


METHODS PAPER

# Clustering of causal graphs to explore drivers of river discharge

Wiebke Günther<sup>1</sup> , Peter Miersch<sup>2</sup>, Urmi Ninad<sup>3,1</sup> and Jakob Runge<sup>1,3</sup>

<sup>1</sup>Institute of Data Science, German Aerospace Center, Jena, Germany

<sup>2</sup>Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

<sup>3</sup>Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany

**Corresponding author:** Wiebke Günther; Email: [wiebke.guenther@dlr.de](mailto:wiebke.guenther@dlr.de)

**Received:** 16 May 2023; **Accepted:** 30 May 2023

**Keywords:** catchment hydrology; causal effect estimation; causal inference; clustering

## Abstract

This work aims to classify catchments through the lens of causal inference and cluster analysis. In particular, it uses causal effects (CEs) of meteorological variables on river discharge while only relying on easily obtainable observational data. The proposed method combines time series causal discovery with CE estimation to develop features for a subsequent clustering step. Several ways to customize and adapt the features to the problem at hand are discussed. In an application example, the method is evaluated on 358 European river catchments. The found clusters are analyzed using the causal mechanisms that drive them and their environmental attributes.

## Impact Statement

This paper discusses how one can classify river catchments based on causal effects between temperature, precipitation, and discharge, that is, the volume of water that leaves the given area over a certain time. The proposed method is applied to 358 European catchments.

## 1. Motivation

In hydrological research, the basic classification of catchments, that is an area where all water drains to a single outlet, is done through analyzing the response of discharge to precipitation input. Here, discharge refers to the volume of water flowing through a river channel per time unit (Turnipseed and Sauer, 2010). However, discharge characteristics are highly heterogeneous, as they depend on catchment characteristics like area, slope, and land cover. Furthermore, catchment behavior is also driven by regional climate and the interaction of hydrometeorological processes, for instance, snow melt, soil moisture, and precipitation events. Historically, such classification or clustering of catchments has been done using a carefully selected subset of their attributes (Wagener et al., 2007). However, many of these attributes are correlated, making it difficult to select a minimal predictive set. Hydrological signatures have also been used as a basis for clustering, leading to clusters that are partially close to ones based on climate behavior (Kuentz et al., 2017; Jehn et al., 2020).

In recent years, the importance of data-driven analysis has been recognized (Peters-Lidard et al., 2017) and catchment classification has also been done using machine learning techniques (Jiang et al., 2022).

However, we note that the tools of causal inference seem to be under-explored to derive relationships between meteorological variables that can serve as a foundation for classifying river catchments. Causal inference algorithms allow, under specific assumptions, to discover and quantify causal relationships from observational data. Moreover, their outputs are inherently explainable by design. This makes them especially suitable for the domain of Earth sciences, since here it is often infeasible to conduct controlled experiments to arrive at causal conclusions (Gnecco et al., 2019; Runge et al., 2019; Samarasinghe et al., 2020).

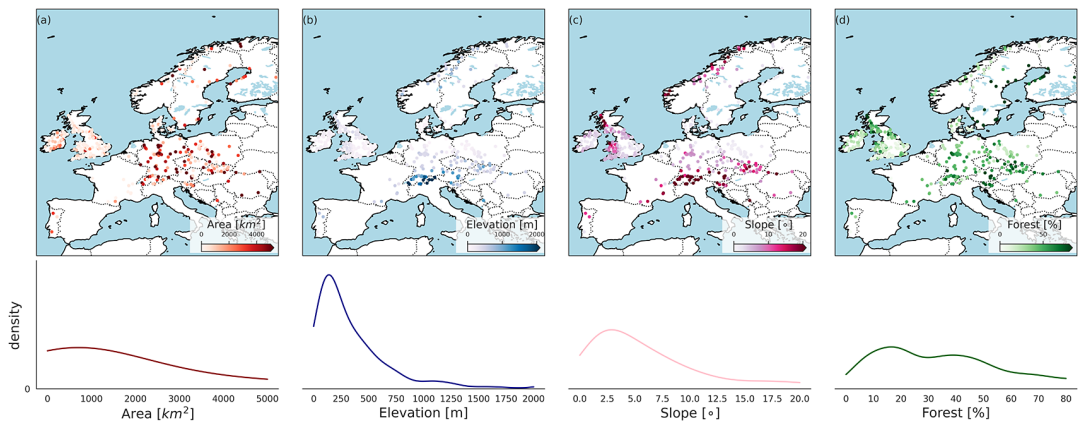
In our analysis, we investigate the impact of temperature and precipitation on observed discharge in European catchments. See [Figure 1](#) for an overview of the considered catchments and their environmental characteristics. We assume linear relationships between the variables, and employ the PCMCI framework by Runge et al. (2019), in combination with expert knowledge, to identify causal relationships. Based on the found causal graphs, we also quantify the causal effect (CE) between the variables based on the path method (Wright, 1934) and Pearl's causal framework (Pearl, 2000) following Runge et al. (2015). Subsequently, we use the estimated CEs as features for clustering using the  $k$ -means algorithm (Lloyd, 1982).

In doing so, we show how methods from causal inference can improve our understanding of discharge-generating mechanisms.

## 2. Method

We want to explore differences in the causal structure of discharge and its drivers across Europe. Our method relies on observations from multiple data sets that have been collected at different locations, that is, on data that is heterogeneous with respect to the environment by which it is influenced. In our application setting, the considered data sets correspond to the catchments, where we observed temperature, precipitation, and discharge. Within each of the  $M$  data sets (i.e., catchments), we have  $N$  observational time series, denoted by the vector  $\mathbf{X}^m$  where  $m$  is the data set index, for variables that are the same across data sets. An observation of variable  $i$  at time point  $t$  within data set  $m$  is then denoted by  $X_t^{i,m}$ . To ease notation, we suppress the data set index in the following. Our analysis comprises three main steps, that is,

1. Estimate a causal graph for each data set (i.e., catchment) separately using the PCMCI algorithm.
2. Extract features of the causal graphs that will be used in the subsequent clustering step. Here, one is rather free in the choice of features, and we will discuss a few different options below. In this work, we focus on features based on linear CE estimation. This step can also be seen as selecting a mapping from the space of graphs into  $V \subset \mathbf{R}^d$  with  $d$  the dimension of the feature space. For instance,  $d$  could be the number of lagged CEs between specific variables  $X^i$  and  $X^j$ . In this space  $V$ , we can use the Euclidean distance for clustering in the next step.



**Figure 1.** Geographic overview and distribution of European catchments characteristics. (a) Area, (b) average elevation, (c) average slope, and (d) forest cover.

3. Finally, employ the  $k$ -means clustering algorithm to find graphs (that correspond to catchments) with similar causal patterns.

### 2.1. Causal discovery

Interdependencies within a system of variables can be conveniently represented by graphical models, where the variables are represented by the nodes of the graph and edges indicate dependence between the respective nodes. Of specific interest to us are causal graphs, where an edge indicates a causal relationship between two nodes. Here, we consider them to be directed acyclic graphs.

To be able to learn such causal graphs from observational data alone, we have to impose some assumptions. In particular, we assume that the causal Markov condition, that is, the independence of error terms driving each subprocess holds, as well as faithfulness which ensures that the graph includes all conditional independencies that hold in the true underlying process. We also assume that all relevant variables are included in the model. This is known as causal sufficiency. We also make the stationarity assumption, that is, that the causal relationships do not change over time. Furthermore, we assume that there are no contemporaneous effects.

In this setting, we apply the PCMCI algorithm to learn the time series graph within each data set. PCMCI is a two-stage causal discovery algorithm based on the framework of constraint-based causal discovery (Spirtes et al., 2000) that is suitable for time series. Its first stage, called  $PC_1$  is based on the PC-stable algorithm that discovers relevant conditions for each of the  $N$  variables by iterative independence testing. These conditions are a superset of the true causal parents with high probability. In stage two, the momentary conditional independence (MCI) test uses the conditions found in stage one to infer a causal link  $X_{t-\tau}^i \rightarrow X_t^j$ , that is, it tests  $X_{t-\tau}^i \perp\!\!\!\perp Y_t \mid \text{pa}(Y_t), \text{pa}(X_{t-\tau}^i)$  where  $\text{pa}(X_t^j)$  denotes the parent superset of  $X_t^j$  found in the first step. Conditioning on the parents of the target  $\text{pa}(X_t^j)$  increases the effect size, and conditioning on  $\text{pa}(X_{t-\tau}^i)$  helps to control for false positives in the highly autocorrelated time series case. For further details on PCMCI, refer to Runge et al. (2019).

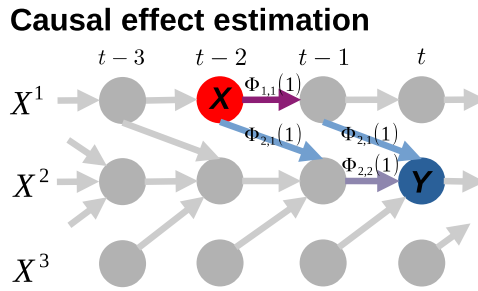
### 2.2. Feature extraction based on CE estimation

Now, we will give further detail on the feature extraction procedure within step 2. The CE of setting  $X_{t-\tau}^i$  to  $x^*$  on  $X_t^j$  is given by  $\Psi_{ji}(\tau) := \frac{\partial}{\partial x^*} \mathbf{E}[X_t^j \mid \text{do}(X_{t-\tau}^i) = x^*]$ . Note that the do-operator refers to a hard intervention on the system of forcing the value of  $X_{t-\tau}^i$  to be  $x^*$ . Following Runge et al. (2015), to estimate the CE from observational data, we fit a linear model of the following form, where we only estimate the coefficients  $\Phi_{ij}$  that correspond to links in our causal graph (see step 1):

$$X_t^j = \sum_{i=0}^{N-1} \sum_{\tau=1}^{\tau_{\max}} \Phi_{ji}(\tau) X_{t-\tau}^i + \varepsilon_t \quad \text{with } \Phi_{ji}(\tau) \neq 0 \text{ only if } X_{t-\tau}^i \text{ is a parent of } X_t^j. \tag{1}$$

One straightforward way of using this approximation of the causal process for clustering, is to directly consider the vector of path coefficients  $(\Phi_{ij}(\tau))_{\tau=1, \dots, \tau_{\max}, i \in V, j \in W}$  for subsets of variables  $V, W \subset \mathbf{X}$  as features. The path coefficient  $\Phi_{ji}(\tau)$  is a standardized version of the corresponding  $\Phi_{ji}(\tau)$  in (1) and represents the change in the expectation of  $X_t^j$  (in units of its standard deviation) induced by raising  $X_{t-\tau}^i$  by 1 standard deviation, while keeping all other parents constant. It, therefore, quantifies the direct effect of  $X_{t-\tau}^i$  on  $X_t^j$ . One potential problem with this approach could be that the frequently (for every not directly linked pair of nodes) appearing zero-values could dominate or skew the clustering results. Graphs with the same nonparental relationships would be in the same cluster even though they exhibit very different behavior in the present links.

Therefore, it might be preferable to use the lagged CEs of the variables in  $V$  onto the variables in  $W$  instead. The matrix  $\Psi(\tau)$  of the standardized CEs between all variables for time lag  $\tau$  can be iteratively computed based on the standardized path coefficients  $\Phi(\tau)$ , as the entry  $\Psi_{ji}(\tau)$  corresponds to the sum over



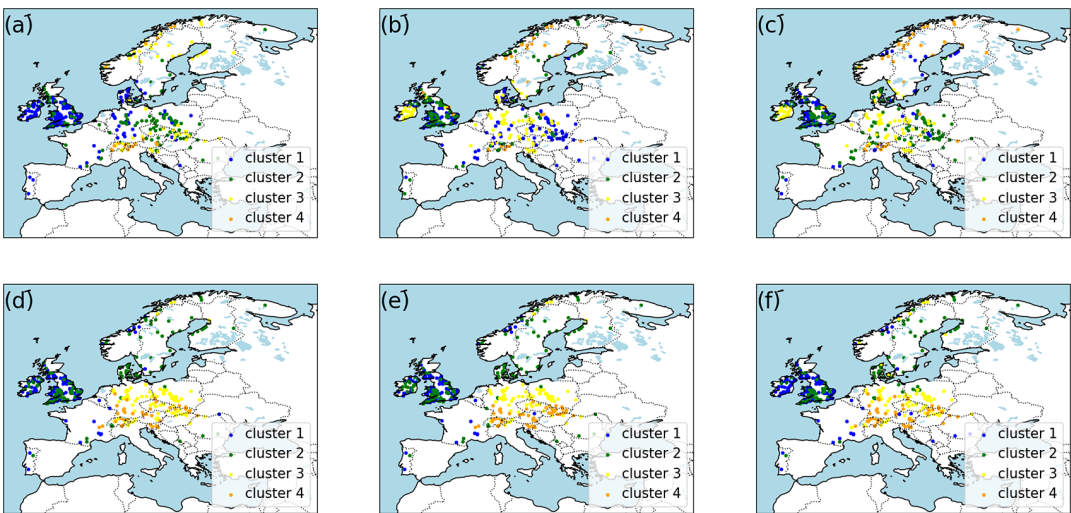
**Figure 2.** Illustration of total causal effect estimation in a time series graph for  $\tau = 2$ . The CE  $\Psi_{ji}(2)$  between  $X_{t-2}^1 =: X$  and  $X_t^2 =: Y$  is computed by summing over the products of path coefficients (link labels) along each path, that is,  $\Psi_{21}(2) = \Phi_{1,1}(1)\Phi_{2,1}(1) + \Phi_{2,1}(1)\Phi_{2,2}(1)$ .

the products of path coefficients along all path between  $X_{t-\tau}^i$  and  $X_t^j$ , see also Figure 2. If we want to consider a large maximal time lag  $\tau_{\max}$ , using  $(\Psi(\tau))_{\tau=1, \dots, \tau_{\max}}$  will give us a high-dimensional feature space.

However, we can reduce its dimension by using aggregated measures. Here, there are various ways in which aggregation is possible. The lagged CEs could be aggregated over the number of variables, or by aggregating in the direction of the lags, for example, by averaging

$$\frac{1}{\tau_{\max}} \sum_{\tau=1}^{\tau_{\max}} \Psi(\tau),$$

or both, as is done in calculating the average causal effect (ACE) or average causal susceptibility (ACS). Please refer to Runge et al. (2015) for the formulas. Note that many different ways of aggregation besides averaging are possible and have to be carefully evaluated within each application context. Additional to averaging, we have included results for maximal lagged CEs, that is, using the features  $(\max_{\tau}(\Psi_{ij}(\tau)))_{i,j=1, \dots, N}$ , in Figure 3.



**Figure 3.** Cluster results for different choices of features. For a brief explanation of these, and other, feature extraction methods see Section 2.2. In the following  $X^1$  denotes temperature,  $X^2$  precipitation and  $X^3$  discharge. The features correspond to (a) average lagged CE of  $X^1$  on  $X^3$ , (b) average lagged CE of  $X^2$  on  $X^3$ , (c) average lagged CE of  $X^i$  on  $X^j$  for all  $i \neq j$ , (d) maximal lagged CE of  $X^i$  on  $X^j$  for all  $i \neq j$ , (e) path coefficients, and (f) ACE of  $X^i$  on the system for all  $i$ .

### 3. Application

Now, we apply our method to the problem of classifying European catchments in terms of their causal interactions between temperature, precipitation, and discharge.

#### 3.1. Data and setup

In this study, we consider 358 gauged catchments across Europe selected based on data availability of daily observational discharge, meteorological observations, watershed boundaries, and morphological catchment characteristics for the study period from 1950 to 2021. Daily observational discharge and watershed boundaries were used from the Global Runoff Data Centre (GRDC) dataset (<https://www.bafg.de/GRDC>, accessed: 21 July 2022). Using the observational  $0.1^\circ$  daily gridded E-OBS dataset (version 26e, Cornes et al., 2018), catchment averaged precipitation and mean surface temperature time series were derived using area-weighted averages of the grid cells within the catchments' boundary. For analysis of the clusters, catchment averaged slope and altitude were derived from the USGS digital elevation model (Earth Resources Observation And Science (EROS) Center, 2017), land cover (impervious, forest, pervious) from the European Space Agency GlobCover (Arino et al., 2012). Deriving the morphological variables from gridded data products was done to allow for comparison to future studies including process-based hydrological models, therefore we restricted our study to catchments where this data is available. We furthermore limit our study to catchments with a minimum of 30 years of continuous discharge records within the study period to ensure sufficient data for our causal inference approach. The catchment areas range between 16 and  $10,000\text{km}^2$ —larger catchments, with the increasing importance of spatial heterogeneity on discharge generation, were omitted. Overall, the selected catchments encompass a large variety of locations, areas, average altitudes, and morphologies, with a cumulation in Great Britain and Ireland due to the availability of watershed boundaries in those regions (Figure 1).

The implementation of our method is based on the Tigramite-package (<https://github.com/jakobrunge/tigramite/tree/master>). For the causal discovery step (step 1 above), we use the PCMCI algorithm. We assume a linear relationship between the variables and therefore use the partial correlation conditional independence test with a 0.05 significance level in the PCMCI algorithm. We also provide the general knowledge that discharge *cannot* cause either temperature nor precipitation to the causal discovery method. We look for causal relationships up to 30 time steps in the past, that is,  $\tau_{\max} = 30$ . Any time slices of samples with missing values are discarded. For step 2, we use the functionality provided in the LinearMediation class of Tigramite. Within step 3, we use the scipy-implementation of the  $k$ -means algorithm with four cluster centers. Results for three and five cluster centers are provided in the Supplementary Material. The choice of this hyperparameter was based on a combination of the Elbow method and the silhouette score (Kaufman and Rousseeuw, 2009). The elbow method corresponds to finding the point after which the sudden drop in average distance from the centroid slows down, see Teoh and Rong (2022) for details. The silhouette score provides a more quantitative measure of how similar and separated the found clusters are. We plotted the average distance from the centroid, the silhouette score, and its average, respectively, over the number of cluster centers. These plots can be found in the Supplementary Material. The results vary slightly over the different features that we used as a basis for the clustering. However, generally speaking, four cluster centers seem to be a good choice that also leads to clusters that are informative and of roughly the same size.

#### 3.2. Results

As one would expect, we find slightly different clusters depending on the causal aspect of the system we are focusing on during the feature extraction phase of our method. We present results for a few different feature choices in Figure 3, more can be found in Supplementary Figure S10. However, some metrics also yield very similar results. We find only minor variation between the clusters for ACE and ACS. Only considering the CE of precipitation on discharge gives us very similar clusters to considering the

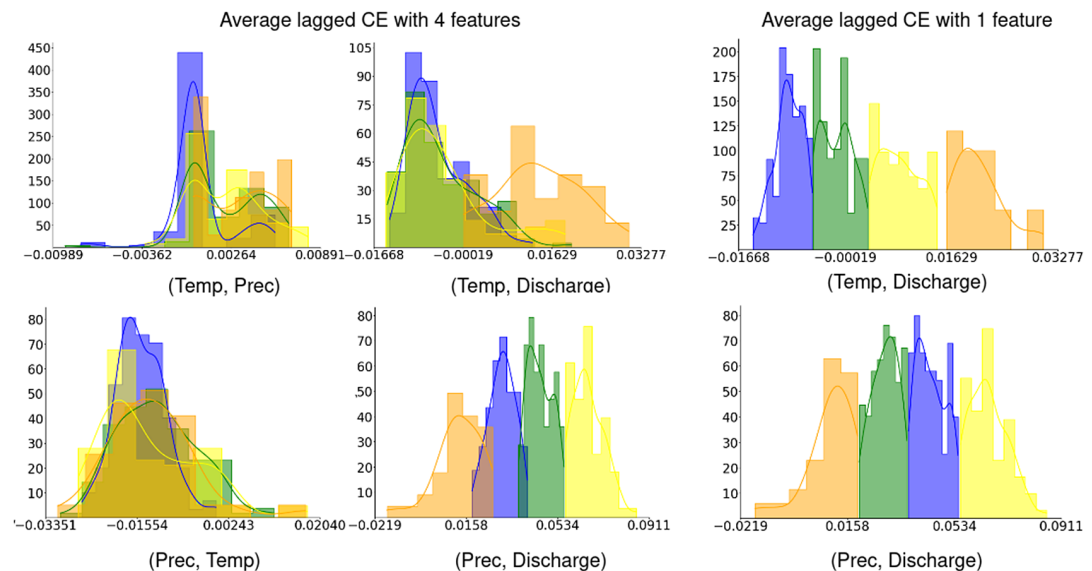
relationships between all variables. Moreover, clusters based on joint CE are almost identical to only considering the maximal CE overall lags for each variable pair.

Supplementary Figure S10 also illustrates the results for an increasing number of clusters. We observe that the clusters tend to become regionally very scattered as the number of clusters grows. This effect is especially apparent in Middle and Western Europe. This is possibly due to the high spatial variability in the European climate and topography.

### 3.3. Distinct causal behavior

The time series graphs that are discovered in step 1 of our method do not differ much across catchments in terms of skeleton, that is, the graph without orientations, and link direction, see Supplementary Figure S10. This is to be expected, since we provided the causal discovery algorithm with substantial expert knowledge. Therefore, we get more regionally varying clusters if we take the variations in strength of the links, or, in other words, of the value of the path coefficients, into account. This behavior is visible in Supplementary Figure S10. In this section, we want to investigate the features that we used for clustering further by having a closer look at the distribution of each component of the feature vectors within each cluster. This will allow us to associate the observed clusters with a specific causal pattern. As an example, we will focus on the average joint CE and some variants of it.

In Figure 4, we observe that for one-dimensional feature spaces, we can directly see that the clusters correspond to a specific range of the feature values. This is shown in the right column of Figure 4 for the one-dimensional features average lagged CE of temperature on discharge, and average lagged effect of precipitation on discharge, respectively. For instance, in the yellow cluster, we see a strong average CE of temperature on discharge. As can be expected, this cluster has the lowest CE of precipitation on discharge. Interestingly, we find that this cluster is also associated with a specific European region, see Figure 3.



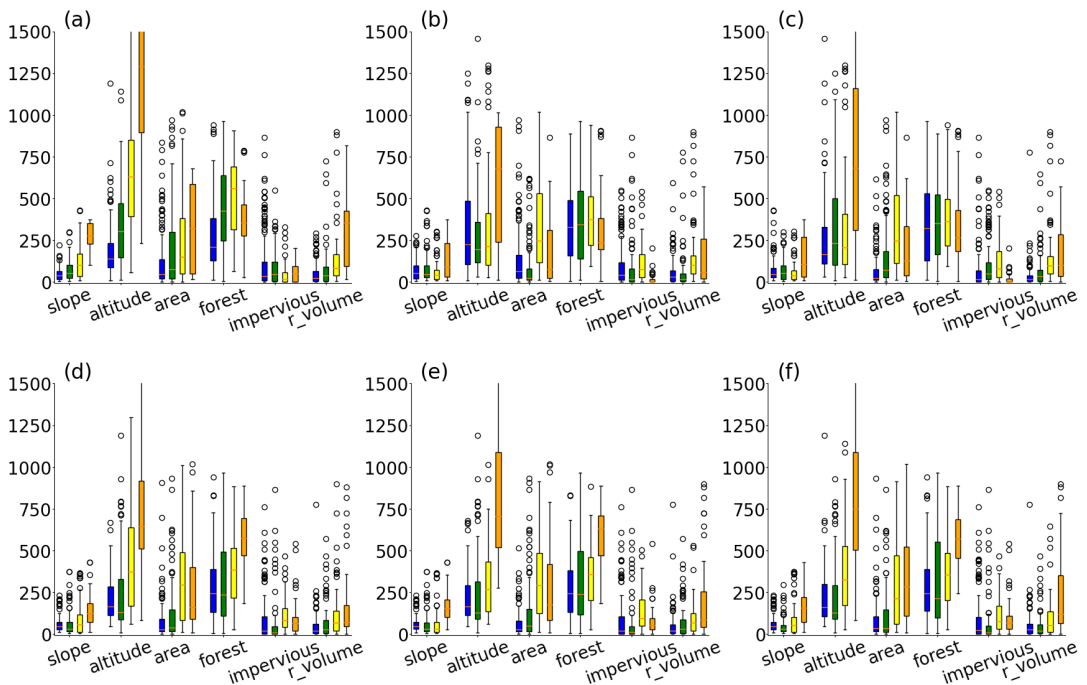
**Figure 4.** Histogram of one-dimensional CE-based features (right column) in comparison to clustering based on feature vectors that include the average joint CE of temperature on precipitation, temperature on discharge, precipitation on temperature, and precipitation on discharge (left and middle column). For instance, the plot in the right upper corner shows the frequency of values of the average lagged CE of temperature on discharge within the different clusters. One can see that in the catchments of cluster 1 (blue), the average lagged CE of temperature on discharge is relatively low in comparison to the other clusters. Colors correspond to the clusters illustrated in Figure 3.

Now, we look at similar plots for the case where we used the four-dimensional feature vector of the average joint CE of temperature on precipitation, temperature on discharge, precipitation on temperature, and precipitation on discharge for clustering (left and middle column in Figure 4). Here, we see that the fourth component almost exactly separates according to the clusters. We suspect that it is dominating the clustering, since the CEs between precipitation and temperature are very low in comparison. More plots can be found in the Supplementary Material.

### 3.4. Distribution of catchment attributes within clusters

We are also interested in analyzing the distribution of environmental variables within each of the found clusters. We illustrate this using boxplots in Figure 5. Also, note that the environmental attributes tend to be strongly correlated. For instance, steeper slope generally corresponds to higher altitude. In Figures 3 and 5, general patterns are also visible across all different choices of features, for example, strong differences in mean altitude between one or two of the found clusters and the remaining ones. We can summarize the differences and similarities with respect to the attributes in the following way:

1. cluster: low altitude, small catchments, western/middle Europe,
2. cluster: low to medium altitude, small, western/middle Europe,
3. cluster: low to medium altitude, larger area, western/middle Europe, and
4. cluster: high altitude, with larger area: seems to be alps, Scandinavian mountains.



**Figure 5.** Distribution of catchment attributes within each cluster. The attributes are slope in degrees times 10, average altitude in m above sea level, area in km<sup>2</sup>, proportion of basin covered by forest, proportion of basin covered by impervious surfaces, and mean annual rain volume in 100km<sup>3</sup>. Colors correspond to the clusters illustrated in Figure 3. In the following  $X^1$  denotes temperature,  $X^2$  precipitation, and  $X^3$  discharge. The features correspond to (a) average lagged CE of  $X^1$  on  $X^3$ , (b) average lagged CE of  $X^2$  on  $X^3$ , (c) average lagged CE of  $X^i$  on  $X^j$  for all  $i \neq j$ , (d) maximal lagged CE of  $X^i$  on  $X^j$  for all  $i \neq j$ , (e) path coefficients, and (f) ACE of  $X^i$  on the system for all  $i$ .

Some of the clusters seem to be strongly related to a specific region, for example, in terms of effect of precipitation on discharge we observe a very clear regional cluster in Ireland. However, it is hard to interpret without further domain knowledge. Moreover, the results could be skewed by the choice of catchments. A large proportion of the catchments are located in Great Britain and Ireland since here watershed boundaries are readily available. This was a limiting factor in our catchment selection because we wanted to be able to infer the catchment size. How this problem can be alleviated in future work is further discussed in the next section.

#### 4. Discussion

In principle, our method is applicable in any situation where there are multiple data sets containing observations of the same variables. This makes our method applicable to a wide range of problems and research areas, not even limited to the domain of Earth science.

Of course, in practice, the availability and quality of data are limiting factors. Therefore, an analogous analysis can be conducted in regions with good public data availability, like North America, whereas it would be more challenging in regions with a lower coverage by weather and gauging stations.

Also, the construction of geographically averaged timeseries and associated environmental attributes might not be as clean-cut as in our application, where the catchment boundaries provide a natural distinction between different data regions. In other questions relevant to the field of Earth Science, the regions to average over might be more arbitrary. This essentially introduces more hyperparameters into the problem. Another potential problem is an insufficient amount of data. For instance, if one is interested in extreme events, like the flood-generating process instead of the discharge-generating process, then there are probably too few events to obtain stable causal discovery and clustering results.

There are still many avenues open for future work. The main next step that we want to take is to only focus on peak discharge events and to see how the drivers of extreme events differ from the baseline behavior of normal fluctuations in discharge. Another major point that we want to explore further is the influence of climate change on the system. So far we have assumed stationarity of the time series but this assumption is violated in a warming climate.

Furthermore, the effect of including more variables into our model has to be investigated. In other words, effects of the possible violation of the underlying causal sufficiency assumption within the PCMC algorithm has to be explored.

Also, further analysis has to be done to distinguish the clusters more, especially clusters 1 and 2. In particular, the imbalance in the dataset of most clusters being located in Great Britain or Ireland has to be addressed. The reason for this is that here a lot of clusters with watershed boundaries are available. However, following the approach presented by Xie et al. (2022), it is possible to infer the watershed boundaries for many more clusters across Europe making them suitable for our analysis. Moreover, the European climate has a high spatial variation that depends on far more factors than have been analyzed by us so far.

#### 5. Conclusion

In this work, we presented how tools from causal discovery and effect estimation can be combined with clustering techniques. The presented approach is very customizable and thus suitable for a wide variety of problems and domains. In an application example, we have explored how causal drivers of discharge, specifically the strengths of their CEs, differ across Europe. We saw how the causal inference methods allow us to find clusters of catchments that can guide domain experts in developing and evaluating hypothesis based on observational data, and how these clusters are affected by different choices in designing the features.

**Acknowledgments.** We thank Jakob Zscheischler and Shijie Jiang for their helpful comments. We also thank the anonymous reviewers whose suggestions helped improve this manuscript.

**Author contribution.** Conceptualization: W.G., P.M., J.R.; Data curation: P.M.; Data visualization: W.G., P.M.; Methodology: W.G., P.M., J.R.; Writing—original draft: W.G., P.M.; Writing—review and editing: W.G., P.M., U.N., J.R. All authors approved the final submitted draft.



**Competing interest.** The authors declare no competing interests exist.

**Data availability statement.** Replication data can be found in the Global Runoff Data Centre (GRDC) data set: [https://www.bafg.de/GRDC/EN/Home/homepage\\_node.html](https://www.bafg.de/GRDC/EN/Home/homepage_node.html).

**Ethics statement.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** W.G. and P.M. were supported by the Helmholtz AI project *CausalFlood* (grant no. ZT-I-PF-5-11). U.N. and J.R. were supported by grant no. 948112 CausalEarth of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program.

**Provenance statement.** This article is part of the Climate Informatics 2023 proceedings and was accepted in *Environmental Data Science* on the basis of the Climate Informatics peer review process.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/eds.2023.17>.

## References

- Arino O, Perez JJR, Kalogirou V, Bontemps S, Defourny P and Van Bogaert E (2012) Global Land Cover Map for 2009 (GlobCover 2009).
- Cornes R, Schrier G, Van den Besselaar E and Jones P (2018) An ensemble version of the EOBS temperature and precipitation data sets. *Journal of Geophysical Research Atmospheres* 123, 9391–9409. <https://doi.org/10.1029/2017JD028200>
- Earth Resources Observation and Science (EROS) Center (2017) Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010). <http://doi.org/10.5066/F7J38R2N>. Available at <https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digitalelevation-global-multi-resolution-terrain-elevation> (accessed 19 November 2020).
- Gnecco N, Nicolai Meinshausen JP and Engelke S (2019) Causal discovery in heavy-tailed models. *The Annals of Statistics* 49(3), 1755–1778.
- Jehn FU, Bestian K, Breuer L, Kraft P and Houska T (2020) Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences* 24(3), 1081–1100.
- Jiang S, Bevacqua E and Zscheischler J (2022) River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences* 26(24), 6339–6359. <https://doi.org/10.5194/hess-26-6339-2022>. Available at <https://hess.copernicus.org/articles/26/6339/2022/> (accessed 20 January 2023).
- Kaufman L and Rousseeuw PJ (2009) *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons.
- Kuentz A, Arheimer B, Huntecha Y and Wagener T (2017) Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences* 21(6), 2863–2879.
- Lloyd S (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2), 129–137.
- Pearl J (2000) *Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
- Peters-Lidard CD, Clark M, Samaniego L, Verhoest NEC, van Emmerik T, Uijlenhoet R, Acheng K, Franz TE and Woods R (2017) Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences* 21(7), 3701–3713. <https://doi.org/10.5194/hess-21-3701-2017>. Available at <https://hess.copernicus.org/articles/21/3701/2017/> (accessed 27 January 2023).
- Runge J, Bathiany S, Bollt E, Camps-Valls G, Coumou D, Deyle E, Glymour C, Kretschmer M, Mahecha M, Munoz J, Nes E, Peters J, Quax R, Reichstein M, Scheffer M, Scholkopf B, Spirtes P, Sun J and Zscheischler J (2019) Inferring causation from time series in earth system sciences. *Nature Communications* 10, 2553. <https://doi.org/10.1038/s41467-019-10105-3>
- Runge J, Nowack P, Kretschmer M, Flaxman S and Sejdinovic D (2019) Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* 5(11), eaau4996.
- Runge J, Petoukhov V, Donges JF, Hlinka J, Jajcay N, Vejmelka M, Hartman D, Marwan N, Paluš M and Kurths J (2015) Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications* 6(1), 1–10.
- Samarasinghe SM, Deng Y and Ebert-Uphoff I (2020) A causality-based view of the interaction between synoptic and planetary-scale atmospheric disturbances. *Journal of the Atmospheric Sciences* 77(3), 925–941.
- Spirtes P, Glymour CN, Scheines R and Heckerman D (2000) *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
- Teoh TT and Rong Z (2022) Clustering. In *Artificial Intelligence with Python*. Singapore: Springer, pp. 213–218.
- Turnipseed DP and Sauer VB (2010) *Discharge Measurements at Gaging Stations*. Technical Report, US Geological Survey.
- Wagener T, Sivapalan M, Troch P and Woods R (2007) Catchment classification and hydrologic similarity. *Geography Compass* 1(4), 901–931.
- Wright S (1934) The method of path coefficients. *The Annals of Mathematical Statistics* 5(3), 161–215.
- Xie J, Liu X, Bai P and Liu C (2022) Rapid watershed delineation using an automatic outlet relocation algorithm. *Water Resources Research* 58(3), e2021WR031129.

**Cite this article:** Günther W, Miersch P, Ninad U and Runge J (2023). Clustering of causal graphs to explore drivers of river discharge. *Environmental Data Science*, 2: e25. doi:10.1017/eds.2023.17