Supplementary Information

S.1 GRDC data

Out of 1182 available stations, 64 are excluded because no boundary data is available, and the catchment area exceeds a size threshold (section S.1.1). Table S1 shows the results of a manual quality assurance procedure, which leads us to exclude another 15 catchments whose data are implausible for pristine catchments (section S.1.2). Section S.1.3 shows the length of the flow record for the 1103 catchments that are used in this study.

S.1.1 Area threshold used to exclude GRDC catchments for which the catchment-averaged climate can not be calculated by using catchment boundaries

Information on catchment boundaries is not available for all catchments in the GRDC data set. If boundaries are available, we use this information to calculate a catchment-averaged climate. Where boundaries are not available, but the catchment is small, we can use the climate at location of the catchment's gauge as representative of the climate in the entire catchment. This section describes the procedure used to determine the catchment size threshold above which we exclude the catchment from use in this study. The GRDC data set provides the location of each catchment's gauge [latitude and longitude coordinates] and the size of the catchment [km^2].

First, we use the square root of catchment area to find the approximate catchment length (Figure S1). The majority of catchments for which no boundary information is available, have an approximated length smaller than the length of 1 grid cell as used in the climate data, and for these no action is necessary. Next, we calculate the correlation length of our three climate indices in the latitude and longitude direction (Figure S2). If a catchment's length is too large compared to the distance until which the climate indices are correlated, we remove the catchment from the data set. Some subjectivity is involved in choosing a threshold level. We need to balance the total number of catchments we can use (which favours keeping larger catchments) and ensuring climatic consistency within the catchment (which favours keeping only small catchments). We have chosen a threshold length of 3 grid cells as an appropriate middle ground. Catchments with an approximated length larger than 3 grid cells (approximately 150 km) are removed from further analysis.

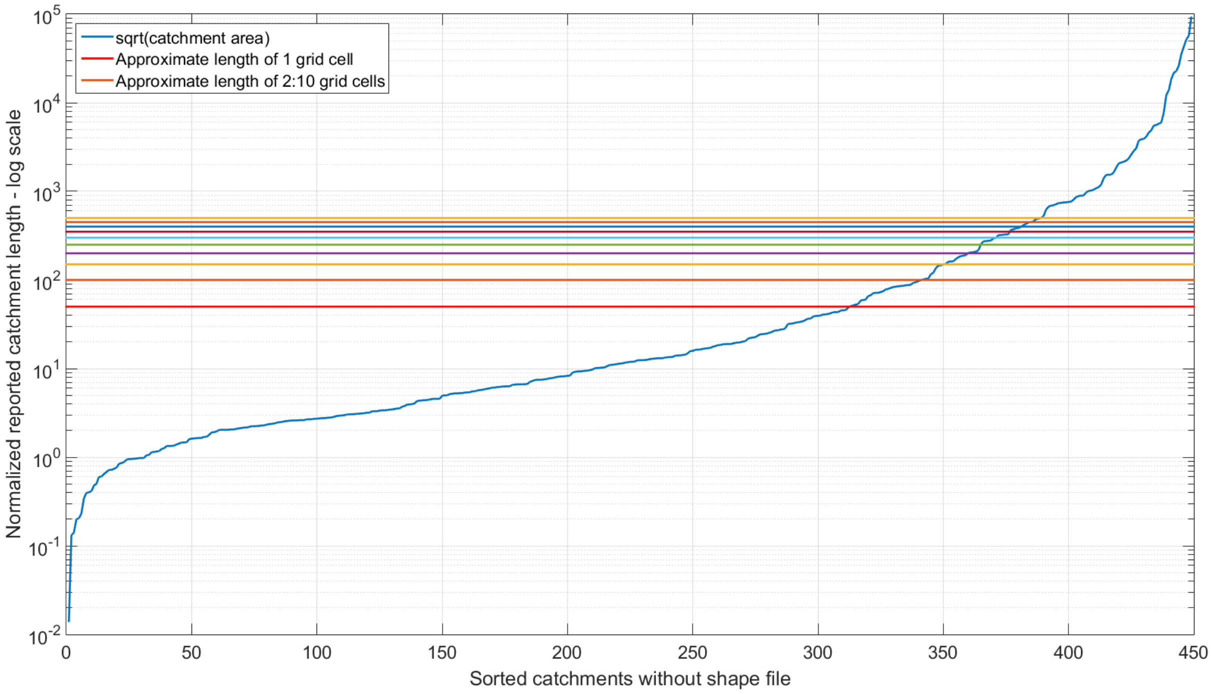


Figure S1: GRDC Pristine Basins without information on catchment boundaries, sorted by approximated catchment length (square root of catchment area). Coloured lines indicated the approximated length of 1 to 10 grid cells as used in the CRU TS climate data. 312 GRDC Pristine Basins have areas smaller than 1 grid cell.

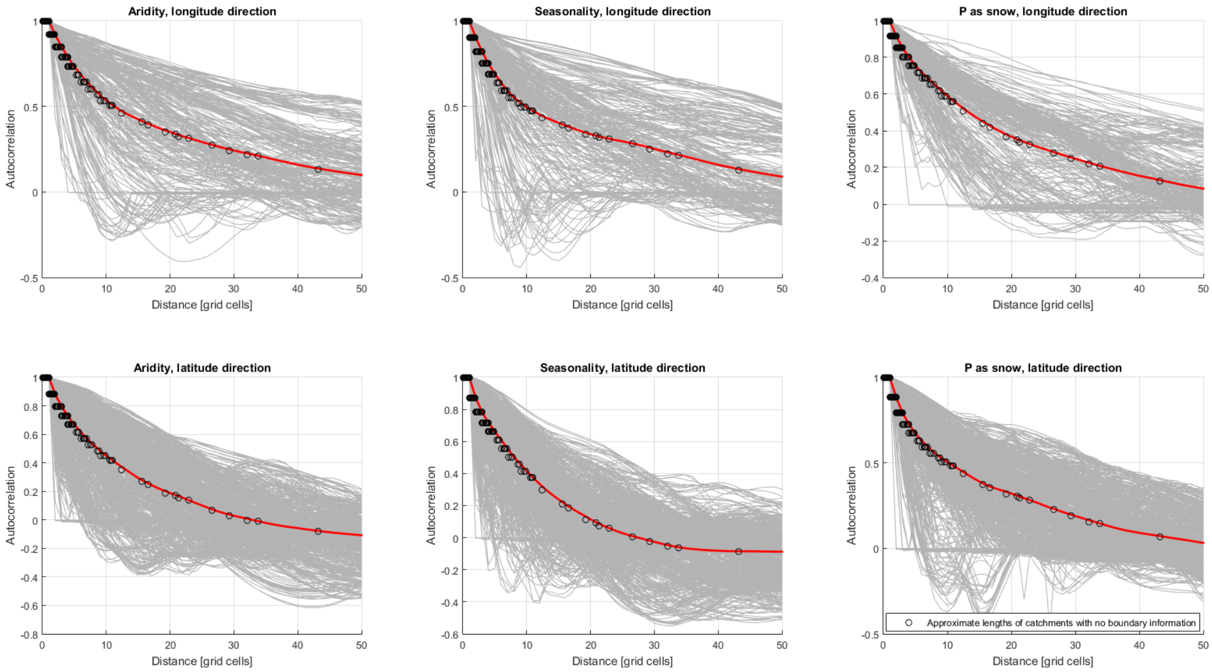


Figure S2: Autocorrelation lengths in longitude (top) and latitude (bottom) direction for three climate indices: average annual aridity (left), aridity seasonality (middle) and fraction of precipitation as snow (right). Circles are the approximate lengths of all GRDC Pristine Basins for which no boundary information is available, matched up with the mean autocorrelation per grid cell distance (red line).

S.1.2 Quality control of flow data

All 1182 Global Runoff Data Centre locations in the "pristine basins" data set have been visually inspected for data errors. Table S1 shows 56 catchments that warranted further investigation and the result of this investigation. We exclude 5 stations due to doubtful data quality, another 5 stations due evidence of hydropower dam construction during the study period, 3 stations due to missing catchment area values and 2 stations due to highly implausible catchment area values. For a further 19 stations we removed part of the time series due to measurement errors (e.g. inexplicable jumps in flow at the start of a new month/year). Two stations are suspected of consistent underestimation of flows during part of the time series, and data were adjusted to better fit the remainder of the time series. Figure S3 summarizes the results of the quality assurance procedure.

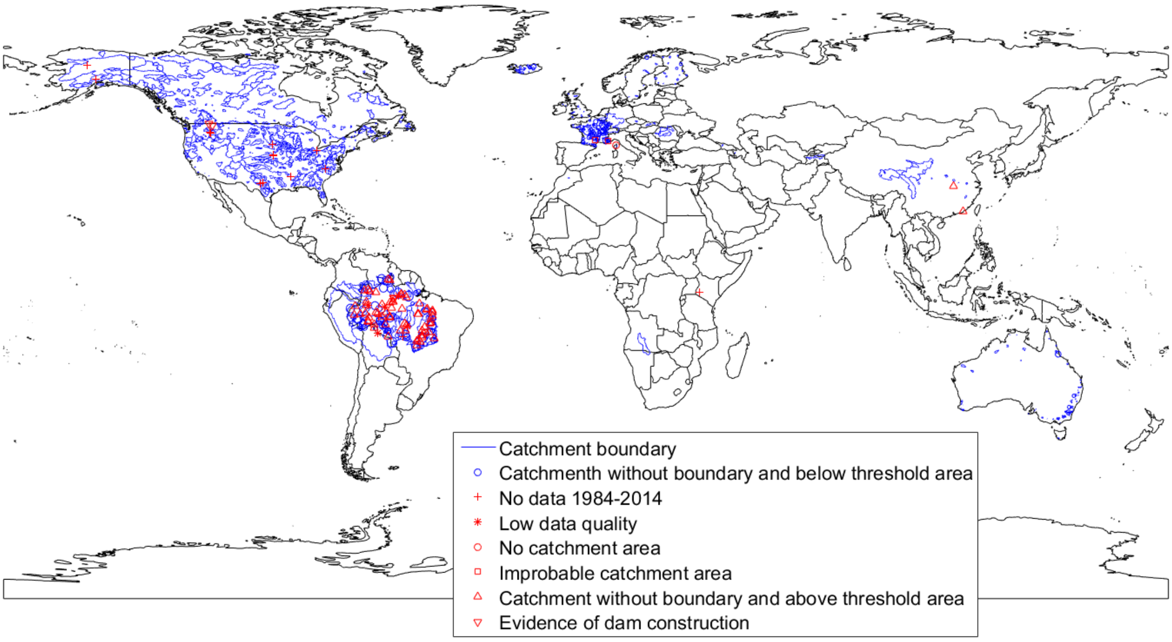


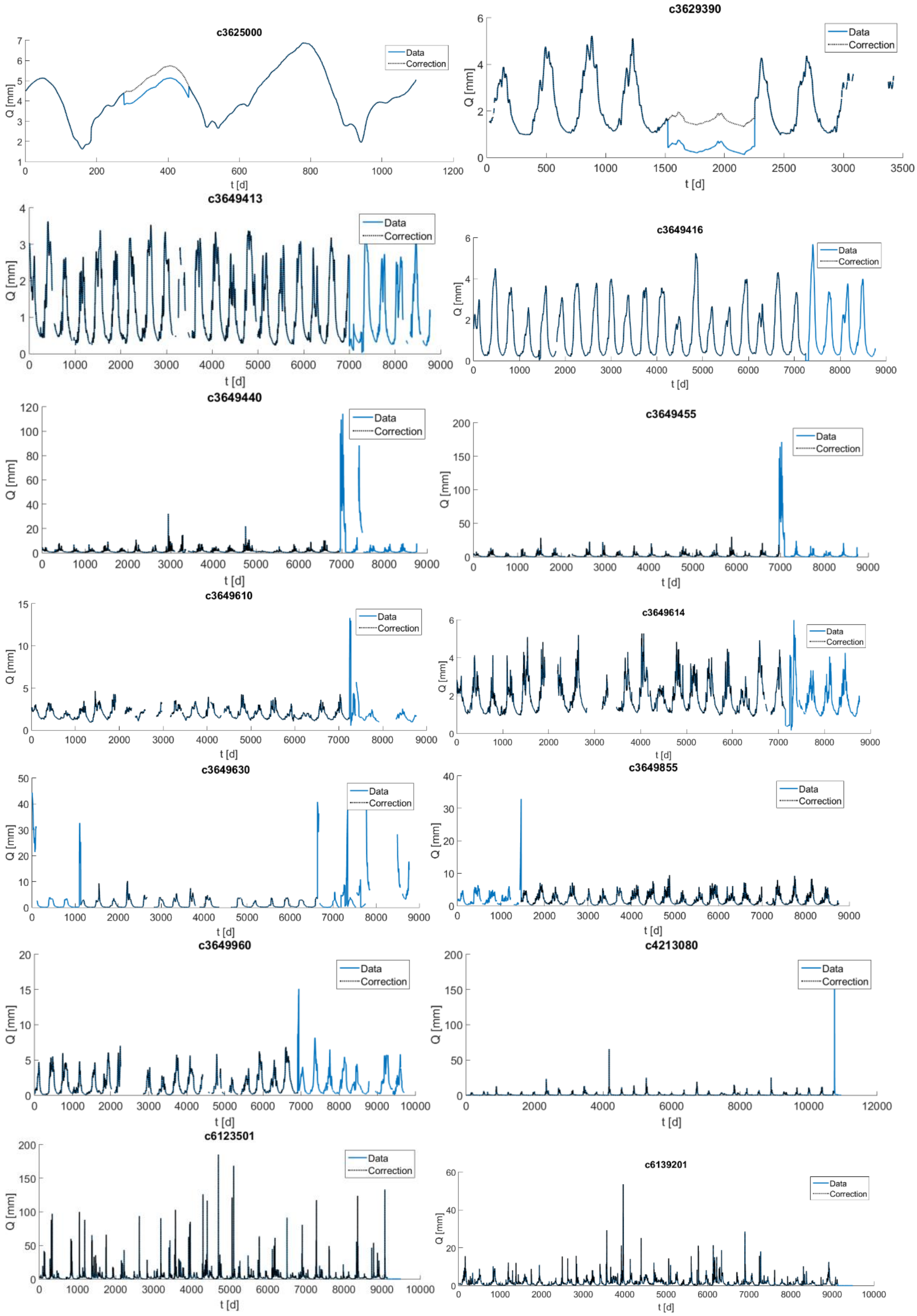
Figure S3: Location and boundaries (where available) of GRDC Pristine Basins used in this study (blue) and Pristine Basins that are removed from further analysis for various reasons (red). Catchments for which no boundary data is available are used if their aproximated length is smaller or equal to a climate correlation threshold and removed from the analysis if larger. The correlation threshold length is set at 3 times the size of a grid cell on which climate data is available, i.e. 0.5° (appendix S.1.1).

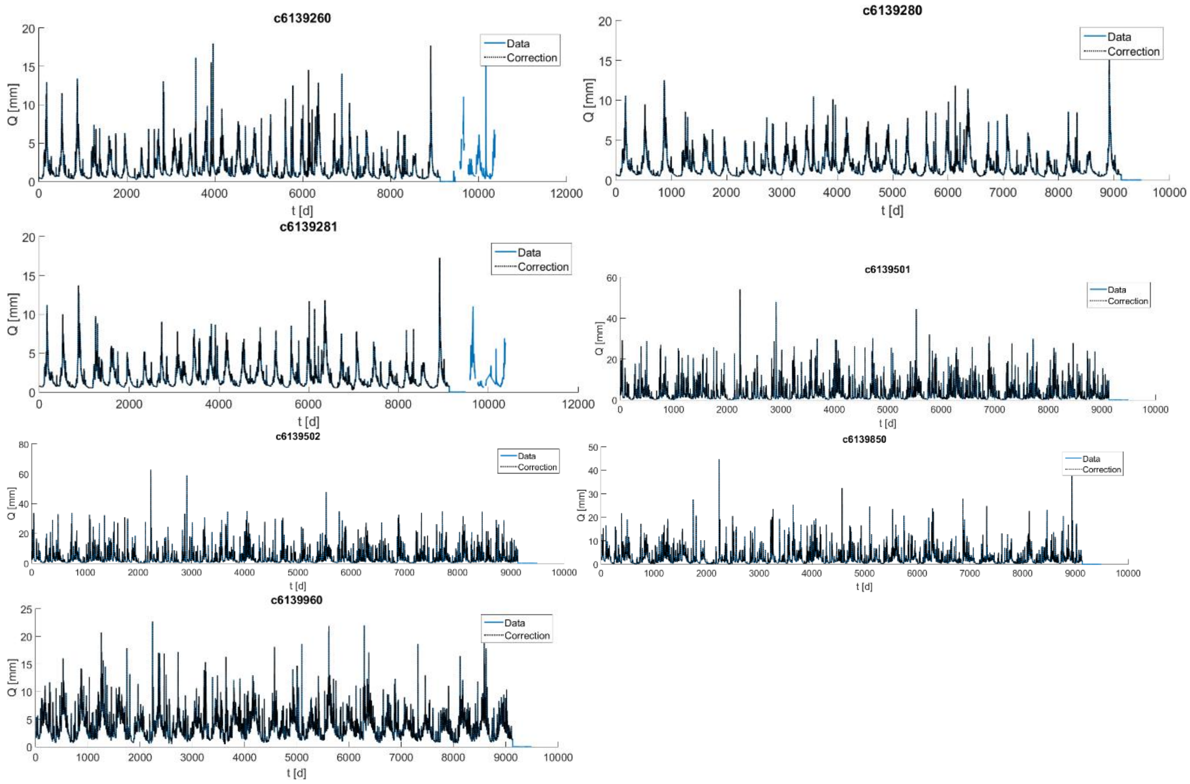
Table S1: GRDC pristine basins that have been investigated further, based on visual inspection of data. Locations with extreme outlier-like peaks have been kept included if additional sources confirmed those peaks to be floods. Extreme peaks with no outside confirmation are treated as data errors.

Catchment ID	Excluded?	Reason
c3624250	Yes	Average flow during first 10 years is 400 m3/s, then a 2-year hiatus in measurements, average flow during last 10 years is 2000 m3/s
c3625000	No, corrected	Amazon river. Days 277:458 are 0.6mm/d lower than rest of record
c3625310	Yes	Amazon river, flows during middle 13 years are factor 5-10 lower than other years
c3627811	Yes	Bolivian river. Very short time series with unexplained drops to zero
c3628300	Yes	Amazon river. Unexplained increase in flow variability for middle 13 years
c3628400	Yes	Amazon river. Flow regime changes drastically due to hydropower dam construction
c3628401	Yes	Amazon river. Flow regime changes drastically due to hydropower dam construction
c3629390	No, corrected	Amazon river. Days 1523:2253 are 1.2mm/d lower than rest of record
c3629800	Yes	Amazon river. Unexplained flow regime changes in middle of data
c3649413	No, corrected	Brazilian river. Sudden drop in flow values, coinciding with start of new month. Possible procedure error. Removed 7000:end

c3649416	No, corrected	Brazilian river. Sudden drop in flow values, coinciding with start of new month. Possible procedure error. Removed days 7245:end
c3649440	No, corrected	Amazon river. Sharp decrease in data quality towards end of series. Removed days 6491:end
c3649455	No, corrected	Amazon river. Sharp decrease in data quality towards end of series. Also mentions dam construction around that period. Removed days 6972:end
c3649461	Yes	No catchment area
c3649465	Yes	No catchment area
c3649610	No, corrected	Amazon river. Sharp decrease in data quality towards end of series. Removed days 7245:end
c3649614	No, corrected	Amazon river. Sharp decrease in data quality towards end of series. Removed days 7153:end
c3649630	No, corrected	Amazon river. Wonky data quality at beginning and end of series. Removed 1:1128 and days 6635:end
c3649855	No, corrected	Amazon river. Sharp decrease in data quality towards beginning of series. Removed suspect data
c3649960	No, corrected	Amazon river. Sharp decrease in data quality towards end of series. Removed days 6910:end
c4102450	No	Alaska river. Shows 3 peaks that look like outliers in Oct-1986, Aug-2006 and Sep-2012. News reports confirm floods at those dates
c4115210	No	Washington river. Shows peak that looks like outlier in Feb-1996. News report confirms a flood.
c4115320	Yes	Montana river. Subject to heavy dam construction. Streamflow record changes drastically for this site around 1970
c4115321	Yes	Idaho river. Subject to heavy dam construction (Libby dam, 1972). Streamflow characteristics change during flow series
c4115322	Yes	Idaho river. Subject to heavy dam construction (Libby dam, 1972). Streamflow characteristics change during flow series
c4118100	No	Nevada river. Shows a peak that looks like an outlier. USGS fact sheet confirms a flood in Jan-1997
c4118105	No	Nevada river. Shows a peak that looks like an outlier. USGS fact sheet confirms a flood in Jan-1997
c4119265	No	Idaho river. Shows outlier-like peak. Various sources confirm a flood event in Jun-2008
c4123255	No	Ohio river. Shows outlier-like peak. Various sources confirm extreme flood in Mar-1997
c4126850	No	Texas river. Peaks are sudden and high. Typical ephemeral stream
c4146230	No	California river. Outlier-like peaks. Typical ephemeral stream
c4146380	No	California river. End of data looks higher than rest of series but no reason to assume errors
c4148070	No	Virginia river, near NC. Outlier-like peak. No sources confirm flood but proximity to NC and matching flood dates imply hurricane-related flood here
c4148110	No	North Carolina river. Outlier-like peak. Various sources confirm Sep-1999 flood due to hurricane
c4148125	No	North Carolina river. Outlier-like peak. Various sources confirm Sep-1999 flood due to hurricane
c4148850	No	Florida river. Low flow variability is high (quick changes), peaks look slower than normal rising limbs. No reason to assume errors, historical data from 1933-2017 looks the same

c4149405	No	Alabama river. Outlier-like peak. News report confirms flood in May-2003
c4149411	No	Mississippi river. Data looks spiky but no reason to assume errors
c4149420	No	Florida river. Outlier-like peak. Weather source confirms flood in Oct-1998
c4149510	No	Florida river. Outlier-like peak. Weather source confirms flood in Oct-1998
c4150310	No	Texas river. Outlier-like peak. News report confirms extreme flood in Oct-1998
c4207750	No	British Columbia river. Outlier like peak. News report confirms extreme flood in Oct-2003
c4213080	No, corrected	Alberta river. Outlier-like peak near end of data. No confirmation found of extreme rain or flow, removed outlier
c5202057	No	New South Wales river. Shows outlier-like peaks. No confirmation of extreme flow but confirmation of extreme rain in the approximate area
c5202065	No	New South Wales river. Shows outlier-like peaks. No confirmation of extreme flow but confirmation of extreme rain in the approximate area
c6123501	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed 9133:end
c6125360	Yes	Catchment area too small, results in unrealistic flows
c6128702	Yes	Catchment area too small, results in unrealistic flows
c6139201	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139260	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139280	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139281	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139501	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139502	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139850	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6139960	No, corrected	France river. Flows drop to (nearly) zero towards end of data, coinciding with start of new year. Possible procedure error. Removed days 9133:end
c6150300	Yes	No catchment area


c3625000
c3629390
c3649413
c3649416
c3649440
c3649455
c3649610
c3649614
c3649630
c3649855
c3649960
c4213080
c6123501
c6139201
Data
Correction
Q [mm]
t [d]



S.1.3 Flow length record

Flow records within the GRDC Pristine Basins data set need to cover a minimum of 20 years, but these do not necessarily overlap with the study period of 1984-2014. Figure S4 shows a histogram of the number of years within the period 1984-2014 available. 1041 catchments (94.3%) used in this study have data records longer than 20 years.

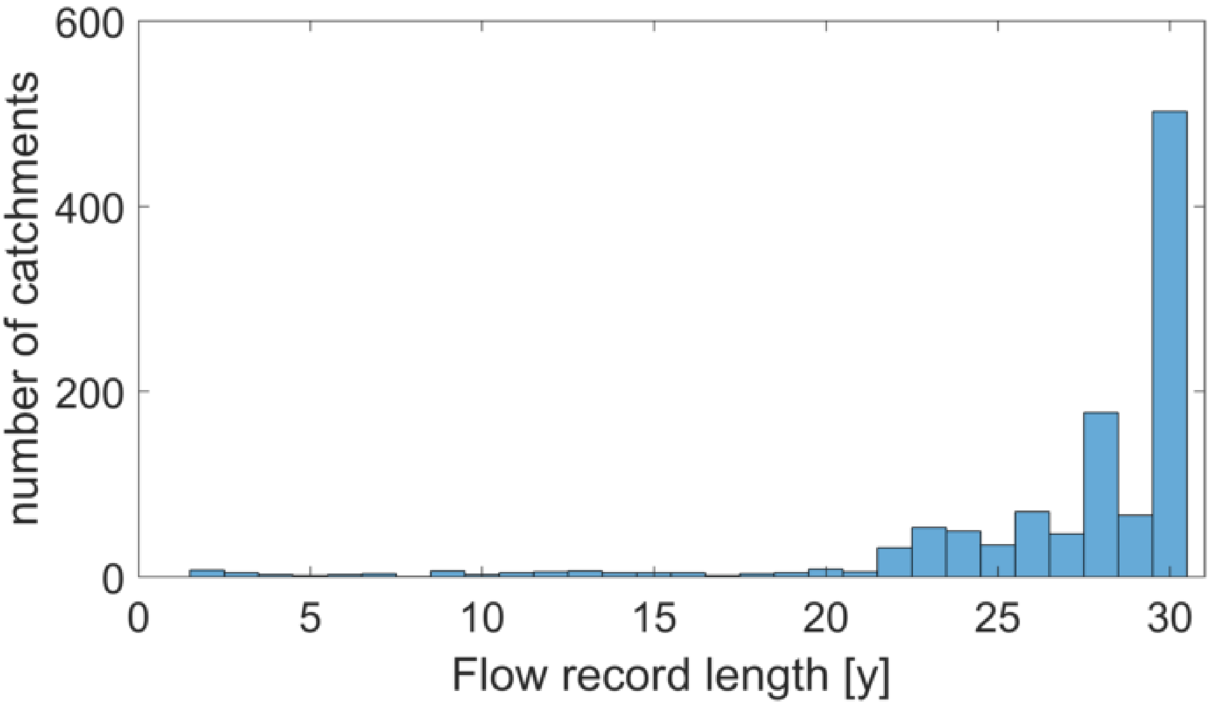


Figure S4: Overview of record length of flow data for the catchments in the GRDC Pristine River basins data set.

S.2 Signature values and significance testing

S.2.1 Procedure

- Context
    - We have calculated 16 signature values for 1103 catchments
    - We have determined the membership of each catchment to each of the 18 possible climate clusters
    - We want to know if there is a statistical difference between the signatures values that are associated with each cluster
- Problem
    - We need a non-parametric method because we are unsure of the distributions that make up the signature values per climate cluster (i.e. we don't know if they are normal, so the t-test shouldn't be used)
    - The Wilcoxon rank test is a suitable test, but can only be used in cases where the samples are unweighted (i.e. all samples belong for 100% to their respective class). The test thus needs to be adapted
- Wilcoxon procedure
    - With the Wilcoxon test, we are testing the null hypothesis $\mu_1 = \mu_2$ against a suitable alternative (Walpole, 1968), with μ_1 being the mean of all observations $x_{i=1:m}$ and μ_2 being the mean of all observations $y_{j=1:n}$
    - The unmodified Wilcoxon test consists of pair-wise comparison of all observations in x_i and y_j. If $x_i > y_j$, a new variable U_x is increased by 1, if $x_i < y_j$, U_y is increased by 1. After all pairs have been compared the p-value for obtaining a given $U = \min(U_x, U_y)$ can be obtained from tables or otherwise, which gives a measure of the likelihood of obtaining a given U value under the null hypothesis. If the p-value is below a certain critical threshold, the null hypothesis can be rejected.
- Solution
    - Modify the Wilcoxon test to account for weighted observations and use bootstrapping to empirically determine the null distribution
    - In this case μ_1 refers to the mean of signatures values belonging to climate cluster 1, consisting of $x_{i=1:m}$ observations with accompanying weights $w_{x,i=1:m}$. Idem for μ_2. During pair-wise comparison, U_x is increased by $w_{x,i} * w_{y,j}$, if $x_i > y_j$. Otherwise, $w_{x,i} * w_{y,j}$ is added to U_y. $U = \min(U_x, U_y)$ can now be tested against an empirical null distribution.
    - To create the empirical null distribution, x_i and y_j are pooled together into a single sample, from which two new uniform random samples are drawn, with replacement, equal in size to x_i and y_j. A U-value is then determined from pairwise comparison of both new samples. This process is repeated *N* times to form an empirical approximation of the null distribution. N is commonly set at 999 (Davison & Hinkley, 1997).
    - Now the U-value of observations can be compared to the empirical distribution of U-values under the null hypothesis, using $p = \frac{r+1}{N+1}$, where *N* is the number of samples and *r* is the number of samples that have a U-value below the U-value calculated for the data (North *et al.*, 2002).
- Illustration of Wilcoxon test, adapted for fuzzy membership

- (a) Distribution of weighted x_i and y_j as observed in the data. The y-axis is meaningless and only used to visualise x_i and y_j better. The U-value for these samples is 204.4
- (b) Uniform random sampling with replacement of x_i and y_j, from a pool made up of x_i and y_j. Visually, the samples are similar to one another, and very different from the pattern seen in (a). The U-value is 697.7
- (c) The average U-value after N-samples levels out at U ≈ 840 and N = 999 seems sufficient.
- (d) Visually comparing the U-value calculated from data with the empirical U-value distribution shows that the null hypothesis can quite probably be rejected. The test statistic $p = \frac{r+1}{N+1} = \frac{0+1}{999+1} = 0.001$ confirms this.

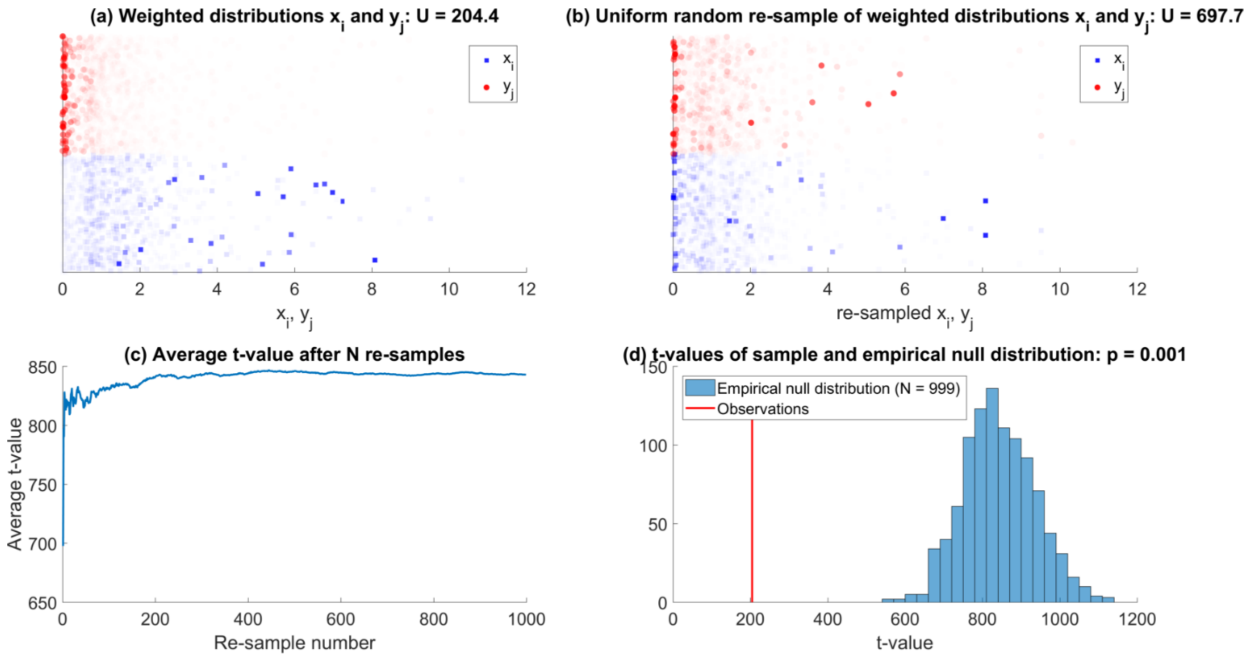


Figure S5: Example of empirical statistical test. (a) Distribution of observed values of the average flow signature for climate clusters 5 (x_i, very wet climate, low seasonality, no frost) and 6 (y_j, very dry climate, low seasonality, no frost). Data points correspond to all 1103 catchments with their transparency dependent on the weight with which they belong to each cluster. The low U-value indicates a large difference in ranks between both samples, which can also be seen from the concentration of high weight values for both climates. (b) Uniform random re-sampling of x_i and y_j, from a sample containing both. The re-sampled distributions are more similar (spread of height weights) and this leads to a higher U-value. (c) The average U-value of the re-sampled distribution after N samples. (d) The resulting distribution of re-sampled U-values (bars) versus the U-value of observations (red line). The observed U-value is very different from the re-sampled distribution, leading to a low empirical p-value and thus a high confidence that the H0 hypothesis $\mu_{xi} = \mu_{yj}$ should be rejected.

S.2.2 Signature values

This section contains plots of the average signature value in each climate cluster and Köppen-Geiger class. Signature values are calculated using daily flow data for each hydrological year available per catchment. This gives up to 30 annual values per catchment (depending on number of available hydrological years), from which the average annual signature value is calculated. Catchment membership to each climate cluster is used to create a weighted average signature value per climate cluster.

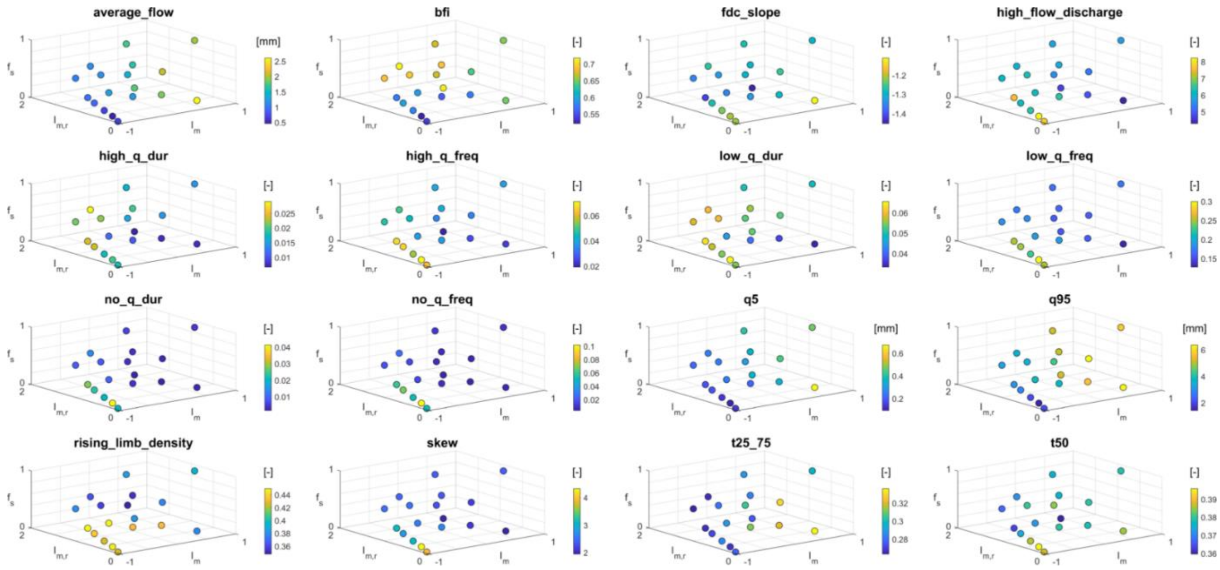


Figure S6: Weighted mean signature values per climate cluster, plotted at the climate cluster centroid location. Values are the weighted mean of average annual signature values per catchment.

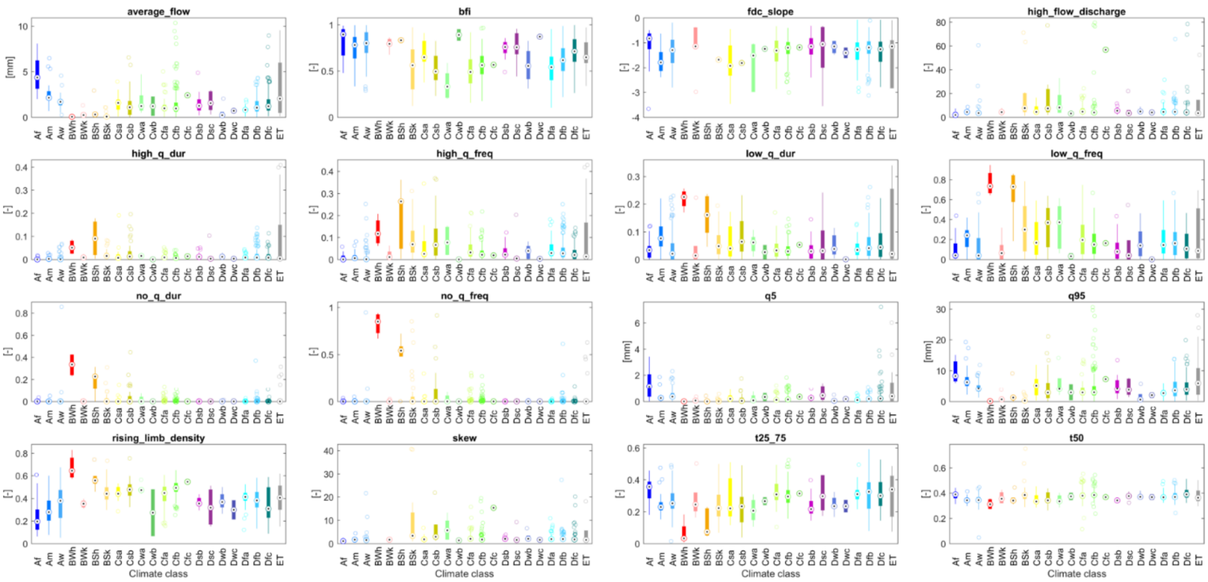


Figure S7: Distribution of the mean of annual signature values per GRDC catchment within each Köppen-Geiger climate. Climate classes without flow values are not shown.

S.2.3 Statistical significance of differences in signature values

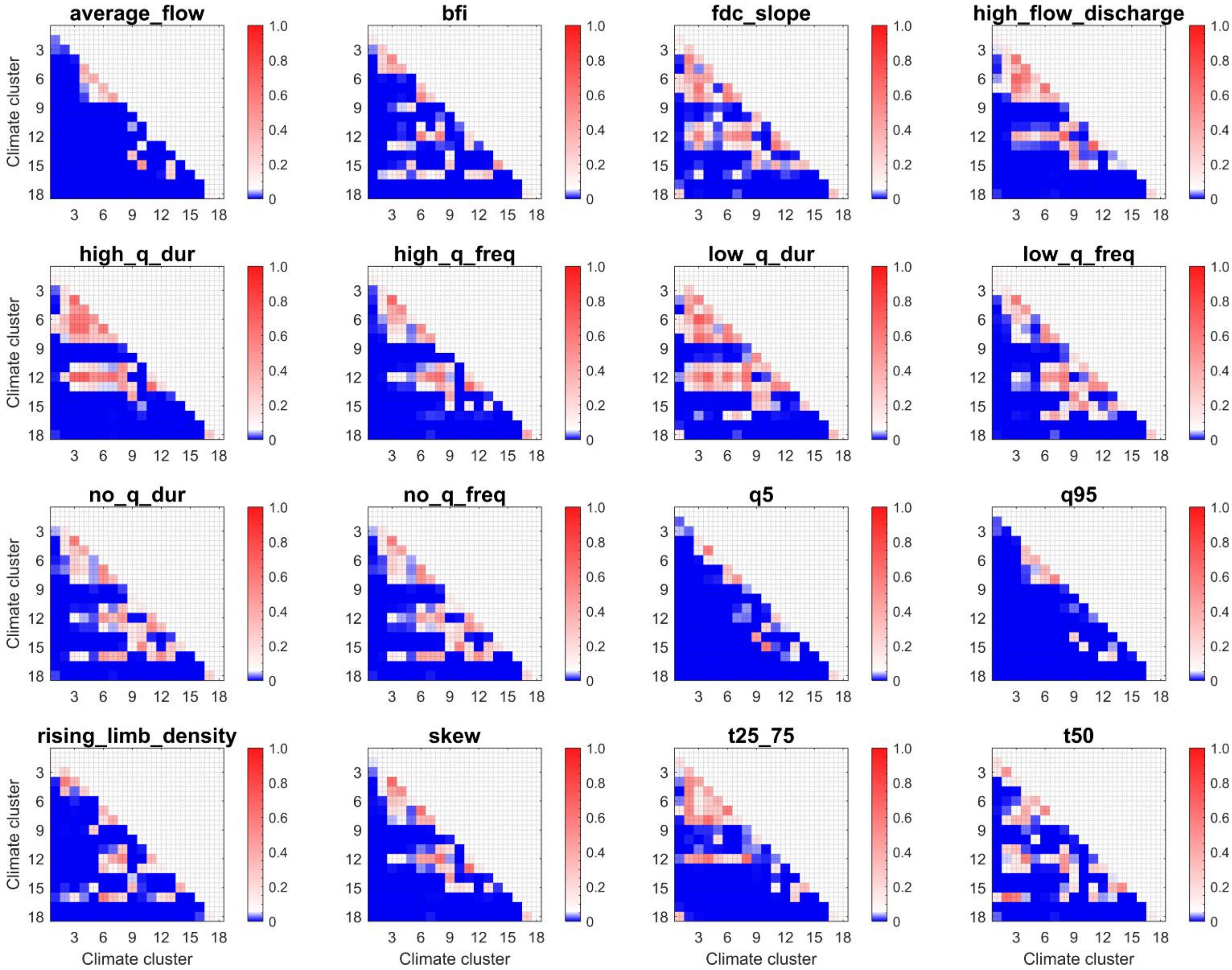


Figure S8: Empirical p-values of differences between annual mean signatures values per climate cluster

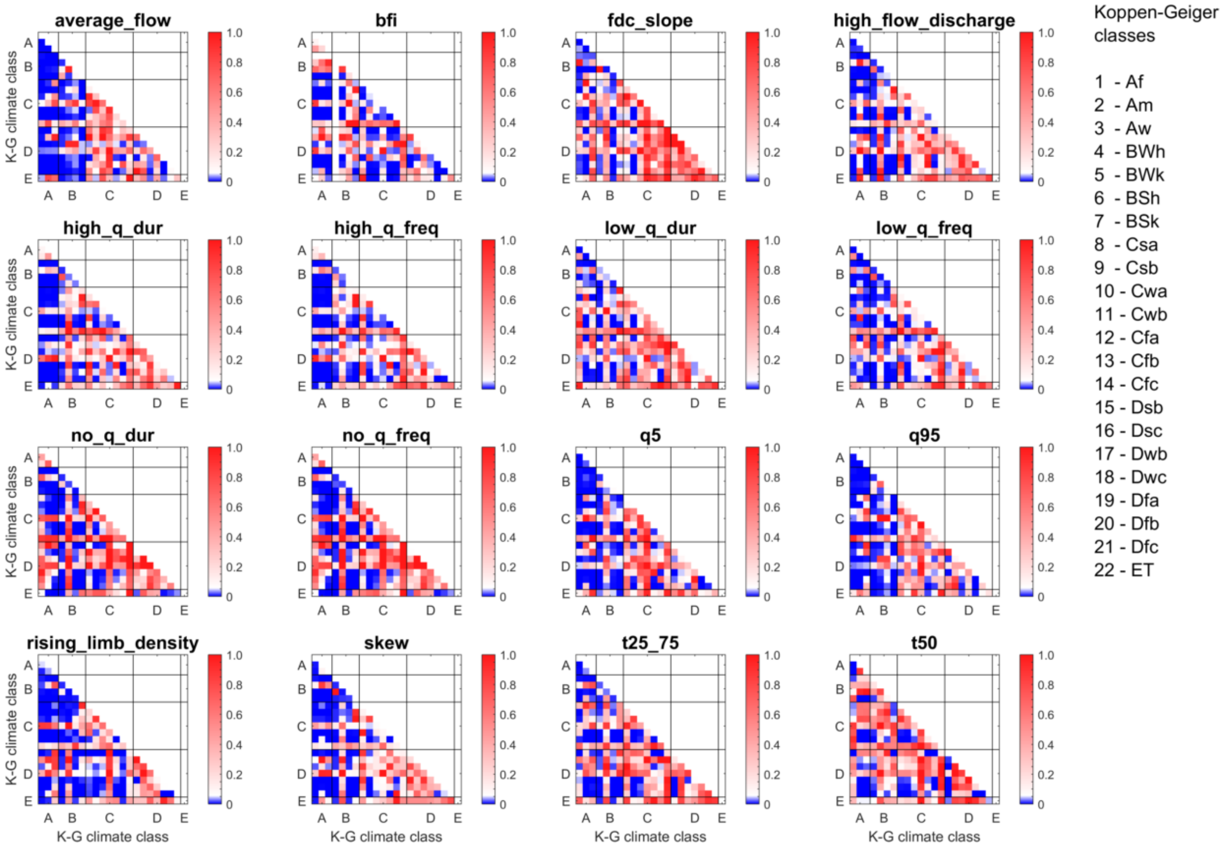


Figure S9: p-values of differences between annual mean signatures values per Koppen-Geiger climate class

S.3 Geographical spread of GRDC catchments per main climate cluster

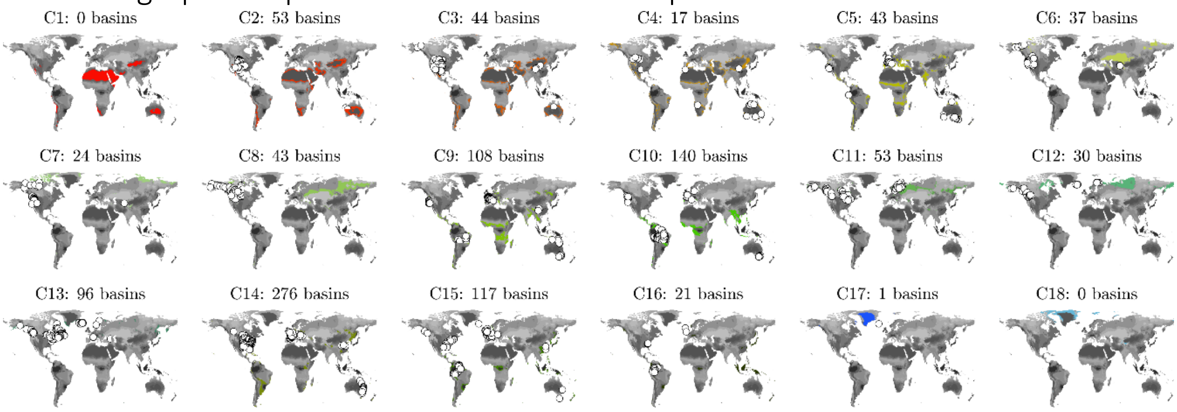


Figure S10: GRDC catchments per main climate cluster (cluster for which each catchment has the highest membership degree). Catchment-averaged climate is used to determine membership degrees where catchment boundary data is available (718 catchments), the climate at the outlet location is used in the remaining cases (385 catchments).