This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869226

doi:10.23689/fidgeo-5812

Report on continental-scale high-resolution modeling of streamflow intermittence

Document authors:	Petra Döll, Mahdi Abbasi and Tim Trautmann (Goethe University Frankfurt am Main, Germany)
Document contributors:	Mathis Messager (INRAE Lyon, France, McGill University, Montreal, Canada) and Bernhard Lehner (McGill University, Montreal, Canada)

Abstract

Quantification of the temporally varying streamflow intermittence at continental scales provides an important basis for evaluating biodiversity, ecosystem functions and ecosystem services in rivers as well as water resources for humans. As streamflow intermittence is often more prevalent in small upstream river reaches than in large downstream rivers, quantification needs to be done with a high spatial resolution. Aggregated to five classes (0, 1-2, 3-15, 16-29, 30-31 no-flow days), the number of no-flow days of approximately 1.5 million river reaches in Europe was estimated for each of the 468 months in the period 1981-2019 using a two-step Random Forest modeling approach. The model was developed based on a custom version of the 15 arc-sec HydroSHEDS drainage direction dataset. Data for 18 predictor variables (on hydrology, climate, physiography, geology, and land cover) as well as daily streamflow observed at 1,915 streamflow gauging stations were prepared as input to the RF model. In addition to upstream drainage area and slope, predictors based on time series of streamflow in 15 arc-sec grid cells were found to be most important for the RF model. These time series were generated by downscaling the 0.5 arc-deg runoff of the global hydrological model WaterGAP (downscaled streamflow is also already available for South America). In Europe but not in South America, the performance of downscaled monthly WaterGAP v2.2e streamflow as compared to streamflow observations is, on average, satisfactory also for small drainage basins of less than 10 km^2. While 99% and 95% of the observed perennial station-months are predicted correctly for the calibration and validation periods, respectively, the RF approach tends to overestimate intermittence.

Considering only the intermittent station-months, the frequency of predicting the correct class among the four classes is about 56% and 47% for the calibration and the validation period, respectively. 9% of all reach-months are simulated to be intermittent. The temporal and spatial patterns of simulated intermittence classes are plausible. The simulated intermittence class in each reach-month will be used by the other DRYvER Work Packages to upscale models developed at the Drying River Network scale.

Keywords

Streamflow intermittence, continental scale, high-resolution, monthly time series, streamflow downscaling, Random Forest

Information Table

PROJECT INFORMATION	
PROJECT ID	**869226**
PROJECT FULL TITLE	**Securing biodiversity, functional integrity and ecosystem services in DRYing riVER networks**
PROJECT ACRONYM	**DRYvER**
FUNDING SCHEME	**Horizon Europe**
START DATE OF THE PROJECT	**1st September 2020**
DURATION	**48 months**
CALL IDENTIFIER	**LC-CLA-06-2019**

DELIVERABLE INFORMATION	
DELIVERABLE No AND TITLE	D1.3 Report on continental-scale high-resolution modeling of streamflow intermittence
TYPE OF DELIVERABLE [1]	R
DISSEMINATION LEVEL [2]	PU
BENEFICIARY NUMBER AND NAME	3 - GU
AUTHORS	Petra Döll, Tim Trautmann, Mahdi Abbasi
CONTRIBUTORS	Mathis Messager, Bernhard Lehner
WORK PACKAGE No	1
WORK PACKAGE LEADER WP LEADER VALIDATION DATE	Jean-Philippe Vidal
COORDINATOR VALIDATION DATE	

[1] **Use one of the following codes:**
R=Document, report (excluding the periodic and final reports)
DEM=Demonstrator, pilot, prototype, plan designs
DEC=Websites, patents filing,press & media actions, videos, etc.
OTHER=Software, technical diagram, etc.
ORDP : Open Research Data Pilot

[2] **Use one of the following codes:**
PU=Public, fully open, e.g. web
CO=Confidential, restricted under conditions set out in Model Grant Agreement
CI=Classified, information as referred to in Commission Decision 2001/844/EC.

Coordinator signature	

Table of Contents

Glossary 6

1 Introduction 7

2 Deriving time series of monthly streamflow in Europe and South America at a spatial resolution of 15 arc-sec by downscaling of 0.5 arc-deg runoff computed by the GHM WaterGAP........ 8

2.1 Principal approach 8

2.2 Description of downscaling mechanisms 9

2.3 Validation 11

2.3.1 Description of the validation dataset 12

2.3.2 Modifications of the HydroSHEDS river network 15

2.3.3 Validation of results for Europe 17

2.3.4 Validation of results for South America 23

3 High-resolution estimation of monthly time series of the number of no-flow days at the continental scale by Random Forest modeling 27

3.1 Overview 27

3.2 Data and pre-processing........ 27

3.2.1 Target........ 27

3.2.2 Predictors 28

3.3 Random Forest modeling approach 31

3.3.1 The RF methodology........ 31

3.3.2 The two-step RF approach 32

3.3.3 Classification performance metrics 33

3.4 RF results and performance 33

3.4.1 Step 1 modeling results 33

3.4.2 Step 2 modeling results 39

3.5 Application of the RF modeling approach to compute monthly time series of flow intermittence (number of no-flow days in five classes) for European HR river reaches during the period 1981-2019 45

3.5.1 Derivation of river reaches ... 45
3.5.2 Simulated intermittence of European river reaches ... 46
4 Conclusions ... 51
5 Shared data ... 52
References ... 53
Appendix ... 56

Glossary

Term	Description
Global hydrological model (GHM)	Global hydrological model that simulates runoff and streamflow on a resolution of 0.5 arc-deg (e.g. WaterGAP)
Low-resolution (LR)	Resolution from which the data are downscaled, determined by the resolution of the GHM WaterGAP (i.e. 0.5 arc-deg)
High-resolution (HR)	Target resolution of the downscaling algorithm, determined by the resolution of the high-resolution river network HydroSHEDs (i.e. 15 arc-sec)
Streamflow (Q)	Streamflow in rivers which is accumulated along the river network
Disaggregated runoff (dR)	Runoff generated in the vertical water balance of the GHM without being routed and disaggregated to high resolution
Net cell runoff (ncR)	Streamflow of LR cell – streamflow of inflowing upstream LR cells, streamflow which is generated or abstracted in this cell
Grid cell Li,Hj	High-resolution grid cell with ID j in low-resolution grid cell with ID i
Drying river network (DRN)	Local river network that is investigated in the project DRYvER with local models and analyses

1 Introduction

Most rivers and streams on Earth cease to flow or dry at least one day per year (Messager et al., 2021). A quantification of streamflow intermittence at larger spatial scales (continental to global) serves as an important basis for evaluating biodiversity, ecosystem functions and ecosystem services in rivers as well as water resources for humans. Due to the restricted number of gauging stations in particular in intermittent streams, such analyses must rely on modeling. Up to now, large-scale quantitative studies on streamflow intermittence either have a coarse spatial resolution of 0.5 arc-deg (ca. 55 km at the equator) and thus cover only the intermittence of large rivers (Döll and Müller Schmied, 2012) or they do not estimate time series of streamflow intermittence but classify river reaches as intermittent or perennial, at a high spatial resolution of 15 arc-sec (Messager et al., 2021). However, headwater stream reaches with small drainage basins are more prone to intermittence than more downstream reaches (Datry et al. 2014, Messager et al., 2021) but cannot be considered in the case of coarse spatial resolutions. Characterization of the temporal structure of flow intermittence (e.g. number of no-flow days, seasonality of intermittence, etc.), beyond the general classification of reaches into perennial and intermittent, is required for supporting analyses of biodiversity and ecosystem functions of intermittent streams and rivers (Datry et al. 2018). To help advance the science and management of freshwater ecosystems globally, new large-scale models of streamflow intermittence are therefore needed that provide information on the frequency, duration and timing of flow cessation across the entire river network, from the headwaters to river mouths.

This report describes the methods and results of modeling streamflow intermittence of river reaches on the continental (European and South American) scale. This work is part of the DRYVER Work Package 1 whose aim is to determine current and future spatiotemporal patterns of streamflow intermittence in river networks at both the continental scale and for selected focal catchments ("Drying river networks" or DRNs).

The specific objective of this project is to produce monthly time series of ecologically relevant indicators of streamflow intermittence for the period 1981-2019, based on data at the high spatial resolution of 15 arc-sec (ca. 500 m). For each reach, streamflow intermittence is quantified as a monthly time series with the number of no-flow days within each month, aggregated to five classes. It is estimated with a random forest (RF) model that uses, as innovative and important predictors, indicators derived from 15 arc-sec monthly streamflow time series that have been generated by downscaling the 0.5 arc-deg output of the global hydrological model WaterGAP 2.2e (described in Chapter 2). In the following, the 15 arc-sec resolution, which corresponds to grid cells of approx. 500 m by 500 m at the equator) will be denoted by HR ("high-resolution") and the 0.5 arc-deg resolution (55 km by 55 km at the equator)) by LR ("low resolution"). 1 LR grid cell contains 14,400 HR grid cells. HydroSHEDS v1 is used as the basic HR drainage direction map (Lehner et al., 2008; www.hydrosheds.org); it was modified for the three focal DRN of the DRYvER project in Finland, Croatia and Hungary (see Section 2.3.2 for details on these modifications). Other hydrological predictors are based on LR output of WaterGAP and characterize, for example, the ratio of groundwater recharge to total runoff. The third group of predictors are static HR environmental indicators such as the drainage area or the aridity index. Chapter 2 presents the methods and results of downscaling LR WaterGAP output to monthly HR streamflow time series for both Europe and South America; it includes validation of the time series against observed monthly streamflow at gauging stations. The methods, validation and results of the RF modeling to estimate the monthly time series of the number of no-flow days per reach are described in chapter 3, only for Europe. We draw conclusions in Chapter 4, followed by technical information on the produced datasets.

2 Deriving time series of monthly streamflow in Europe and South America at a spatial resolution of 15 arc-sec by downscaling of 0.5 arc-deg runoff computed by the GHM WaterGAP

2.1 Principal approach

The principal approach for downscaling information computed by a low-resolution (LR, 0.5 arc-degree, LR) global hydrological model (GHM) to high-resolution (HR, 15 arc-sec) streamflow is based on Lehner and Grill (2013). The idea is to first map LR runoff (a component of the vertical water balance of the GHM) to a high-resolution (HR) river network as “disaggregated runoff” dR, by assuming that runoff (in eq. water height) is the same in all HR grid cells within an LR grid cell. To integrate the additional information from the routing routine of the LR GHM, in particular about the impact of surface water bodies and human water use on streamflow, the HR runoff is corrected by taking into account the difference between the LR streamflow. The correction term is applied in a spatially weighted way to the HR grid cells. The weights are based on HR streamflow obtained by accumulating dR along the HR flow direction assuming that the bigger a stream gets the more it is impacted. Finally, HR streamflow is estimated by accumulating the corrected HR runoff along the flow direction.

This principal approach is expressed as

$$Q_{Li,Hj} = flowacc\left(dR_{Li,Hj} + C_{Li} * W_{Li,Hj}\right) \tag{1}$$

$$C_{Li} = \sum_{j=1}^{14400} dR_{Li,Hj} - ncR_{Li} \tag{2}$$

$$W_{Li,Hj} = \frac{flowacc\left(dR_{Li,Hj}\right)}{\frac{1}{14400} * \sum_{j=1}^{14400} flowacc(dR_{Li,Hj})} \tag{3}$$

where $Q_{Li,Hj}$ is streamflow of HR cell j located within LR cell i, C_{Li} is the correction term applied to all HR cells within LR cell i (Eq. 2), and $W_{Li,Hj}$ is the correction weight (Eq. 3). Flowacc() represents the flow-accumulated variable and ncR_{li} represents net cell runoff of LR grid cell i, which can be calculated as streamflow of LR grid cell i minus streamflow of all direct upstream LR grid cells.

This principal approach, which was originally developed to compute HR mean monthly streamflow (i.e. long-term mean streamflow in the 12 calendar months) was adapted and extended to obtain monthly time series of HR streamflow and to meet the requirements for the intended modeling of streamflow intermittence. This is described in the following section.

2.2 Description of downscaling mechanisms

Calculating HR disaggregated runoff $dR_{Li,Hj}$ as the HR basis for streamflow estimation

Monthly time series of R_{Li} are first computed as the sum of LR surface runoff (R_s in Müller Schmied et al., 2021 , p. 1045) and an LR groundwater discharge to surface water bodies ((Q_g in Müller Schmied et al., 2021, p. 1046) computed by WaterGAP are used as the basic input to compute $dR_{Li,Hj}$ (Eq. 1). The sum of both variables represents the not routed outflow of the vertical water balance of the land area of each LR grid cells to the surface water bodies in the LR grid cells. In the original method of Lehner and Grill (2013) solely total from land (the sum of surface runoff and groundwater recharge) was used to calculate dR. In the approach presented here, the groundwater storage impact on groundwater discharge to the stream and thus monthly HR streamflow Is taken into account.

The sum of LR surface runoff and groundwater discharge is interpolated to an intermediate resolution of 0.1 arc-deg using an inverse distance interpolation with a power of 2, a radius of 1.8 arc-deg, and taking into account a maximum of 9 low-resolution data points. This interpolation is also needed because there are areas that are covered by the HR but not the by LR river network. This is the case in coastal regions and at full lake cells of the LR river network. This interpolated 0.1 arc-deg grid is then disaggregated to the high resolution by repeating the values for all high-resolution grid cells.

Redistribution of water storage modifications in big lakes and reservoirs

In WaterGAP, reservoirs with a maximum storage capacity of at least 0.5 km³, regulated lakes having a maximum storage capacity of at least 0.5 km³ or an area of more than 100 km² and lakes with a minimum area of 100 km² are calculated as so-called „global" surface water bodies (Müller Schmied et al., 2021) that receive water not only from the runoff generated within the cell but also from upstream streamflow. They may spread over more than one LR grid cell and their water balance is calculated in the assigned outflow cell. Thus the net cell runoff (ncR_{li}) of this outflow grid cell includes the net cell runoff generated by the "global" surface water bodies that need to be redistributed to all LR riparian grid cells and their HR cells. This is done by calculating the change of the "global" surface water body storage from the month before to the subject month. This amount is reduced from the net cell runoff of the outflow cell and redistributed in an area-weighted way to all LR riparian cells. So that every LR cell has its net cell runoff from "global surface water bodies" assigned based on the area of the global surface water bodies in this cell. Then these LR values are applied to those HR cells which are covered by polygons of global surface water bodies. As those HR grid cells have different pixel areas the correction values are calculated area-weighted.

Correction for differing characteristics of the river networks in rivers with large catchments

The presented algorithm uses GHM output, which was simulated using the DDM30 (Döll & Lehner, 2002) river network, and downscales it to the high-resolution river network of HydroSHEDS (Lehner et al., 2008). Given the different spatial resolutions and the generation process, the river networks differ locally in their characteristics. One major difference is that the HydroSHEDS river network contains endorheic sinks that are not covered by the DDM30 river network. Those local endorheic sinks cannot be covered by the LR DDM30 due to their subgrid nature. The correction term C_{Li} may thus be underestimated in magnitude. With such a subgrid endorheic sink covering half of an LR cell the original correction term $C_{Li,org}$ is applied to all HR cells in the LR cell but only half of it affects the mainstream for which the C_{li} was calculated.

The following presented correction mechanism, developed in the original method of Lehner and Grill (2013), aims particularly to correct for those subgrid endorheic sinks but also covers general deviations in catchment areas due to the resolution. The correction mechanism is applied for rivers with an upstream area of at least 50,000 km² and a reasonable accordance of drainage areas of DDM30 and HydroSHEDS. For rivers with catchment areas between 50,000 and 100,000 km^2 they are allowed to differ up to 20% and for rivers with catchment areas of > 100000 km^2, they are allowed to differ up to 50 %. These criteria are necessary as locally the river networks can diverge strongly, especially in the headwater areas, but those deviations are equalized later downstream and additional correction would worsen the results. Furthermore, it is necessary to avoid overcorrection due to mismatching river networks (e.g. in HydroSHEDS a LR grid cell contains only a tributary to a mainstream, whereas the corresponding LR grid cell of DDM30 contains the mainstream). However, these criteria limit the corrected size of subgrid endorheic sinks to 50,000 km² and the correction is effective solely in rivers with large catchments.

For those LR cells within the criteria range, the original correction term $C_{Li,org}$ (Eq. 2) is extended by an additional correction term. This modification of the correction term C_{LI} is calculated by comparing the original net cell runoff of the LR cell with the net cell runoff which is generated in the largest river with the preliminarily corrected streamflow $Q_{Li,Hj}$ (Eq. 1). The following equation is only valid for large-river cells as defined above.

$$C_{Li} = C_{Li,org} + (\, ncR_{Li}^{LR} - ncR_{Li}^{HR}) \tag{4}$$

The HR net cell runoff representation of LR grid cell Li (ncR_{Li}^{HR}) is calculated as the streamflow of the HR grid cell with the maximum upstream area in Li minus the corresponding streamflow values of direct upstream LR grid cells.

To avoid overcorrection with this modification approach due to gaps caused by the large river definition the modifications are propagated downstream.

Shifting of correction terms to downstream HR grid cells

GHMs simulate the processes of the vertical water balance (i.e. runoff generation) and route the water along the river network for each 0.5 arc-degree grid cell. The simulated LR streamflow represents the properties of the largest river in the LR grid cell. The correction term is applied to HR cells by weighting it by streamflow (Eq. 3), to apply those corrections more strongly to the larger river systems for which they were calculated. To even focus correction more strongly on large rivers, an additional modification of correction terms is introduced. It shifts the corrections partially downstream along the LR river network if the upstream area of the HR cell in the downstream cell is at least 90% of the source cell. This criterion guarantees that the correction values are solely shifted to larger streams and so that shifting in LR cell with mismatching river networks is avoided. The fraction of the correction term that is shifted downstream is computed as

$$fr_{Li}^{shift} = \begin{cases} \dfrac{(2 * \, upA_{Li,down}^{Max} - upA_{Li}^{Max})}{2 * \, upA_{Li,down}^{Max}}, & upA_{Li,down}^{Max} > 0.9 \, * \, upA_{Li}^{Max} \\ 0, & upA_{Li,down}^{Max} \leq 0.9 \, * \, upA_{Li}^{Max} \end{cases} \tag{5}$$

with $upA_{Li,down}^{Max}$ representing the maximum HR upstream area in the downstream grid cell and upA_{Li}^{Max} representing the maximum HR upstream area in the evaluated LR grid cell Li. Following this approach, the modified correction term consists of the part which is not shifted downstream and the parts which originate from the shifted correction terms from direct upstream cells ($C_{Li,upj}$) (eq. 6).

(6)	$C_{Li} = C_{Li,org} * \left(1 - fr_{Li}^{shift}\right) + \sum_{j=1}^{n} \left(C_{Li,upj} * fr_{Li,upj}^{shift}\right)$

Negative and extreme correction values

Given the principal approach of comparing ncR to generated dR and additional modification of this correction term, negative values of streamflow or extreme correction values may be calculated. This can especially happen, if the LR and HR river networks are not aligned or their upstream areas are differing too much. To avoid extreme over-correction values a threshold of 0.001 m³ per second per km² upstream area for the final correction term is applied. Furthermore, negative streamflow values which may originate from side effects of the correction mechanisms are not accumulated along the river network. That means that in the flow accumulation algorithm a negative streamflow value can turn positive if streamflow is coming from upstream cells but a negative streamflow value is not propagated along the river network. Furthermore, in the final step, all remaining negative streamflow values are set to zero.

Technical implementation

The technical implementation can be made available upon request. A software compilation was developed in Python to run the downscaling algorithm. Furthermore, a set of Python scripts (with ArcPy dependency) has been developed to preprocess necessary static data.

The following static data are necessary to run the downscaling algorithm:

Data	Description
flow_dir_15s_by_continent.gdb	HydroSHEDS flow directions [Esri flow direction codes]
pixel_area_skm_15s.gdb	HydroSHEDS area of HR grid cells [km²]
flowdir_30min.tif	DDM30 flow directions [Esri flow direction codes]
landratio_correction.tif	Ratio of land area to continental area [-]
orgDDM30area.tif	Area of LR cells of WaterGAP [km²]
pixareafraction_glolakres_15s.tif	Ratio of HR cell on global surface water bodies [-]

2.3 Validation

The downscaling methodology is applied to WaterGAP 2.2e simulations from 1901 – 2019 forced with GSWP3-W5E5. The resulting time series of monthly streamflow data is compared to observed streamflow at gauging stations for all months with available observations. The number of compared monthly streamflow values in the focal DRN is provided in Table 1-A in the Appendix. The performance is measured with the modified Kling Gupta efficiency (KGE) statistics after (Kling et al., 2012, Eq. 8) with its components correlation coefficient (r), bias ratio (β) and variability ratio (γ). KGE can range from -Inf to 1; values below -1 are not shown with higher values indicating higher performance. A value comprised between -0.41 and 1 indicates that the model performs better than simply using the mean of the observed data to predict the values (Knoben et al., 2019). Furthermore, the Nash Sutcliffe efficiency of logarithmic values of observed and simulated streamflow (logNSE) is used as a metric focusing on low flow conditions (Oudin et al., 2006), which are assumed to be important for modeling intermittency.

2.3.1 Description of the validation dataset

The downscaling results are validated against a compiled dataset of gauging stations with time series data from different sources. The station dataset compiled by (Messager et al., 2021) is the main component of this dataset and is extended by i) gauging stations with significant anthropogenic influence (e.g. dam upstream) or that underwent a change from being perennial to intermittent or vice-versa over the course of the streamflow record, which were originally excluded in the dataset of Messager et al. (2021), and ii) data from stations in the DRNs, which were provided within the context of the DRYvER research project. Furthermore, the original daily time series data were obtained from the original data sources to estimate the days without flow per year.

The stations are mapped in a semi-automatic process to the river network of HydroSHEDS ensuring that the HydroSHEDS river network reproduces the characteristics of the watershed (Messager et al., 2021)(i.e. the upstream area should not deviate more than 10% of the value which is reported in the station metadata). Furthermore, the stations must hold at least 120 months of data after removing months with missing values.

For Europe, the compiled dataset consists of 1816 stations, from which 861 stem directly from the Global Runoff Data Centre (GRDC, 2015), 904 from databases used to produce the Global Streamflow Indices and Metadata (GSIM) archive (Do et al., 2018; Gudmundsson et al., 2018), and 51 from local data providers within the DRYvER context (Sauquet, 2020). From those stations, 320 have at least 5 consecutive days of streamflow below 0.001 m³/s (1 L/s) following the definition of the SMIRES meta-database (Sauquet, 2020). We are using this classification for describing purposes in the following description of the characteristics of the gauging stations. The distribution of upstream area and average discharge shows that most gauging stations have an upstream area of 10 to 2000 km² and the average discharge is between 1 and 1000 m³/s (see Figure 1). This reflects the general tendency to measure rather bigger and perennial streams as those are important for flood forecasting or inland water transport.

The spatial distribution of the gauging stations used for validation is shown in Figure 2. An accumulation of validation stations is found in Germany, Austria, Switzerland, France, Spain, and the UK. As non-perennial classified stations are mainly found in arid regions (e.g. in Spain or Cyprus) but also are found in other climates (e.g. in western France).

For South America, the compiled dataset consists of 808 stations, from which 290 stem directly from GRDC, and 518 from GSIM monthly time series. At 31 GRDC stations, streamflow records with at least 5 consecutive days of streamflow below 0.001 m³/s (1 L/s) are found. For GSIM stations, a different criterion was used because only monthly indices are available: 106 stations with at least one monthly MIN7 index (minimum 7-day average streamflow) below 0.001 m³/s were thus classified as intermittent. In total 137 of the 808 stations are classified as non-perennial for describing purposes. The distribution of upstream area and average discharge shows that most gauging stations have an upstream area of 100 to 10000 km² and the average discharge is between 10 and 1000 m³/s (see Figure 3). Comparing the validation datasets in South America and Europe the stations in South America tend to have bigger catchments and higher average streamflow records. The validation dataset in Europe holds also more than double the number of stations used for South America.

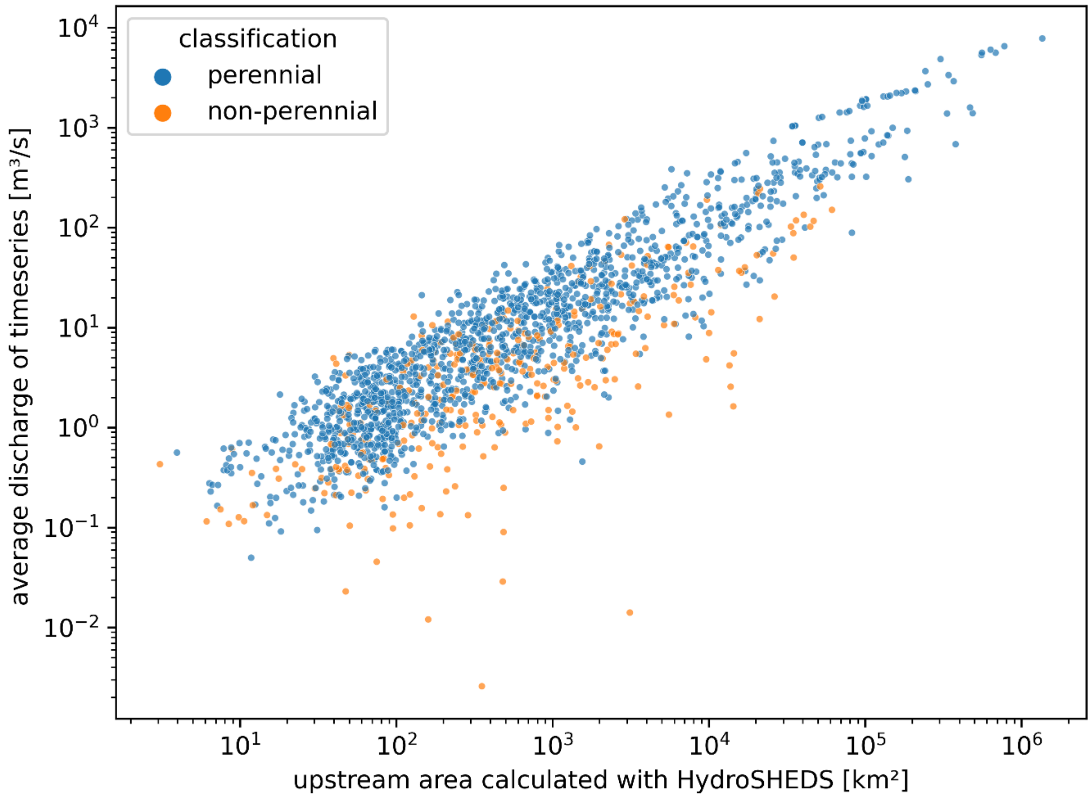

Figure 1: Upstream area and average discharge of gauging stations used for validation of downscaling monthly time series of streamflow in Europe

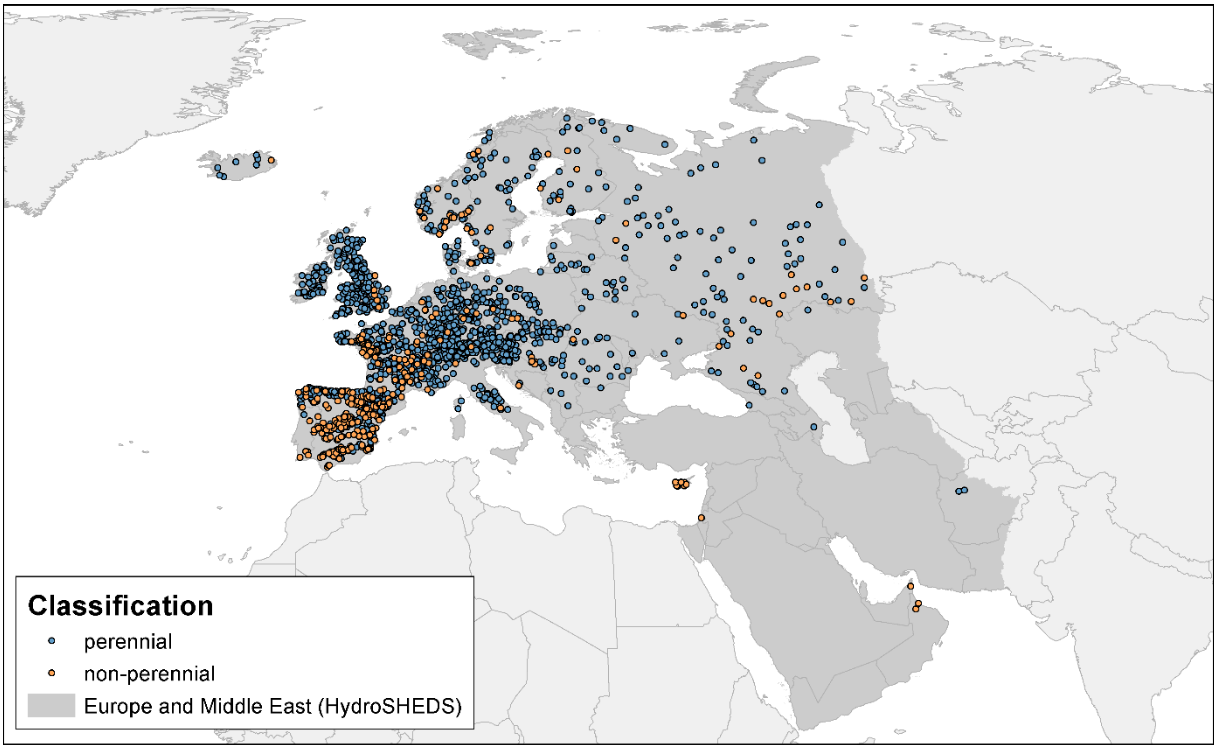

Figure 2: Spatial distribution of gauging stations in Europe used for validation of developed high-resolution monthly streamflow time series

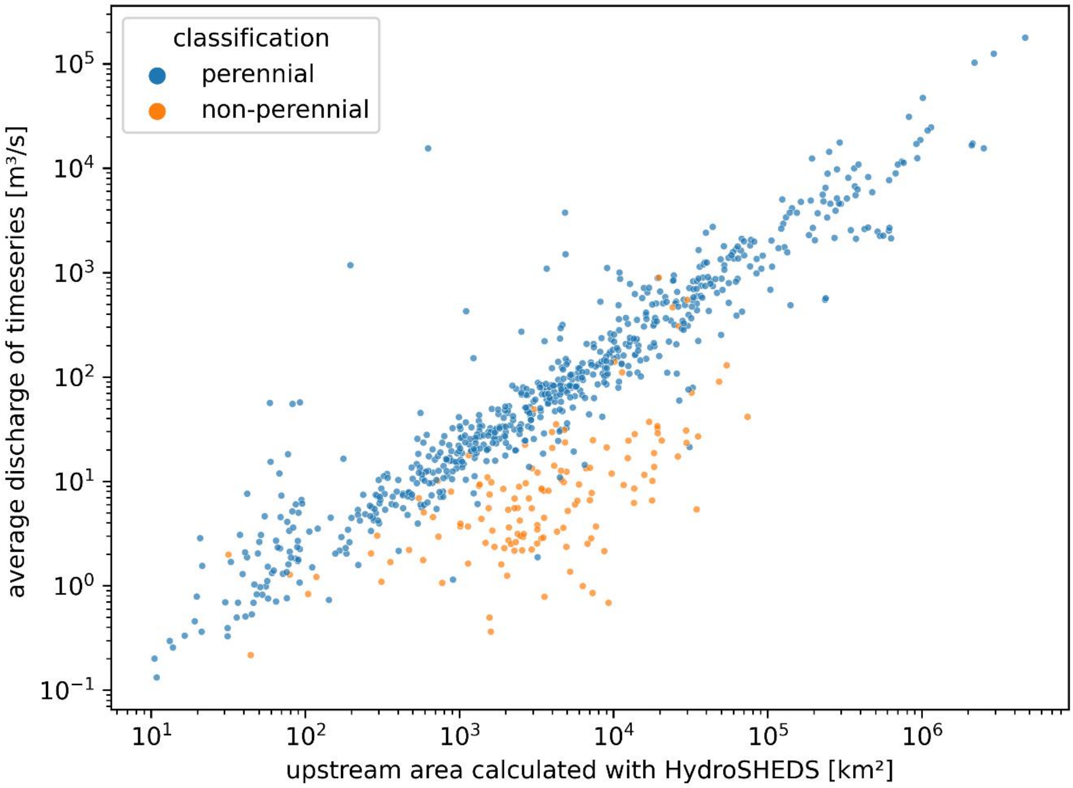

Figure 3: Upstream area and average discharge of gauging stations used for validation of downscaling monthly time series of streamflow in South America

The validation stations in South America are mainly located in Brazil (see Figure 4), with a regional focus in the regions Southeast and Northeast. No validation stations are located in Chile, Peru, Uruguay, and Paraguay. The majority of as non-perennial classified stations are found in the Northeast region of Brazil.

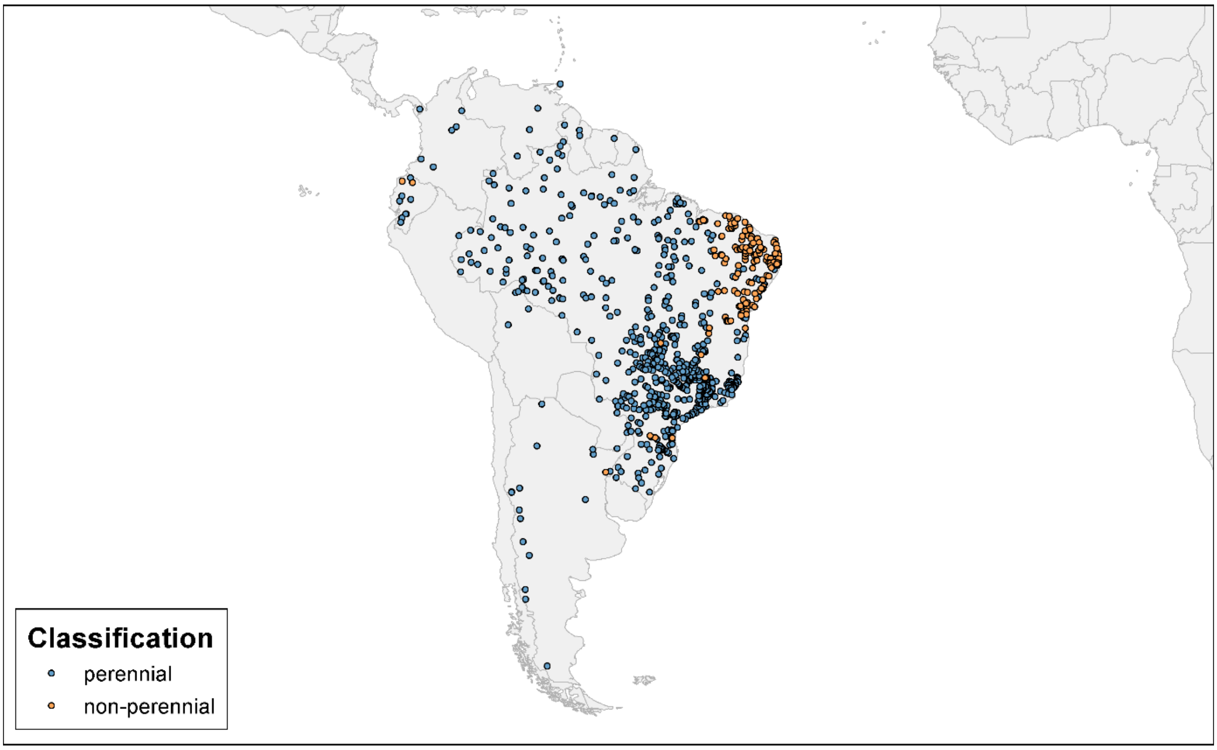

Figure 4: Spatial distribution of gauging stations in South America used for validation of developed high-resolution monthly streamflow time series

2.3.2 Modifications of the HydroSHEDS river network

A comparison of the results of the focal DRN models and the continental scale models is necessary for the validation of the performance of the flow intermittence modeling. The river network of HydroSHEDS used on the continental scale differs significantly in three focal DRNs, so modifications of the HydroSHEDS network are necessary.

The focal DRN Finland has the largest deviations among all focal DRNs (Figure 5). This is explainable by the coarser resolution of the DEM (Hydro1k) used for the delineation of HydroSHEDS river network north of 60° longitude, where the finish focal DRN is located. A big part in the east of the focal DRN is connected to the main river network. Furthermore, in the west and the southwest of the study area, the catchment is extended via modifications of the flow directions. However, the agreement between the river networks of the focal DRN and modified HydroSHEDS is still the worst in comparison with the other focal DRNs.

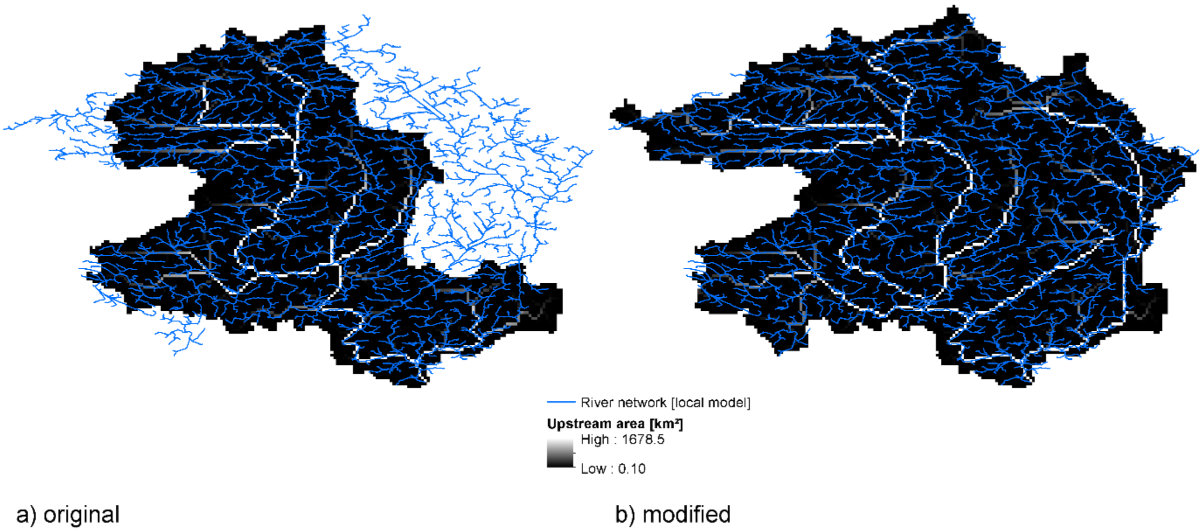

Figure 5: Modifications of HydroSHEDS river network and visualized with upstream area calculated with unmodified (a) and modified (b) river network for the focal DRN Finland.

The river network in the Hungarian focal DRN also needs modifications to make the results comparable (Figure 6). The catchment outlet which drains into the Drava river is moved to the west and the river network is modified to cover the missing subcatchment in the original river network in HydroSHEDS.

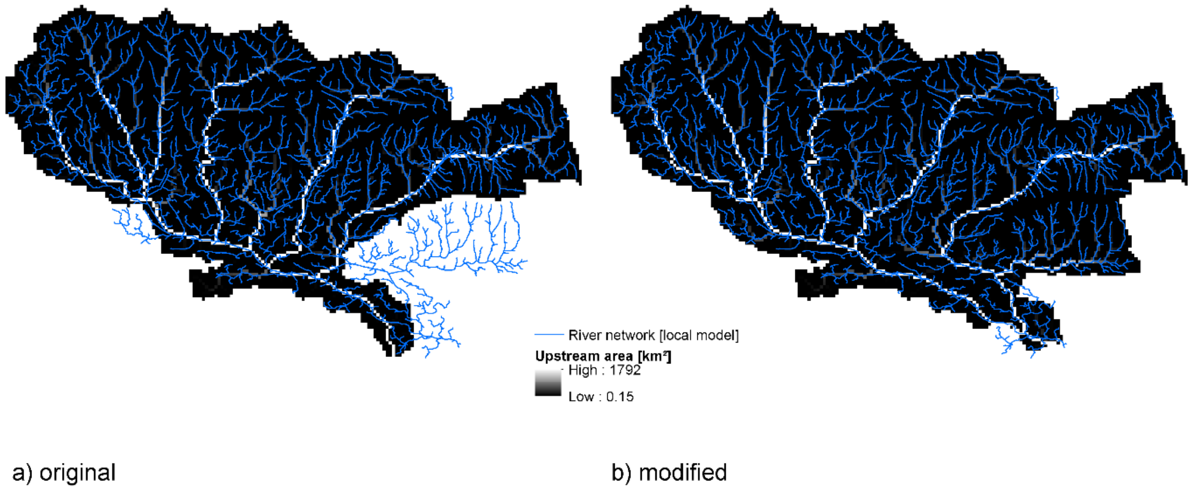

Figure 6: Modifications of HydroSHEDS river network and visualized with upstream area calculated with unmodified (a) and modified (b) river network for the focal DRN Hungary.

In the focal DRN in Croatia two inland sinks are present in the HydroSHEDS river network, which do not correspond to the river network of the local model (Figure 7). Furthermore, a small subcatchment in the north-west part of the focal DRN is removed to generate a comparable river network.

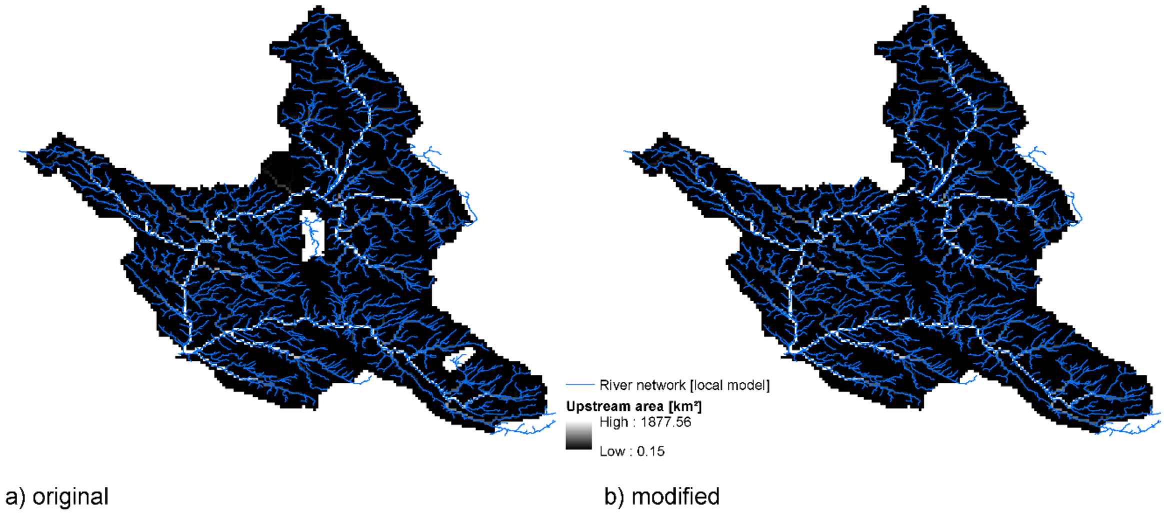

Figure 7: Modifications of HydroSHEDS river network and visualized with upstream area calculated with unmodified (a) and modified (b) river network for the focal DRN Croatia.

2.3.3 Validation of results for Europe

In Europe, the simulated time series reach a median KGE value of 0.57 overall gauging stations, showing a good agreement between simulated and observed monthly time series of streamflow. The KGE values are best for gauging stations with a catchment area of over 10,000 km² as those catchments cover several LR grid cells and thus the simulation performance of the LR GHM is maintained with the downscaling method. The resulting median KGE values get smaller with catchment sizes as the sub-grid processes, which cannot be simulated with the downscaling algorithm, gain importance (Figure 8). Spatial clusters of negative KGE values indicating bad downscaling or simulation performance can be identified in the southwestern part of Spain, Iceland, Italy, and Cyprus (Figure 9). The performance of simulated downscaled streamflow is dependent on I. the performance of the LR GHM, II. the quality of the HR river network in the catchment and III. the downscaling method. For the mentioned regions the LR GHM results can be identified as a major cause for the poor performance of downscaled streamflow time series. Iceland, Cyprus, and parts of Italy are not calibrated within the GHM, and in the affected region in Spain, the model performance is rather bad even with calibration (Müller Schmied et al., 2021).

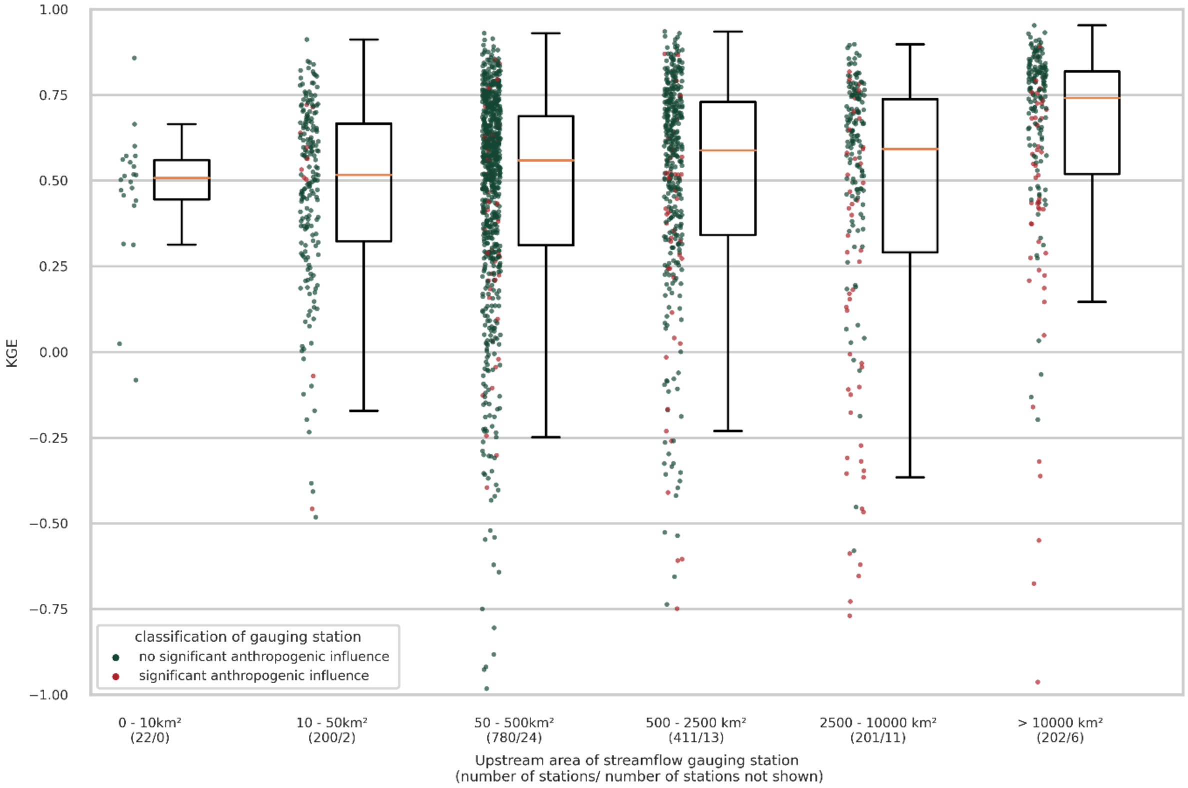

Figure 8: KGE values over all validation gauging stations in Europe grouped by upstream area

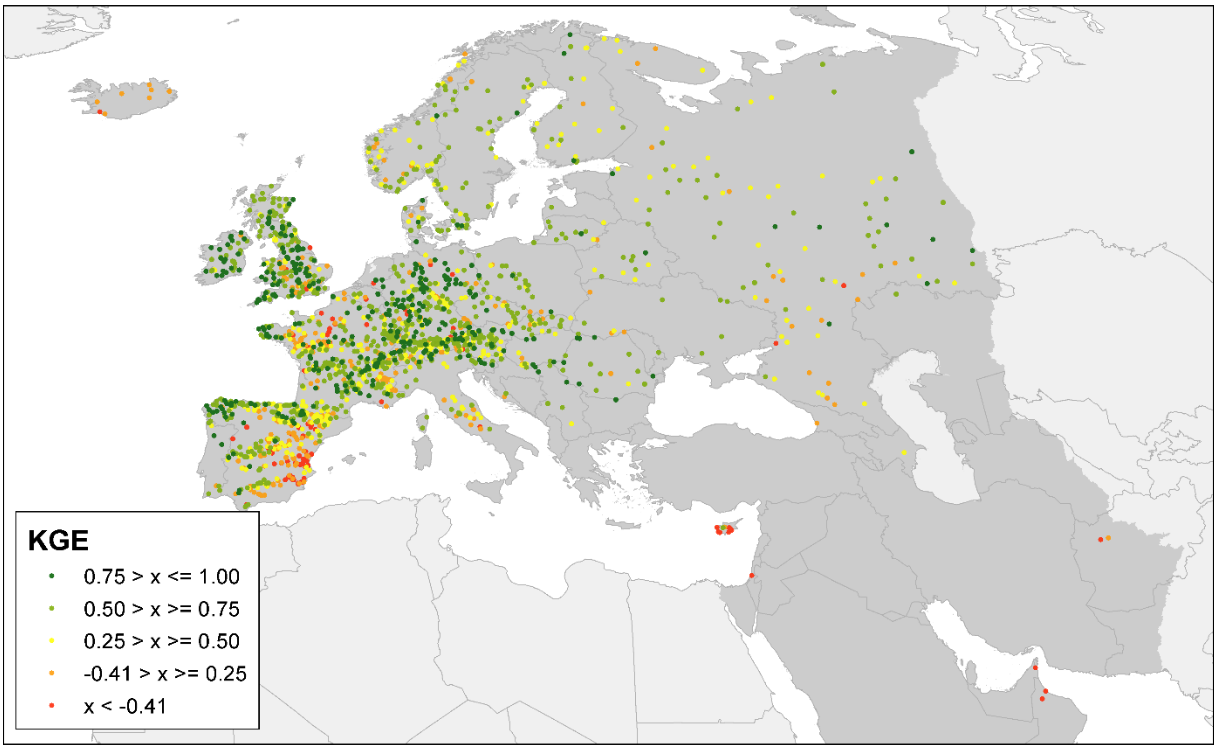

Figure 9: KGE values for all validation gauging stations in Europe

In general, the components of the KGE (r, β, and γ) reveal that with decreasing catchment area the downscaled simulated streamflow tends to underestimate the magnitude and variability of the streamflow (Figure 11 and Figure 12). However, the correlation coefficient values show only small variation throughout the catchment area groups (Figure 10).

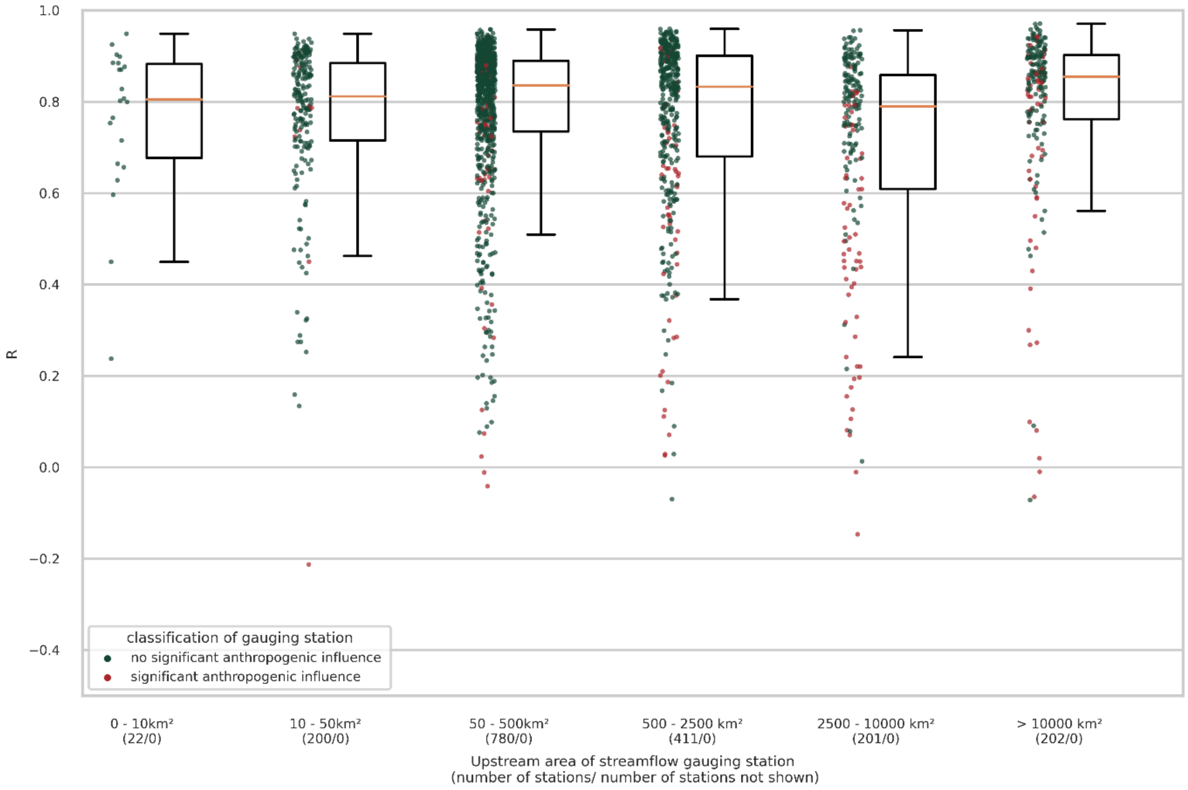

Figure 10: Correlation coefficient (r) over all validation gauging stations in Europe grouped by upstream area

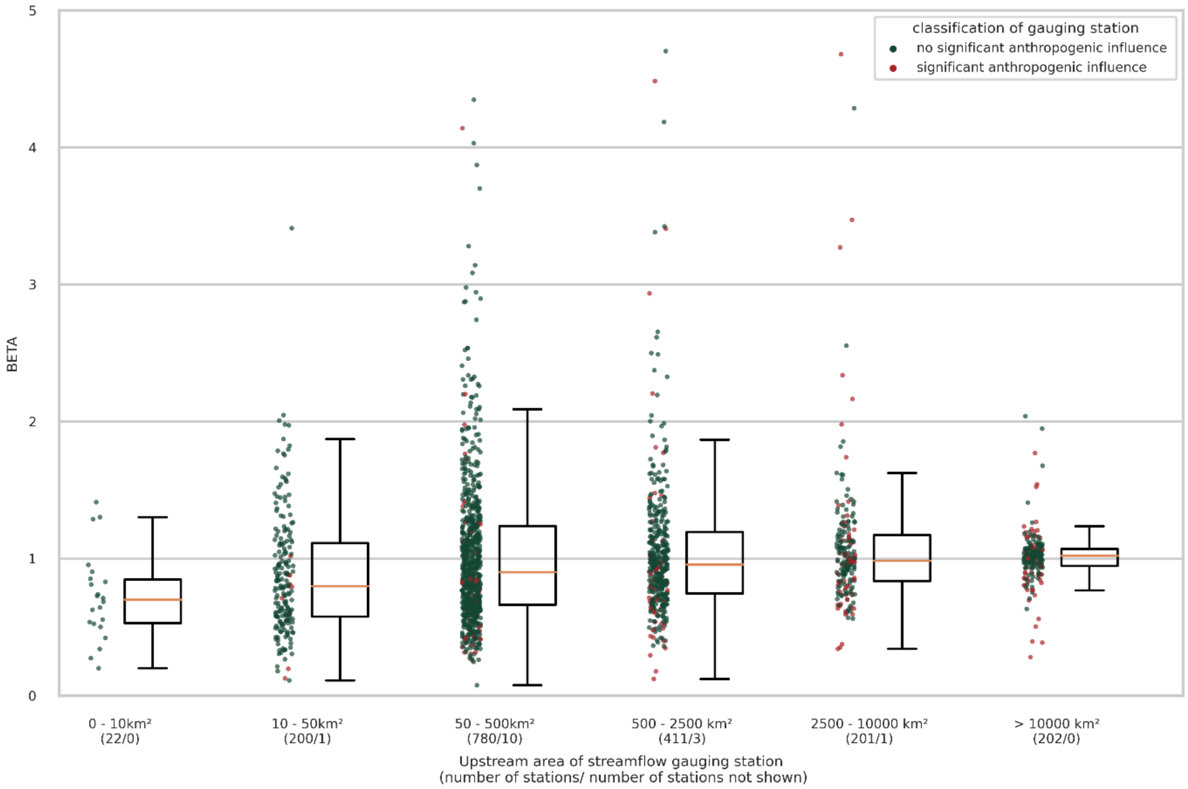

Figure 11: Bias ratio (β) overall validation gauging stations in Europe grouped by upstream area

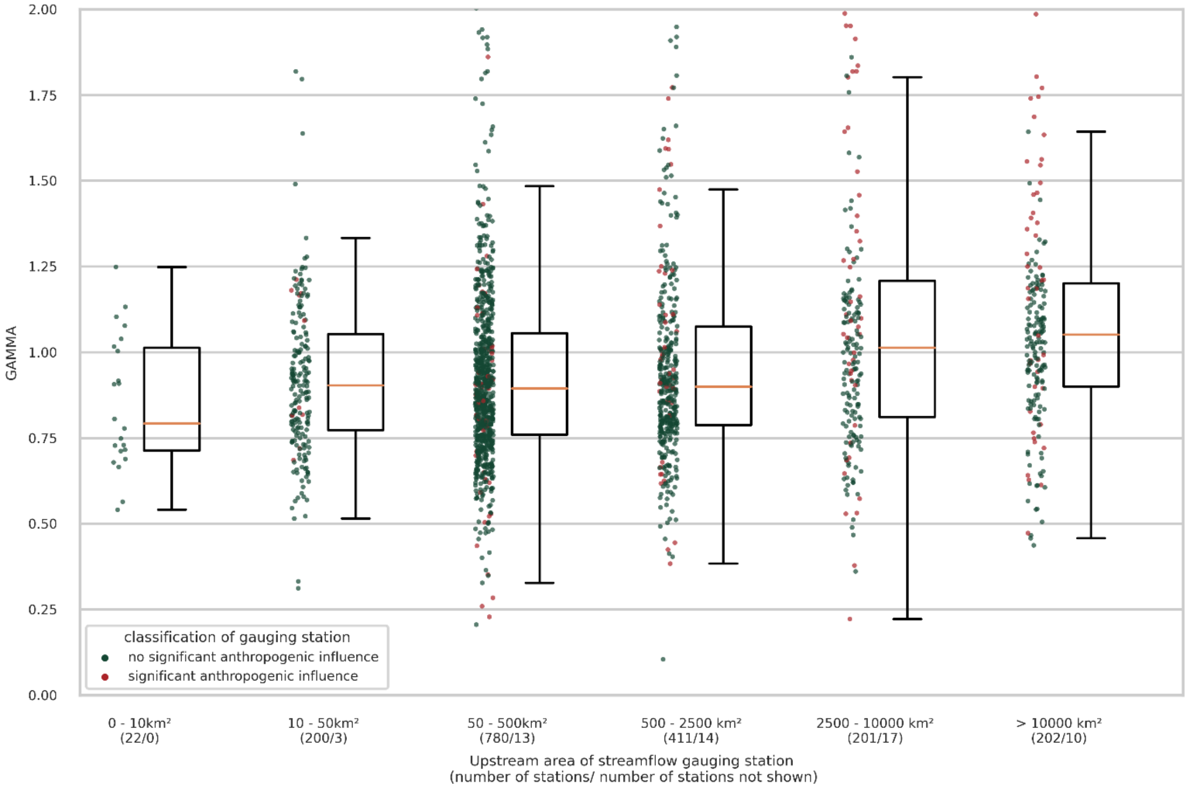

Figure 12: Variability ratio (γ) overall validation gauging stations in Europe grouped by upstream area

Overall validation gauging stations in Europe we find a logNSE median value of 0.39 in contrast to an NSE median value of 0.44. This relatively small difference between logNSE and normal NSE shows that the downscaling method performs comparable well under low flow conditions compared to normal flow conditions. Gauging stations with more than 10,000 km² catchment area perform best evaluated with the logNSE and the values decrease with catchment sizes (Figure 13). However, the number of stations in the groups with different catchment areas is not equally distributed.

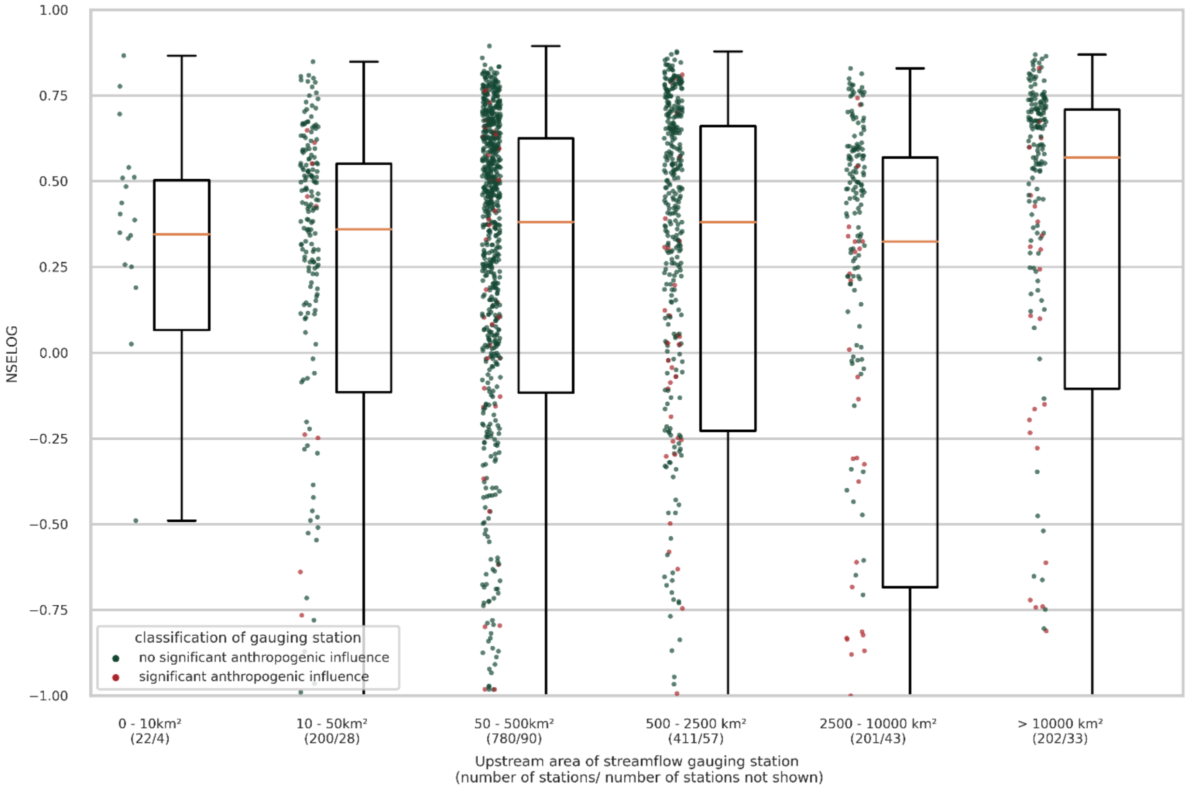

Figure 13: Logarithmic Nash Sutcliffe efficiency (logNSE) overall validation gauging stations in Europe grouped by upstream area

In the European focal DRNs, 44 gauging stations showed a good agreement of catchment areas, which is necessary for validation. Besides some outliers, the simulated downscaled time series of monthly streamflow show good performance overnall DRNs (Figure 14). However, the streamflow under low flow conditions at several stations in Croatia and Hungary cannot be simulated well (Figure 15). In the Appendix, the detailed performance values of all DRN stations are shown.

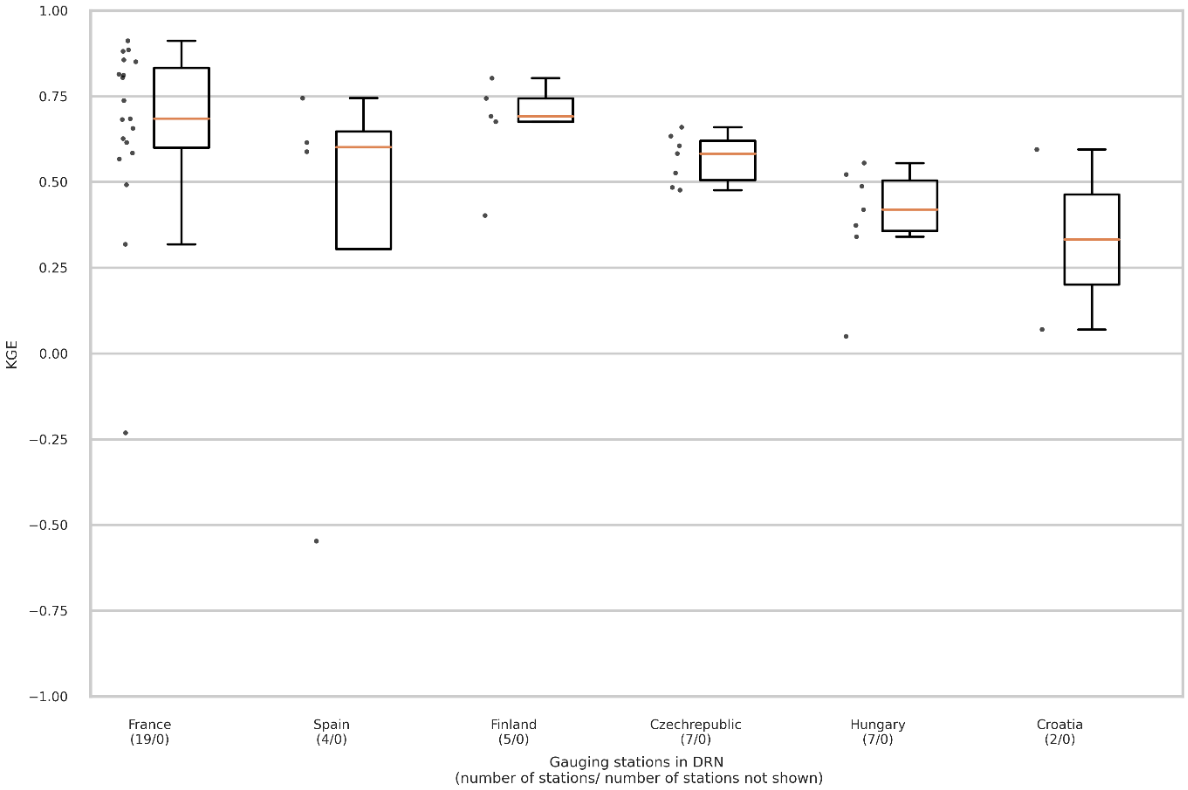

Figure 14: KGE values of gauging stations in DRNs

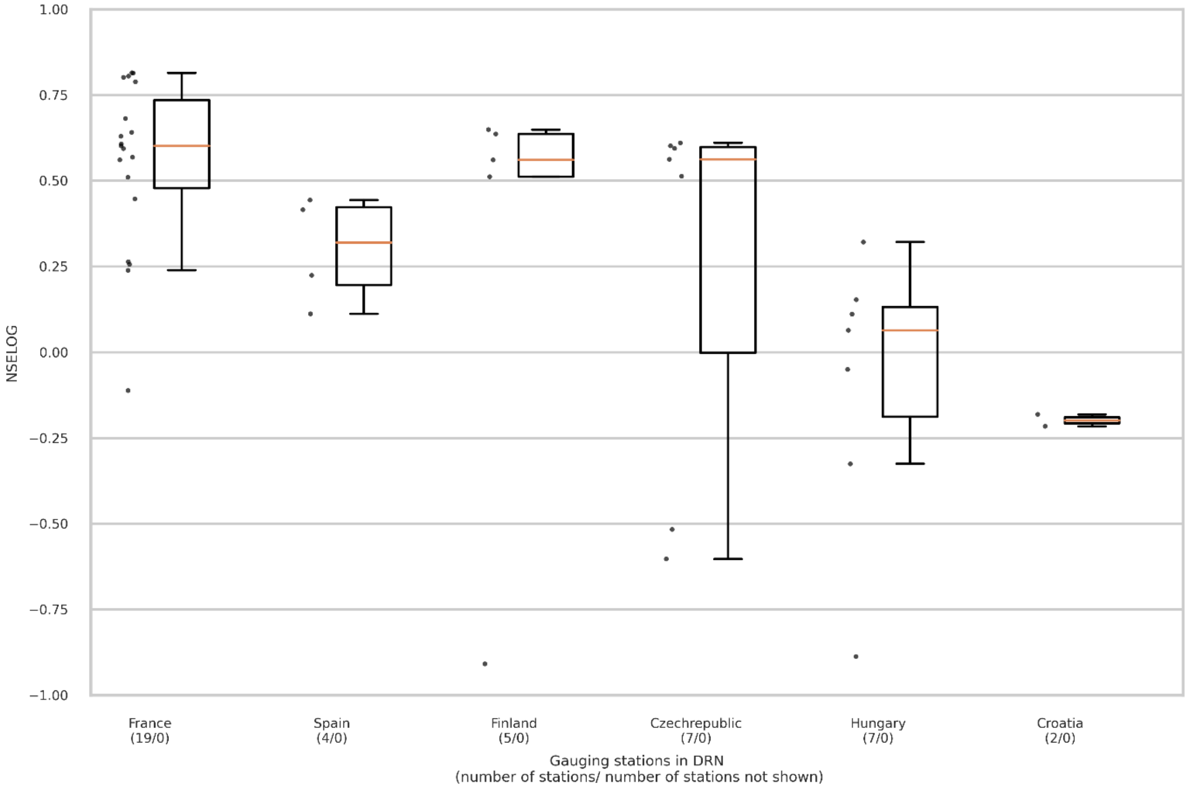

Figure 15: Logarithmic Nash Sutcliffe efficiency (logNSE) of gauging stations in DRNs

2.3.4 Validation of results for South America

In South America, the performance of the developed HR streamflow time series is, with a median KGE value of 0.29, significantly lower than in Europe. Especially in the upstream area classes below 10,000 km² low KGE values are reached (Figure 16). Most stations that show bad performance (KGE values below -0.41) are located in the western part of the hydrographic macroregion “Paraná” and the hydrographic macroregions “Eastern Northeast Atlantic” and “San Francisco” (Figure 17). Those regions show also bad streamflow performance in LR GHM results (Müller Schmied et al., 2021, Fig. 8). The components (r, β and γ) reveal that the majority of stations with less than 500 km² catchment area the streamflow is underestimated (β<1), whereas the majority of stations with catchment areas of more than 2500 km² the streamflow is overestimated (Figure 19). The variability of the observations is not well captured by the developed HR streamflow time series, which again may be traced back to the relatively poor performance of LR GHM streamflow (Müller Schmied et al., 2021, Fig.8). When focusing on the low flows the majority of the stations reach a negative logNSE (Figure 21). However, the majority of stations with a catchment area of more than 10,000 km² show a rather good performance.

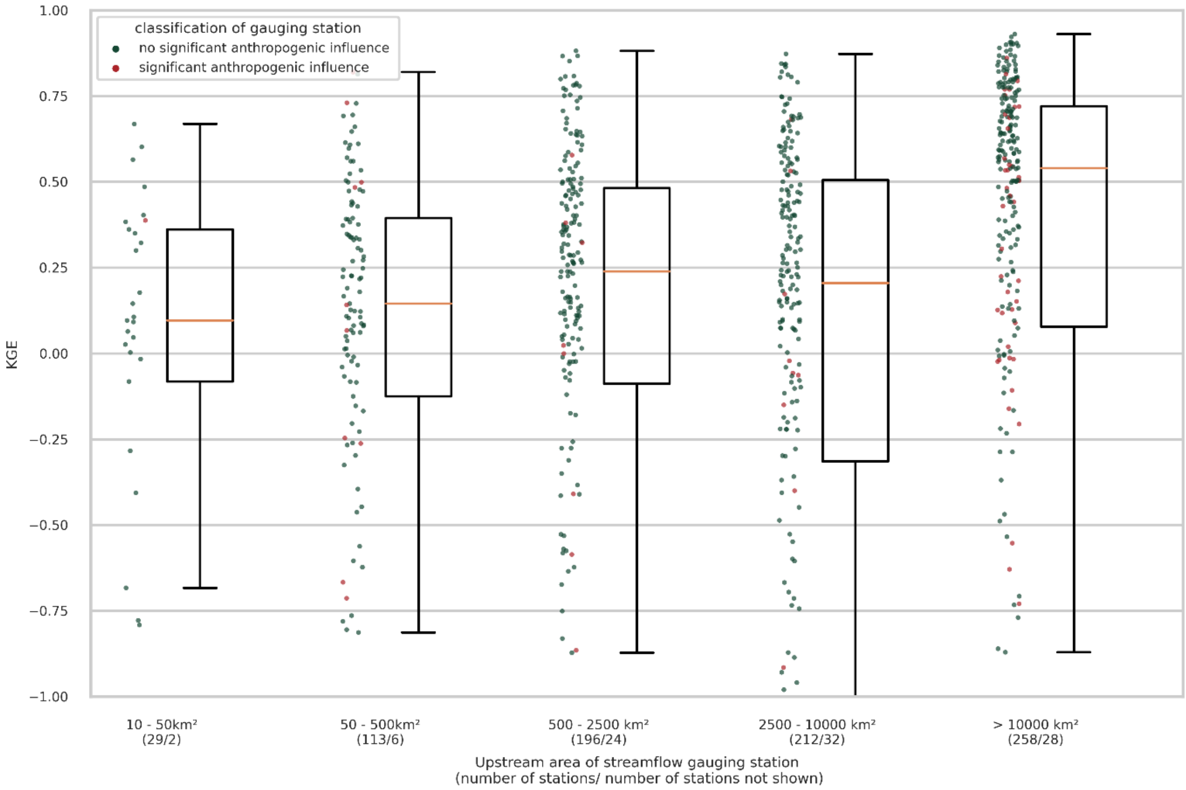

Figure 16: KGE values over all validation gauging stations in South America grouped by upstream area

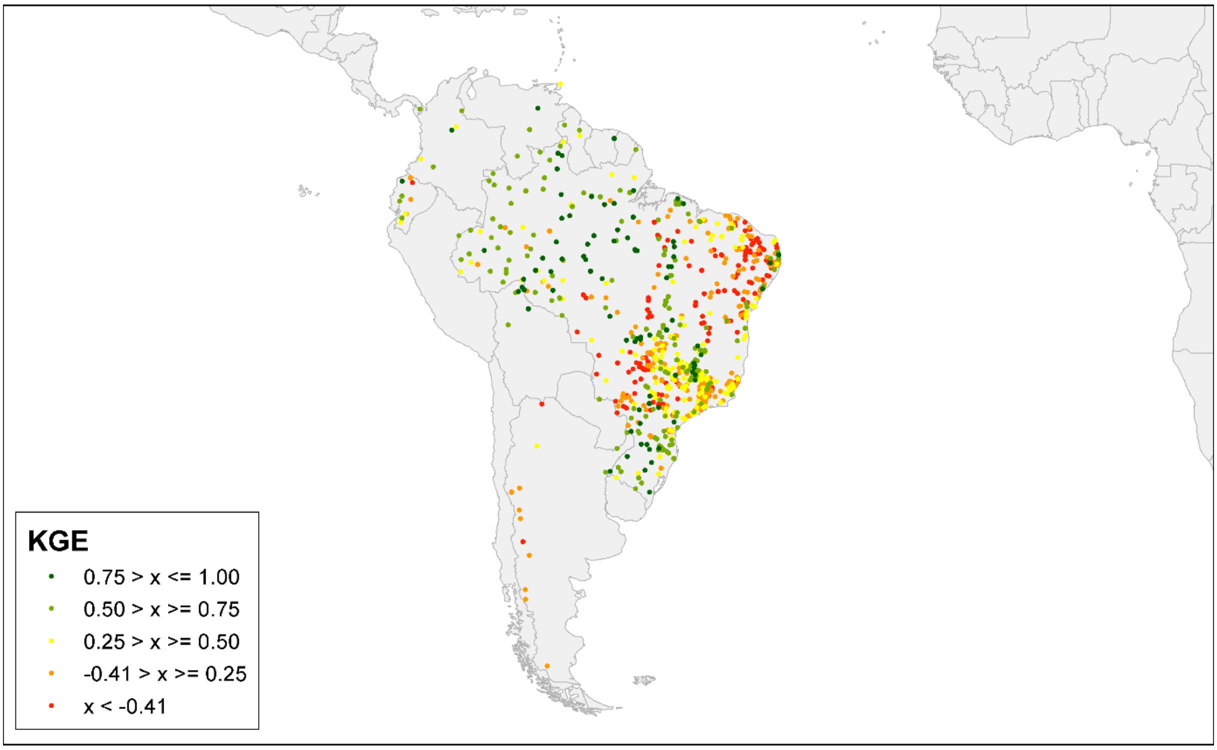

Figure 17: KGE values for all validation gauging stations in South America

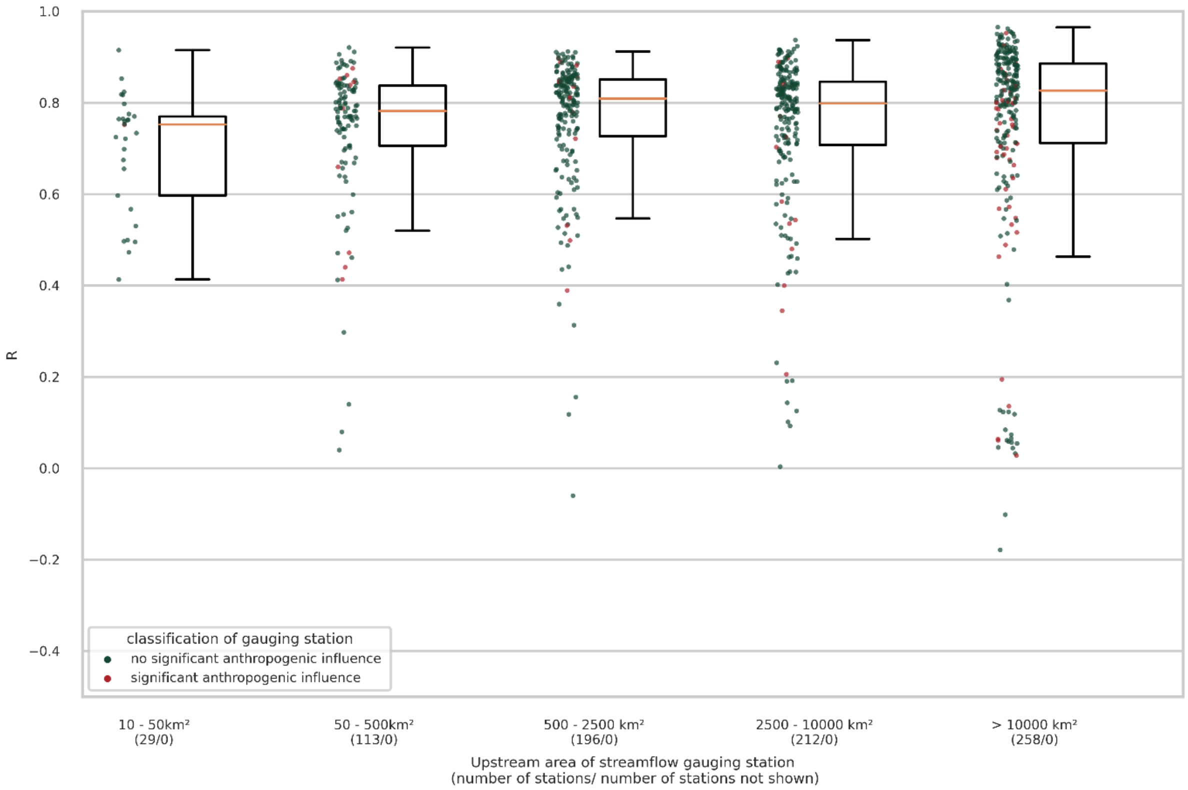

Figure 18: Correlation coefficient (r) over all validation gauging stations in South America grouped by upstream area

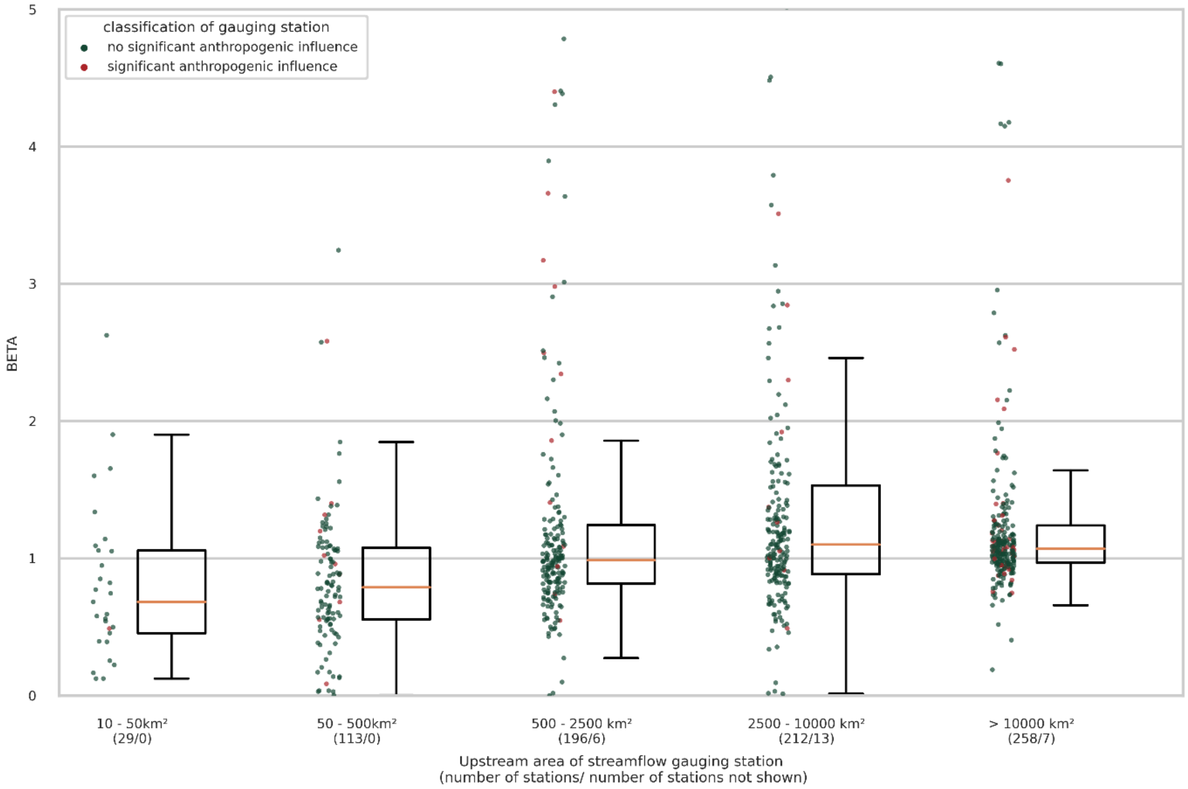

Figure 19: Bias ratio (β) overall validation gauging stations in South America grouped by upstream area

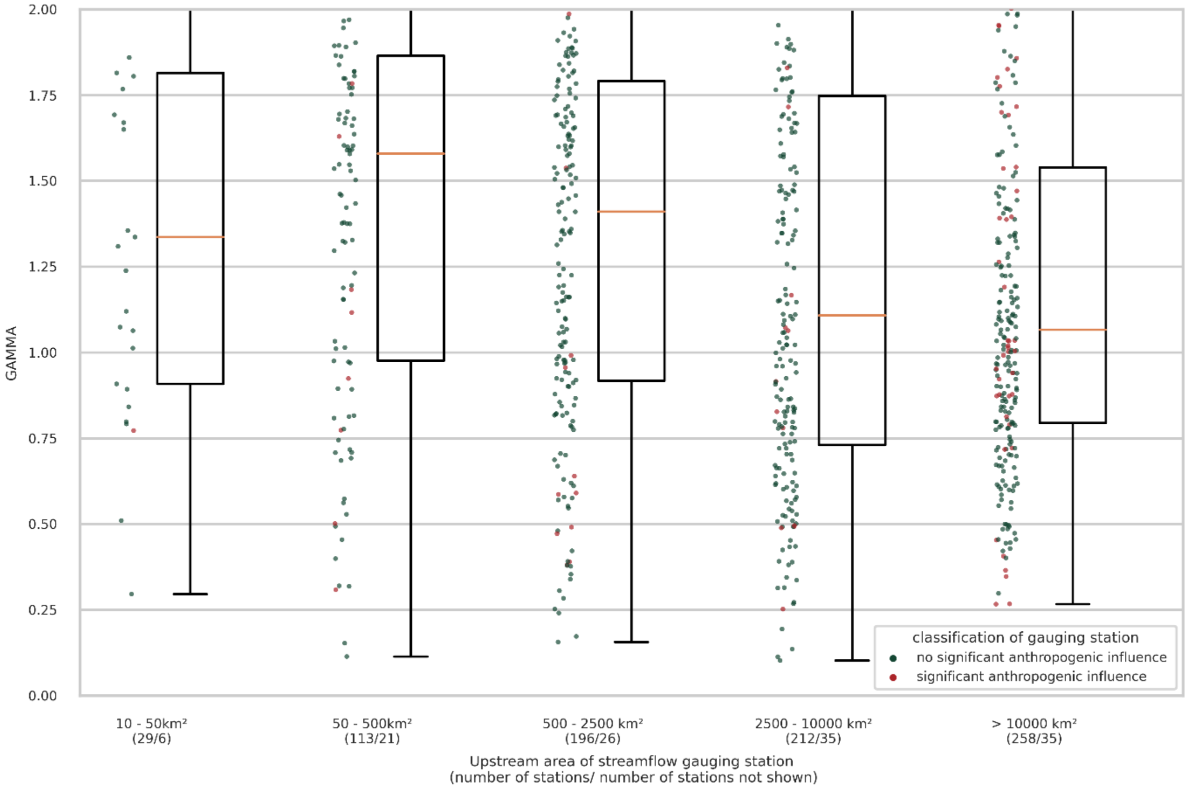

Figure 20: Variability ratio (γ) overall validation gauging stations in South America grouped by upstream area

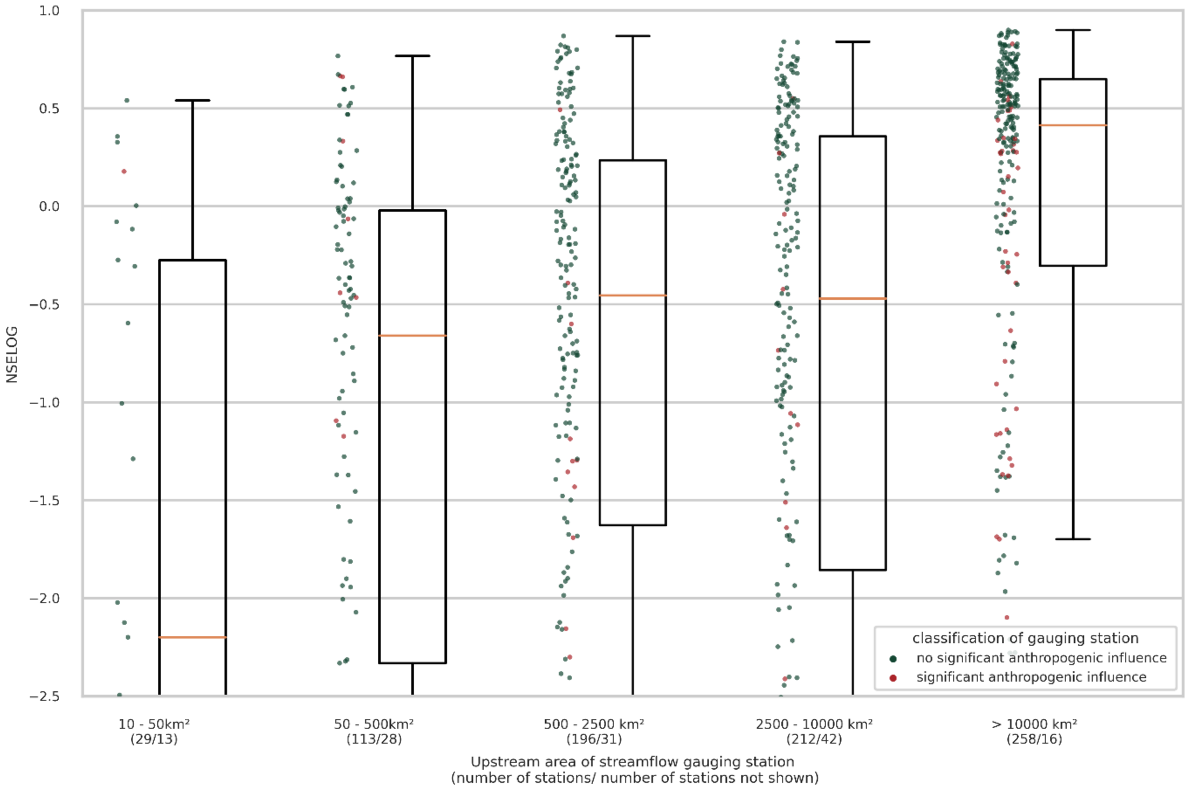

Figure 21: Logarithmic Nash Sutcliffe efficiency (logNSE) overall validation gauging stations in South America grouped by upstream area

3 High-resolution estimation of monthly time series of the number of no-flow days at the continental scale by Random Forest modeling

3.1 Overview

We developed a novel approach based on random Forest (RF) modeling to produce high resolution reach-scale estimates of the spatiotemporal patterns of streamflow intermittence across all of Europe. In addition, process-based models are notoriously unsuccessful at predicting low flows and streamflow intermittence at large scale due to the complexity of the mechanisms involved, such as groundwater-surface water interactions (Döll et al. 2016; Zaherpour et al. 2018). Machine learning (ML) approaches like RF, by contrast, do not require knowledge on the processes underlying the phenomenon of interest and are thus a promising tool to produce hydrological predictions at this combination of resolution and scale. ML methods identify the pattern underlying a phenomenon based on observations, i.e. they learn from the observations and then can make predictions. ML methods are grouped into supervised learning, unsupervised learning and reinforcement learning Supervised learning approaches are a promising tool for estimating streamflow intermittence at a high spatial resolution without the need for high-resolution hydrological modeling, which is infeasible at the continental scale. The supervised learning method Random Forest (RF) is well suited for both classification and regression tasks (Breiman et al., 2001), and thus for relating observations of streamflow intermittence with diverse high-resolution data (predictors) to predict HR streamflow intermittence. RF modeling has already been used for hydrological classification problems, i.e. for predicting classes of hydrological characteristics, including intermittence (e.g., for the world: Messager et al., 2021; Australia: Bond and Kennard, 2017; France: Snelder et al., 2013).

As input data for the RF model, we combined daily streamflow observations (target observations), the updated 15 arc-sec HydroSHEDs drainage direction map (Section 2.3.2), the HR downscaled monthly streamflow time series of WaterGAP, other LR WaterGAP data and other static HR physiogeographic data (Section 3.2) to estimate the target, the monthly time series of the number of no-flow days. We used an efficient two-step RF approach that addresses the highly unbalanced target observations, which are dominated by months without no-flow days, to determine the relationship between these predictors of intermittence and the observed monthly values of no-flow days (Section 3.3). The results of the two-step RF modeling approach are described and validated in Section 3.4. We then used the updated HydroSHEDs drainage direction map to generate more than 1.5 million river reaches for which the RF modeling approach was then applied, resulting in a monthly time series of the number of no-flow days (aggregated to five classes) for each river reach, covering the period 1981-2019 (Section 3.5). The spatial extent of the RF modeling is Europe without Russia and Turkey.

3.2 Data and pre-processing

Obtaining and manipulating the observational data for the target variable and the data for the many predictor variables is an important and time-consuming step in the RF modeling process.

3.2.1 Target

The target, or response variable, of the RF models developed in this study is the number of no-flow days per month in 15-arc cells. Observation-based Information on this variable, from which the RF model can be generated, can only be derived from continuous daily observations of streamflow at gauging stations. The gauging stations need to be co-registered with the updated HydroSHEDS drainage direction map (section 2.3.2) to identify to which grid cell the station data are to be related. We collected such observations from the GRDC and GSIM databases (Section 2.3.1). Altogether 1918 GRDC and GSIM stations with daily streamflow are available for Europe. However, most of the GRDC and GSIM stations are on perennial streams, without any no-flow days, which also reflects the result of a global-scale analysis of Krabbenhoft et al. (2022) that intermittent river reaches are underrepresented by existing streamflow gauging stations. We therefore used the SMIRES meta-data on gauging stations in 19 European countries with flow intermittence (Sauquet, 2020) to directly ask the national streamflow data providers for the daily streamflow time series of the gauging stations listed in the SMIRES table. In this way, we obtained daily streamflow time series of a total of 375 SMIRES gauging stations. Please note that these stations have not been included in the validation of HR downscaled WaterGAP streamflow (Section 2.3.1).

From all of these stations, stations suitable for deriving target observations were selected. In the first step, we checked the correct location of the SMIRES gauging stations by comparing the upstream area given in the meta-data with those of HydroSHEDS. The GRDC and GSIM stations had already been quality-checked by Messager et al. (2021; Section 2.3.1). If the drainage areas deviated by more than 10% the stations were manually relocated to a suitable grid cell with a deviation of less than 10%. If this was not possible, the station was excluded from RF modeling. For the remaining stations, we excluded all station-months that did not have streamflow observations for all days of the month and only kept all stations that had at least 18 station-months of daily streamflow data. To identify no-flow days, we called all days with a streamflow of not more than 1 l/s a no-flow day. Finally, we computed, as the target of the RF modeling, the number of no-flow days per month and station, i.e. per station-months, for all the remaining stations. The maximum period with observed no-flow days per station-month is 1981-2019 (i.e., station-months before or after this period of 468 months were omitted from subsequent analyses).

Streamflow data from a total of 1,915 stations were used for setting up the European RF model, i.e. for calibrating it, and for validating its results. For each station, the first two-thirds of all months with observations within the analysis period were used for model calibration and the last one-third for validating the model. In the case of full data, the calibration and validation period is 1981-2006 (312 months) and 2007-2019 (156 months), respectively. The total number of station-months with the observed number of no-flow days that was entered into the RF model, i.e. the number of station-months of the calibration period, was 371,550, corresponding to an average of more than 16 years per station. The respective numbers for model validation are 188,156 station-months or more than 8 years. While 3.8% of the station-months were intermittent, i.e. contained at least one no-flow day, during the calibration period, this number increased very slightly to 3.9% during the validation period. During the calibration period, 24% of the stations had at least one no-flow day, while the value decreased to 18% for the validation period, the decrease being caused by the shorter validation period as compared to the calibration period.

3.2.2 Predictors

Monthly time series of hydrological indicators as well as static physiographic characteristics were considered as predictors for building the statistical RF model of the number of no-flow days per month and river reach at the continental scale. The 18 predictors include monthly time series for the period 1981-2019 at two spatial resolutions (15 arc-sec HR and 0.5 arc-deg LR), both of which are based on WaterGAP output (and input) as well as static variables with a resolution of 15 arc-sec. The HR hydrological predictors are all derived from the HR downscaled monthly time series of streamflow (Section 2). Streamflow is converted into area-specific streamflow by dividing it by the upstream area. The static HR predictors are selected from the set of important potential predictors from Messager et al. (2021) but also include two karst-related predictors derived from a newer dataset on karst occurrence (WOKAM dataset) and an updated version of Global Aridity Index (Version 3; Zomer et al., 2022). The predictors entered into the RF modeling are listed in Table 1. The values of these predictors were assembled for each of the 371,550 station-months that were available for setting up the model (model calibration), i.e. for the 1,915 HR grid cells that contain a gauging station. In the case of the HR streamflow predictors and the static predictors karst_status, the value for the HR grid cell for which the number of no-flow days is to be predicted (target cell) was used as a predictor. In the case of all other predictors, the values in the upstream areas of the target grid cell as defined by the updated HR HydroSHEDS drainage direction map were aggregated to be predictive for flow intermittence at the target cell. In the case of the LR predictors, the value of the predictor was assumed to be the same in all HR cells within the LR cell.

After finalizing the RF model, we found that for one out of the 1915 gauging stations (in France), the applied drainage direction area was too small, leading to an overestimation of the HR time series of area-specific streamflow. By mistake, one of the 18 predictors of Table 1, the glacier area fraction in the upstream drainage basin glacier_fraction, was not considered in step 1 (but in step 2 where its importance was found to be very low, see Figure 29). In addition, we found that the LR indicator runoff_dvar is conceptually flawed as 1) WaterGAP results in many daily runoff values of zero due to its runoff generation algorithm and 2) it is not meaningful to divide a runoff value by drainage area. Fortunately, the predictor has rather low importance (ranks 8 and 10 in step 1 and step 2, respectively, Figures 24 and 29).

Table 1. Description of the predictors used in RF modeling, with their abbreviations, units and sources. Specific streamflow is streamflow divided by upstream drainage area. Predictors marked with an asterisk (*) were averaged across the total drainage area upstream of the reach pour point.

Category	Predictor type	Predictor	Abbrevation (unit)	Source
Hydrology	Monthly time series HR (15 arc-sec)	Monthly specific streamflow	Q $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology		Interannual variability of monthly specific streamflow, per calendar month, in terms of standard deviation	Q_iav_sd $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology		Interannual variability of monthly specific streamflow, per calendar month, in terms of coefficient of variation	Q_iav_cv (-)	Downscaled WaterGAP 2.2e
		Minimum monthly specific streamflow of the past 12 months	Q_min_p12 $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology		Mean monthly specific streamflow of the past 12 months	Q_mean_p12 $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology		Minimum monthly specific streamflow of the past 3 months	Q_min_p3 $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology		Mean monthly specific streamflow of the past 3 months	Q_mean_p3 $(\mathrm{m}^3\ \mathrm{sec}^{-1}\mathrm{km}^{-2})$	Downscaled WaterGAP 2.2e
Hydrology	Monthly time series LR (0.5 arc-deg)	Daily variability (max – min) of runoff from land per month divided by upstream drainage area*	runoff_dvar $(\mathrm{mm\ s}^{-1}\mathrm{km}^{-2})$	WaterGAP 2.2e
Hydrology		Diffuse groundwater recharge to runoff from land ratio	gwr_to_runoff_ratio (-)	WaterGAP 2.2e
Climate		Number of wet days*	wet_days (days/ mon)	WaterGAP 2.2e
Climate	Static HR (15 arc-sec	Global aridity index (P/PET)* (long-term average per calendar month	P_to_PET_ratio (1/10000)	Global Aridity Index v2[1]
Land cover		Potential natural vegetation classes, dominant class in upstream drainage basin (range 1-15)	pot_nat_vegetation (-)	EarthStat[2]
Land cover		Land cover, dominant class in upstream drainage basin(range1-22)	land_cover (-)	GLC2000[3]
Land cover		Glacier area fraction in upstream drainage basin	glacier_fraction(%)	GLIMS[4]
Physiography		Drainage area	drainage_area (km^2)	HydroSHEDS[5]
Physiography		Terrain slope*	slope (10^{-2} deg)	EarthEnv-DEM90[6]
Soil+Geology		Fraction of karst area (karst area / drainage area) in upstream drainage basin	karst_fraction (%)	WOKAM[7]
Soil+Geology		Occurrence of karst (1 if karst, 0 if not)	karst_status (-)	WOKAM[7]

1: (Zomer et al. 2022); 2: (Ramankutty and Foley 1999); 3: (Bartholomé and Belward 2005); 4: (GLIMS & NSIDC); 5: (Lehner et al. 2008); 6: (Robinson et al. 2014); 7: (Chen et al. 2017)

3.3 Random Forest modeling approach

3.3.1 The RF methodology

In the late 20th century, a new group of ML methods emerged that are nowadays called decision tree algorithms, such as the Classification and Regression Tree (CART). Exploiting tree-based methods alone sometimes leads to overfitting (good performance with calibration data while much worse on unseen data) due to the complexity and non-linear underlying pattern in the data. To overcome this problem, Breiman et al. (2001) combined a random predictor selection at each tree node (Amit et al., 1997) with a bootstrap aggregation (bagging) ensemble method (Breiman, 1996), which resulted in the popular and effective ML method "Random Forest". In RF, which focuses on variance reduction, the trees are trained randomly on different parts of the training dataset consisting of records containing the observed value of the target and all the pertaining predictors. Thus, many uncorrelated decision trees (i.e. a forest) are built by relating different sub-sets of predictors and the target observations. In the case of classification trees, which were used in our study, each tree is generated by using a threshold for a predictor to split, in the first step, each random subset (node) into two further subsets (nodes at the tree level below the first node) that are more homogeneous with respect to the target variable than the node above. In this way, the so-called impurity of the node/subset is reduced, where the impurity is quantified by the GINI importance that decreases with increasing homogeneity of the subsets. Each split results in two nodes, each of which has a smaller number of records than the node above. Each tree is grown until the number of records in a node reaches a prescribed minimum node size, i.e. the number of records that should not be underrun. The output of the RF for a set of predictors is the class selected by most trees. Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees. For a detailed description and interpretation of RF please refer to Tyralis et al., (2019), Louppe (2014) and Hastie et al., (2008).

The RF also results in information about the importance of the different predictors for the target, in the form of the Mean Decrease in Impurity (MDI), GINI Impurity and Actual Impurity Reduction (AIR) predictor importance metric. The average impurity reduction over all trees in the forest indicates the importance of each predictor. The higher the AIR, the more important the predictor. Partial dependence plots show the probability of a predicted class as a function of the values of an individual predictor, holding the rest of the predictors at their respective mean values; the flatter (horizontal) the line, the less important it is which value the predictor takes.

Similar to other ML methods, RF has a set of hyperparameters. The hyperparameters of the RF method are 1) the fraction of the original training data that is randomly sampled without replacement to construct each tree (alpha), 2) the number of predictors that are sampled from the full set of predictors and considered by each tree when splitting a node (MTRY) and 3) the minimum number of observations that a terminal node can contain, which influences the depth of the trees. We did not tune the number of trees in the forest because the performance of RF has been demonstrated to monotonically and asymptotically increase with the number of trees, so that it should be set as high as computationally possible. However, alpha, MTRY and the minimum node size were adjusted during training to optimize performance. Estimation of these two hyperparameters can be done by cross-validation or metaheuristic optimization algorithms like Particle Swarm Optimization (PSO). In this study, we used cross-validation. We applied the ranger package of Wright and Ziegler (2017) for RF modeling; it is a fast implementation of RF suited for high-dimensional data.

3.3.2 The two-step RF approach

There are many months without any no-flow day in the target observations;datasetsuch severe imbalance in training data could bias the resulting model (Japkowicz and Stephen, 2002). Therefore, a sequential approach to statistical modeling was taken. In step 1, an RF model is built to achieve a binary classification of each month as either intermittent (with at least one no-flow day) or perennial, with a probability threshold of 0.50. In the second step, another RF model was generated for only those station-months that were observed to be intermittent. For those, we developed in step 2 another classification model for predicting the number of no-flow days per month, distinguishing the four classes 1-2, 3-15, 16-29 and 30-31 no-flow days, after having evaluated also a model with six classes (1-2, 3-8, 9-15, 16-22, 23-29 and 30-31 no-flow days per month). Figure 22 provides the workflow of the two-step approach, including the RF model setup (calibration) and the validation of the RF models. The two RF models are then applied sequentially to simulate monthly time series of no-flow days per month for river reaches. Each river reach and month (reach-month) is characterized by its number of no-flow days, aggregated to five classes: 0, 1-2, 3-15, 16-29 and 30-31 no-flow days (Section 3.5).

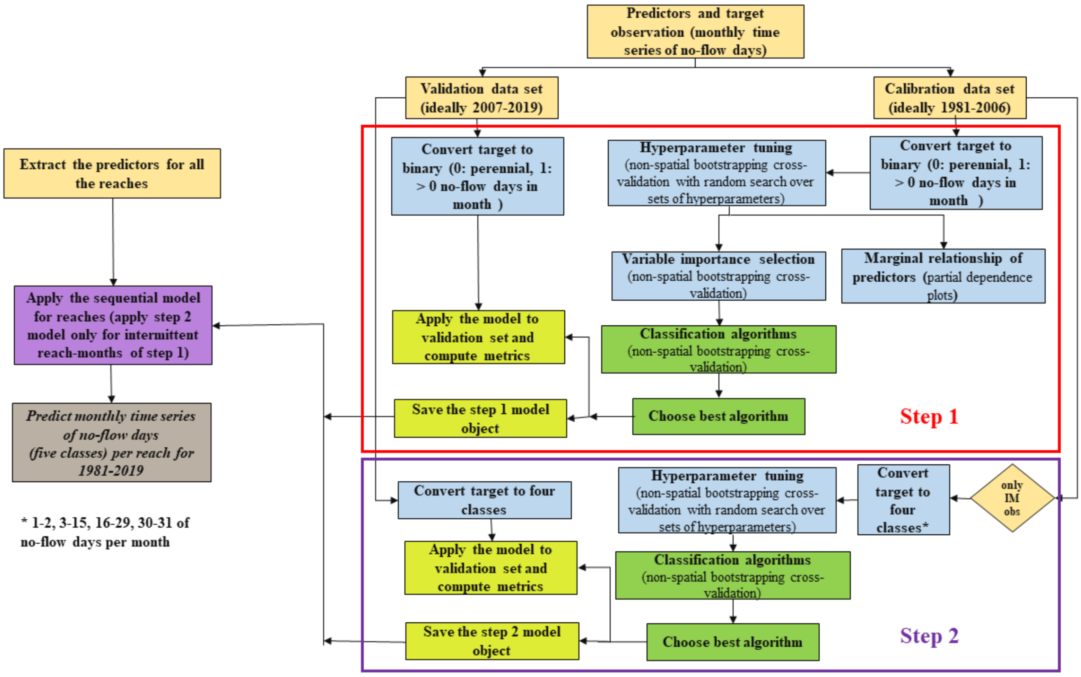
Predictors and target observation (monthly time series of no-flow days)
Validation data set (ideally 2007-2019)
Calibration data set (ideally 1981-2006)
Extract the predictors for all the reaches
Convert target to binary (0: perennial, 1: > 0 no-flow days in month)
Hyperparameter tuning (non-spatial bootstrapping cross-validation with random search over sets of hyperparameters)
Convert target to binary (0: perennial, 1: > 0 no-flow days in month)
Variable importance selection (non-spatial bootstrapping cross-validation)
Marginal relationship of predictors (partial dependence plots)
Apply the sequential model for reaches (apply step 2 model only for intermittent reach-months of step 1)
Apply the model to validation set and compute metrics
Classification algorithms (non-spatial bootstrapping cross-validation)
Predict monthly time series of no-flow days (five classes) per reach for 1981-2019
Save the step 1 model object
Choose best algorithm
Step 1
Convert target to four classes
Hyperparameter tuning (non-spatial bootstrapping cross-validation with random search over sets of hyperparameters)
Convert target to four classes*
only IM obs
* 1-2, 3-15, 16-29, 30-31 of no-flow days per month
Apply the model to validation set and compute metrics
Classification algorithms (non-spatial bootstrapping cross-validation)
Save the step 2 model object
Choose best algorithm
Step 2

Figure 22: Workflow of simulating monthly time series of streamflow intermittence on river reaches, i.e. the number of no-flow days per month in five classes.

We addressed the problem of imbalanced data (many more perennial station-months than intermittent ones in step 1 and a large number of station-months with 30-31 no-flow days in step 2) not only by the sequential approach, but also by applying, in step 1, standard oversampling (oversampling the minor class 10 times) and, in step 2, the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002), whereby the minority classes are oversampled by creating „synthetic" examples rather than by over-sampling with replacement. Minority classes were oversampled such that, for each minority class, the number of training observations in that class was 80% of the number of observations in the majority class.

Hyperparameter tuning in steps 1 and 2 was carried out by evaluating random ensembles of the combinations of MTRY and minimum node size. 60 and 96 ensemble members were evaluated in steps

1 and 2, respectively. We applied a resampling method, where cross-validation (CV) was done three times with different sub-subsets of the records. In each CV round, the training dataset is split randomly into k parts and then the model is trained k times on k-1 parts and evaluated on other fold for k times, each time with a different combination of parts. In step 1, k was 4, in step 2, k was 5. The hyperparameter set leading to the highest Bacc value (see next section) was then used to set up the RF model, to counter the problem of overfitting and consequently bias and variance errors. For step 1, the optimal values for MTRY and minimum node size were 8 and 7, respectively; the corresponding values for step 2 are 2 and 6. The run time for step 1 was 10 days and only 7 hours for step 2.

3.3.3 Classification performance metrics

Classification metrics are the standard for evaluating the performance of an RF classification model. They quantify how well the predicted class fits to the observed class. We evaluated the balanced accuracy Bacc, precision, recall, sensitivity, specificity, and F-score. The metrics are explained in Figure 23 for the case of a binary classification as done in step 1. For the step 2 RF modeling, the metrics were first calculated separately for each class (6 or 4), and then averaged over the classes. In a strongly unbalanced dataset such as the dataset in our study, Bacc provides the best indication of how well the classification performs.

		Actual	
		Positive Non-perennial	Negative perennial
Predicted	Positive Non-perennial	True Positive (TP)	False Positive (FP)
	Negative perennial	False Negative (FN)	True Negative (TN)

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$Balanced\ accuracy = \frac{sensitivity + specificity}{2}$$

$$F - score = \frac{2 * Recall * Precision}{(Recall + Precision)}$$

Figure 23: Binary confusion matrix and classification metrics in case of two classes only.

3.4 RF results and performance

3.4.1 Step 1 modeling results

The AIR of the 17 predictors of the intermittence state of each station-month (with at least one no-flow day or perennial) indicate that the upstream drainage area of each streamflow observation station and the average slope in the upstream area are the two most important predictors for the intermittence state (Figure 24). Importance ranks 3 to 5 were assigned by the RF model to three predictors derived from the HR downscaled monthly WaterGAP streamflow, 1) the streamflow of the current month, 2) the interannual variability of monthly streamflow expressed in standard deviation and 3) the minimum streamflow during the previous 12 months.

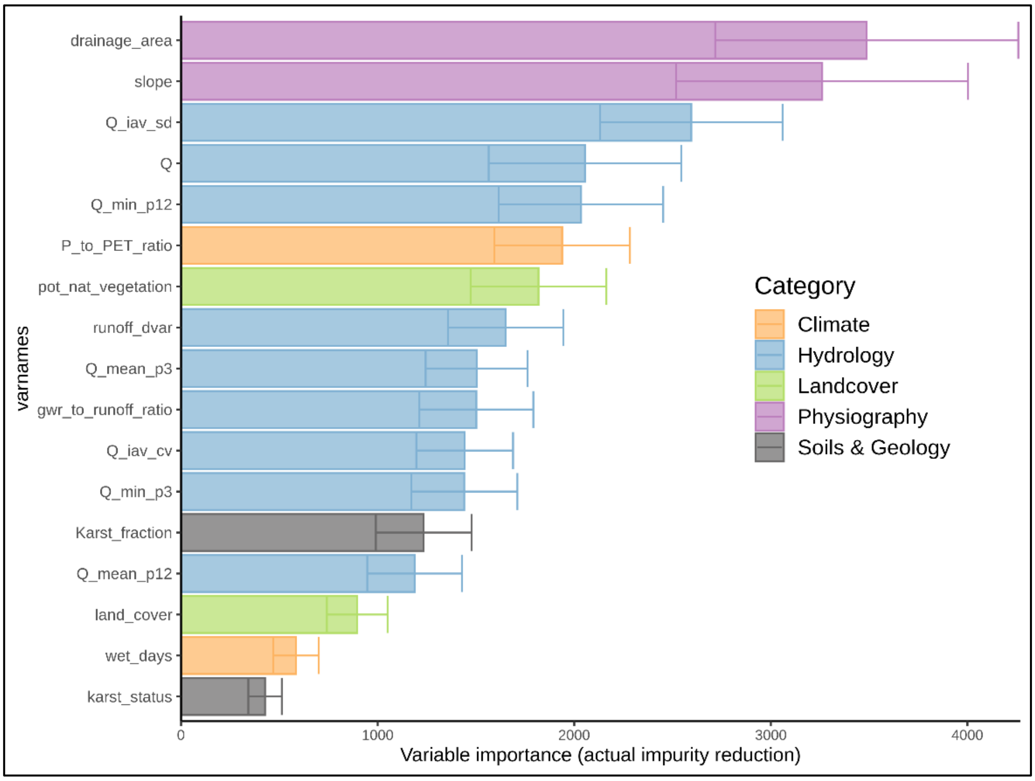

Figure 24: Ranked importance of the predictors as quantified by AIR for predicting perennial and intermittent station-months (step 1 RF model)

The partial dependence plot for the drainage area shows that the probability that a station-month is classified as intermittent decreases as expected with increasing drainage area, for drainage areas up to approximately 30,000 km^2 (Figure 25). The plot for slope, however, indicates an unexpected negative correlation between the probability of being an intermittent month and the slope; the probability is computed to be higher for flat terrain, while hydrological expertise expects more intermittency in steeper terrains. The negative correlation can be explained by the spatial distribution of the gauging stations; gauging stations in steeper terrain are those in the mountainous regions along the Spanish Atlantic coast, the Pyrenees and the Alps, wet regions with large runoff. In the case of all three HR hydrological indicators, the probability of at least one no–flow day in the month decreases with increasing values up to a certain predictor value and then shows no further dependence for higher predictor values, which also is in accordance with expectations. The distribution of the standard deviation of monthly streamflow per calendar month is dominated by the heterogeneity of the mean streamflow itself. The partial dependence plot for interannual variability as expressed by the coefficient of variation (Q_iav_cv) shows the expected behavior, with the intermittence probability increasing with increasing Q_iav_cv in the range of approx. 0.1<Q_iav_cv<1.8 (Figure 25). With 0.86, the balanced accuracy Bacc is quite high for the calibration and does not deteriorate much for the validation data (Table 2). The same is true for the other metrics.

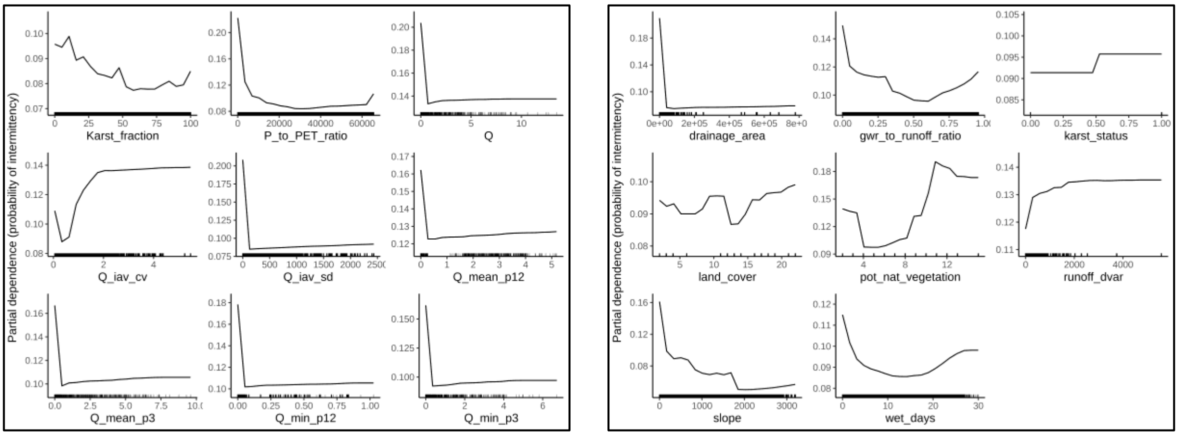

Figure 25: Partial dependence plot showing the probability of a station-month of being classified as intermittent as a function of the values of the 17 predictors (Step 1 RF model).

Table 2: Classifying station-months as either perennial or intermittent (at least 1 no-flow day): classification performance metrics of the step 1 model.

	Bacc	Recall	Precision	Sensitivity	Specificity	F-score
Calibration	0.86	0.75	0.56	0.75	0.98	0.64
Validation	0.83	0.70	0.50	0.70	0.97	0.58

Figure 26 shows maps of the ratio of predicted to observed intermittent months (with at least one no-flow day) per gauging station, for both the calibration and validation periods. For reference, the percentage of observed intermittent months per gauging station is shown, too, in Figure 26. Stations without any observed intermittent months dominate (grey dots in Figure 26). For most other stations, less than 30% of the months are observed to be intermittent (Figure 26a, b) in both the calibration and validation periods. Most stations with more than 30% intermittent months are located in Spain; here, the percentage of intermittent months is lower in the validation period than in the calibration period. The step 1 model both over- and underestimates the observed percentage of intermittent months, with the largest discrepancies (dark blue: zero predicted months and at least one observed intermittent month, red: zero predicted months and at least one observed intermittent month) occurring at stations with a zero or small observed or predicted intermittent months (compare Figure 26c,d to Figure 26a,b).

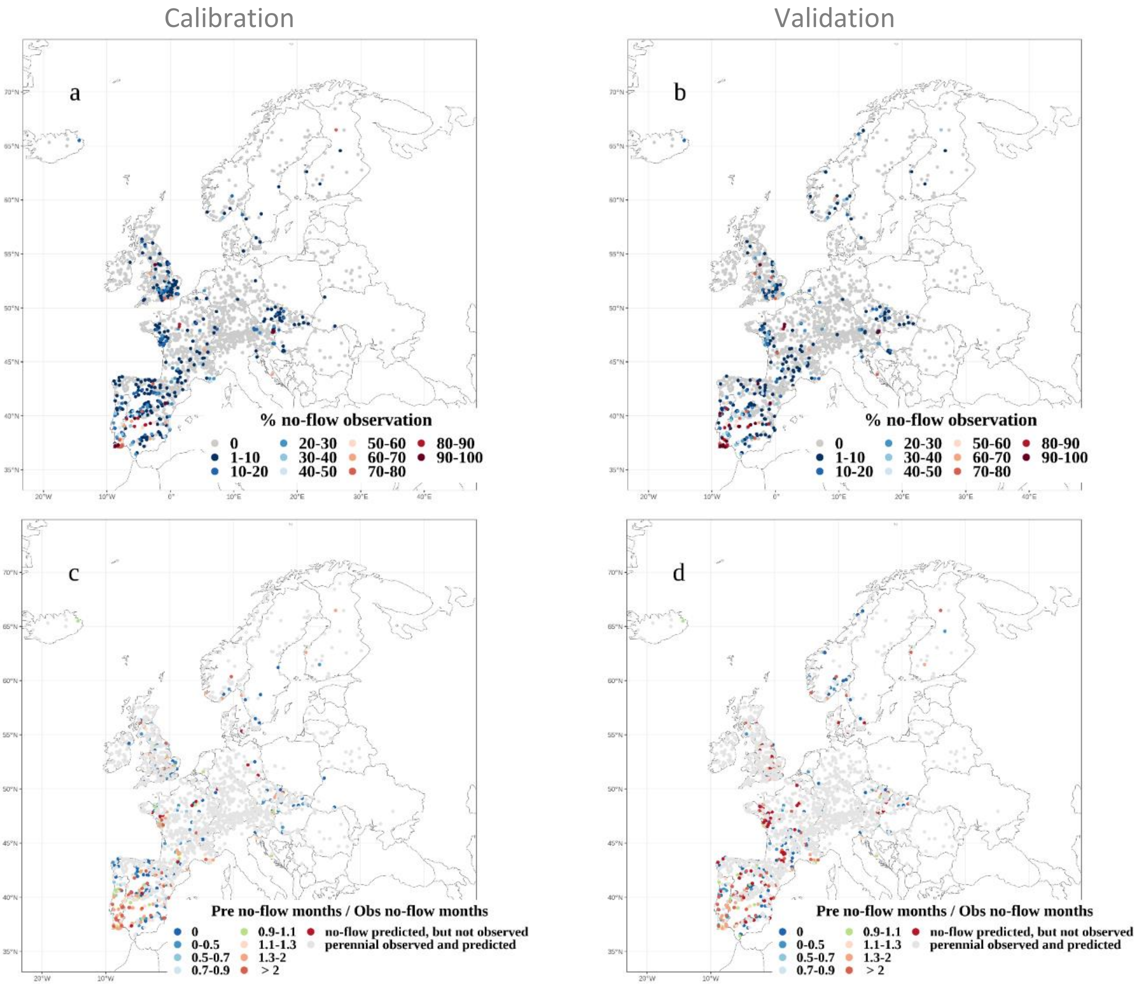

Figure 26: Percentage of observed intermittent months (with at least one no-flow day) per gauging station for the calibration (a) and validation (b) periods, as well as the ratio of predicted to observed intermittent months per gauging station for the calibration (c) and validation (d) periods.

Of the observed completely perennial 1452 stations without any intermittent month during the calibration period, 1437 are correctly predicted to be completely perennial, i.e. 98.9% (Table 3). However, only 72.3% of the 463 observed intermittent stations (with at least one intermittent month) are correctly predicted to be intermittent. 97.6% of all observed perennial station-months are correctly predicted, but only 75.2% of the observed intermittent station-months. About 8500 perennial station months are wrongly predicted to be intermittent, while about 3500 intermittent station months are wrongly predicted to be perennial. For the validation, the fraction of correctly predicted perennial months decreases only slightly to 97.1 but the fraction of correctly predicted intermittent months decreases more strongly to 69.8%. As described in Section 3.1, the percentage of observed intermittent months did not change much between the calibration and the validation period.

Considering the seasonal variations of streamflow intermittence in Europe, the highest number of observed intermittent months occurs in September-October-November (SON), the second highest in June-July-August (JJA), and the third highest in March-April-May (MAM). December-January-February (DJF) has the lowest number of intermittent months but not much less than MAM (Table 3). The percentage of correctly predicted intermittent months is highest in SON and JJA, somewhat higher than in DJF and MAM in the calibration period but more than 10% higher than in DJF and MAM in the validation period. Comparing the decrease in performance regarding intermittent station-months between the calibration and the validation period, the decrease was very low in SON and JJA (less than 3%) and rather high (almost 10%) in DJF and MAM.

Table 3. Number of observed perennial (P) and intermittent (I) stations and station-months (bottom numbers) and of correctly simulated perennial and intermittent station and station-months (top numbers). Information on station-months is provided for all months (4th column) and the four seasons December to February (DJF), March to May (MAM), June to August (JJA) and September to November (SON).

		Number of stations	Number of station-months	DJF	MAM	JJA	SON
Calibration	$\frac{P_{sim}}{P_{obs}}$	$\frac{1437}{1452}=98.9$	$\frac{348884}{357377}=97.6$	$\frac{88226}{89932}=98.1$	$\frac{91590}{93475}=98$	$\frac{85390}{87879}=97.2$	$\frac{83678}{86091}=97.2$
	$\frac{I_{sim}}{I_{obs}}$	$\frac{335}{463}=72.3$	$\frac{10661}{14173}=75.2$	$\frac{1829}{2588}=70.7$	$\frac{1960}{2646}=74.1$	$\frac{3349}{4331}=77.3$	$\frac{3523}{4608}=76.4$
Validation	$\frac{P_{sim}}{P_{obs}}$	$\frac{1480}{1555}=95.1$	$\frac{175975}{181201}=97.1$	$\frac{43868}{44823}=97.9$	$\frac{44832}{45860}=97.8$	$\frac{43247}{44849}=96.4$	$\frac{44028}{45669}=96.4$
	$\frac{I_{sim}}{I_{obs}}$	$\frac{246}{360}=68.3$	$\frac{5108}{7315}=69.8$	$\frac{945}{1516}=62.3$	$\frac{1030}{1590}=64.8$	$\frac{1435}{1931}=74.3$	$\frac{1641}{2278}=74.5$

That the step 1 RF model tends to overestimate the occurrence of intermittent station-months is not only seen in Table 1 but also in Figure 27. The Nash-Sutcliffe efficiency (NSE), which is both sensitive to bias and correlation, shows satisfactory values, with NSE = 0.67 for the calibration period and NSE=0.51 in the validation period. NSE can range from –∞ to 1 (perfect fit). An NSE value of zero is considered to be a benchmark for the skill of the model, as with NSE = 0, the mean of the data has the same skill as the model. The overestimation of intermittent months dominantly occurs at the outlet of relatively small drainage basins, with upstream areas of less than 10 km^2 in the calibration period and with less than 5 km^2 in the validation period (Figure 28). However, only 47 out of the 1915 stations have drainage areas of less than 10 km^2. In the validation period, streamflow intermittence is also overestimated for the largest basins, with areas of more than 10,000 km^2.

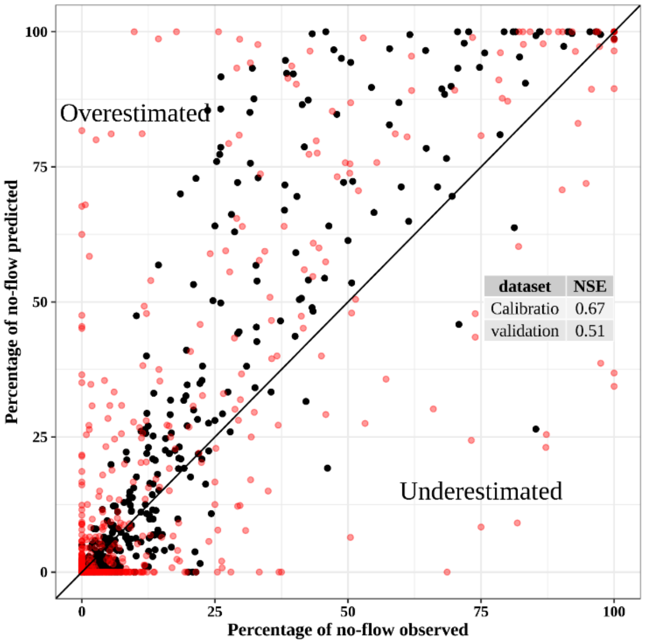

Figure 27: Comparison of the observed and simulated percentage of intermittent months for both the calibration (black) and the validation (red) periods. Each point represents a streamflow gauging station.

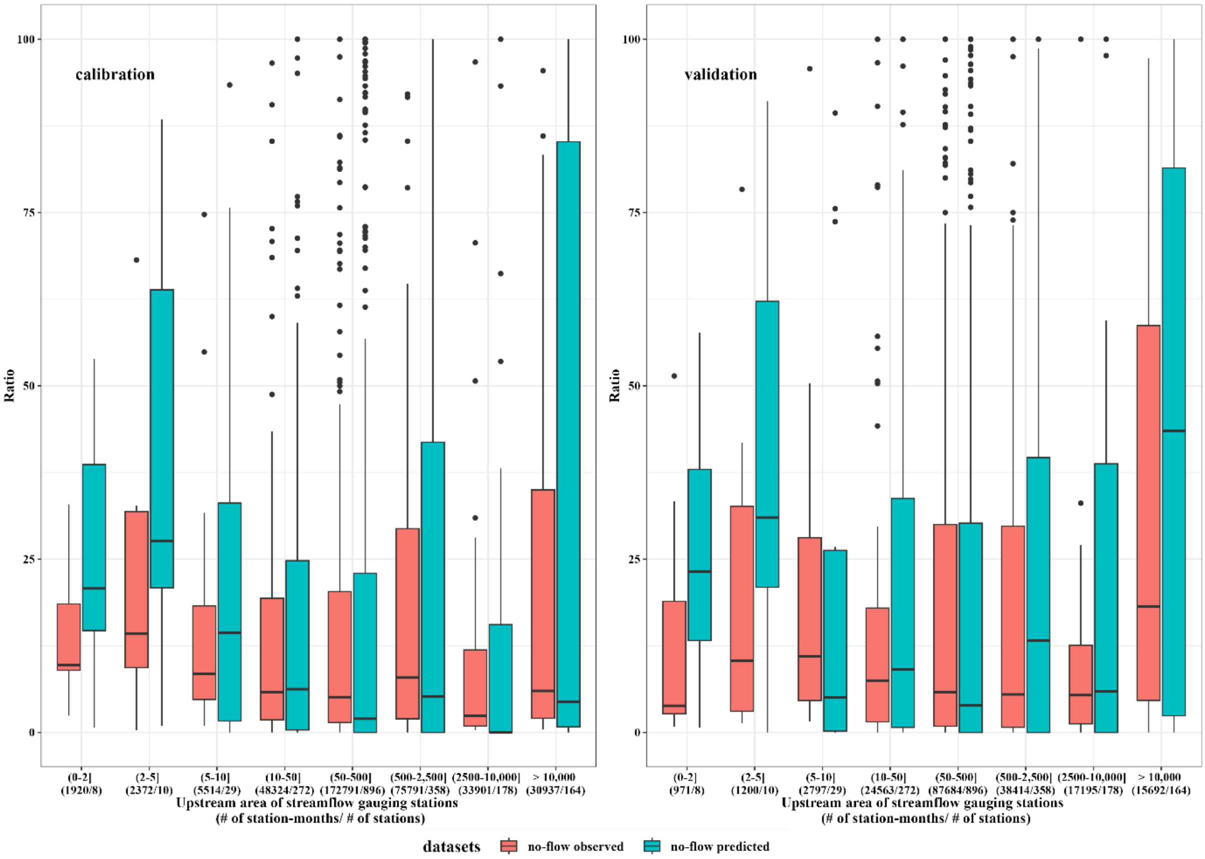

Figure 28: Performance of the step 1 RF model as a function of upstream drainage area [km^2] of the streamflow gauging stations. The box plot shows the fractions of all station-months in a drainage area class that are observed (red) or simulated (green) as intermittent. The values below the upstream area show the number of station-months/number of stations included in the drainage area class. Only stations that are observed or predicted to have at least one intermittent month are included in the figure.

The differences in the occurrence of intermittence between different years can be simulated quite well, as seen by the rather high correlation between observed and predicted annual numbers of the number of intermittent months (Figure 29c,d). This means that the step 1 model can identify dry and wet years well. The poor NSE values (Figure 29a,b) are likely due to the overall bias of the RF model towards too much intermittency.

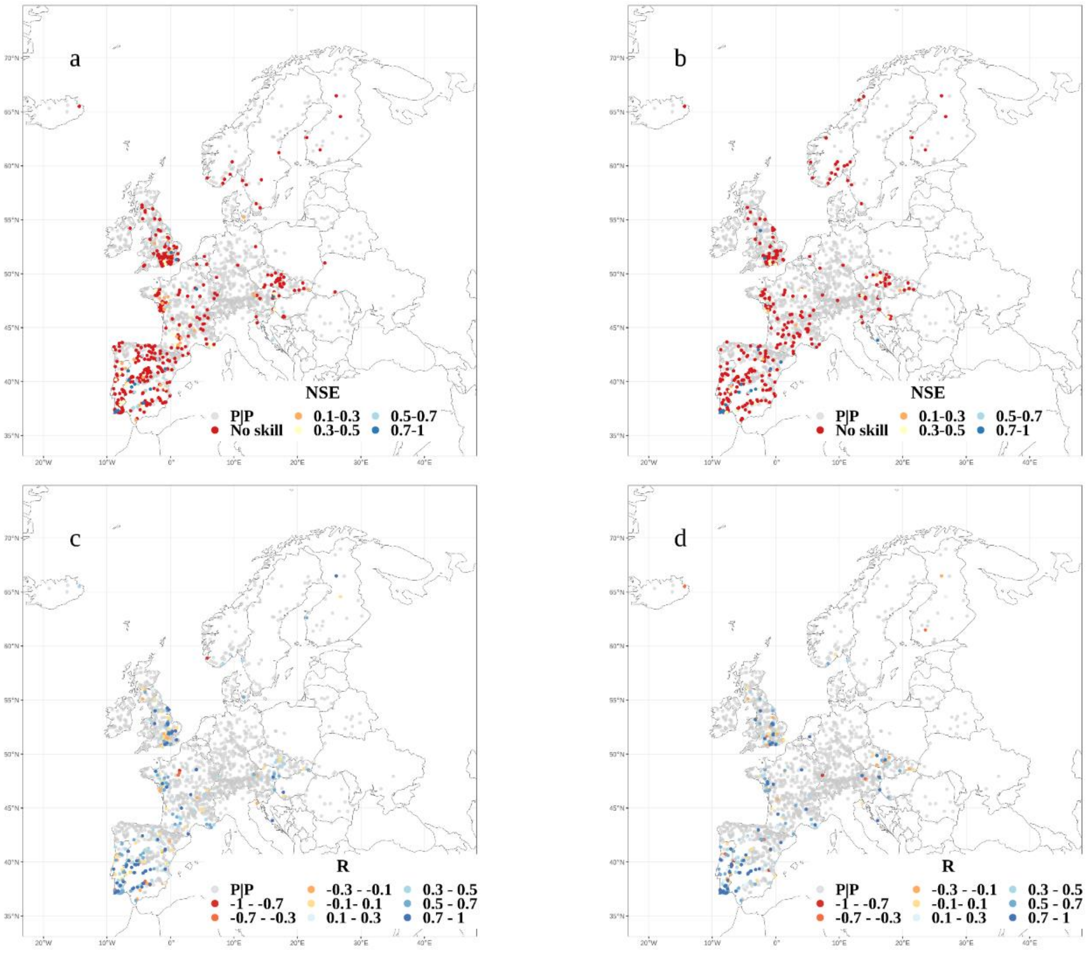

Figure 29. NSE (a, b) and correlation (c, d) for the annual time series of the number of intermittent months per year for the calibration (a, c) and the validation (b, d) (step 1 RF model).

3.4.2 Step 2 modeling results

Monthly specific streamflow (HR), average slop upstream, and drainage area are the three most important predictors of the number of no-flow days in the station-months aggregated to four classes based on AIR (Figure 30). These three predictors are also the most important predictors for the step 1 model, albeit in a different order (Figure 24). The fourth rank is taken by the static HR predictor aridity (PET_to_P_ratio), while five predictors derived from HR specific streamflow take the following ranks.

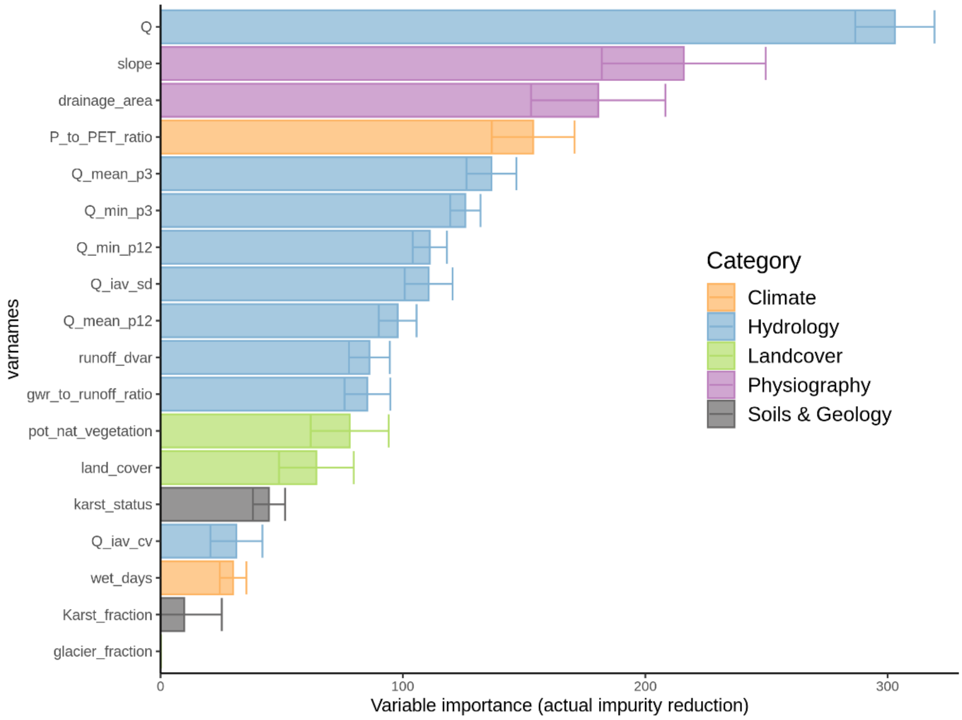

Figure 30: Ranked importance of the predictors as quantified by AIR for predicting the number of no-flow days per station-month in the case of observed intermittent station-months (step 2 RF model).

The classification performance is expected to decrease with an increasing number of classes. When comparing the performance metrics for four and six classes of the number of no-flow days in observed intermittent station-months, the balanced accuracy in the case of six classes is not much worse that in the case of four classes (Table 4) and can still be considered satisfactory. However, precision and F-score degrade significantly in the case of six classes. The confusion matrices for the two classification options clearly show that six classes would lead to an exceedingly high number of wrong predictions (Figure 31b). While in the case of a predicted class n, the majority of the observations are within class n for both classifications the majority of all observations is in the correct class only for four classes (Figure 31a, b). Therefore, the step 2 RF modeling for four classes was used to generate the model output dataset. For 56% of all intermittent station-months, the correct class (1-4) is assigned. Classes 2 and 4 dominate both the prediction and the observations. However, the step 2 RF model for four classes tends to predict too many 30-31 no-flow days (class 4 in Figure 31) and not enough 1-2 no-flow days. Model performance degrades for the validation data, where the correct class can only be identified for 47% of the intermittent station-months (Figure 31c).

Table 4. Classification performance metrics of the step 2 RF model with the adopted six classes of monthly flow intermittence as compared to the performance metrics for four classes, both for the calibration and validation periods.

	Bacc	Recall	precision	sensitivity	specificity	F-score
Six classes: calibration	0.62	0.36	0.39	0.36	0.88	0.36
Six classes: validation	0.58	0.29	0.31	0.29	0.87	0.28
Four classes: calibration	0.67	0.49	0.55	0.49	0.84	0.50
Four classes: validation	0.60	0.38	0.51	0.38	0.81	0.35

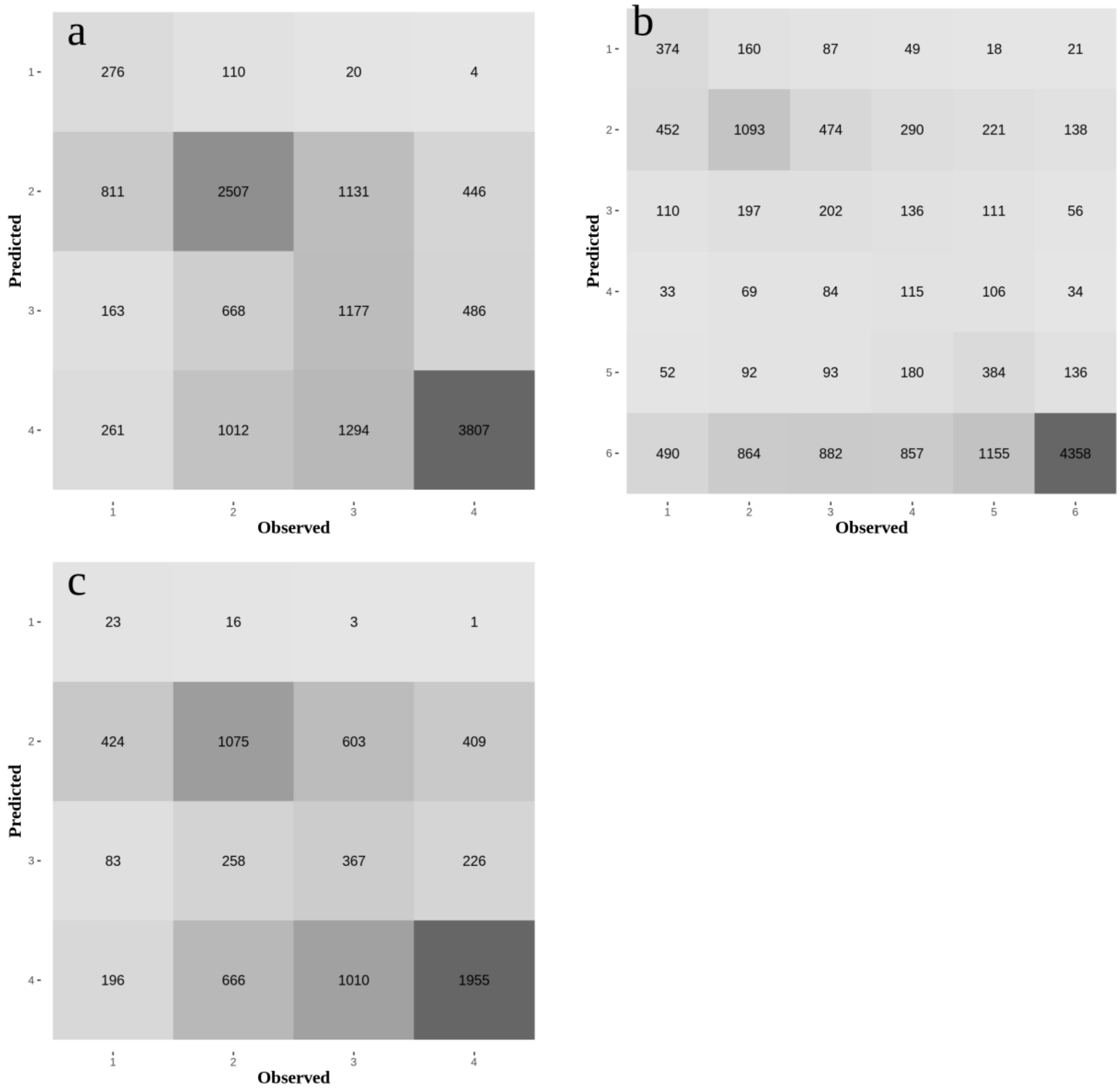

Figure 31: The confusion matrix of predicting four classes of no-flow days per station-month (a) or six classes (b) for the calibration period and four classes for the validation period (c) (step 2 RF model).

To understand the spatial distribution of observed flow intermittence and step 2 RF model performance, Figures 32 and 33 show the percentage of months that belong to one of the four classes for the calibration and validation datasets, respectively. Most of the gauging stations that have 30-31 no-flow days for more than half of the months are located in Spain and Great Britain. The percentage of intermittent months at each gauging station that was classified correctly in the four classes varies strongly between stations (Figure 34). During the calibration period, at most stations more than half of the intermittent months at each station are classified to belong to the correct class. The performance degrades for the validation dataset, when often less than half of the intermittent months are classified correctly.

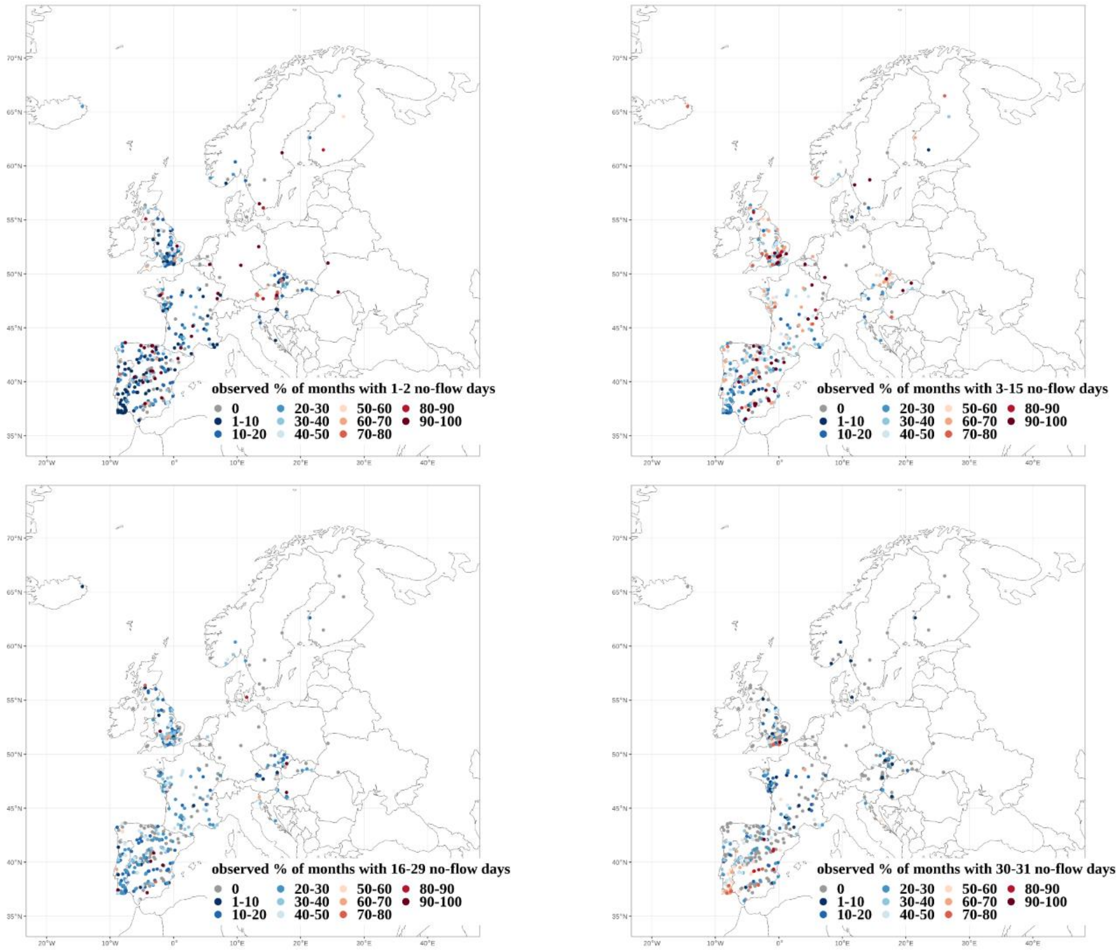

Figure 32: Percentage of intermittent months with observations of the four intermittence classes at gauging stations in the calibration dataset.

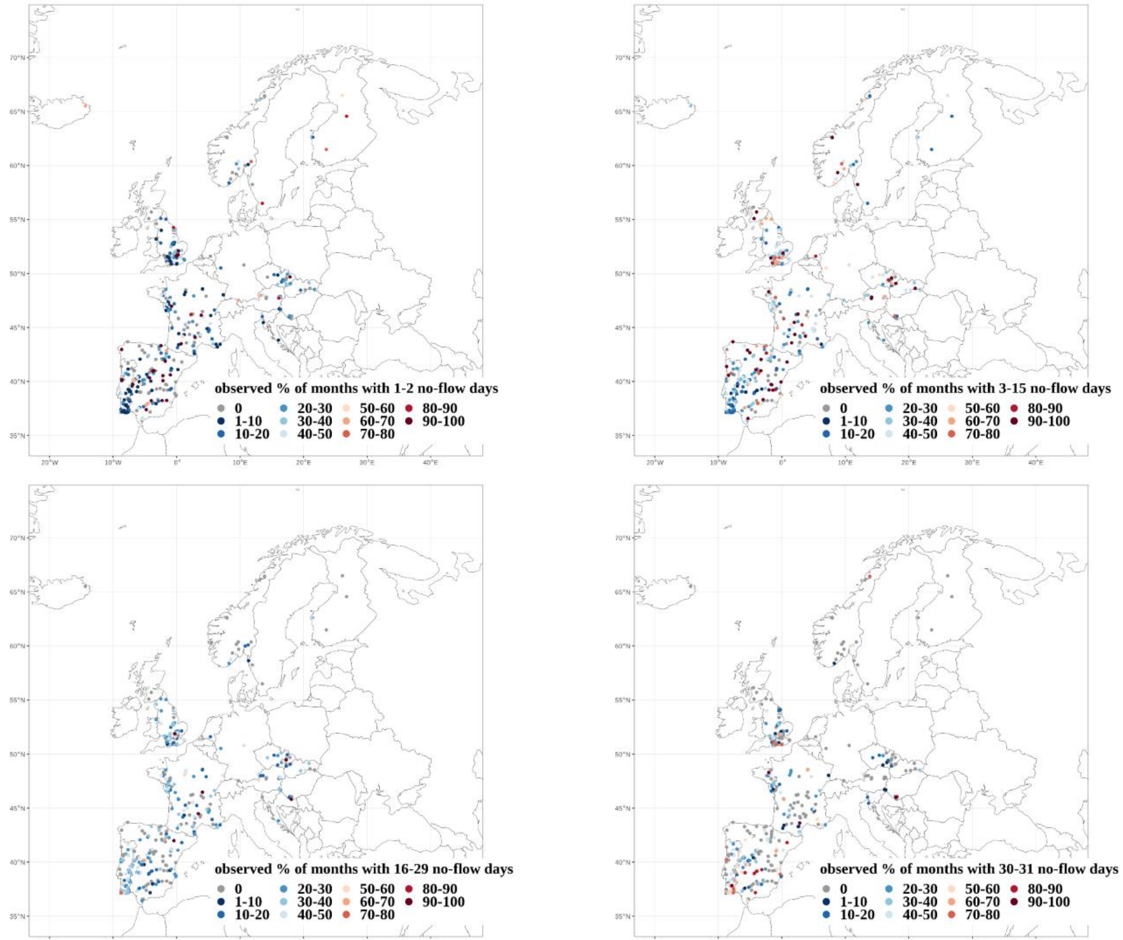
observed % of months with 1-2 no-flow days
observed % of months with 3-15 no-flow days
observed % of months with 16-29 no-flow days
observed % of months with 30-31 no-flow days
0
1-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
70°N
65°N
60°N
55°N
50°N
45°N
40°N
35°N
20°W
10°W
0°
10°E
20°E
30°E
40°E

Figure 33: Percentage of intermittent months with observations of the four intermittence classes at gauging stations in the validation dataset.

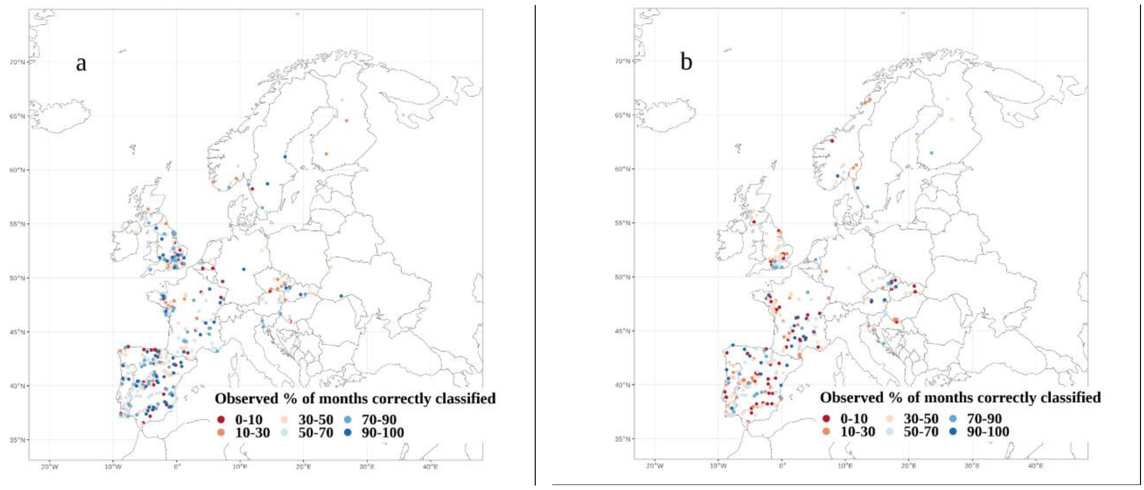

Figure 34: The percentage of intermittent months at streamflow gauging stations that is classified correctly into the four classes for the calibration (a) and validation (b) periods.

The ability of the step 2 RF model to predict monthly time series of the number of no-flow days (aggregated to four classes, i.e. a time series of the values 1, 2, 3 and 4) is rather high when considering the performance metric polychoric correlation coefficient (suitable for quantifying the correlation of ordinal categorical variables), which only considers the phase of the temporal variability and not the absolute value for the calibration period (Figure 35c). However, correlation during the validation period degrades strongly for many station (Figure 35d). The mean bias of the time series is, for most gauging stations, positive, i.e. the RF model tends to overestimate the number of no-flow days both for the calibration (Figure 35a) and the validation periods (Figure 35d). Overestimation is increased during the validation period in Spain but decreased in the Czech Republic.

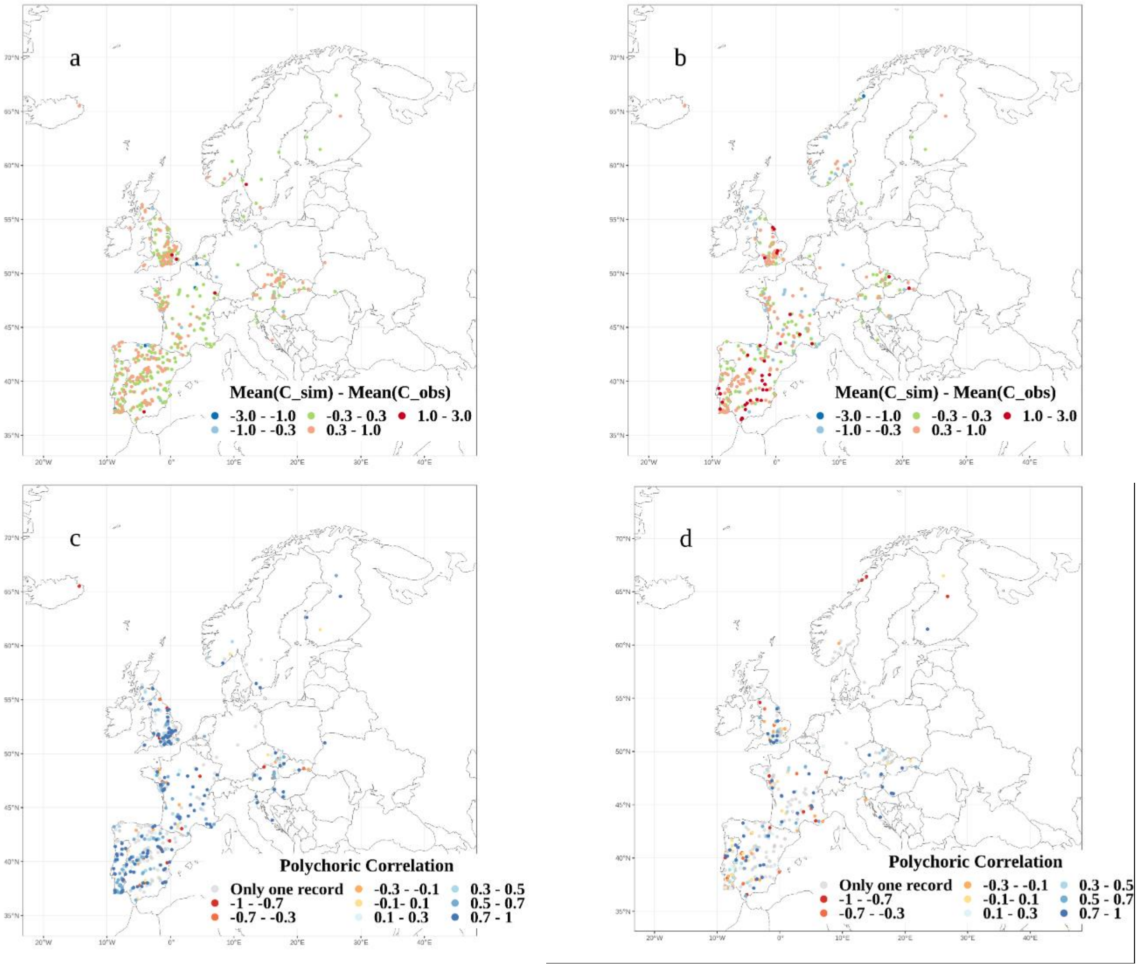

Figure 35. Performance of monthly time series of classes (1, 2, 3 or 4) for the intermittent months at each gauging station (step 2 RF model): Bias expressed as simulated mean value minus observed mean value (positive: overestimation of no-flow days) (a: calibration period; b; validation period) as well as polychoric correlation coefficient (c: calibration period, d: validation period).

3.5 Application of the RF modeling approach to compute monthly time series of flow intermittence (number of no-flow days in five classes) for European HR river reaches during the period 1981-2019

Application of the RF model presented in Section 3.4 was done by first applying the step 1 model to the river reaches identified as described in Section 3.5.1. Then, the step 2 model was applied for all river reaches and months (reach-months) that are predicted to be intermittent (at least one no-flow day) by the step 1 model. Finally, the intermittence status of each reach-month was quantified by the five classes 0-4, corresponding to 0, 1-2, 3-15, 16-29 and 30-31 no-flow days. All predictors used for generating the RF models were used for model application as all were found to be significant ($p<0.05$).

3.5.1 Derivation of river reaches

We had planned to apply the RF model for each HR grid cell covering Europe and South America. However, we found that this would be computationally too expensive regarding both computation time and data storage. Only for Europe (without Russia), there are 73 million HR grid cells and 468 months (1981-2019) such that more than 34 billion predictions of the intermittence status would have to be computed. Therefore, we decided to apply the RF model to predict the intermittency status of river reaches. Predictions are then made for the most downstream HR grid cells of each river reach and are assumed to represent the mean conditions over the whole river reach. Although the HydroRIVERS dataset (Linke et al. 2019) already provides a seamless digital representation of the river network at 15 arc-sec, we produced a new river network dataset described herein to (i) account for the drainage direction modifications implemented in the three DRNs, and (ii) to extend the coverage of the river network to smaller rivers and streams compared to HydroRIVERS to suit the purpose of other DRYVER Work Packages. The river network was delineated following the same overall procedure as HydroRIVERS, but using a different set of input datasets that were created as part of the downscaling of the global hydrological model WaterGAP 2.2e (section 2): the drainage area upstream of every cell and the long-term mean annual discharge in each cell. Those two input datasets differ from the original ones used in producing HydroRIVERS because (i) they were both produced using the custom drainage direction dataset modified in the three DRNs for DRYvER, (ii) the streamflow dataset was produced with a different version of the global hydrological model (WaterGAP 2.2d was used for deriving HydroRIVERS) and (iii) the downscaling procedure of the hydrological model. The streams/rivers were defined to start at all pixels with an upstream drainage area of more than 2 km^2 (instead of 10 km^2 for HydroRIVERS) or at a grid cell where the mean annual downscaled HR streamflow of WaterGAP 2.2e during the period 1981-2019 exceeds 0.03 m^3/s (instead of 0.1 m^3/s in HydroRIVERS). Decreasing the threshold for streamflow to 0.02 m^3/s would lead to artefactual "aggregates" of streams in wet areas. The total number of reaches in Europe (without Russia) is 1,533,471, such that the European streamflow intermittence dataset will contain a total of 717,664,428 reach-months covering the period 1981-2019.

In the six focal DRNs of the DRYvER project in Europe, the newly derived river network covers more of the first-order streams included in the high-resolution stream networks provided on the DRYvER server than HydroRIVERS (using the catchment area and estimated discharge at the middle point of first-order streams in the DRNs as reference). The coverage percentages increase as follows: focal DRN Finland: from 6% to 12%, Hungary: from 20 % to 25%, Czech Republic: from 29 % to 45%, Croatia: from 32% to 53%, Spain: from 32% to 55% and France: from 35% to 60%.

Note that the river reaches as derived from the drainage direction dataset may not correspond to actual river reaches. In particular, river reaches (and therefore the streamflow intermittence status) are also identified in actual lakes and man-made reservoirs. Users of the developed streamflow intermittence dataset for reaches may need to mask out simulated reaches that are actually within lakes and reservoirs. Finally, while the extension of the river network to smaller river reaches (compared to HydroRIVERS) implies an increase in precision, it also implies a decrease in reliability and accuracy of the newly delineated streams (in terms of their spatial representation) due to the uncertainties in the underpinning global geometric and hydrologic data.

3.5.2 Simulated intermittence of European river reaches

90.87% of the more than 717 million reach-months in Europe during 1981-2019 are simulated as perennial, e.g., they have no days without streamflow (Table 5). This is a larger fraction than the fraction in the calibration or validation datasets, where more than 96% of the station-months were observed to be perennial, and more than 94.5% were predicted to be perennial. Only a negligible number of reach-months is simulated to have 1-2 no-flow days, which is very different from the observed station-months in this class, but less different from the predicted station-months. The intermittence class 3-15 no-flow days is the second-largest class both for the reaches and the gauging stations. More than 3% of the reach-months are simulated to be in this class, while less than 1% are simulated to be in the class 16-29 no-flow days. More than 5% of the reach-months are simulated to be in the class 30-31 no-flow days, which is much larger than the observed or predicted values for the gauging stations.

Table 5. Occurrence of the five intermittence classes in reach-months in Europe (without Russia and Turkey) during 1981-2019 (last column) as compared to the occurrence at the gauging stations used to set up the RF model (first four columns), where the fraction of all station-months with observed and simulated classes is provided for both the calibration and the validation dataset. The percentage values for the predicted station-months relate to the total station-months with observations; the step 2 model predicting the four classes with no-flow days was set up only for the station-months that are observed to be intermittent.

Class	Station-months Calibration		Station-months Validation		Reach-months (1981-2019)
	Observed	Predicted	Observed	Predicted	Predicted
Perennial	357,377	352,396	181,201	178,182	652,151,841
	96.19%	*94.84%*	*96.11%*	*94.52%*	*90.87%*
1-2 no-flow days	1,511	410	726	43	2,176
	0.41%	*0.11%*	*0.39%*	*0.02%*	*0.00%*
3-15 no-flow days	4,297	4,895	2015	2,511	22,884,683
	1.16%	*1.32%*	*1.07%*	*1.33%*	*3.19%*
16-29 no-flow days	3,622	2,494	1983	934	5,557,991
	0.97%	*0.67%*	*1.05%*	*0.50%*	*0.77%*
30-31 no-flow days	4,743	6,374	2591	3,827	37,067,737
	1.28%	*1.72%*	*1.37%*	*2.03%*	*5.17%*
Total	371,550	366,569	188,516	185,497	717,664,428
	100%	*98.66%*	*100%*	*98.40%*	*100%*

The spatial pattern of simulated intermittence over Europe is provided in Figure 36, showing for each reach the fraction of intermittent months with at least one no-flow day in the 468 months of 1981-2019. The highest fractions occur in Spain, while mountainous areas (with a high average slope) show perennial conditions, maybe because most gauging stations used for model set-up are in wet mountain regions (Section 3.4.1).

As expected, intermittent conditions exist much more widely in the summer than in the winter. The monthly time series of the frequency of the intermittence class perennial, aggregated over all reaches in Europe, peaks, in most years, in January and reaches its lowest value in July (Figure 37). The other intermittency classes peak in July-September. There is interannual variability even aggregated over all of Europe, with e.g. dry winters in 1992 and 2012. There is no predicted temporal trend of decreasing or increasing intermittence at the pan-European scale.

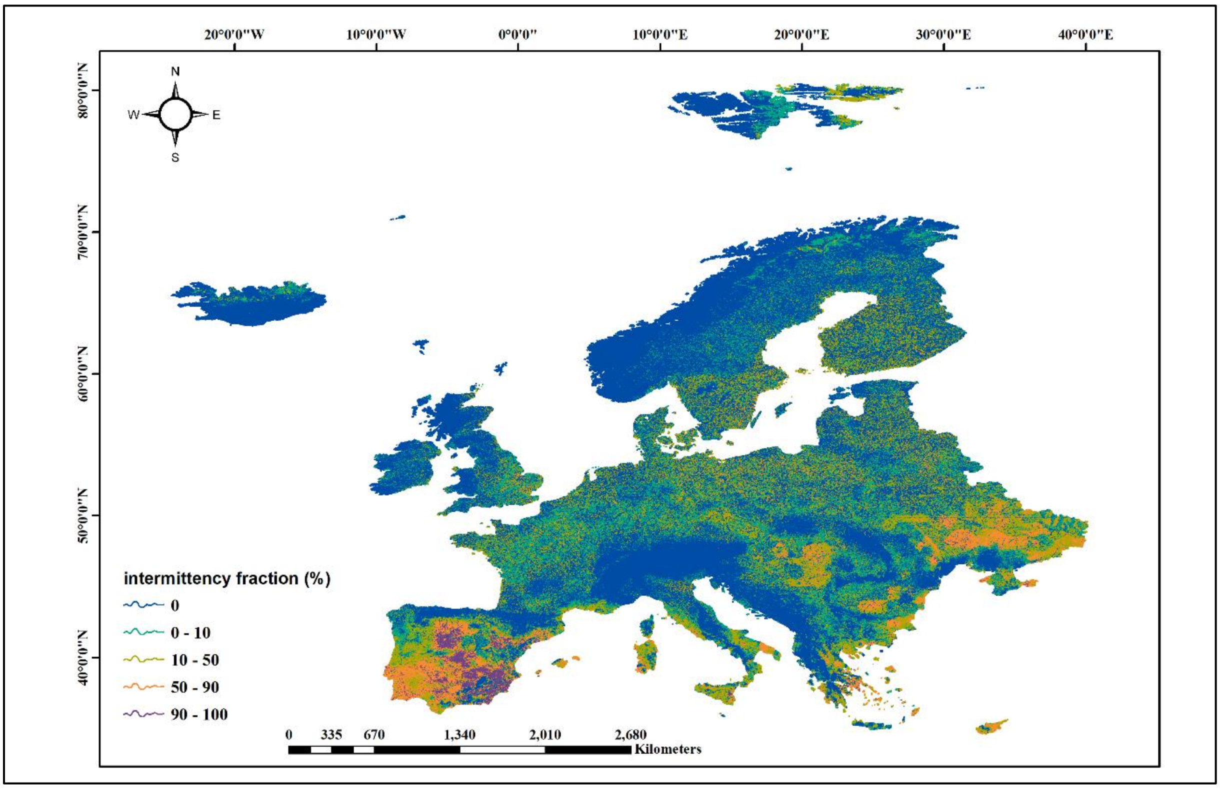

Figure 36. Percentage of months with at least one no-flow day (output of application of step 1 RF model) during the period 1981-2019 (map).

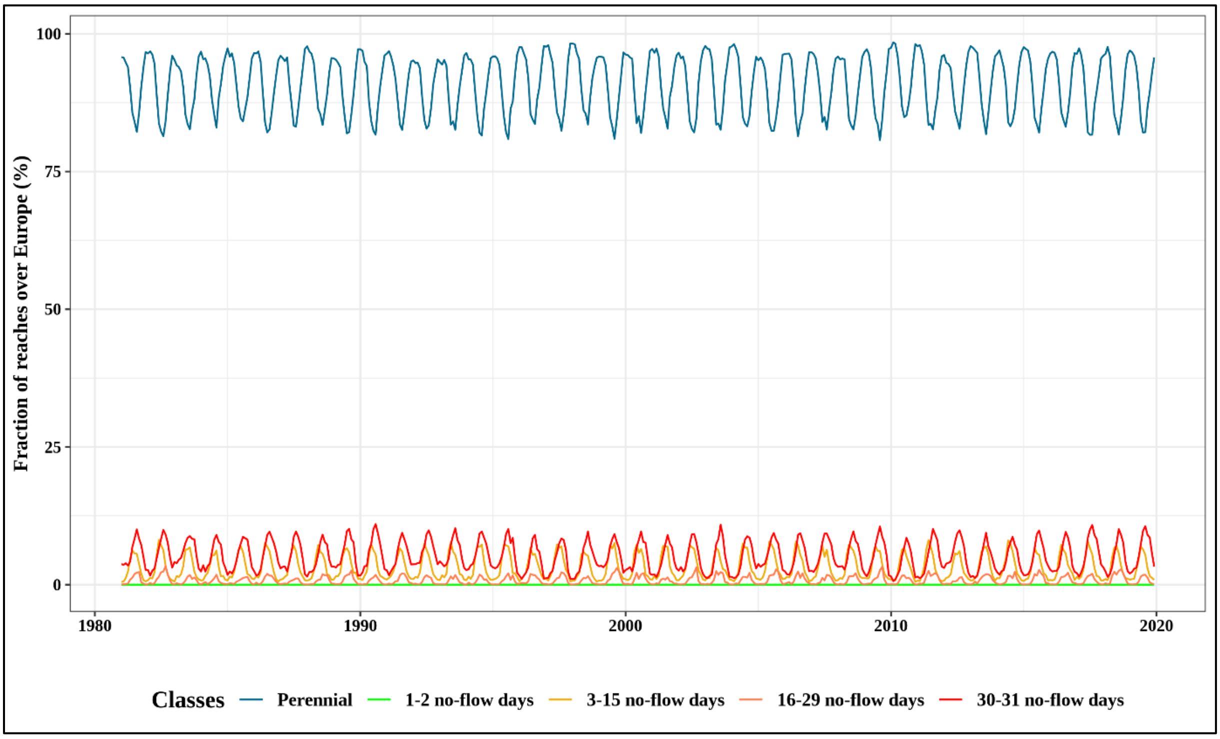

Figure 37. Monthly time series of the percent of all European reaches in the five intermittence classes 1981-2019 (5 colored lines)

The seasonality of intermittence in Europe can also be seen on the intermittence maps (with five classes) for January 2019 and July 2019 (Figure 38). Still, the river reaches in large areas in Spain are simulated to be totally dry even in January 2019. In July 2019, reaches in a very large part of the Iberian peninsula and parts of Southern and Southeastern Europe are simulated to be totally dry, while even in Northern Europe (Finland, South Sweden) many reaches with 3-15 no-flow days are simulated.

As expected from general principles and the RF model, the number of no-flow days per month decreases with increasing upstream drainage area (Figure 39 and Table 6; Table 6 shows the same values as those that are depicted in Figure 39). However, this tendency is clearer in the observations at the gauging stations than in the simulated reaches. The reason for the lower number of no-flow days for the reaches with an upstream drainage area below 2 km^2 (as compared to the next higher size class) is due to the definition of reaches (Section 3.5.1). Reaches with drainage areas below 2 km^2 were only generated if the mean annual streamflow exceeded a certain threshold, such that the smallest class only contains reaches in relatively wet areas. The two drainage area size classes 2-5 km^2 and 5-10 km^2, which make up 46% of all reaches in Europe (Figure 39), have approximately the same intermittence class frequencies. About 6-7% of the reaches are in class 4 (30-31 no-flow days) and roughly the same percentage in class 2 (3-15 no-flow days). About 1.7% are in class 3 (16-29 no-flow days, while here, like in all size classes, the percentage of reaches with 1-2 no-flow days is negligible (Table 6). The frequency distribution is very different for all larger drainage area size classes, in particular regarding the occurrence of class 2, which drops to less 0.8% in the drainage area size class 10-50 km^2 and to less than 0.1% in all larger size classes. In the two size classes 10-50 km^2 and 50-500 km^2, which account for 36% of all reaches, the frequency of months with 30-31 no-flow days remains at the high level of 6-7%, while most of the other reaches are perennial (Table 6). Reaches with upstream areas of more than 500 km^2 are significantly less intermittent. Less than 0.3% of the reach-months are simulated to be intermittent if drainage areas exceed 2500 km^2.

Figure 39 also shows how different the distribution of upstream drainage areas is between the streamflow gauging stations used for deriving the RF model and the reaches for which we simulate intermittence. While more than half of the reaches have a drainage area of less than 10 km^2, this is only the case for 2.5% of the gauging stations. This discrepancy certainly decreases the reliability of simulated intermittence for reaches with small drainage areas.

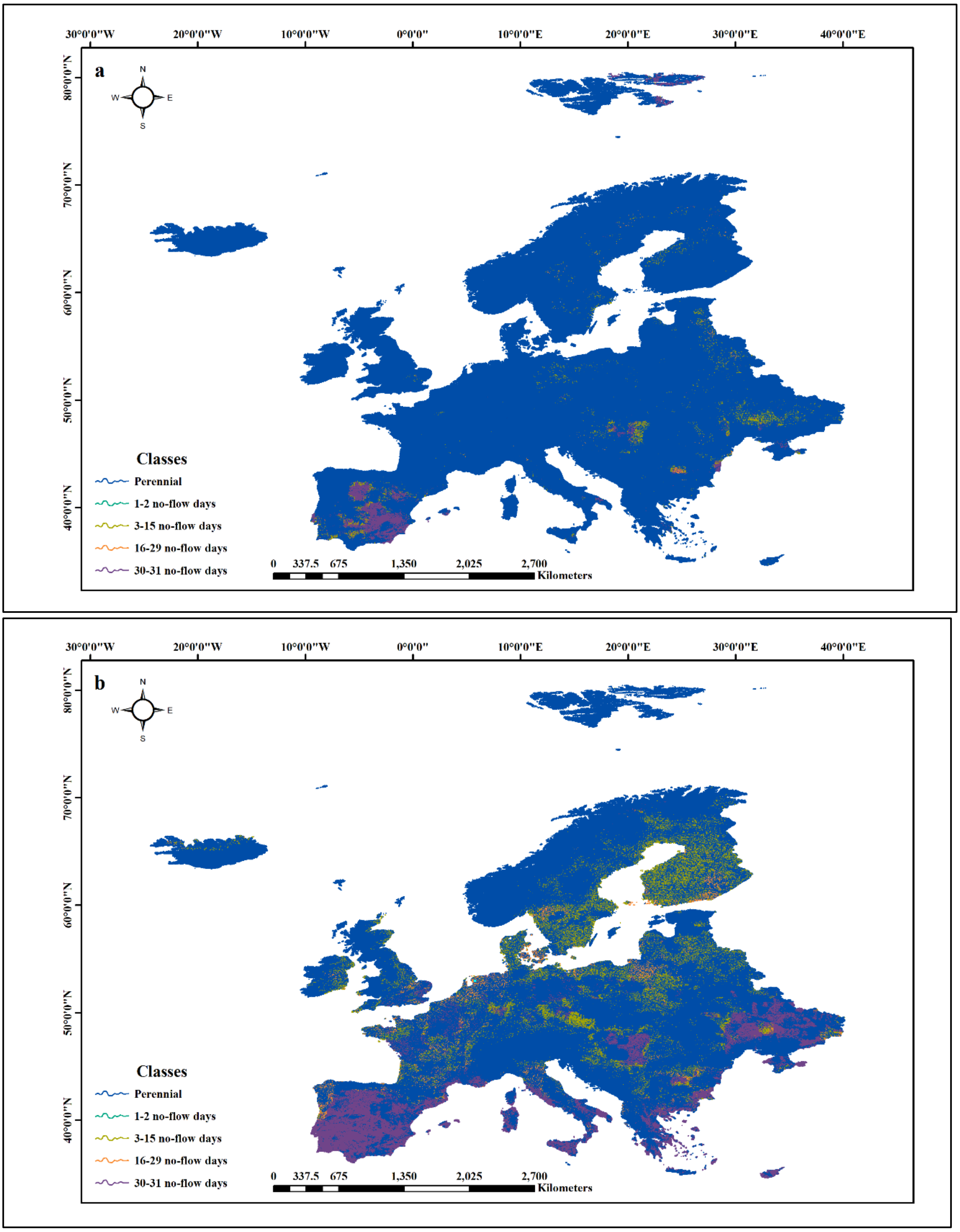

Figure 38. Number of no-flow days, in five classes, in January 2019 (a) and July 2019 (b) (maps)

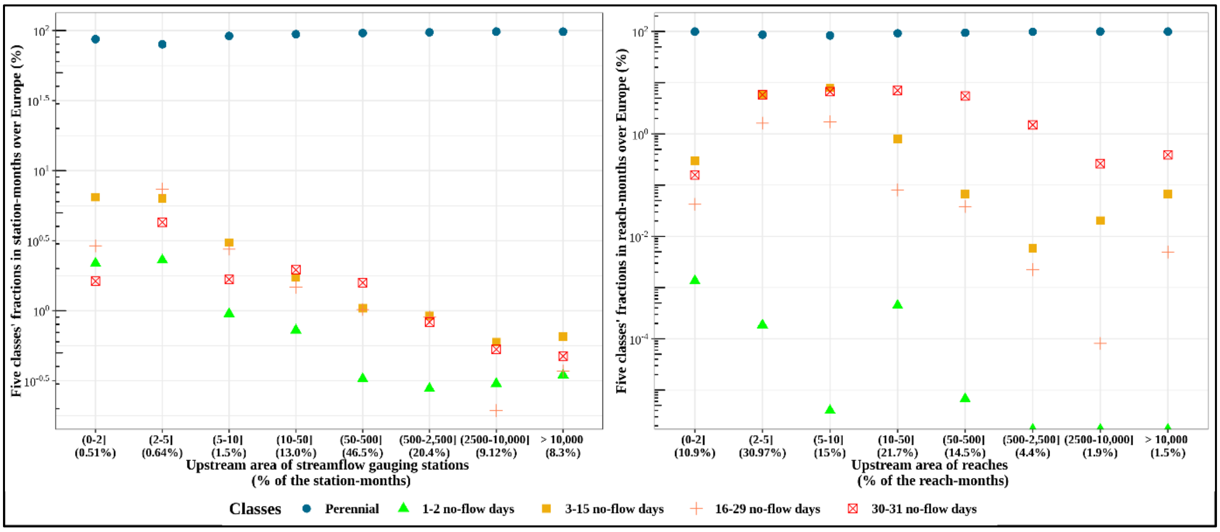

Figure 39. Percent of observed station-months (left) and reach-months in 1981-2019 (right) in the five classes as a function of upstream drainage area [km^2] of the streamflow gauging stations or the reach. Ccalibration and validation data are combined, and the observed fraction is shown. The percentage of the station-months or reach-months contained in each drainage area size class is provided below each size class. In total, 560,066 station-months and 717,664,428 reach-months are considered. Note that the y-axes for the stations and the reaches are scaled differently.

Table 6. *Percent of observed station-months (for both calibration and validation) and reach-months (1981-2019) in the five classes as a function of upstream drainage area [km^2] of the streamflow gauging stations or the reach. Classes: 0: perennial, 1: 1-2 no-flow days, 2: 3-15 no-flow days, 3: 16-29 no-flow days, 4: 30-31 no-flow day. Calibration and validation data are combined, and the observed fraction is shown. In total, 560,066 station-months and 717,664,428 reach-months are considered. This table contains the same values as Figure 39.*

Upstream area [km^2]	Percent of station-months in classes 0-4					Percent of reach-months in classes 0-4				
	0	1	2	3	4	0	1	2	3	4
(0-2]	86.8	2.18	6.47	2.91	1.63	99.5	0.001	0.30	0.04	0.16
(2-5]	79.7	2.30	6.35	7.36	4.28	86.6	0.0002	6.00	1.62	5.81
(5-10]	91.6	0.95	3.10	2.76	1.70	83.6	0.00	8.00	1.73	6.69
(10-50]	94.1	0.73	1.73	1.50	2.00	92.1	0.0005	0.81	0.08	7.05
(50-500]	96.0	0.33	1.00	1.00	1.60	94.4	0.00	0.07	0.04	5.50
(500-2500]	97.1	0.28	0.92	0.90	0.83	98.5	0.00	0.006	0.002	1.50
(2500-10000]	98.4	0.30	0.60	0.19	0.53	99.7	0.00	0.02	0.00	0.26
>10000	98.2	0.35	0.66	0.37	0.48	99.5	0.00	0.08	0.005	0.40

4 Conclusions

The quality of the 0.5 arc-deg (LR) runoff of the global hydrological model WaterGAP appears to be, at least for Europe, good enough to be spatially downscaled by a new downscaling algorithm to 15 arc-sec (HR) grid cells, deriving, for each 0.5 arc-deg grid cell, monthly streamflow time series for 14,400 grid cells. While HR streamflow can certainly not reflect HR anthropogenic interferences with streamflow including small reservoirs, weirs and water abstractions, impacts of reservoirs and water abstractions are taken into account at the LR by WaterGAP. Predictors derived from the HR streamflow were found to be important for estimating, by a two-step RF modeling approach, streamflow intermittence in Europe, compared to intermittence based on daily time series of observed streamflow. We found that it is possible to predict, for each gauging station, the number of no-flow days per month in five classes (0, 1-2, 3-15, 16-29, 30-31 days). However, the model tends to overestimate the number of no-flow days. While wet and dry years can be reasonably distinguished by the RF model, the monthly intermittence class can only be simulated reasonably for the calibration period. The RF modeling approach resulted in predictions of the monthly time series of five intermittence classes at more than 1.5 million reaches in Europe during the period 1981-2019. As the performance of the HR streamflow in South America is lower, and there does not exist a dedicated dataset of daily streamflow for intermittent streams, which could be collected in this study for Europe based on metadata from the SMIRES initiative, we expect a worse simulation of intermittence in South America than in Europe.

The spatial and temporal patterns of simulated streamflow intermittence in Europe are roughly plausible. Further comparisons with the intermittence results for the focal DRNs and discontinuous observations of no-flow days (e.g., from the datasets dryrivers, crowdwater and onde) could be used to validate the predictions and will be performed at a later time, after some model improvements. Model improvements may include 1) using streamflow data during the whole period 1981-2019 to set up the RF model, 2) adjusting predictor selection (e.g., slope and aridity index), 3) undersampling of months with 30-31 no-flow days in step 2 of the RF modeling and 4) choosing different intermittence classes. Model uncertainty is and will remain high, in particular for small reaches, due to the 0.5 arc-deg resolution of the WaterGAP model (including the impossibility to consider the impact of local river modifications such as weirs, small reservoirs and water abstractions) and the very low number of daily streamflow observations for small upstream reaches. In addition, the counterintuitive negative correlation of slope and the low frequency of intermittent station-months, both of which are due to the distribution of gauging stations, are expected to negatively impact the RF model results.

This study shows the value of simulating intermittence at a high spatial resolution instead of at the resolution of global hydrological models, where one grid cell covers approximately 2500 km^2. According to the simulation results, less than 0.3% of the reach-months are intermittent if the upstream drainage area of the reach is larger than 2500 km^2, while about 15% of the reaches with drainage areas below 10 km^2 are intermittent. The intermittence at scales below 2500 km^2 would remain undetected without the downscaling of 0.5 arc-deg WaterGAP model output.

5 Shared data

Three datasets are available on the DRYvER server (folder “continental-scale high-resolution modeling of streamflow intermittence2), the first two in the folder “subtask1.4.1_downscaling”, the third on in the folder “subtask1.4.2_statistical_modeling”.

1 Dataset “**Monthly time series of streamflow in the period 1981-2019 in 15 arc-sec grid cells, for Eurasia**”. Unit: m^3/s. Folder name: DS_flow_WaterGAP_eurasia

2 Dataset “**Monthly time series of streamflow in the period 1981-2019 in 15 arc-sec grid cells, for South America**”. Unit: m^3/s. Folder name: DS_flow_WaterGAP_SAmerica

Both datasets contain 468 NetCDF files, one per month for the period 1981-2019

3 Dataset “**Monthly time series of five intermittence classes 1981-2019 in reaches, for Europe without Russia and Turkey**”. Class values: 0, 1, 2, 3, 4.

0: no no-flow day in month

1: 1-2 no-flow days in month

2: 3-15 no-flow days in month

3: 16-29 no-flow days in month

4: 30-31 no-flow days in month

Folder name: intermittence_classes_reaches_Europe

The dataset is in the format of a geodatabase. In addition, the results for the six European focal DRNs are provided as six shapefiles.

References

Amit, Y., Geman, D., Wilder, K. (1997). Joint induction of shape features and tree classifiers. IEEE transactions on pattern analysis and machine intelligence, 19(11), 1300-1305.

Bond, N. R. & Kennard, M. J. (2017). Prediction of hydrologic characteristics for ungauged catchments to support hydroecological modeling. Water Resour. Res. 53, 8781–8794.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research. 16, 321-357.

Chen Z, Auler AS, Bakalowicz M, Drew D, Griger F, Hartmann J, Jiang G, Moosdorf N, Richts A, Stevanovic Z, Veni G, Goldscheider N. 2017. The World Karst Aquifer Mapping project: concept, mapping procedure and map of Europe. Hydrogeology Journal 25: 771–785.

Datry, T., Boulton, A.J., Bonada, N., Fritz, K., Leigh, C., Sauquet, E., Tockner, K., Hugueny, B., Dahm, C.N. (2018). Flow intermittence and ecosystem services in rivers of the Anthropocene. The Journal of Applied Ecology, 55, 353-364. https://doi.org/10.1111/1365-2664.12941.

Datry, T., Larned, S.T., Tockner, K. (2014). Intermittent rivers: A challenge for freshwater ecology. BioScience, 64(3), 229–235. https://doi.org/10.1093/biosci/bit027.

Do, H. X., Gudmundsson, L., Leonard, M., Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. Earth System Science Data, 10(2), 765–785. https://doi.org/10.5194/essd-10-765-2018

Döll, P., Douville, H., Güntner, A., Müller Schmied, H., Wada, Y. (2016). Modelling Freshwater Resources at the Global Scale: Challenges and Prospects. Surveys in Geophysics 37, 195–221. https://doi.org/10.1007/s10712-015-9343-1.

Döll, P., Müller Schmied, H. (2012). How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis. Environ. Res. Lett., 7 (1), 014037 (11pp). https://doi.org/10.1088/1748-9326/7/1/014037.

Döll, P., Lehner, B. (2002). Validation of a new global 30-min drainage direction map. *Journal of Hydrology*, *258*(1-4), 214–231. https://doi.org/10.1016/S0022-1694(01)00565-0.

GLIMS & NSIDC (2012). Global land ice measurements from space (GLIMS) glacier database, v1. National Snow and Ice Data Center (NSIDC), https://doi.org/10.7265/N5V98602.

Global Runoff Data Centre. In-situ river discharge data (World Meteorological Organization, accessed 15 May 2015); https://portal.grdc.bafg.de/applications/public.ht ml?publicuser=PublicUser#dataDownload/Home

Gudmundsson, L., Do, H. X., Leonard, M., Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. Earth System Science Data, 10(2), 787–804. https://doi.org/10.5194/essd-10-787-2018.

Hastie, T., Tibshirani, R., Friedman, J. (2008). Random forest. In The Elements of Statistical Learning. https://doi.org/10.1007/b94608_15

Japkowicz, N., Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449. https://doi.org/10.3233/ida-2002-6504

Kling, H., Fuchs, M., Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. Journal of Hydrology, 424-425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Krabbenhoft, C.A. et al. (2022). Assessing placement bias of the global river gauge network. Nature Sustainability, 5, 586-592. https://doi.org/10.1038/s41893-022-00873-0

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Lehner, B., Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15), 2171–2186. https://doi.org/10.1002/hyp.9740

Lehner, B., Verdin, K., Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data. Eos, Transactions American Geophysical Union, 89(10), 93. https://doi.org/10.1029/2008eo100001

Louppe, G. (2014). Understanding random forests: From theory to practice. PhD dissertation, University of Liege, Belgium. arXiv preprint arXiv:1407.7502.

Messager, M. L., Lehner, B., Cockburn, C., Lamouroux, N., Pella, H., Snelder, T., et al. (2021). Global prevalence of non-perennial rivers and streams. Nature, 594(7863), 391–397. https://doi.org/10.1038/s41586-021-03565-5

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., & Niemann, C., et al. (2021). The global water resources and use model WaterGAP v2.2d: model description and evaluation. Geoscientific Model Development, 14(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C. (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resources Research, 42(7). https://doi.org/10.1029/2005WR004636

Ramankutty N, Foley JA. 1999. Estimating historical changes in global land cover: croplands from 1700 to 1992. Global Biogeochemical Cycles 13: 997–1027.

Robinson N, Regetz J, Guralnick RP. 2014. EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. ISPRS Journal of Photogrammetry and Remote Sensing 87: 57–67.Sauquet, E. (2020). Science and Management of Intermittent Rivers and Ephemeral Streams. A META-DATABASE OF AVAILABLE HYDROLOGICAL DATA FROM GAUGING STATIONS WITH ZERO-FLOW EVENTS IN THE PARTICIPATING COUNTRIES. Retrieved from https://www.smires.eu/wp-content/uploads/2020/02/D1-Metadata.pdf

Snelder, T. H. et al. (2013). Regionalization of patterns of flow intermittence from gauging station records. Hydrol. Earth Syst. Sci. 17, 2685–2699.

Tyralis, H., Papacharalampous, G., Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water (Switzerland), 11(5). https://doi.org/10.3390/w11050910

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. 10.18637/jss.v077.i01.

Zaherpour, J., et al. (2018). Worldwide Evaluation of Mean and Extreme Runoff from Six Global-Scale Hydrological Models That Account for Human Impacts. Environmental Research Letters 13, no. 6 (June 12, 2018): 065015. https://doi.org/10.1088/1748-9326/aac547.Zomer RJ, Xu J, Trabucco A. 2022. Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. Scientific Data 2022 9:1 9: 1–15.

Appendix

Table 1-A: Gauging stations in DRNS and performance of downscaled streamflow in comparison to observations

DRN	Station ID	Longitude	Lattitude	Upstream area (HydroSHEDS)	Upstream area (Station data)	NSE	logNSE	KGE	R	GAMMA	BETA	number of months compared
Croatia	7005	16.1869	44.0410	330.3	NA	-0.03	-0.18	0.59	0.69	1.23	1.12	228
Croatia	7242	16.0498	43.8381	438.9	NA	0.28	-0.22	0.07	0.85	0.65	1.85	207
Czechrepublic	367000	17.2609	49.5773	3099.7	3323.6	0.22	0.59	0.66	0.81	1.09	1.27	468
Czechrepublic	341000	16.9013	50.0884	104.1	96.5	0.18	-0.60	0.48	0.70	1.08	0.58	420
Czechrepublic	421800	17.5190	48.8776	68.8	65.8	0.60	0.61	0.61	0.79	0.67	0.94	359
Czechrepublic	413000	17.5013	49.1332	7920.2	7890.3	0.21	0.56	0.58	0.86	1.00	1.39	470
Czechrepublic	345000	16.9114	50.0409	351.7	349.8	0.25	-0.52	0.53	0.72	1.04	0.62	450
Czechrepublic	426000	16.9886	48.6870	9916.7	9721.8	-0.07	0.51	0.48	0.87	1.01	1.51	254
Czechrepublic	403000	17.3987	49.3023	7013.6	7013.3	0.31	0.60	0.63	0.86	1.01	1.34	469
Finland	2100300	25.0030	60.6281	67.5	65.0	0.06	-0.91	0.40	0.75	1.21	0.50	215
Finland	2101700	24.9843	60.2378	1678.5	1667.0	0.64	0.56	0.68	0.85	0.79	0.81	942
Finland	2101520	25.0839	60.3174	332.0	348.0	0.73	0.51	0.74	0.89	0.90	0.79	612
Finland	2101220	24.8651	60.2963	1209.6	1311.0	0.68	0.65	0.69	0.88	0.78	0.82	689
Finland	2104900	24.7589	60.3322	206.3	208.0	0.75	0.64	0.80	0.89	0.95	0.85	123
France	V2114010	5.8689	46.7562	106.5	117.0	0.73	0.57	0.63	0.92	0.79	1.30	527
France	V2945210	5.1968	45.9395	33.0	33.0	-0.27	0.26	0.49	0.76	1.01	1.45	165
France	V2035010	5.9578	46.6237	44.0	46.0	-0.48	0.26	0.32	0.85	0.93	1.66	727
France	V2814030	5.3271	46.0642	343.6	349.0	0.69	0.51	0.57	0.95	0.81	1.38	564
France	V2420560	5.8732	46.3781	148.0	164.0	0.40	0.24	0.58	0.80	1.15	0.67	276
France	V2012010	5.9558	46.7239	193.7	210.0	0.70	0.61	0.74	0.92	1.05	0.76	470
France	V2924010	5.4393	45.9482	243.4	232.0	0.83	0.81	0.88	0.91	0.92	0.97	592
France	V2322010	5.6656	46.3975	1075.2	1120.0	0.85	0.81	0.89	0.93	1.00	0.91	970

France	V2414030	6.0160	46.5311	86.3	85.0	0.68	0.64	0.81	0.84	1.04	0.91	262
France	V2414010	5.8678	46.4165	212.6	216.0	0.56	0.59	0.68	0.84	0.96	0.73	629
France	V2444020	5.7062	46.3627	573.6	650.0	0.46	0.45	0.62	0.87	1.04	0.64	582
France	V2942010	5.2340	45.9064	3507.3	3630.0	0.80	0.68	0.86	0.90	1.09	0.94	730
France	V2712010	5.3364	46.0471	2644.7	2760.0	0.74	0.60	0.80	0.88	1.10	0.88	729
France	V2505020	5.5405	46.1363	91.4	92.0	0.79	0.81	0.81	0.92	0.96	0.84	372
France	V2814040	5.4464	46.3019	200.2	193.0	0.72	0.63	0.66	0.95	0.87	1.31	467
France	V2814020	5.3904	46.1120	319.2	324.0	-0.38	-0.11	-0.23	0.95	0.65	2.18	618
France	V2206010	5.7753	46.6481	48.4	49.0	0.83	0.79	0.91	0.91	1.01	0.98	747
France	V2030410	5.9402	46.7040	182.2	196.0	0.58	0.56	0.68	0.87	1.01	0.71	217
France	V2202010	5.7667	46.6858	670.8	650.0	0.82	0.80	0.85	0.92	1.01	0.87	661
Hungary	7	18.1240	46.0380	170.1	162.0	-0.14	-0.89	0.37	0.54	1.37	0.79	166
Hungary	4	17.7330	45.9820	412.7	417.0	-0.39	0.06	0.42	0.66	0.93	1.47	758
Hungary	5	17.9280	45.9900	193.8	190.0	0.21	0.11	0.34	0.72	0.68	1.50	295
Hungary	17	17.9770	46.0430	116.3	112.0	0.37	0.32	0.56	0.65	0.76	1.14	520
Hungary	1	18.0815	45.8180	1184.7	11.4	0.14	0.15	0.49	0.61	0.79	1.27	642
Hungary	6	17.8050	46.0820	164.4	162.0	-0.72	-0.33	0.05	0.35	0.67	1.61	384
Hungary	3	18.0940	45.8260	580.8	593.0	-0.27	-0.05	0.52	0.58	1.20	1.12	733
Spain	HS11	-5.4074	36.4704	599.3	NA	0.48	0.44	0.59	0.72	1.01	0.70	254
Spain	HS9	-5.4520	36.4282	230.3	245.0	0.77	0.42	0.62	0.89	0.76	1.28	241
Spain	HS103	-5.3384	36.5785	461.0	NA	-2.20	0.11	-0.55	0.84	0.91	2.54	139
Spain	HS13	-5.2467	36.5674	160.9	NA	0.51	0.22	0.74	0.81	1.02	1.18	243

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 869226

This document reflects only the author's view. The Agency and the Commission
are not responsible for any use that may be made of the information it contains.