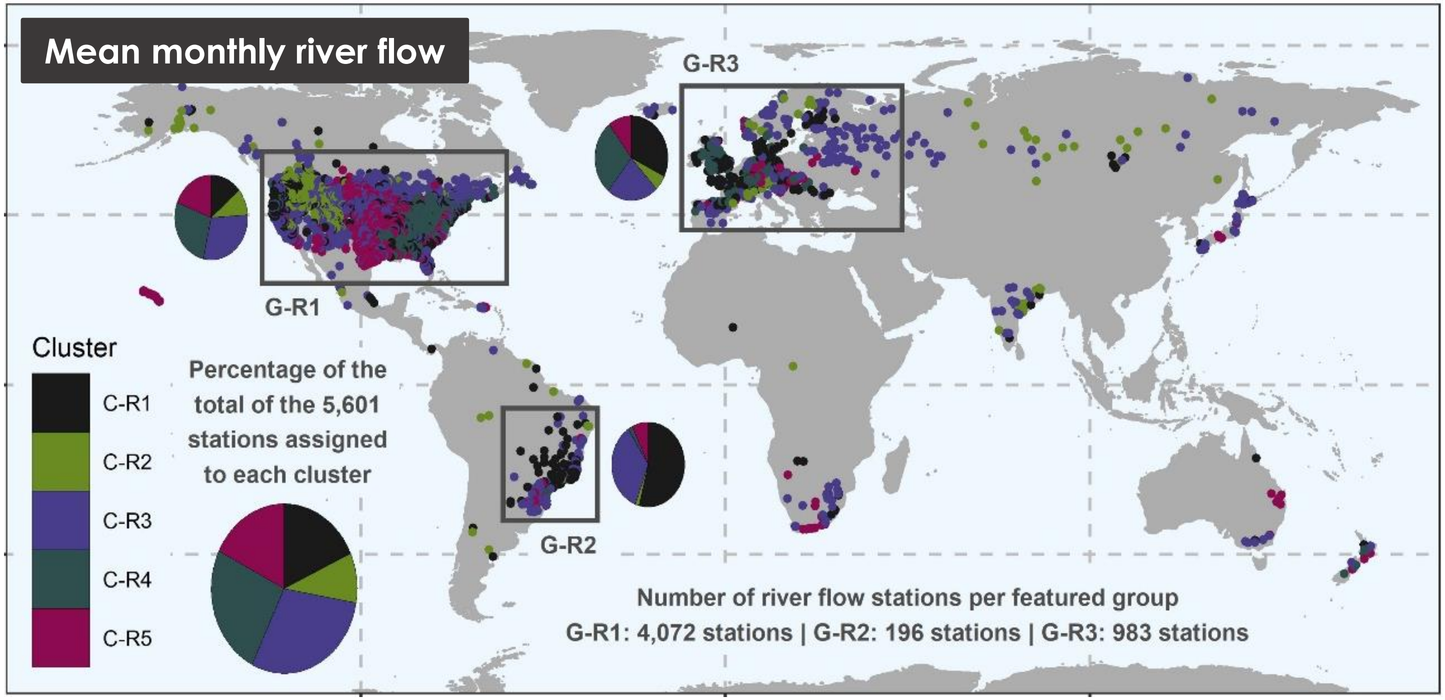


The clusters have been obtained by implementing a close variant of the method by Papacharalampous et al. (2021).

Further reading: Papacharalampous et al. (2022b)

Forecastability comparisons across hydroclimatic clusters

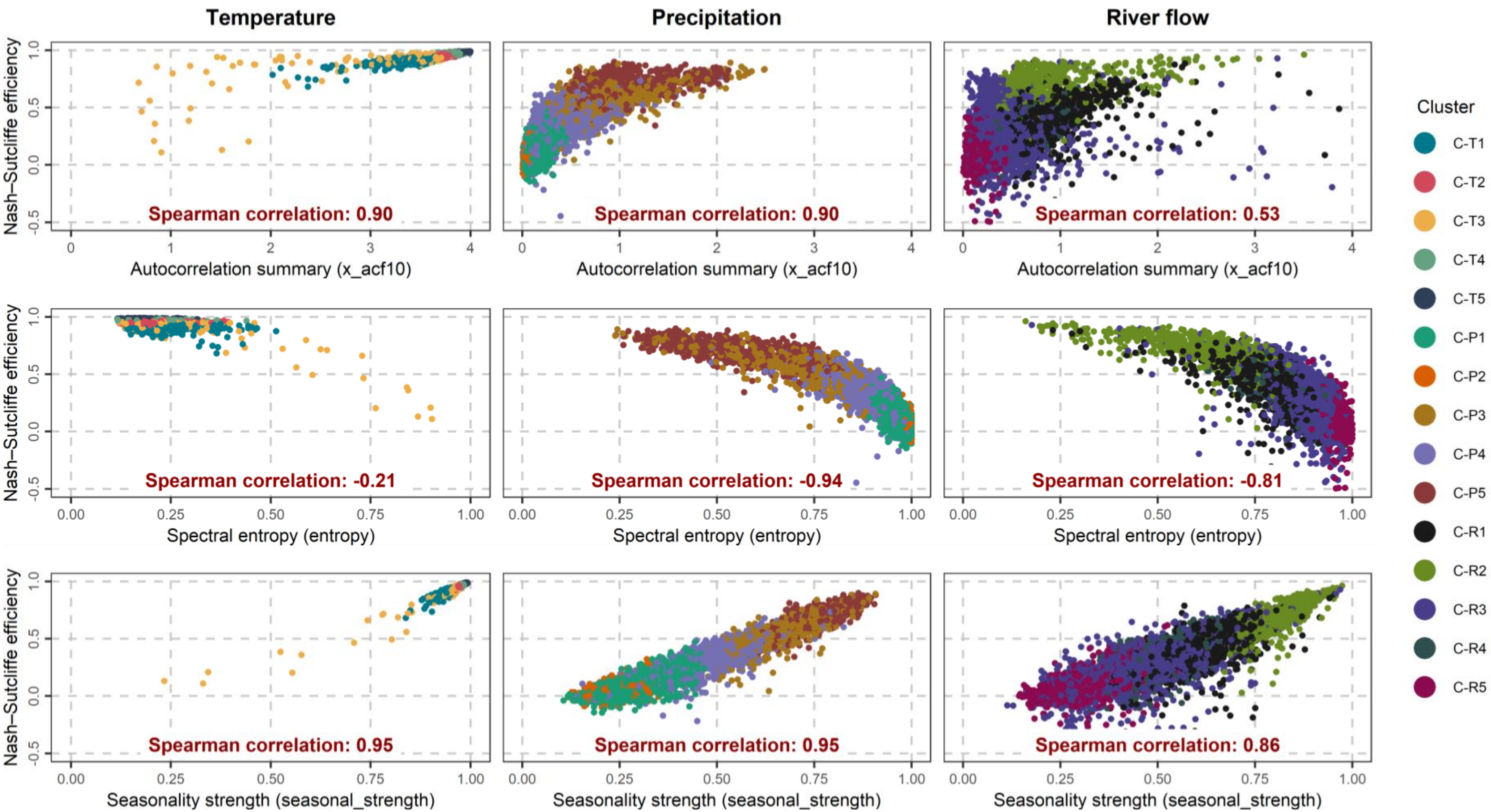
River flow time series forecastability in terms of Nash-Sutcliffe efficiency in the different clusters



The clusters have been obtained by implementing a close variant of the method by Papacharalampous et al. (2021).

Further reading: Papacharalampous et al. (2022b)

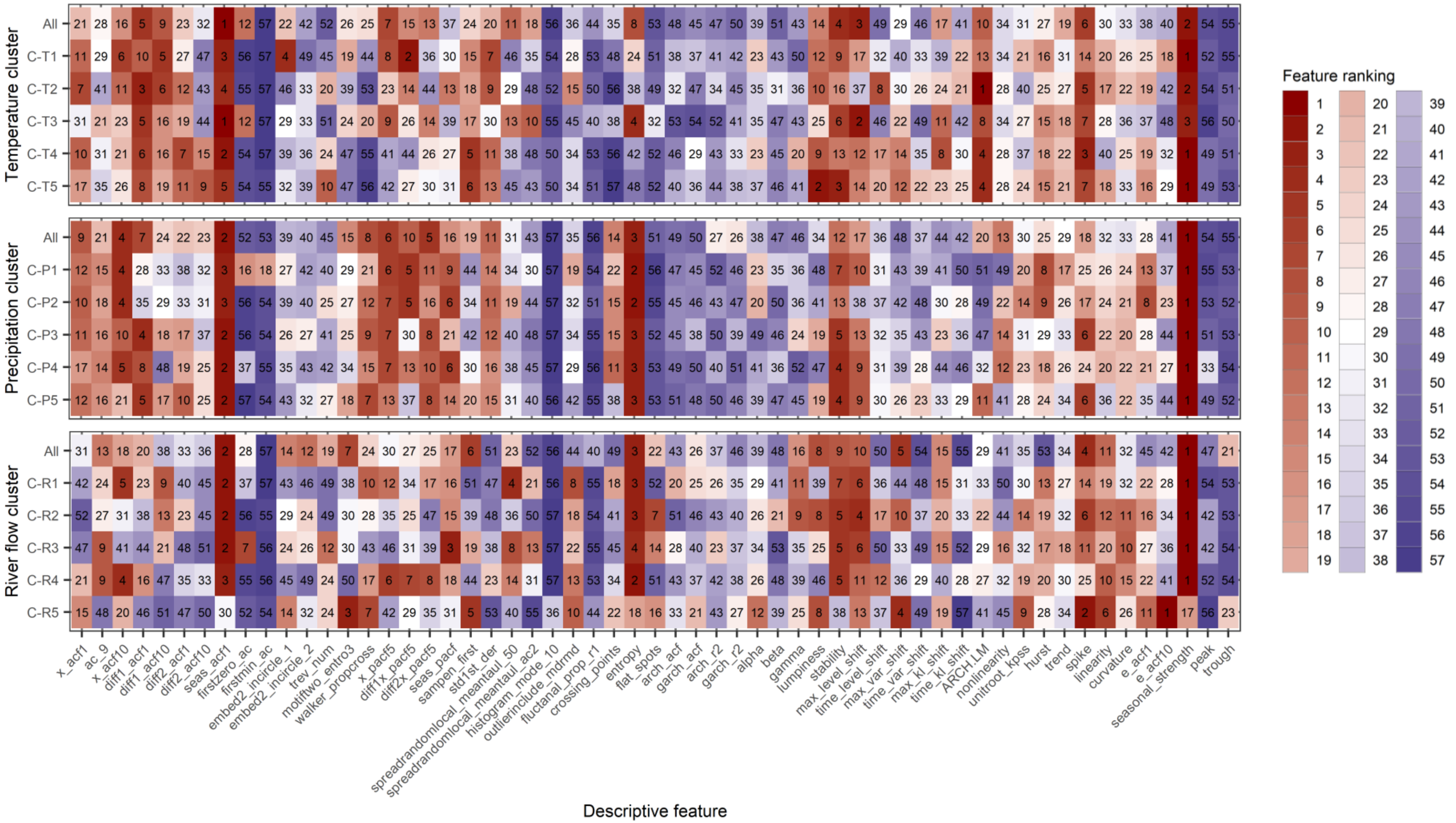
Forecastability comparisons across hydroclimatic clusters



The clusters have been obtained by implementing a close variant of the method by Papacharalampous et al. (2021).

Further reading: Papacharalampous et al. (2022b)

Forecastability comparisons across hydroclimatic clusters



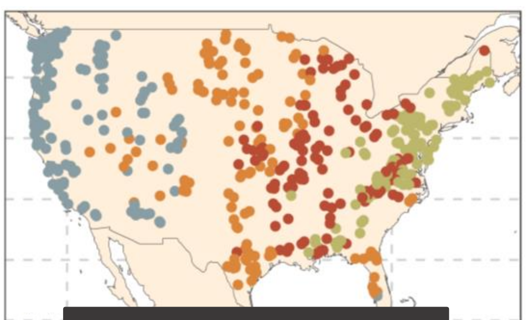
The clusters have been obtained by implementing a close variant of the method by Papacharalampous et al. (2021).

Further reading: Papacharalampous et al. (2022b)

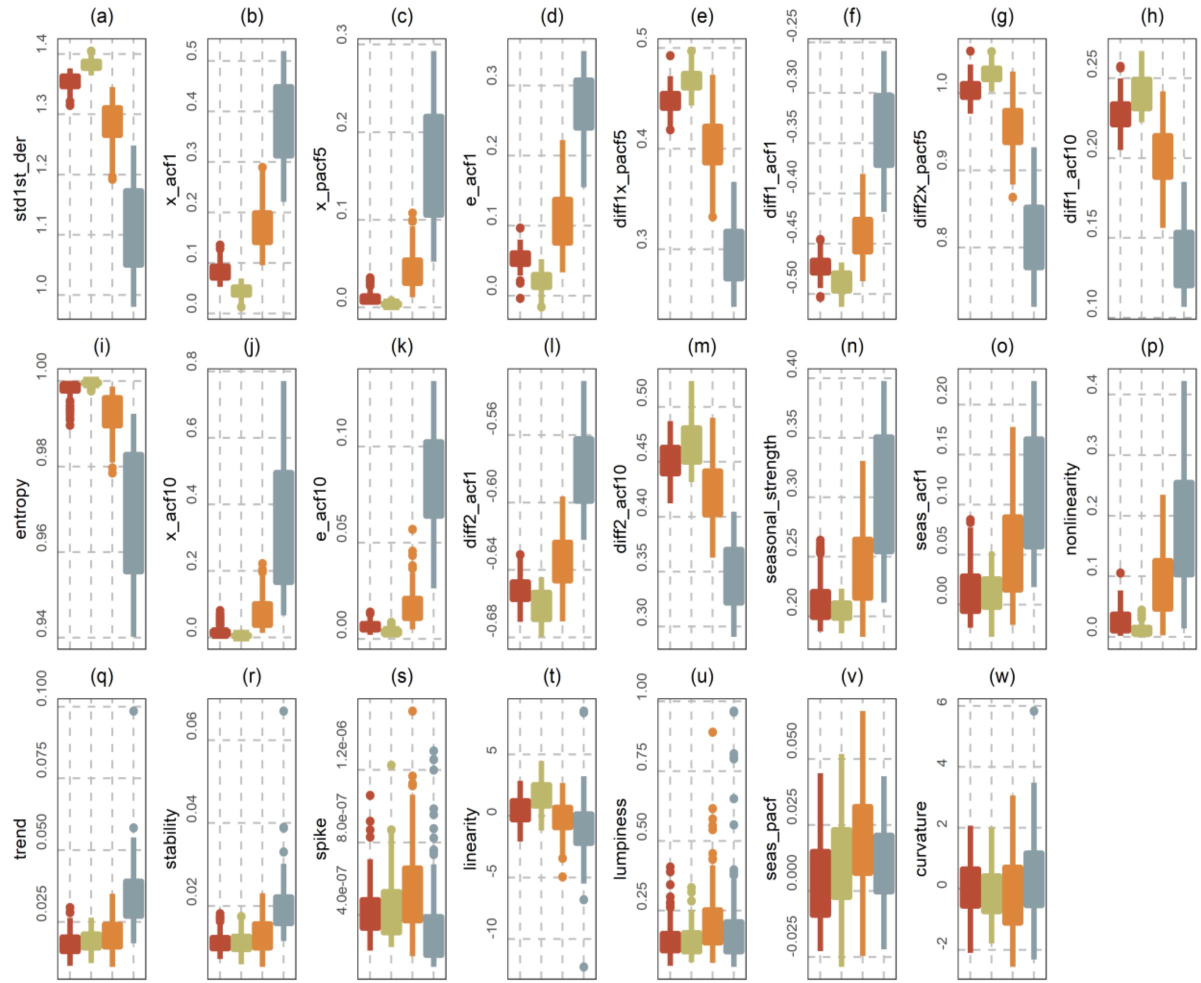
Hydroclimatic clusters based on 23 time series features

A compilation of 23 features for hydroclimatic time series analysis at multiple time scales

`x_acf1, x_acf10, diff1_acf1, diff1_acf10, diff2_acf1, diff2_acf10, seas_acf1, x_pacf5, diff1x_pacf5, diff2x_pacf5, seas_pacf, std1st_der, entropy, lumpiness, stability, nonlinearity, trend, spike, linearity, curvature, e_acf1, e_acf10, seasonal_strength`



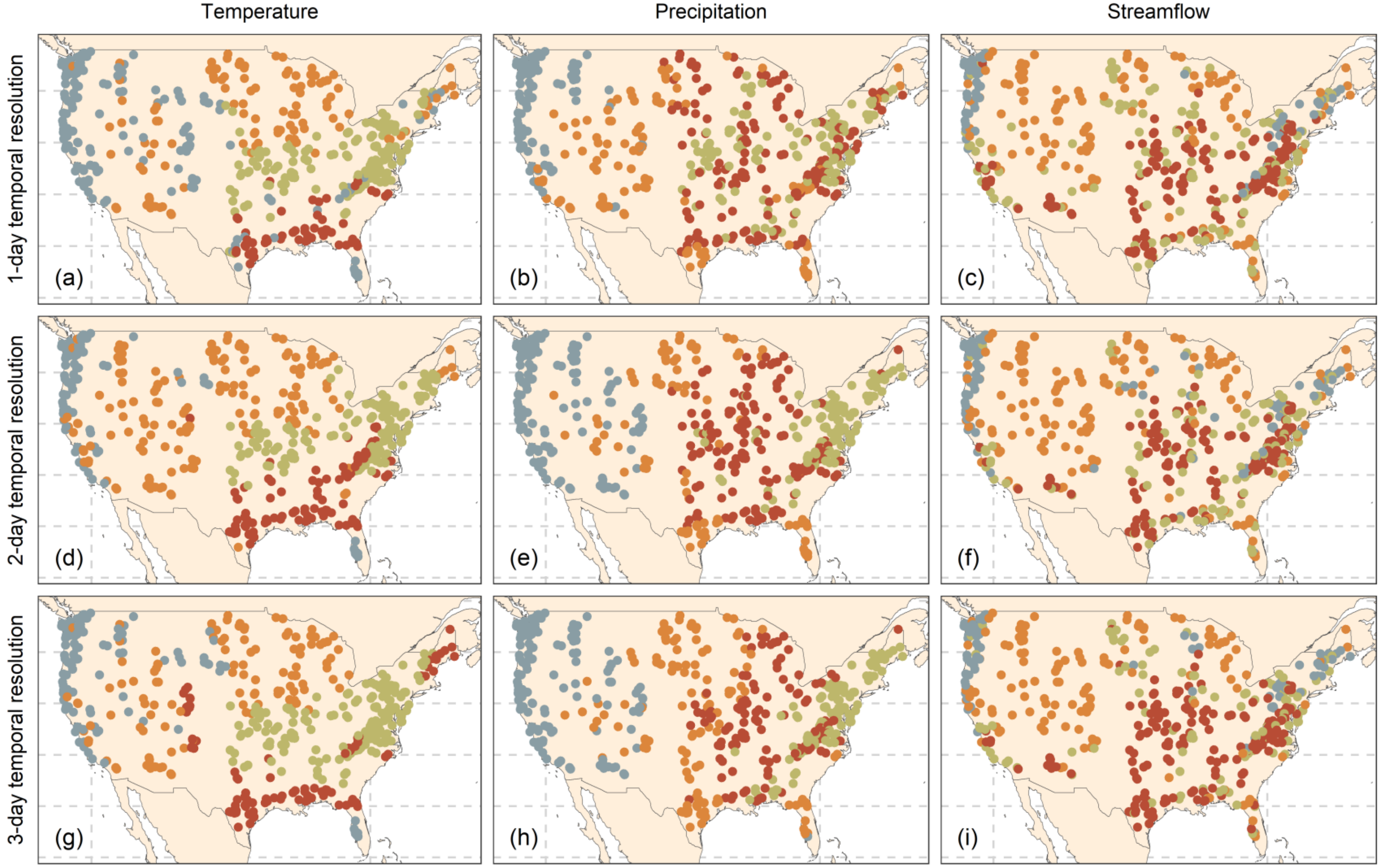
3-day temporal resolution



The feature importance in clustering decreases as we move from (a) to (w).

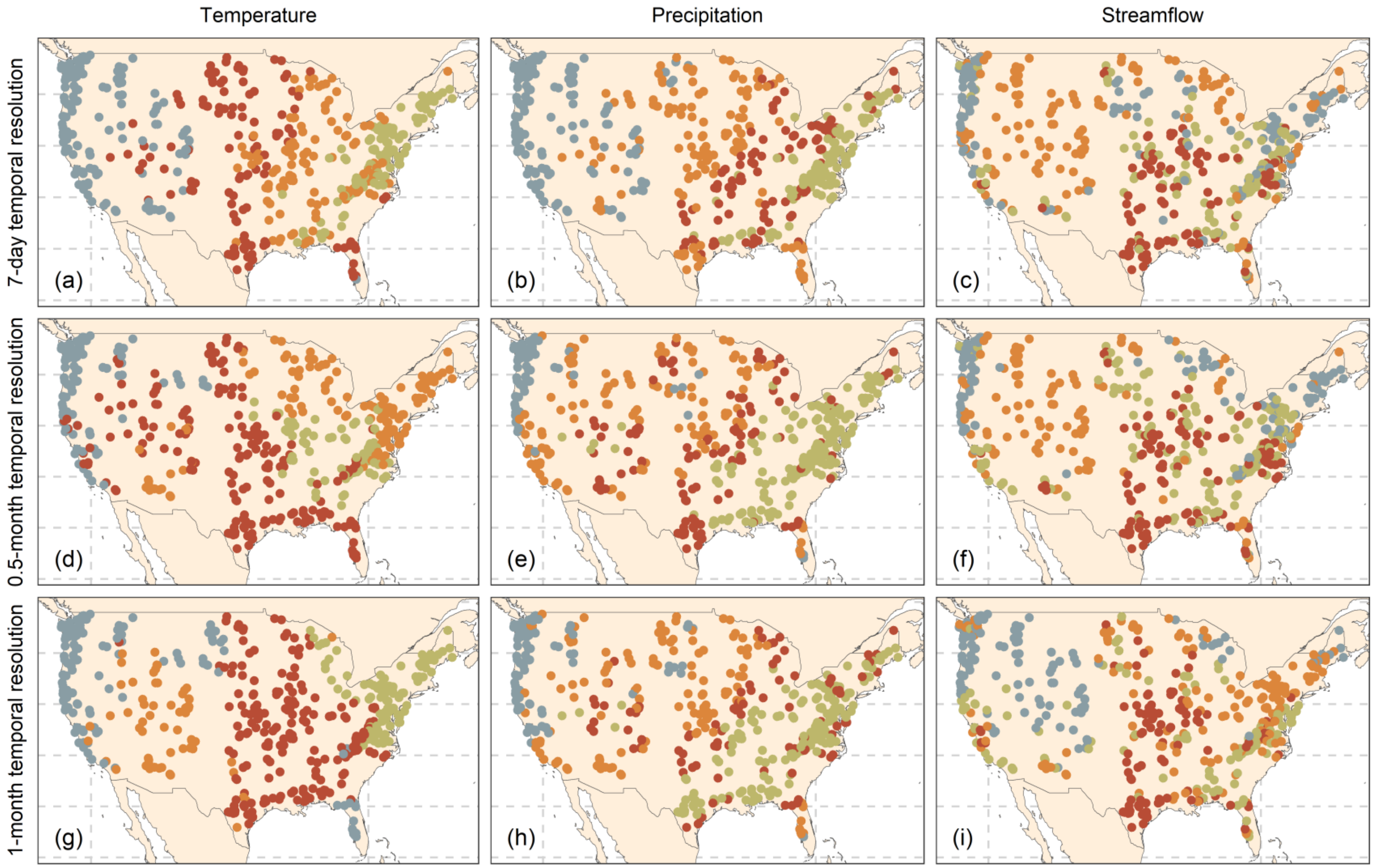
Further reading: Papacharalampous et al. (2022a)

Hydroclimatic clusters at multiple time scales



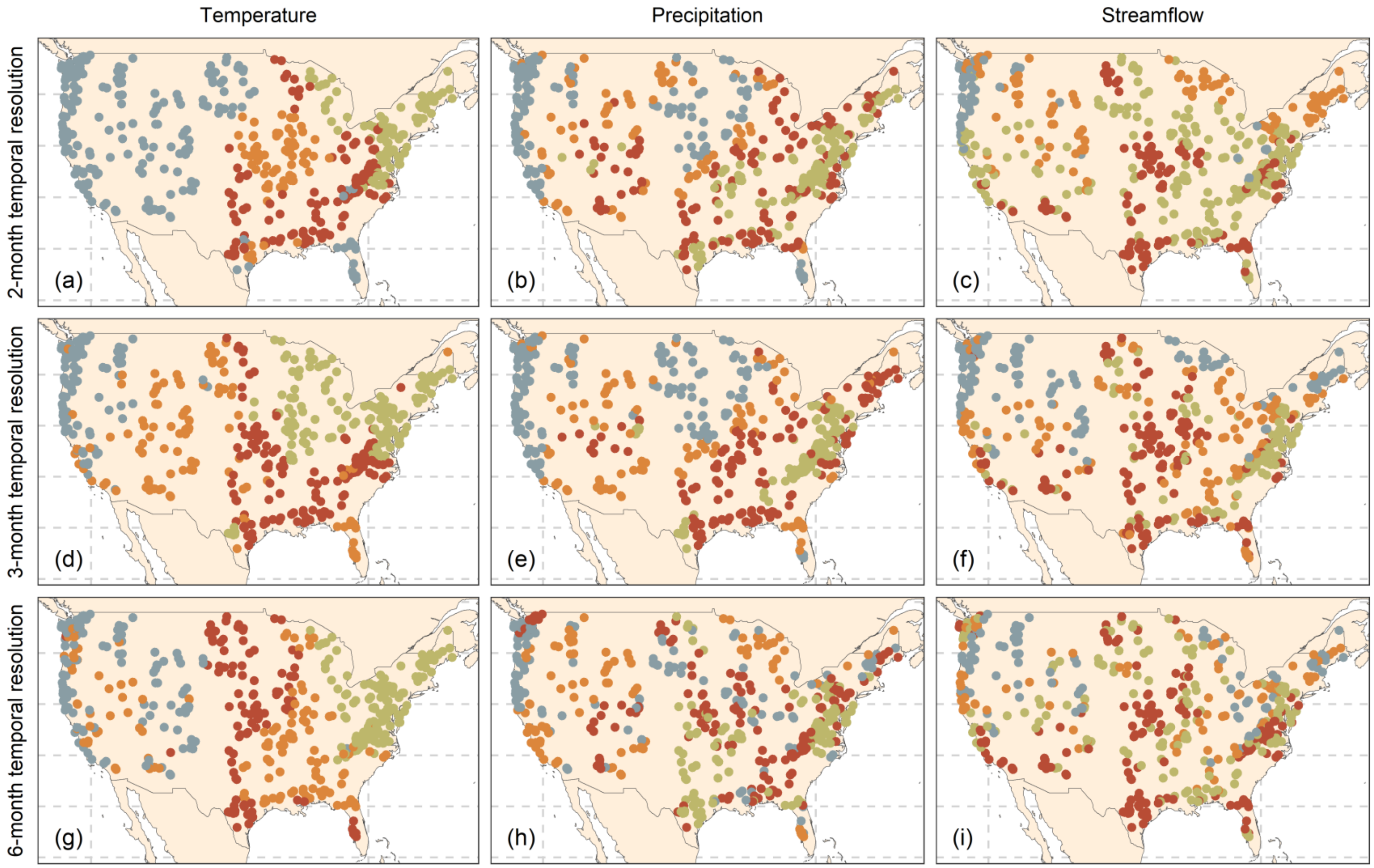
Further reading: Papacharalampous et al. (2022a)

Hydroclimatic clusters at multiple time scales



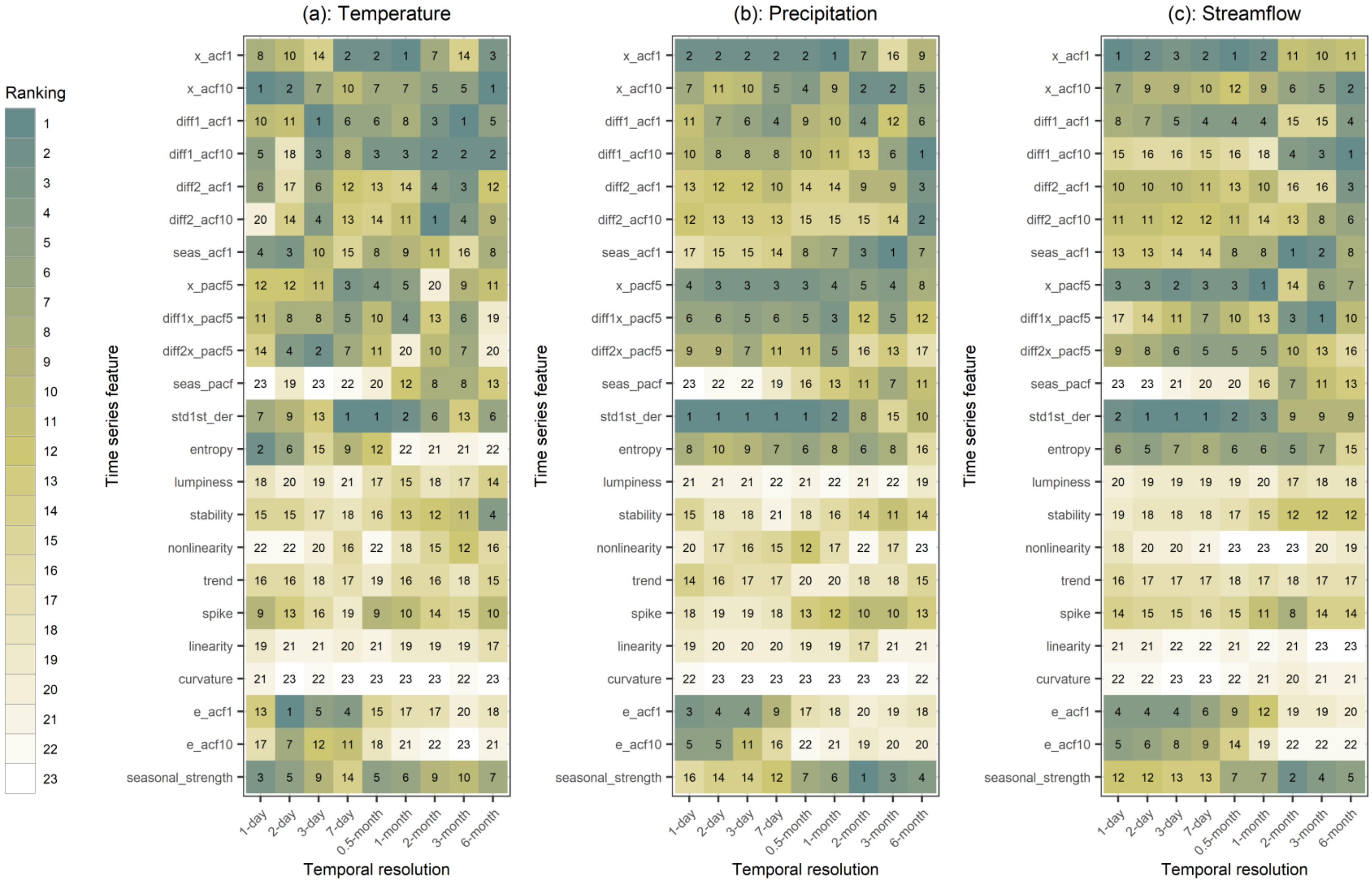
Further reading: Papacharalampous et al. (2022a)

Hydroclimatic clusters at multiple time scales



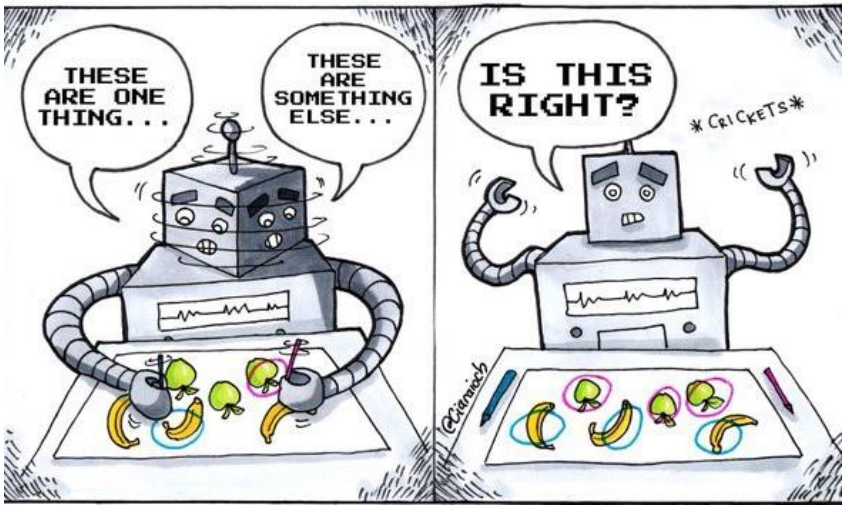
Further reading: Papacharalampous et al. (2022a)

Feature importance in clustering at multiple time scales



Further reading: Papacharalampous et al. (2022a)

Summary, discussion and take-home messages



- One way for improving clustering performance is finding **new informative features** to cluster upon.
- Therefore, Papacharalampous et al. (2021) proposed to cluster hydroclimatic time series by exploiting the concept of **massive feature extraction**.
- This concept is new in the field, although time series features are of fundamental and practical interest in stochastic (statistical) hydrology (see, e.g., the central themes, concepts and directions provided by Montanari et al. 2013).
- The usefulness of the new approach in hydroclimatic time series clustering was demonstrated through a variety of **global-scale** and other **large-scale investigations** (Papacharalampous et al. 2021, 2022a,b).
- These investigations were conducted for **temperature**, **precipitation** and **streamflow** variables at **several temporal scales**.

Summary, discussion and take-home messages

- Indeed, there are numerous **time series features** whose computation is meaningful for various hydroclimatic variables and at various temporal scales with **minimal adaptations** (e.g., the time series features in Papacharalampous et al. 2021, 22022a,b).
- The general purpose character of the proposed clustering methods differentiates them notably from signature-based clustering methods (e.g., from Jehn et al. 2020).
- An even more substantial difference with other clustering methods in hydrology (e.g., with the methods by Hall and Blöschl 2018; Jehn et al. 2020; Fischer and Schumann 2021) is the consideration of **both interpretable and less interpretable features** in the clustering under the proposed central concept.
- In fact, the application of **explainable machine learning** showed that features from either of the above categories can be important in hydroclimatic time series clustering (Papacharalampous et al. 2021, 22022a,b).

Explainable machine learning

Figure source: https://www.analyticsinsight.net/a-beginners-guide-to-four-principles-of-explainable-artificial-intelligence

Summary, discussion and take-home messages

- More generally, a **massive** and collective examination of **hydroclimatic features** is necessary for understanding **hydroclimatic variability**, **change** and **predictability**.
- Particular focus on a single feature or a single feature category (e.g., on trends) could be misleading in hydroclimatic time series analysis contexts.
- A few **limitations** characterize the to-date applications of the proposed approach to hydroclimatic time series clustering and suggest open themes for **future research**.
- Indeed, this approach could be coupled with external methods for identifying an **optimal number of clusters**.
- It could also be applied with other **algorithms** (e.g., with **boosting**; for its theoretical properties, see Tyralis and Papacharalampous 2021, Section 3).
- Lastly, it could exploit information from **additional time series features**.

The main concept behind boosting

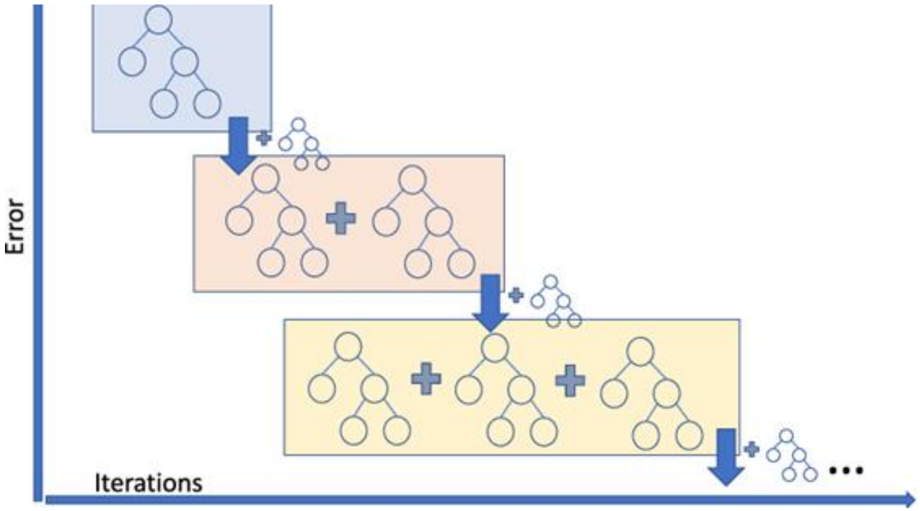


Figure source: https://medium.com/analytics-vidhya/what-is-gradient-boosting-how-is-it-different-from-ada-boost-2d5ff5767cb2

References

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

Fischer S, Schumann AH (2021) Regionalisation of flood frequencies based on flood type-specific mixture distributions. Journal of Hydrology X 13:100107. doi:10.1016/j.hydroa.2021.100107

Fulcher BD (2018) Feature-based time-series analysis, in: Dong G, Liu H (Eds) Feature Engineering for Machine Learning and Data Analytics. CRC Press, pp. 87–116

Fulcher BD, Jones NS (2014) Highly comparative feature-based time-series classification. IEEE Transactions on Knowledge and Data Engineering 26(12):3026–3037. doi:10.1109/TKDE.2014.2316504

Fulcher BD, Little MA, Jones NS (2013) Highly comparative time-series analysis: The empirical structure of time series and their methods. Journal of the Royal Society Interface 10(83):20130048. doi:10.1098/rsif.2013.0048

Hall J, Blöschl G (2018) Spatial patterns and characteristics of flood seasonality in Europe. Hydrology and Earth System Sciences 22(7):3883–3901. doi:10.5194/hess-22-3883-2018

Hyndman RJ, Wang E, Laptev N (2015) Large-scale unusual time series detection. 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, pp. 1616–1619. doi:10.1109/ICDMW.2015.104

Hyndman RJ, Kang Y, Montero-Manso P, Talagala T, Wang E, Yang Y, O'Hara-Wild M (2020) tsfeatures: Time Series Feature Extraction. R package version 1.0.2. https://CRAN.R-project.org/package=tsfeatures

Jehn FU, Bestian K, Breuer L, Kraft P, Houska T (2020) Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. Hydrology and Earth System Sciences 24(3):1081–1100. doi:10.5194/hess-24-1081-2020

Kang Y, Hyndman RJ, Smith-Miles K (2017) Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting 33(2):345−358. doi:10.1016/j.ijforecast.2016.09.004

Kang Y, Hyndman RJ, Li F (2020) GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. Statistical Analysis and Data Mining: The ASA Data Science Journal 13(4):354–376. doi:10.1002/sam.11461

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18−22

Montanari A, Young G, Savenije HHG, Hughes D, Wagener T, Ren LL, Koutsoyiannis D, Cudennec C, Toth E, Grimaldi S, et al. (2013) "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022. Hydrological Sciences Journal 58(6):1256–1275. doi:10.1080/02626667.2013.809088

Papacharalampous GA, 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. doi:10.1016/j.scitotenv.2020.144612

Papacharalampous GA, Tyralis H, Markonis Y, Hanel M (2022a) Hydroclimatic time series features at multiple time scales. arXiv:2112.01447

Papacharalampous GA, Tyralis H, Pechlivanidis IG, Grimaldi S, Volpi E (2022b) Massive feature extraction for explaining and foretelling hydroclimatic time series forecastability at the global scale. Geoscience Frontiers 13(3):101349. doi:10.1016/j.gsf.2022.101349

Tyralis H, Papacharalampous GA (2021) Boosting algorithms in energy research: A systematic review. Neural Computing and Applications 33:14101–14117. doi:10.1007/s00521-021-05995-8

Tyralis H, Papacharalampous GA, 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. doi:10.3390/w11050910

Wang X, Smith K, Hyndman RJ (2006) Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 13:335–364. doi:10.1007/s10618-005-0039-x