Streamflow intermittence in Europe: Estimating high-resolution monthly time series by downscaling of simulated runoff and Random Forest modeling

Petra Doell¹, Mahdi Abbasi¹, Mathis Loic Messager², Tim Trautmann¹, Bernhard Lehner², and Nicolas Lamouroux³

¹Goethe University Frankfurt
²McGill University
³INRAE

December 21, 2023

Abstract

Knowing where and when rivers cease to flow provides an important basis for evaluating riverine biodiversity, biogeochemistry and ecosystem services. We present a novel modeling approach to estimate monthly time series of streamflow intermittence at high spatial resolution at the continental scale. Streamflow intermittence is quantified at more than 1.5 million river reaches in Europe as the number of no-flow days grouped into five classes (0, 1-5, 6-15, 16-29, 30-31 no-flow days) for each month from 1981 to 2019. Daily time series of observed streamflow at 3706 gauging stations were used to train and validate a two-step Random Forest modeling approach. Important predictors were derived from time series of monthly streamflow at 73 million 15 arc-sec (~500 m) grid cells that were computed by downscaling the 0.5 arc-deg (~55 km) output of the global hydrological model WaterGAP, which accounts for human water use. Of the observed perennial and intermittent station-months, 97.8% and 86.4%, respectively, are correctly predicted. Interannual variations of the number of intermittent months at intermittent reaches are satisfactorily simulated, with a median Pearson correlation of 0.5. While the spatial prevalence of intermittent reaches is underestimated, the number of intermittent months is overestimated in dry regions of Europe where artificial storage abounds. Our model estimates that 3.8% of all European reach-months and 17.2% of all reaches were intermittent during 1981-2019, predominantly with 30-31 no-flow days. Although estimation uncertainty is high, our study provides, for the first time, information on the continent-wide dynamics of intermittent rivers and streams.
Streamflow intermittence in Europe: Estimating high-resolution monthly time series by downscaling of simulated runoff and Random Forest modeling

Petra Döll¹,²*, Mahdi Abbasi¹*, Mathis Loïc Messager³,⁴, Tim Trautmann¹, Bernhard Lehner⁴, Nicolas Lamouroux³

¹Institute of Physical Geography, Goethe University Frankfurt, Frankfurt/Main, Germany
²Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F) Frankfurt, Frankfurt/Main, Germany
³INRAE, UR RiverLy, Lyon-Villeurbanne, France
⁴Department of Geography, McGill University, Montreal, Canada

Corresponding author: Petra Döll (p.doell@em.uni-frankfurt.de)

*Equal contribution

Key Points:

- Streamflow intermittence at more than 1.5 million European reaches was estimated for every month during 1981-2019.
- 18.7% of the European river network length and 3.8% of all reach-months are intermittent, predominantly with 30-31 no-flow days.
- 15 arc-sec monthly streamflow obtained by downscaling the output of a global hydrological model serves as input to Random Forest modeling.
Abstract
Knowing where and when rivers cease to flow provides an important basis for evaluating riverine biodiversity, biogeochemistry and ecosystem services. We present a novel modeling approach to estimate monthly time series of streamflow intermittence at high spatial resolution at the continental scale. Streamflow intermittence is quantified at more than 1.5 million river reaches in Europe as the number of no-flow days grouped into five classes (0, 1-5, 6-15, 16-29, 30-31 no-flow days) for each month from 1981 to 2019. Daily time series of observed streamflow at 3706 gauging stations were used to train and validate a two-step Random Forest modeling approach. Important predictors were derived from time series of monthly streamflow at 73 million 15 arc-sec (~500 m) grid cells that were computed by downscaling the 0.5 arc-deg (~55 km) output of the global hydrological model WaterGAP, which accounts for human water use. Of the observed perennial and intermittent station-months, 97.8% and 86.4%, respectively, are correctly predicted. Interannual variations of the number of intermittent months at intermittent reaches are satisfactorily simulated, with a median Pearson correlation of 0.5. While the spatial prevalence of intermittent reaches is underestimated, the number of intermittent months is overestimated in dry regions of Europe where artificial storage abounds. Our model estimates that 3.8% of all European reach-months and 17.2% of all reaches were intermittent during 1981-2019, predominantly with 30-31 no-flow days. Although estimation uncertainty is high, our study provides, for the first time, information on the continent-wide dynamics of intermittent rivers and streams.

Plain Language Summary
Even in wet climates, small streams can seasonally dry up. In drier areas, large rivers might not carry water for weeks or months. However, as streamflow observations are lacking for most drying rivers, we know little about when, where, and how long rivers experience such a streamflow intermittence that is crucial for both river life and human water supply. We developed and applied a novel approach to estimate, for the first time, the temporal dynamics of streamflow intermittence across European rivers and streams, including small ones. This approach combines the output of a global hydrological model with streamflow observations and other data. We refined the global model output available for 50 km cells to monthly streamflow in 500 m cells. We then applied a machine learning model to predict the number of days without water flow in each month during the period 1981-2019 for over 1.5 million river segments. We found that 17% of all European segments and 4% of all months at all segments experienced at least one day without flow. In the future, the model will be used to estimate the impact of climate change on streamflow intermittence.

1. Introduction
It has recently been estimated that most rivers and streams on Earth have reaches that naturally cease to flow or dry at least one day per year (Messager et al., 2021). Natural streamflow intermittence is most prevalent in semi-arid and arid regions, where it may occur even in large rivers, but it is also widespread in smaller headwater streams across humid regions. For example, 25–40% of the total length of streams and rivers in France are estimated to be intermittent (Snelder et al., 2013). In most basins, the likelihood and degree of streamflow intermittence, i.e., the fraction of no-flow days, increases with decreasing mean streamflow or upstream area (Datry et al., 2014;
Waterways can also cease to flow without being dry due to freezing conditions, so intermittent streams are a significant feature of cold landscapes as well (Buttle et al., 2012; Shanafield et al., 2021). Anthropogenic alterations of the natural flow regime resulting, for example, from human water abstractions or the operation of artificial reservoirs, can increase or decrease the number of no-flow days (Richter, 1997).

While streamflow intermittence can be monitored by measuring streamflow at gauging stations, these measurements come with numerous limitations (Zimmer et al., 2020) and only cover a very small part of all reaches, being particularly sparse where intermittent conditions prevail (Sauquet et al., 2021b; Krabbenhoft et al., 2022). In addition, streamflow observations are insufficient to derive projections of future changes in intermittence due to anthropogenic drivers, including climate change as well as artificial reservoirs and land and water use (Sauquet et al., 2021a; Döll & Müller Schmied, 2012). Therefore, comprehensive analyses of streamflow intermittence and its effects on water resources for humans and other biota require a modeling approach.

Large-scale modeling of streamflow intermittence is necessary for assessments of biodiversity, ecosystem functions and ecosystem services of rivers and streams at national to global scales. Until now, however, continental- or global-scale modeling studies on streamflow intermittence have either provided a static classification of river reaches into intermittent or perennial at high spatial resolution (15 arc-sec, ca. 500 m; Messager et al., 2021) or time series of intermittent streamflow conditions at a low spatial resolution (0.5 arc-deg, ca. 50 km; Döll & Müller Schmied, 2012). Messager et al. (2021) used Random Forest modeling to estimate which river reaches cease to flow at least one day per year or for at least thirty days per year; this was achieved for 23.3 million km of mapped rivers and streams across the globe (except Antarctica) whose long-term average naturalized discharge exceeds 0.1 m³/s. Despite its fine resolution, such a static classification of reaches as either perennial or intermittent fails to characterize the temporal structure of flow intermittence (e.g., the number of no-flow days or seasonality of intermittence) which is required for analyzing the biodiversity and ecosystem functions of intermittent streams and rivers (Datry et al., 2018). By contrast, daily streamflow time series simulated by global hydrological models such as the WaterGAP model used in Döll and Müller Schmied (2012) do represent the temporal dynamics of streamflow intermittence. However, these coarser models overlook headwater stream reaches with small drainage basins, which are more prone to intermittence than larger downstream reaches and comprise the majority of global river length (Messager et al., 2021).

Simulating daily streamflow in small headwater streams requires small computational grid cell sizes (e.g., 500 m or less). Such small grid cells can easily be implemented in hydrological models if the drainage basin of study is small (Mahoney et al., 2023). However, this is not feasible across large geographic extents like entire continents or the world due to the lack of high-resolution climate data at these scales and computational constraints resulting from the large number of small (high-resolution) grid cells (Bierkens et al., 2015, Döll et al., 2016). A 0.5 arc-deg grid cell, typical for global hydrological models, contains 14,400 individual 15 arc-sec grid cells; in Europe alone (without Russia and Turkey), for example, there are about 73 million 15 arc-sec cells. Furthermore, hydrological models are often less successful in simulating low flows than mean flows (Zaherpour et al., 2018). Most hydrological models are process-based, i.e., they attempt to estimate water storage and fluxes across the different compartments of the terrestrial part of the hydrological cycle with sets of mathematical equations (Telteu et al., 2021). However, a satisfactory process-based simulation of low-flow, and particularly no-flow conditions, is very difficult even at small scales.
in part because the simulation of two-way exchange flows between surface water bodies and groundwater bodies requires coupling of a hydrological model with a gradient-based groundwater model (Döll et al., 2016). To help advance the science and management of freshwater ecosystems globally, new approaches are thus needed to produce large-scale high-resolution models of streamflow intermittence that provide information on the frequency, duration and timing of flow cessation across the entire river network, from the headwaters to river mouths.

Machine learning methods such as Random Forest (RF) have the advantage over process-based models that they do not require detailed knowledge of the processes underlying the phenomenon of interest and are thus a promising tool to produce large-scale high-resolution predictions of no-flow conditions. However, to achieve temporally explicit predictions, these models require temporally explicit predictors. The respective strengths of global hydrological models and machine learning methods can hence be combined by using the dynamic output of the former as an input predictor for the latter to achieve large-scale high-resolution modeling of the temporal structure of streamflow intermittence.

Here, we present such a combined modeling approach for computing monthly time series of streamflow intermittence conditions at the continental scale for river reaches that can be defined with a spatial resolution of 15 arc-sec. Our RF modeling approach combines temporally explicit predictor variables derived from the low resolution (LR, 0.5 arc-deg) state-of-the-art global hydrological model WaterGAP 2.2e (Müller Schmied et al., 2021) with several high-resolution (HR, 15 arc-sec) static predictor variables (e.g., drainage area and irrigated area). As part of this approach, WaterGAP LR output is spatially downscaled to derive HR monthly time series of streamflow. While all predictors used in the model are based on globally available data, the approach was developed using time series of daily streamflow observed at 3706 gauging stations throughout Europe (resulting in more than 1 million station-month with information on the number of no-flow days). It was then applied to estimate streamflow intermittence in Europe.

Section 2 presents the data and methods of this study. In section 3, the downscaled HR monthly streamflow time series are compared to observations at all gauging stations that were used to set up and calibrate the RF model. RF model performance and results of the RF application are presented in section 4. Section 5 provides validation and discussion of the streamflow intermittence modeling approach, while conclusions are drawn in section 6.

2. Methods and data

Below, we first explain the downscaling method applied to derive HR time series of streamflow from the LR output of the global hydrological model WaterGAP (Section 2.1). We then describe the compilation of the dataset of observed daily streamflow in Europe that was used for both validating the HR streamflow and for deriving the target data of the RF modeling approach (Section 2.2). This is followed by the description of the RF modeling approach, which consists of two sequential RF models (Section 2.3) and the definition of European river reaches for which model predictions are made (Section 2.4). In Section 2.5, performance metrics are explained.

The hydrographic dataset applied throughout this study is the global HR drainage direction map of HydroSHEDS v1 (Lehner et al., 2008; www.hydrosheds.org). This dataset represents, for each 15 arc-sec grid cell on land, the direction in which water would flow from that cell to its neighboring cells given topography. It serves to downscale LR outputs from WaterGAP, to co-
register streamflow gauging stations, to delineate river reaches for which the RF model produces predictions and to quantify predictors that are aggregated over the upstream areas of stations and reaches. In this study for Europe, HydroSHEDS was modified in three drainage basins (each about 200 km²) in Finland, Hungary and Croatia due to their use as case study basins in the related DRYvER project (see Döll et al., 2023a and https://www.dryver.eu/about/case-studies.

2.1. Downscaling of LR WaterGAP output to obtain time series of monthly HR streamflow

2.1.1. WaterGAP

WaterGAP is a global water resources and use model that covers all continents except Antarctica (see Müller Schmied et al., 2021, for details). It computes time series of water use for irrigation, livestock, manufacturing, cooling of thermal power plants and households, distinguishing groundwater and surface water sources. It also simulates water flows (e.g., evapotranspiration and runoff) and water storages (e.g., in soil and groundwater), taking into account the impact of net abstractions from groundwater and surface water bodies as well as of artificial reservoirs.

However, only operation of the globally largest 1109 artificial reservoirs (including regulated lakes) is simulated explicitly in WaterGAP, while smaller reservoirs only add to the fraction of each LR cell that is made up by of so-called local lakes, thus affecting evapotranspiration and flow dynamics in a very coarse way only. In WaterGAP, daily water flows and storages of 10 storage compartments are simulated in each LR grid cell. Total runoff from land is partitioned into fast (surface) runoff and groundwater recharge. Surface runoff from within a grid cell reaches surface water bodies (wetlands, lakes, reservoirs and rivers) on the same day, while groundwater recharge flows from the soil into the groundwater, which then releases groundwater discharge to surface water bodies as a function of groundwater storage. Only one river is assumed to exist within each LR grid cell, and the streamflow computed by WaterGAP refers to the outflow from the LR grid cell to the next downstream grid cell, which is prescribed by the LR drainage direction map DDM30 (Döll & Lehner, 2002). Groundwater discharge to surface water bodies may become zero in case of groundwater depletion, but loss of streamflow to the groundwater cannot be simulated. The LR WaterGAP output used in this study was computed by forcing version 2.2e of WaterGAP with the climate data set GSWP3-W5E5 (Müller Schmied et al., 2023a) for the time period 1901-2019. The model was calibrated against long-term mean annual streamflow observations observed at 1509 gauging stations globally (with a drainage area of at least 9,000 km²) by adjusting 1-3 model parameters.

2.1.2. Downscaling approach

A number of approaches for generating time series of high-resolution streamflow from the output of global hydrological models were recently developed (Lin et al.; 2019, Kallio et al., 2021; Chuphal and Mishra, 2023). Our approach for downscaling the LR output of a global hydrological model to HR streamflow is based on the conceptual framework developed by Lehner and Grill (2013) which was globally applied, for example, in Linke et al. (2019). In this study, we generalized and adapted the approach, including some simplifications, to enable a computationally efficient generation of HR time series of monthly streamflow. As a distinct feature, the downscaling approach does not simply disaggregate and then route the sum of LR surface runoff and groundwater recharge (i.e., total runoff from land) along the HR river network, as this would disregard water retention in the groundwater and in surface water bodies, evaporation from surface water bodies and as well as human water use. Instead, our approach uses both surface runoff and...
groundwater discharge estimates from the LR model and projects the results onto the HR river network using geospatial interpolation methods. Considering the original LR groundwater discharge estimates allows for better representation of HR streamflow variability because it takes into account the storage capacity of surface water bodies and groundwater aquifers that smoothen and delay the streamflow signal. Further corrections take into account the LR net cell runoff of WaterGAP, which, in addition to surface runoff and groundwater discharge, takes into account the dynamics of surface water bodies and human water use.

Here, we only describe the core elements of the downscaling method; for details see Supplementary Information Text S1. The sum of LR monthly surface runoff and groundwater discharge (expressed as specific volume flow per unit area, i.e., m$^3$ s$^{-1}$ km$^{-2}$) is first interpolated from 0.5 arc-deg to an intermediate resolution of 0.1 arc-deg to avoid abrupt changes in streamflow at the edges between LR cells. This is performed using an inverse distance interpolation with a power of 2 and taking into account the nearest 9 LR data points. A maximum interpolation radius of 1.8 arc-deg is allowed to extend data into areas where land cells are represented in the HR hydrography but not in the LR river network. This is the case in coastal regions and in missing cells within large lakes of the LR model. In the next step, the 0.1 arc-deg values are disaggregated to the 15 arc-sec HR grid cells by assigning the same 0.1 arc-deg value to all respective 15 arc-sec cells, assigning null values to HR cells outside of the continental boundaries of HydroSHEDS.

This disaggregated “runoff” $dR_{i',Hj'}$ of HR cell $j$ located within LR cell $i$, expressed as volume flow, is then corrected to integrate information from the routing routine of the LR model, in particular about the impact of surface water bodies and human water use on streamflow. Finally, HR streamflow $Q_{i',Hj'}$ in each month is computed by accumulating the sum of $dR_{i',Hj'}$ and a weighted correction term along the HR drainage direction, with

$$Q_{i',Hj'} = \text{flowacc}(dR_{i',Hj'} + C_{i',Hj'} \times W_{i',Hj'})$$  \hspace{1cm} (1)

$$C_{i} = \text{ncR}_{i} - \sum_{j=1}^{14400} dR_{i,Hj}$$  \hspace{1cm} (2)

$$W_{i,Hj} = \frac{\text{flowacc}(dR_{i,Hj})}{\sum_{j=1}^{14400} \text{flowacc}(dR_{i,Hj})}$$  \hspace{1cm} (3)

where $C_{i}$ is correction term for each LR cell $i$ (Equation 2), $W_{i,Hj}$ is the correction weight to apply the correction term for each HR cell $j$ within LR cell $i$ (Equation 3), flowacc() represents the flow-accumulated variable computed by summing the values of all HR grid cells ($L_{i';Hj'}$) upstream of and including cell $L_{i,Hj}$. The net cell runoff of LR grid cell $i$, $\text{ncR}_{i}$, is calculated as the streamflow of LR grid cell $i$ minus the streamflow of all upstream LR grid cells, and is corrected for the different land masks (i.e., ocean-continent boundaries) of WaterGAP and HydroSHEDS.

The GHM WaterGAP computes streamflow not only by routing surface runoff and groundwater discharge. It additionally considers the impacts of reservoirs, lakes and wetlands as well as human abstractions of groundwater and surface water within each LR grid cell. These impacts are included in $\text{ncR}_{i}$. This is why the disaggregated runoff is corrected by $C_{i}$, i.e., the difference between the LR $\text{ncR}_{i}$ and the sum of HR $dR_{i,Hj}$ within the LR cell (Equation 2). For example, if human water use leads to a streamflow reduction in the LR cell, net cell runoff is lower than the sum of surface runoff and groundwater discharge ($\text{ncR}_{i} < \sum_{j=1}^{14400} dR_{i,Hj}$), such that $C_{i}$ becomes negative and HR streamflow is reduced, too.
Further correction terms are added to $C_L^i$ as computed by Equation 2 in specific grid cells, such as HR endorheic sinks, outflow cells of lakes and reservoirs, or cells containing large rivers with a drainage area of more than 50,000 km² (see Supporting Information Text S1). The final correction term is applied in a spatially weighted way to the HR grid cells. With the help of the correction weight $W_{L,H,i}$, a greater share of the total correction occurs in the downstream HR grid cells within each LR cell, which reflects the assumption that downstream HR cells are more affected by surface water bodies and human water use than upstream cells within the LR. The downscaling approach was implemented in Python.

2.2. Compilation and processing of measured streamflow for the computation of target observations and for the validation of simulated HR streamflow

Long-term historical information on the number of no-flow days per month in waterways, the target for the RF modeling, can only be derived consistently from continuous daily observations of streamflow at gauging stations. We collected most of these observations from the Global Runoff Data Centre (GRDC; https://www.bafg.de/GRDC) and the Global Streamflow Indices and Metadata archive (GSIM; Do et al., 2018; Gudmundsson et al., 2018), the largest existing global repositories of streamflow gauging station data. Altogether, daily streamflow records for 2930 GRDC and GSIM stations are available through these datasets for Europe. However, most of the GRDC and GSIM stations are on perennial streams, without any no-flow days in their record, which reflects the global underrepresentation of streamflow gauging stations on intermittent river reaches (Krabbenhoft et al., 2022). Therefore, we used metadata on gauging stations with flow intermittence in 19 European countries from the SMIRES meta-database (Sauquet, 2020) to obtain daily streamflow time series directly from national streamflow data providers for 375 additional gauging stations listed in the database. As flow intermittence in Europe is most prevalent in Mediterranean regions, we additionally retrieved daily streamflow data for a total of 55, 648 and 1031 gauging stations from governmental websites for Corsica (https://www.sandre.eaufrance.fr/), Italy (http://meteoniardo.altervista.org/) and Spain (https://ceh.cedex.es/anuarioaforos/demarcaciones.asp), respectively.

From this compiled streamflow dataset, records suitable for deriving target observations were selected for subsequent analyses. We first checked whether each gauging station was correctly located on the updated 15 arc-sec HydroSHEDS drainage direction map by comparing the upstream area given in the metadata with the upstream area of the HR cell where the station was located. Confirmatory checks also involved inspecting high-resolution satellite imagery and comparing the river and station names provided in the metadata to topographic maps (ESRI ArcGIS basemaps). If the drainage areas deviated by more than 10%, the stations were manually relocated to a suitable HR grid cell with a deviation of less than 10% and/or associated to a river or stream with the correct name in topographic maps (if provided in the metadata). If this was not possible, the station was excluded from the RF modeling. For the remaining stations, we excluded all station-months with any missing or suspicious daily flow values following the approach of Gudmundsson et al. (2018). We then excluded all stations that had less than 36 station-months of daily streamflow data. Finally, we labeled all days with a mean streamflow of 0.001 m³ s⁻¹ or less as no-flow days and computed, as the target of the RF modeling, the number of no-flow days per month and station (i.e., per station-month). The maximum period with observed no-flow days and streamflow per station-month is 1981-2019 (468 months).

In total, data on streamflow at 3706 stations during 1981-2019 were used for calibrating and
validating the RF models, corresponding to 1,166,944 station-months (26 years of useable data per
station on average). While 2.8% of the station-months were intermittent, 24% of the stations had
at least one no-flow day. In addition, daily streamflow values were aggregated to monthly values
for the same station-months to serve for the validation of the simulated HR streamflow (Section
2.1).

2.3. Random Forest modeling approach

2.3.1. Overview

The supervised machine learning method RF is well suited for both classification and regression
tasks (Breiman et al., 2001). RF modeling has already been used for hydrological classification
problems, i.e., for predicting classes of hydrological characteristics including intermittence
(global: Messager et al., 2021; Australia: Bond and Kennard, 2017; France: Snelder et al., 2013).
Tyralis et al. (2019) provide a review of RF methods with a focus on hydrological applications.

With less than 3% of all observed station-months in our European streamflow dataset being
intermittent (Section 2.2), the dataset of target observations used for training the model is highly
imbalanced, which can severely bias the resulting predictions (Japkowicz and Stephen, 2002). To
mitigate this problem, two RFs are set up sequentially in our modeling approach. The first RF is
developed to predict months with and without no-flow days (intermittent station-month and
perennial station-months, respectively) in a binary way. The second RF is trained only with data
for intermittent station-months to predict the number of no-flow days in four classes. The two
calibrated RFs were then successively applied to predict the occurrence of five intermittence
classes (0, 1-5, 6-15, 16-29 and 30-31 no-flow days per month) for each of the 468 months from
1981 to 2019 at more than 1.5 million river reaches in Europe (without Russia and Turkey, see
Section 2.4 for the definition of reaches).

In this study, we used a derivative of the standard RF algorithm for making probabilistic
predictions of class membership (Malley et al., 2012), which is included in the ‘ranger’ R package
(Wright & Ziegler, 2017) that we used for the RF modeling. The ‘ranger’ R package is a fast
implementation of RF suited for high-dimensional data (Tyralis et al., 2019). The two consecutive
RF models were trained and optimized by cross-validation, i.e., calibrated, by relating observations
of the number of no-flow days per station-month at streamflow gauging stations (target of RFs) to
23 predictors, 9 of them temporally-explicit (i.e., dynamic). For the RF training, each streamflow
station is assigned to a HR grid cell. For RF predictions, each European river reach is assigned to
the HR cell containing its downstream end. In the following two sections, the compilation and
processing of the predictor variables and the 2-step RF modeling approach are explained.

2.3.2. Predictors

A total of 23 predictor variables were used in both RFs (Table 1). We selected predictors based on
their potential causal influence on streamflow intermittence. In contrast to multiple regression
analysis, RF can leverage information from highly correlated predictors while producing unbiased
predictions (Tyralis et al., 2019). Five HR dynamic predictors were derived from the WaterGAP
HR monthly streamflow time series and indicate the streamflow conditions in each HR stream
reach (represented by its respective most downstream HR cell) in the current month and past 3 and
12 months. Prior to computing these predictors, streamflow was converted into area-specific
streamflow by dividing it by the drainage area of the HR cell (i.e., the area of the upstream drainage
basin). This is done because the spatial variation of streamflow is mainly due to the area that drains
to the point of the streamflow observation and streamflow intermittence is expected to depend on area-specific streamflow. In addition, drainage area of the HR cell was included as a static predictor (Table 1). Of the 2 LR dynamic predictors, one is the ratio of diffuse groundwater recharge to total runoff from land for each month as computed by WaterGAP; a higher ratio of delayed groundwater discharge is expected to decrease the likelihood of no-flow days. The other LR dynamic predictor is the average number of days with substantial precipitation (i.e., >2.5 mm/d) per month according to the WaterGAP climate forcing; a low number of days with substantial precipitation in a month may lead to an increased likelihood of no-flow days. For both LR dynamic predictors, the average value over the upstream basin of each HR grid cell was computed assuming that the values in all upstream HR cells are identical within a given LR cell. The 5 HR and 2 LR dynamic predictors vary between the 468 months of the study period. Three of the 16 static HR predictors vary with the calendar month, the two predictors that quantify the interannual variability of monthly streamflow and the aridity index, which is included as the long-term mean values for the 12 calendar months. Interannual variability was computed from the HR monthly time series of area-specific streamflow as either the standard deviation or the coefficient of variation of all streamflow values of each of the 12 calendar months for the period 1981-2019 (Table 1).

In addition to drainage area, the other 12 HR static predictors include the dominant potential natural and actual land cover class in the upstream basins and the average slope, glacier fraction and lake fraction in the upstream basin. These static HR predictors are selected from the set of globally important predictors from Messager et al. (2021). Additional static predictors include five suspected anthropogenic drivers of streamflow intermittence and two newly developed karst-related predictors derived from the World Karst Aquifer Map (WOKAM) of karstifiable areas (Table 1). In the case of the static predictors of karst status and degree of regulation, the value for the HR grid cell for which the number of no-flow days is to be predicted (target cell) is used as a predictor. In the case of the anthropogenic drivers irrigated area fraction and human population, two sets of predictor values are taken into for each: one set of values computed by aggregating over the (total) upstream basin and the other computed by aggregating over the immediate upstream basin, which only encompasses upstream HR grid cells that drain directly into the respective stream reach (see Linke et al., 2019, for additional descriptions of these spatial units).

To train the RF models, the values of these predictors were assembled for each of the 1,166,944 station-months for which daily streamflow observations are available, i.e., for the 3706 HR grid cells that contain a gauging station. For model application, the predictor values for each reach (i.e., for the most downstream HR grid cell of each reach) were computed to predict the occurrence of one of the five intermittence classes for each reach-month.
### Table 1

*Predictors Used in RF Modeling, with their Abbreviations, Units and Data Sources*

<table>
<thead>
<tr>
<th>Category</th>
<th>Predictor type</th>
<th>Predictor</th>
<th>Abbreviation (unit)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrology</td>
<td>Monthly area-specific</td>
<td>Monthly area-specific streamflow</td>
<td>Q (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Minimum monthly area-specific</td>
<td>Minimum monthly area-specific streamflow of the past 12 months</td>
<td>Q(<em>{\text{min}</em>\text{p12}}) (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Monthly time series HR</td>
<td>Mean monthly area-specific streamflow of the past 12 months</td>
<td>Q(<em>{\text{mean}</em>\text{p12}}) (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Minimum monthly area-specific</td>
<td>Minimum monthly area-specific streamflow of the past 3 months</td>
<td>Q(<em>{\text{min}</em>\text{p3}}) (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Mean monthly area-specific</td>
<td>Mean monthly area-specific streamflow of the past 3 months</td>
<td>Q(<em>{\text{mean}</em>\text{p3}}) (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Monthly time series LR</td>
<td>Ratio of diffuse groundwater recharge to runoff from land, mean over ub(^a)</td>
<td>gwr_to_runoff_ratio (-)</td>
<td>WaterGAP 2.2e</td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td>Number of wet days, mean over ub(^b)</td>
<td>wet_days (days mon(^{-1})/100)</td>
<td>WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Interannual variability of monthly area-specific streamflow, per calendar month, in terms of standard deviation</td>
<td>Interannual variability of monthly area-specific streamflow, per calendar month, in terms of standard deviation</td>
<td>Q(<em>{\text{iav}</em>\text{sd}}) (m(^3) s(^{-1}) km(^{-2}))</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Interannual variability of monthly area-specific streamflow, per calendar month, in terms of coefficient of variation</td>
<td>Interannual variability of monthly area-specific streamflow, per calendar month, in terms of coefficient of variation</td>
<td>Q(<em>{\text{iav}</em>\text{cv}}) (-)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Climate</td>
<td>Static HR</td>
<td>Aridity index (long-term average P/PET), per calendar month, mean over ub(^a)</td>
<td>P(<em>{\text{to}</em>\text{PET}}) ratio (1/10000)</td>
<td>Global-AL_PET_v3(^c)</td>
</tr>
<tr>
<td>Land cover</td>
<td>Potential natural vegetation classes (ranges: 1-15), spatial</td>
<td>Potential natural vegetation classes (ranges: 1-15), spatial</td>
<td>pot_nat_vegetation (-)</td>
<td>EarthStat(^d)</td>
</tr>
</tbody>
</table>
Land cover

- Land cover classes (ranges: 1-22), spatial majority in ub
- Land cover classes (ranges: 1-22), spatial majority in ub

Land cover

- Glacier area fraction in ub
- Glacier area fraction in ub

Physiography

- Drainage area (km²)
- Drainage area (km²)

Physiography

- Terrain slope, mean over ub
- Terrain slope, mean over ub

Geology

- Fraction of karst area in ub
- Fraction of karst area in ub

Geology

- Occurrence of karst (1 if karst, 0 if not) at HR grid cell
- Occurrence of karst (1 if karst, 0 if not) at HR grid cell

Anthropogenic drivers

- Fraction of area equipped for irrigation in ub
- Fraction of area equipped for irrigation in ub

- Fraction of area equipped for irrigation in iub
- Fraction of area equipped for irrigation in iub

Population density in ub

- Population density in ub
- Population density in ub

Population density in iub

- Population density in iub
- Population density in iub

Degree of regulation (total upstream artificial reservoir storage volume / annual streamflow volume) at HR grid cell

- Degree of regulation (total upstream artificial reservoir storage volume / annual streamflow volume) at HR grid cell
- Degree of regulation (total upstream artificial reservoir storage volume / annual streamflow volume) at HR grid cell

Lakes

- Fraction of lake area in ub
- Fraction of lake area in ub

Note: Area-specific streamflow is streamflow at the HR grid cell divided by upstream drainage area. The units are those for the data sets used as input to the RF modeling, in which the integer values were partly multiplied by 10, 100 or 1000 to increase the precision. ub: HR upstream basin. iub: HR immediate upstream drainage basin, refers to all the HR grid cells that drain directly into the respective stream reach. Zomer et al. (2022). Ramankutty and Foley (1999). Bartholomé and Belward (2005). GLIMS & NSIDC (2012). Lehner et al. (2008). Robinson et al. (2014). Chen et al. (2017). Siebert et al. (2015). Bondarenko et al. (2020). Lehner et al. (2011). Messager et al. (2016).
2.3.3. Two-step RF modeling approach

The first RF model in our approach (Figure 1, step 1) results in a binary classification of station-months as either intermittent or perennial, whereas the second RF model (step 2) was only applied to intermittent station-months and classified them into four ordinal intermittent classes: 1-5, 6-15, 16-29 and 30-31 no-flow days per month. We performed a classification into four classes based on a previous study with a two-step RF model (with less target observations and slightly different predictors) where the performance for six classes was not satisfactory (Döll et al., 2023a). The four classes were defined such that they are informative for biodiversity and ecosystem function studies while keeping the number of observations per class approximately balanced. Following model training and validation for each of the RFs (Figure 1, left-hand side), we sequentially applied the calibrated models (right-hand side) to predict monthly streamflow intermittence for all reaches in Europe derived from the HR drainage direction map (see Section 2.4).

Figure 1. RF modeling workflow for simulating monthly time series of streamflow intermittence on river reaches, i.e., the number of no-flow days per reach-month in five classes (0, 1-5, 6-15, 16-29, 30-31 no-flow days). Each of the two RF models is first set-up, by calibrating it such that the observed targets are best simulated; this includes the tuning of three hyperparameters in a non-spatial cross-validation (left-hand side of schematic). The intermittence status of each month and reach in Europe during 1981-2019 is calculated with the two calibrated RFs by applying first the step 1 RF for all reach-months and then the step 2 RF for all intermittent reach-months (right-hand side of schematic).

Despite implementing the two-step approach, class imbalance persists in each step of the modeling process, with many more perennial station-months than intermittent ones in step 1 and a relatively
large number of station-months with 30-31 no-flow days in step 2. Therefore, we applied standard
oversampling of the minor class (intermittent) in step 1 by a factor of 34.68, the ratio of perennial
to intermittent months. In step 2, the three minor classes were oversampled such that, for each
minority class, the number of training observations in that class was equal to the number of
observations in the majority class (30-31 no flow days).

Unlike hydrological models, RF models, which are ensembles of classification trees, do not have
parameters that represent properties of the modeled system. However, they have hyperparameters
that determine how the RF algorithm exactly works, which can be tuned to maximize predictive
accuracy while minimizing overfitting. RF hyperparameters are 1) the sample fraction, i.e., the
fraction of the training data that is randomly sampled without replacement for generating each
individual tree, 2) the number of predictors that are sampled from the full set of predictors and
considered by each tree when splitting each node (MTRY) and 3) the minimum number of
observations that a terminal node can contain, which influences the depth of the trees (i.e., when
tree construction stops). Model performance increases asymptotically with the number of decision
trees. In this study, the number of decision trees was set to 800 to limit run times.

In each step, the RF was tuned and evaluated by twice-repeated three-fold nested cross-validation.
Nested cross-validation, a resampling method that combines two levels of cross-validation loops
(outer and inner loops) separates hyperparameter tuning in the inner loop from model performance
evaluation in the outer loop (Bischl et al. 2012). In each loop, cross-validation uses different
portions of the data to iteratively test and train a model on the different subsets of the data. A three-
fold cross-validation means that the RF is trained with a random selection of two-thirds of the
samples (training data), each sample consisting of the predictors and the target for one station-
month. The predictive accuracy of the model is then evaluated with the remaining third of the
samples (testing data). In a twice-repeated three-fold nested cross-validation, there are six rounds
of cross-validation in total with different training and test data. Hyperparameter tuning for RF
step1 and step2 was performed through 15 and 55 unique combinations of hyperparameter,
respectively. For each round, hyperparameters were tuned by evaluating the performance of 15
and 55 unique combinations of hyperparameters in the case of the step 1 RF and step 2 RF,
respectively.

Model validation (Sections 4.1 and 4.2) was done using the results of the six rounds of cross-
validation, i.e., the results of the six RF models with an optimal combination of hyperparameters
as determined by the inner loop. For each station-months, the two predicted probabilities of
pertaining to a certain class were averaged and the class was assigned. The threshold for assigning
the perennial or intermittent class was set to a probability of 50%, consistent with our efforts to
balance the training data set.

The RF showing the highest balanced accuracy (BACC; Section 2.5) across all six rounds was
used for model application (Section 4.3). This resulted in a calibrated RF model consisting of the
best-performing step 1 RF and the best-performing step 2 RF. For step 1, the optimal values for
sample fraction, MTRY and minimum number of observations for the terminal node were 0.25, 4
and 2, respectively; the corresponding values for step 2 were 0.75, 6 and 10, respectively.

We computed the relative contribution of predictors to the predictive ability of the model, in the
form of the Actual Impurity Reduction (AIR) predictor importance metric. The higher the AIR,
the more important the predictor. The role of predictor variables was also evaluated with partial
dependence plots, which depict the marginal relationship between each predictor variable and the
probability of a predicted class while holding the rest of the predictors at their respective mean values. Using 20 processors (Intel Xeon silver 4114 2.2 GHz) in parallel, the run time for setting up the step 1 RF was about 14 days, and 14 hours for setting up the step 2 RF.

2.4. Definition of stream reaches for model application

It would be computationally too expensive to estimate the streamflow intermittence status for all HR grid cells in Europe, regarding both computation time and data storage. With 73 million HR grid cells across Europe and 468 months (1981-2019), more than 34 billion predictions would have to be computed. Therefore, we applied the two RF models sequentially to predict the streamflow intermittence status of river reaches rather than individual grid cells. Predictions are made for the most downstream HR grid cell of each river reach and are assumed to represent the mean conditions over the whole river reach.

River reaches at the HR resolution of 15 arc-sec are available in HydroSHEDS (HydroRIVERS, Lehner and Grill, 2013, https://www.hydrosheds.org/products/hydrorivers) but they insufficiently cover headwater streams for the purpose of our study (Döll et al., 2023a); in addition, we had slightly modified the HydroSHEDs drainage direction map. Therefore, river reaches were newly generated from the modified HydroSHEDs HR drainage direction map by applying the following delineation thresholds: streams were defined to start at all HR grid cells with an upstream drainage area of more than 2 km² (instead of 10 km² in 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 m³/s (instead of 0.1 m³/s in HydroRIVERS). Decreasing the threshold for streamflow to 0.02 m³/s would lead to potential "aggregates" of multiple streams in one grid cell in wet areas. Using these delineation thresholds, the resulting number of reaches in Europe is 1,533,471, with an average reach length of 2.0 km (standard deviation 1.7 km), representing a total stream network length of 3.06 million km. Accordingly, the European data set of monthly streamflow intermittence status contains a total of 717,664,428 reach-months covering the period 1981-2019.

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 delineated inside the boundaries of lakes and artificial reservoirs. Users of the streamflow intermittence dataset may therefore need to mask out simulated reaches as appropriate.

2.5. Performance metrics

As the observation data were strongly imbalanced, we evaluated model performance through the cross-validation of the two RFs based on the BAlanced ACCuracy (BACC). BACC provides a better indication of the classification performance of imbalanced models than raw accuracy (the percentage of correctly classified observations). In the binary case of step 1, BACC is the mean of sensitivity and specificity, with

\[
sensitivity = \frac{TP}{(TP+FN)}
\]

\[
specificity = \frac{TN}{(TN+FP)}
\]

where TP: true positive, FN: false negative, TN: true negative and FP: false positive, resulting from the confusion matrix (Figure S1). In the multiclass case of step 2, we follow the definition of Urbanowicz & Moore (2015) whereby the mean of sensitivity and specificity is calculated for each
of the four classes and then averaged over the classes.

Model performance was also evaluated with the Nash-Sutcliffe efficiency (or model efficiency), a traditional performance metric in hydrological modeling. It provides an integrated measure of model performance concerning mean values and variability and is computed as

\[
NSE = 1 - \frac{\sum_T (\text{sim}(t) - \text{obs}(t))^2}{\sum_T (\text{obs}(t) - \mu_{\text{obs}})^2}
\]

where \(\mu_{\text{obs}}\) is the mean of observations across all time steps; \(\text{sim}(t)\) and \(\text{obs}(t)\) refer to the simulated and observed values respectively, at time-step \(t\) of a total number of time steps \(n\). NSE can range from \(-\infty\) to 1; a value of 0 indicates that the model performs no better than simply using the mean of the observed data to predict the values, and a value of 1 indicates perfect agreement between the observed and modeled values.

3. Validation of HR time series of monthly streamflow in Europe

Comparing the downscaled HR monthly streamflow time series to the monthly time series of observed streamflow at the 3706 gauging stations across Europe yielded a median NSE value of 0.41; NSE exceeds 0 for 69% of the stations, and 25% of stations exceed the value of 0.64 which indicates a relatively good performance. When NSE is computed with the logarithm of streamflow, which puts a larger weight on low-flow months of interest for intermittence, NSE exceeds 0 for 63% of stations and 0.57 for 25% of stations. This shows that streamflow during the low-flow months is also estimated reasonably well. However, the performance of simulated HR streamflow is very poor in most of Spain, where human activities strongly impact streamflow (Figure 2). Although the impact of artificial reservoirs as well as groundwater and surface water abstractions are simulated by WaterGAP, the coarse resolution of the original model calculations (at LR grid cells) prevents the identification of the specific locations of these impacts in the downscaling procedures. Also, the HR location of natural surface water bodies, i.e., lakes and wetlands, is not explicitly taken into account in the downscaling method, causing potential misallocation of their attenuating effects on HR streamflow. Furthermore, other anthropogenic disturbances such as weirs are not accounted for in the original WaterGAP estimates. A poorer performance of HR streamflow in strongly altered streams is therefore due to both downscaling constraints and the difficulty of simulating human impacts at the LR resolution.
The performance of the downscaling algorithms can be assessed by comparing the NSE values at gauging stations with different upstream areas (Figure 3). LR streamflow as computed by WaterGAP is generally only compared to streamflow observed at gauging stations with upstream areas of more than 10,000 km$^2$, as a single LR grid cell can cover more than 2,500 km$^2$ (Müller Schmied et al. 2021). The high uncertainty of the global climate datasets used as the input of WaterGAP also inherently limits model performance for smaller basins. The performance of simulated streamflow does not decrease much with decreasing upstream area of the gauging stations (Figure 3a). For example, the median NSE for drainage basins larger than 10,000 km$^2$ is 0.51, while it is only slightly lower at 0.38 for the smallest drainage basins with areas below 2 km$^2$. The median NSE of logarithmic streamflow decreases from 0.40 for the basins larger than 10,000 km$^2$ to 0.14 for basins smaller than 2 km$^2$ (Figure 3b). When interpreting the NSE values, it should be noted that the stations are not equally distributed between the different catchment area classes; for instance, there are less than 100 stations with an upstream area of less than 10 km$^2$. Furthermore, this relationship between predictive performance and catchment area is not consistent among stations on intermittent and perennial waterways. In the case of intermittent stations (n=885), there is a decline in NSE values from basins with upstream areas of less than 50 km$^2$ to basins larger than 10,000 km$^2$ (Figure S2), whereas the opposite is true for perennial stations.
While small intermittent basins smaller than 2 km\(^2\) are characterized by a median NSE of 0.49 (median NSE for log streamflow: 0.21), the large basins over 10,000 km\(^2\) show a very poor performance with a median NSE of less than 0 (Figure S3). This might be due to the difficulty of simulating the impact of reservoir operations on intermittence. Considering the size class of 50-500 km\(^2\), which include the most stations of both intermittent (>100 stations) and perennial types (>1000 stations), the median NSE is 0.23 for intermittent stations and 0.43 for perennial stations (Figures S2 and S3).

**Figure 3.** NSE of monthly streamflow time series (left) and of the logarithm of monthly streamflow time series (right) for all 3706 streamflow stations with observations, grouped in size classes of the upstream area of the streamflow gauging stations. The boxes indicate the 25\(^{th}\), 50\(^{th}\) (median) and 75\(^{th}\) percentiles, the whiskers indicate the 5\(^{th}\) and 95\(^{th}\) percentiles of the samples. The blue lines of the violin plot show the smoothed distribution of the data points. The “number of stations not shown” indicates the number of stations with an NSE of less than -1.

### 4. RF modeling results

#### 4.1. Model validation

**4.1.1. Step 1 RF**

The cross-validation of the calibrated step 1 RF resulted in a BACC of 0.92. Of all perennial station-months, 98% were correctly identified as perennial, i.e., without any no-flow day (Table 2). Consequently, 25,496 (2%) of all perennial station-months were erroneously identified as intermittent. Of the intermittent station-months, 86% were correctly identified as intermittent, i.e.,
4,463 intermittent months were wrongly identified as perennial. Thus, the step 1 RF tends to overestimate the occurrence of intermittent months in absolute terms. In Europe, streamflow intermittence is more prevalent in the summer (JJA) and in the fall (SON) than in winter (DJF) and spring (MAM), and this is also the case for the number of predicted intermittent months (Table 2). A higher percentage of intermittent station-months, about 88%, was correctly identified as intermittent in JJA and SON than in the other two seasons (Table 2).

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Number of Observed and Correctly Simulated Perennial and Intermittent Stations and Station-months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of stations</td>
</tr>
<tr>
<td></td>
<td>All</td>
</tr>
<tr>
<td>Correctly simulated as perennial</td>
<td>2806</td>
</tr>
<tr>
<td>Observed as perennial</td>
<td>2821</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Correctly simulated as intermittent</td>
</tr>
<tr>
<td>Observed as intermittent</td>
<td>885</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Observed (bottom numbers) and correctly simulated (top numbers). Information on station-months is provided for all months and the four seasons December to February (DJF), March to May (MAM), June to August (JJA) and September to November (SON).

The overestimation of intermittent months mainly occurs at stations that are both observed and simulated to be intermittent, i.e., stations that have at least one no-flow day in the whole period 1981-2019, as only 15 perennial gauging stations, scattered throughout Europe, were erroneously predicted to be intermittent (dark red symbol in Figure 4b). Thus, 99.5% of all 2,821 stations observed to be perennial were correctly simulated to be perennial (Table 2, grey symbols in Figure 4b). The 885 gauging stations with at least one intermittent month, i.e., 24% of all stations considered in this study, are particularly concentrated on the Iberian Peninsula, Sardinia and Cyprus (Figure 4a), where gauging stations commonly recorded more than 20% of intermittent months. Elsewhere, almost all intermittent stations have less than 20%, and mostly less than 10%, of intermittent months. No intermittence is observed in winter months in the northern parts of Scandinavia, even though no-flow conditions are commonly reported in these climates because of dry conditions, the storage of precipitation as snow, and freezing (Buttle et al., 2012). Intermittence was not even observed at a station on a northern Norwegian island with a small drainage area of 19 km². Intermittence at the stations in western Finland occurs only in the summer and only at
stations with small upstream areas. The two intermittent stations in northern Sweden are located downstream of large artificial reservoirs.

Over a third of intermittent stations (334 out of 885) were wrongly simulated to be perennial by the step 1 RF (dark blue dots in Figure 4b); these stations are distributed across Europe with no clear spatial clustering. Many of these stations are located on streams that normally flow year-round but that exceptionally dried, for example, during a severe drought. Indeed, these intermittent stations that were wrongly classified as perennial have a median of only 2 intermittent months across their entire record (range: 1-19 months), while the 551 correctly classified intermittent stations have a median of 35 months (range 2-431 intermittent months).

When considering only the 885 intermittent stations, the median and mean percent of observed intermittent months are 5.6% and 15.8%, respectively. Whereas 86% of all observed intermittent stations months (28,244 out of 32,707) are correctly predicted to be intermittent, 11% of observed perennial station-months at intermittent stations (25,398 out of 233,195) are wrongly predicted to be intermittent. This resulted in a general overestimation of the total share of station-months at intermittent stations. While 13% of all station-months at intermittent stations are observed to be intermittent (and 11% correctly predicted as such), 21% are predicted to be intermittent. The overestimation is concentrated in regions with a relatively high prevalence of intermittence, i.e., large parts of the Iberian Peninsula, Sardinia and Cyprus (compare Figures 4a and 4b), where intermittent months are often overestimated by a factor of more than 2 (Figure 4b). The main suspected reasons for this overestimation are the poor ability of the downscaled streamflow estimates (Figure 2) and the RF model to capture the strong human impacts on streamflow dynamic in large parts of Spain as well as Cyprus (not Sardinia). In these semi-arid regions, a multitude of small and large dams as well as water transfers by canals often make naturally intermittent streamflow perennial (Chiu et al., 2017). Even though some large reservoirs are considered when computing LR net cell runoff used to estimate HR streamflow, simulation of reservoir outflow is very uncertain already at LR. In addition, information on reservoirs, weirs or canals in the individual HR cells within each LR is not taken into account in the streamflow downscaling approach. The reservoirs included in the computation of the static HR predictor by the predictor dor (degree of regulation by upstream dams; Lehner et al., 2011) (Table 1) only takes into account a subset of the actual reservoirs and misses small ones.

Interannual variability of the number of intermittent months per year is simulated quite satisfactorily, in particular for gauging stations in southern Spain (Figure 4c). Considering all 885 intermittent stations, the median Pearson correlation coefficient between the observed and predicted annual time series of the number of intermittent months is 0.50. Thus, the step 1 RF is able to capture the interannual variability of climatic conditions. That said, the corresponding NSE values (i.e., based on the annual time series; not shown) are below zero at almost all stations due to the strong overall overestimation of intermittent months.
Figure 4. Percentage of observed intermittent months (with at least one no-flow day) per gauging station for all observations during 1981-2019 (a), ratio of the number of predicted months to the number of observed intermittent months (P: perennial, I: intermittent) (b) and Pearson correlation of the annual time series of the number of intermittent months (c), as simulated by the step 1 RF model.

As expected, gauged streams in smaller drainage basins are both observed and simulated to be more strongly intermittent than larger drainage basins, especially in the two smallest drainage basin classes 0-2 km² and 2-5 km² (Figure S4). However, intermittent months are also most overestimated in these size classes; the predicted median proportion of intermittent months for these stations is twice the observed median of about 13%. For drainage basins larger than 2500 km², on the contrary, the step 1 RF tends to underestimate the already low percentage of intermittent months (though it strongly overestimates intermittence for a few basins, too; Figure S4).
4.1.2. Step 2 RF

The target of the step 2 RF are the observations of the number of no-flow days, in four classes (1: 1-5, 2: 6-15, 3: 16-29, 4: 30-31 no-flow days) in observed intermittent months. At most intermittent gauging stations, class 1 (1-5 no-flow days) dominates, whereas class 4 (30-31 no-flow days) dominates in many stations with more than 10% of intermittent months, in particular in the central and southern part of the Iberian Peninsula and in Cyprus (Figure S5). With a BACC of 0.67 (averaged over the four classes) in the cross-validation of the calibrated step 2 RF, the classification performs satisfactorily. More than three quarters of the station-months with observed class 4 (30-31 no-flow days) are correctly classified, and almost half of the station-months with 1-5 and 16-29 no-flow days are correctly classified (Figure 5). Although the model exhibits weaker performance for station-months with 6-15 observed no-flow days, these observations are still more likely to be correctly classified than pertaining to any of the three other classes. Classification performance is highest for the class with most observations, 30-31 no-flow days, as can be expected in RF modeling. In total, 54% of the 32,707 station-months are classified into the correct observed class, and of the wrongly classified observations, 70% are predicted to belong to neighboring classes (Figure 5).

![Confusion matrix of predicting four classes of no-flow days per station-month. The top number in each box shows total number of station-months belonging to the observed and simulated intermittence class, the bottom number the percent of the total number of station-months that are observed to be in the intermittence class (step 2 RF model).](image)
The percentage of intermittent months that are correctly classified into the four classes shows no spatial pattern across Europe (Figure 6a), although the overestimation of no-flow days is most pervasive in Spain where the number of observed no-flow days is already high (red in Figure 6b). The step 2 RF tends to overestimate the number of no-flow days in the intermittent station-months where the step 1 RF also overestimates the number of intermittent months (e.g., in many stations on the Iberian Peninsula). The bias shown in Figure 6b correlates weakly with the ratio of predicted to observed intermittent months shown in Figure 4b, with a Pearson correlation coefficient of 0.11. The correlation between the monthly time series of observed and simulated intermittence classes, as measured by the Spearman rank correlation coefficient, is positive for most gauging stations, and larger than 0.3 for 38% of stations (Figure 6c). This correlation analysis does not include the perennial months at a station. The overall performance of the monthly time series of five classes, with class 0 for perennial months, reflects the combined performance of the step 1 and step 2 RFs and thus the overall RF modeling approach used for estimating streamflow intermittence for all reach-months in Europe. These correlation values, shown in Figure 6d, are much higher than the correlation for just the intermittent months; values larger than 0.9 dominate. The median Spearman rank correlation coefficient for the monthly time series of the five intermittence classes is 0.81, with 90% of the stations exceeding a value of 0.58 and 14% of the stations exceeding a value of 0.99.
Figure 6. Percentage of intermittent months that are correctly classified into the four classes (1: 1-5 no-flow days per month, 2: 6-15, 3: 16-29, 4: 30-31) by the step 2 RF at each of 885 gauging stations with at least 1 no-flow day in their record (a), bias expressed as simulated mean class number (1 through 4) minus observed mean class number (green: correct average classification, red: overestimation of no-flow days, blue: underestimation of no-flow days) (b), and Pearson correlation coefficient for the monthly time series of simulated and observed intermittence classes, for four classes 1-4 (c) and five classes 0-5, with class 0: 0 no-flow day (d). All correctly classified perennial stations were omitted from the maps and would show a correlation coefficient of 1.

4.2. Importance of predictors and dependence of predicted class on predictor values

All 23 predictors were found to be significant at the p-value = 0.05 level. The relative importance of the 23 predictors differs strongly between step 1 RF (identifying whether a station-month is
intermittent) and step 2 RF (identifying the number of no-flow days in intermittent months, in four classes) (Figure 7). However, two predictors computed from the downscaled HR monthly streamflow, namely the monthly area-specific streamflow (Q) and the mean of the area-specific streamflow of the previous three months (Q\_mean\_p3), are among the five most important predictors in both RFs. Both are predicted to be negatively correlated to the probability of intermittence, as was expected (Figure S6).

The most important predictor in the step 1 RF is the size of the drainage basin of the streamflow gauging station (Figure 7), with the probability of intermittence decreasing with increasing size up to a drainage area of about 20,000 km² (Figure S6). Terrain slope (slope) and the precipitation to potential evapotranspiration ratio (P\_to\_PET\_ratio) show a similar importance in step 1 and take up ranks 4 and 5, respectively. The partial dependence plots for the step 1 RF show, for all but 2 of the 23 predictors, correlations between the predictor and the likelihood of intermittence that are expected by hydrologists. For example, 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 for Q\_iav\_cv > 0.4. Exceptions to this correspondence between model predictions and hydrological understanding include the terrain slope (slope) and, albeit less conclusively, the degree of regulation (dor) (Figure S6). Steeper slopes across the upstream drainage area are expected to make intermittence more likely (Šarauškienė et al., 2020) due to a decrease in the fraction of runoff that recharges groundwater and thus a decrease in baseflow, but the RF predicted the opposite correlation. This 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, i.e., wet regions with large runoff. As for the degree of regulation, artificial reservoirs can make streams either more perennial or more intermittent, depending on reservoir management (e.g., for hydroelectricity, irrigation, flood control) and river type (Datry et al., 2023). Here, the step 1 RF showed that increased regulation was associated with greater levels of intermittence (Figure S6). A likely reason for this correlation is that many stations downstream of large dams in our training dataset were located in dry areas like Spain, where intermittence is common and flow regulation by reservoirs is associated with extensive water withdrawal (Sabater & Tockner, 2009). This predictor’s importance in RF 1 is very low (Figure 7), so the impact of this counterintuitive relationship on model predictions is minor.

In the step 2 RF, all of the five most important predictors are dynamic predictors. They include four HR predictors derived from the downscaled WaterGAP output (Figure 7). In addition to the highest ranking Q and Q\_mean\_p3, the minimum area-specific streamflow over the previous three months (Q\_min\_p3) and the mean area-specific streamflow over the previous 12 months (Q\_min\_p12) are among the five most important predictors. The LR predictor of the number of wet days per month is ranked second in importance.
Figure 7. Predictor importance for step 1 RF (left) and step 2 RF (right). The higher the impurity reduction, the larger the relative importance of a predictor. The higher absolute values for the step 1 RF are due to the larger number of station months available as target. Error bars show the standard deviation across the six cross-validation training sets calculated for both the step 1 and step 2 RFs. The relatively larger error bars for the step 2 RF are due to considering four classes instead of only two in the step 1 RF. Dynamic HR predictors are indicated by * and dynamic LR predictors by +.

4.3. Predicted time series of monthly streamflow intermittence status of stream reaches in Europe

In total, 96.2% of the approximately 718 million reach-months at more than 1.5 million stream reaches in Europe are simulated as perennial in the period 1981-2019 (Table 3). 82.2% of the stream reaches and 81.3% of the European network length of 3.06 million km are simulated to never have experienced a no-flow day during this period. Reaches with intermittent months are simulated to exist in almost all European countries, but high percentages of intermittent months are prevalent on the Iberian Peninsula, Sardinia and Cyprus and also occur in southern Italy and
Greece (Figure 8). Large regions with low fractions of intermittent months exist in France but also in Finland, Belarus and Ukraine.

Table 3
Occurrence of the Five Intermittence Classes at the Gauging Stations and All Reach-months in Europe

<table>
<thead>
<tr>
<th>Class</th>
<th>Station-months</th>
<th>Reach-months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Predicted</td>
</tr>
<tr>
<td>0: Perennial</td>
<td>1,134,237</td>
<td>1,113,204</td>
</tr>
<tr>
<td></td>
<td>97.20%</td>
<td>95.39%</td>
</tr>
<tr>
<td>1: 1-5 no-flow days</td>
<td>5,643</td>
<td>5,248</td>
</tr>
<tr>
<td></td>
<td>0.48%</td>
<td>0.45%</td>
</tr>
<tr>
<td>2: 6-15 no-flow days</td>
<td>6,030</td>
<td>5,338</td>
</tr>
<tr>
<td></td>
<td>0.52%</td>
<td>0.45%</td>
</tr>
<tr>
<td>3: 16-29 no-flow days</td>
<td>8,634</td>
<td>8,484</td>
</tr>
<tr>
<td></td>
<td>0.74%</td>
<td>0.73%</td>
</tr>
<tr>
<td>4: 30-31 no-flow days</td>
<td>12,400</td>
<td>13,637</td>
</tr>
<tr>
<td></td>
<td>1.06%</td>
<td>1.17%</td>
</tr>
<tr>
<td>Total</td>
<td>1,166,944</td>
<td>1,145,911</td>
</tr>
</tbody>
</table>

Note. In this study, Europe does not include Russia and Turkey. The gauging stations represent those which were used to set up the RF model, where the fraction of all station-months with observed and simulated classes is provided. In each column, the first row shows the total number of stations or reaches in the class and the second row shows the percentage in the class. The percentage values for the reach-months relate to the total number of reach months (468) during 1981-2019; and for the station-months, to the number of station-months with observations. As the step 2 RF model predicting the four classes of no-flow days was set up only for the station-months that are observed to be intermittent, the predicted class percentages do not add up to 100%.
The predicted prevalence of perennial conditions across reaches is similar to the observed prevalence in streamflow gauging stations where 97% of the observed station-months and 76% of the stations are perennial. As drainage area is the most important predictor for a station-month being perennial or intermittent, with small basin size leading to a higher probability of intermittence, it is surprising that a higher percentage of reaches is simulated to be perennial as compared to the gauging stations. Reaches with small upstream basins of less the 50 km² comprise 78% of all reaches, whereas only 12% of gauging stations have such small basins (Table 4). This highlights the importance of the interplay of all predictors of the step 1 RF and may be affected by our voluntary addition of intermittent data in observations.

**Figure 8.** Percentage of months with at least one no-flow day for European stream reaches during the period 1981-2019.
Table 4

<table>
<thead>
<tr>
<th>Upstream area [km²]</th>
<th>Gauging stations</th>
<th>Reaches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Observed (m³ s⁻¹)</td>
<td>Mean Predicted (DSSᵃ) (m³ s⁻¹)</td>
</tr>
<tr>
<td>(0-2]</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>(2-5]</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>(5-10]</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td>(10-50]</td>
<td>0.79</td>
<td>0.58</td>
</tr>
<tr>
<td>(50-500]</td>
<td>4.55</td>
<td>3.25</td>
</tr>
<tr>
<td>(500-2500]</td>
<td>17.29</td>
<td>15.65</td>
</tr>
<tr>
<td>(2500-10000]</td>
<td>56.91</td>
<td>57.73</td>
</tr>
<tr>
<td>&gt;10000]</td>
<td>512.31</td>
<td>544.30</td>
</tr>
</tbody>
</table>

ᵃDownscaled streamflow.

The fraction of reach-months with 30-31 no-flow days (3.4%) is much higher than the corresponding fraction of the stream-months that are observed and predicted to occur at the streamflow gauges (1.1%; Table 3). This is not due to the much higher prevalence of reaches with small upstream basins than of stations with such small basins (Table 4) as also in each drainage area size class, the fraction of months with 30-31 no-flow days is larger for the reach-months than for the station-months (Table 5). Both station observations and reach predictions agree that the likelihood of perennial months increases and the likelihood of 30-31 no-flow days decreases with increasing size of the drainage basin (Table 5). The exception are the smallest reaches with an upstream area of 2 km² or smaller, because we only generated such small reaches from the 15 arc-sec drainage direction map where the mean annual downscaled HR streamflow during the period 1981-2019 exceeds 0.03 m³/s (section 2.4); this explains the high fraction of perennial months in the smallest size class. One reason for the higher prevalence of the class 30-31 no-flow days for the reach-months as compared to the station-months in all size classes between 2 km² and 500 km² may be that the average streamflow for all reach-months of a certain size class is smaller than for the gauges (both observed and predicted) (Table 4). This discrepancy likely led to more dry months, because streamflow is the most important predictor in the step 2 RF (Figure 7). At the same time, the fraction of perennial reach-months, which is determined by the step 1 RF, is also
higher than the fraction of perennial station-months in each size class, such that the other three
intermittence classes are predicted to be very rare among the reach-months. The reason for this is
unknown, but one aspect to consider may be that streamflow is not the most important predictor
in the step 1 RF (Figure 7).

Table 5

Percent of Observed Station-months and Predicted Reach-months (1981-2019) in the Five
Intermittence Classes

<table>
<thead>
<tr>
<th>Upstream area [km²]</th>
<th>Observed station-months in classes 0-4 (%)</th>
<th>Predicted reach-months in classes 0-4 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(0-2]</td>
<td>87.30</td>
<td>2.57</td>
</tr>
<tr>
<td>(2-5]</td>
<td>88.48</td>
<td>1.93</td>
</tr>
<tr>
<td>(5-10]</td>
<td>94.18</td>
<td>1.11</td>
</tr>
<tr>
<td>(10-50]</td>
<td>95.85</td>
<td>0.59</td>
</tr>
<tr>
<td>(50-500]</td>
<td>96.83</td>
<td>0.34</td>
</tr>
<tr>
<td>(500-2500]</td>
<td>98.33</td>
<td>0.24</td>
</tr>
<tr>
<td>(2500-10000]</td>
<td>99.32</td>
<td>0.10</td>
</tr>
<tr>
<td>&gt;100000</td>
<td>98.85</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Note. These values are represented as a function of upstream drainage area [km²] of the streamflow gauging stations
or the reach. Classes: 0: perennial, 1: 1-5 no-flow days, 2: 6-15 no-flow days, 3: 16-29 no-flow days, 4: 30-31 no-flow
day. In total, 1,166,944 station-months and 717,664,428 reach-months are considered.

The actual number of perennial months in reaches with upstream areas of 2-50 km², the dominant
upstream area classes listed in Table 4, may even be higher as the step 1 RF tends to underestimate
the fraction of perennial station-months (Figure S4). However, the number of streamflow gauging
stations for that class, in particular in the size class under 10 km², is rather small (Table 4).

The prevalence of intermittence across the European river network shows a clear seasonal and
interannual variability. While 97.6-99.8% of the European reaches are perennial in January and
February, this is the case for only 89.6-93.4% in August and September (Figure 9). There is no
overall trend over the whole 39-year period, but seasonal minima and maxima of the fraction of
perennial reaches decreased from 2013 to 2019, while the opposite is true for the fraction of months
with 30-31 no-flow days (Figure 9). The southern European countries of Portugal, Spain, Italy,
Greece and Cyprus have a much higher seasonal range of the fraction of perennial reaches; in July
to August, only about 70% of the reaches are perennial, while in winter, it is close to 90-99%, depending on the year (Figure S8a). In the Scandinavian countries Norway, Sweden and Finland, the (very low) occurrence of intermittent conditions is larger in the second half of the study period, but the highest level of intermittence occurred in 1996, related to unusually low precipitation (Figure S7b). As an illustration of the spatial distribution and seasonality of streamflow intermittence, the European maps for streamflow intermittence in January and August 2019 are shown in Figure S8.

Figure 9. Monthly time series of the percent of all European stream reaches in the five intermittence classes for the period 1981-2019.

5. Discussion

In this study, we chose to use all daily streamflow observations available for the study period to set up the RF model, to obtain a robust model based on the maximum amount of information. A
temporal validation of the models with independent data was not conducted. The RF model of Döll et al. (2023a), with less streamflow observations and slightly different predictor variables and intermittence class definitions, was trained with data for a calibration period that encompassed, for each gauging station, the first two-thirds of the available observed months, while the rest was left for independent validation; 99% and 95% of the perennial station-months were predicted correctly for the calibration and the validation period, respectively. Considering only the intermittent station-months (i.e., predictions from the step 2 RF), the frequency of predicting the correct class decreased from 56% in the calibration period to 47% in the validation period. In this study, we validate our predictions with a data set of visually observed intermittence for France. Then, we discuss the challenges of deriving continental-scale high-resolution estimates of monthly streamflow and streamflow intermittence.

5.1. Validation of streamflow intermittence predictions using ONDE observations for France

We used observations from the French national river drying observatory (ONDE, 2020; Beaufort et al., 2018) to validate our predictions for 2,865 reaches and 148,004 reach-months in France of whether each reach-month was intermittent (with at least one no-flow day) or perennial, i.e., the step 1 RF. The ONDE network consists of a stable set of approximately 3,300 sites on river and stream reaches of Strahler orders under five, which have been inspected since 2012 by trained public staff from the French Biodiversity Office (OFB in French), at least monthly between May and September with the objective of identifying all drying events. If either the status “no visible flow” or “dried out” was assigned in any month, we considered the reach-month to be observed as intermittent. Considering that its objective is to track intermittence in mostly headwater streams, the ONDE dataset has a much higher percentage of intermittent reaches and reach-months than the European streamflow gauging station dataset used to set up the RF model. While 61% of the reaches and 15% of the reach-months are intermittent in the ONDE dataset, only 24% of the European gauging stations and 2.8% of the station-months are intermittent. Considering only French gauging stations, the respective values are 38% and 3.5%. About 73% of ONDE reaches have a drainage area of less than 50 km², which is similar to the fraction of European reaches in the size class (Table 4), whereas this is the case for only 12% in the European data set of gauging stations.

Compared to the ONDE data, the step 1 RF model underestimates the number of intermittent reach-months (Figure 10d), whereas it tends to overestimate the number of intermittent station-months relative to the 3706 European streamflow stations. With a BACC of 0.53, only 8% of the intermittent reach-months in ONDE are correctly identified (Figure 10c). Underestimation occurs in all size classes, increasing from an underestimation of, on average, 4 months for upstream areas of less than 10 km² to an underestimation of 6-7 months for basins between 10 km² and 2,500 km². Considering whether reaches are intermittent or perennial, only 23% of the intermittent reaches were correctly predicted as such, compared to 62% for the European stations used to set up the step 1 RF (Table 2). Our RF model achieves a balanced accuracy of only 0.54 (Figure 10a) in its binary classification of ONDE reaches, while the global static RF model of naturally intermittent reaches of Messager et al. (2021) yielded a slightly higher value of 0.59 (Figure 10b). The spatial pattern of agreement of the static global model is less patchy than that of our model. The global model predicts intermittence to occur in large contiguous areas, as it is mainly driven by larger-scale climatic predictors whereas our dynamic European model is strongly driven by small-scale streamflow characteristics. In addition, our model is based on more streamflow gauging stations.
For unknown reasons, our RF model cannot predict intermittent reaches along the Mediterranean coast, which differs from the static global model. If the threshold for perennial conditions is increased to a probability of 75%, which does not lead to a decrease of BACC, 91% of the intermittent reaches would be correctly identified, but then 85% of the perennial reaches would be incorrectly predicted as intermittent.

Figure 10. Comparison of simulated intermittence of reaches and reach-months with the ONDE data set of visually observed intermittence. Correspondence between the simulated and observed intermittence state of reaches for our RF model (a) and the RF model of Messager et al. (2021) (b), percent of correctly classified reach-months in our model (c) and ratio of predicted to observed intermittent months in our model (d).
5.2. Challenges of continental-scale high-resolution estimation of streamflow and streamflow intermittence

Given the amount, spatial resolution and uncertainty of available data, it is very challenging to achieve a good prediction of HR streamflow intermittence for all of Europe. One reason is that continental and global-scale streamflow simulations for relatively large rivers represented by LR models often strongly differ from observations. This mismatch stems from large-scale models not being calibrated in a basin-specific manner against observed daily streamflow, as is done with small-scale models. In addition, their input data are coarser and usually less accurate than those of small-scale models. Here, using an advanced downscaling algorithm, the output of a LR hydrological model (Figure 11a, b) was downscaled, by a factor of 14,400, to generate monthly time series of streamflow at 15 arc-sec resolution (Figure 11c). It is encouraging that these simulated HR streamflow time series show skill for most streamflow stations with upstream areas smaller than an LR cell, even down to upstream areas of less the 5 km² (Figures 2 and 3). However, the number of evaluated gauging stations with such small upstream areas was very small (Figure 3). The estimated HR streamflow time series enabled, together with other predictors, the estimation of HR streamflow intermittence. A comparison of Figures 11b and 11d, which shows LR and HR intermittence, respectively, for a part of France illustrates the strongly increased information content of the European HR streamflow intermittence dataset as compared to an LR intermittence estimation.
Figure 11. Illustration of downscaling of LR WaterGAP output to the HR stream network of HydroSHEDS and the resulting resolution-dependent characterization of intermittence. Panels show LR (0.5 arc-deg) grid cells with the sum of surface runoff and groundwater discharge (the main input to the downscaling algorithm) (a), LR reaches with their intermittence status (b), HR (15 arc-sec) grid cells with downscaled streamflow (c) and HR reaches with intermittence status in 5 classes (d). The figure shows the situation for the example of August 2003. In c and d, the locations of the streamflow gauging stations used for validation of downscaled streamflow and as target for the RF model are added.
It is very difficult to judge the realism or plausibility of the predicted reach intermittence. The model validation against an independent data set for France (section 5.1) indicates a severe underestimation of observed intermittent reach-months, while the comparison of predicted intermittence to streamflow observations used for setting up the RF model indicates that the RF model overestimates intermittence, particularly for the relatively dry regions of Europe such as large parts of the Iberian Peninsula. The latter may be explained by the suspected anthropogenic perennialization of streamflow by many small and large dams that have been constructed to make water available even in periods of low or no flow. Still, the BACC for predicting intermittent station-months (0.92) was very good. We found that the RF model can simulate well the interannual variability of the number of intermittent months at the streamflow gauging stations (Figure 4c), which is an important positive characteristic if the model is to be used for assessing the impact of drought conditions or climate change. The partial dependence plots for the step 1 RF show that the model identifies tendencies in the probability of a station-month being intermittent that agree with expert expectations (except for terrain slope), which increases our trust in the derived RF. Moreover, the correlation between the observed and predicted monthly time series of the five intermittence classes is high at most intermittent stations (Figure 6d), which indicates a good representation of the seasonality of streamflow intermittence. Averaged over all station-months with available intermittence observations, there is no bias in the prediction of the five intermittence classes per size class of upstream area as compared to observations (Table 3). However, the prediction of the number of no-flow days per reach-month in four classes must be considered to be less reliable than the prediction of a reach-month as either intermittent or perennial, as indicated by the lower BACC of 0.67 for the step 2 RF.

In this study, we estimate that 18.7% of the European stream network length were intermittent in the period 1981-2019, while the global model of Messager et al. (2021) predicts a value of 17.1% for the our European study area. However, these values cannot be compared directly for various reasons. Our river network includes smaller headwater streams than the global model (representing 12.4% of the European river network used in this study; see Section 2.4) and the definition of intermittent reach is slightly different (global model: 1 no-flow day per year, our model: 1 no-flow day during the period 1981-2019). In addition, the global model aimed to predict natural intermittence by excluding heavily influenced gauging stations, relying on naturalized hydrology for the period 1971-2000. Still, our model predicts a similar prevalence of intermittent reaches in Europe.

The prevalence of intermittence across European rivers and streams by this study, with 17.8% of intermittent reaches and 3.8% of the reach-months, is much lower than in the study of Döll et al. (2023a), with values of 39.6% and 9.1%, respectively, even though the same HR streamflow estimates were used in the RF modeling. Even though some predictors (related to irrigation, population and the degree of regulation by reservoirs) were added and one (daily streamflow variability) removed in this study, we attribute this strong discrepancy to the different observations of the RF target variables, which were derived from daily time series of streamflow observations. We explicitly tried to obtain streamflow from dry areas and small streams, with a higher likelihood of intermittence, and added data from Cyprus and Italy (for Sardinia and the Po basin, but time series for the rest of Italy were shorter than our inclusion threshold of 36 months), but we could not obtain in time any data for, e.g., Greece, Albania and Bulgaria. The data set was rather extended by stations for more humid regions such as Scandinavia, the three Baltic states, Poland and Belarus, most of which are perennial (comp. Figure 4a). When the streamflow observations data set was
extended from the one used by Döll et al. (2023a), i.e., from 1915 gauging stations to the 3706 stations in this study (see Table 4 for station numbers per drainage area class), the additional stations had a smaller fraction of intermittent months than the original data set. In this way, we have further biased the target data set and therefore may have caused a biased streamflow intermittence prediction for the reaches. Still, we expect that almost doubling the amount of target observations as compared to the study of Döll et al. (2023a) increased the reliability of the RF models. The additional streamflow data and predictors are informative because the BACC for the step 1 RF increased from 0.85 in Döll et al. (2023a) to 0.92 in this study (while the BACC for the step 2 RFs are the same). However, the fit to the ONDE data as measured by the BACC for the identification of intermittent reaches remained the same as that of the step 1 RF model by Döll et al. (2023a). The comparison of the European streamflow intermittence estimation by Döll et al. (2023a) and in this study shows the major impact of available target observations on RF modeling results.

Upstream area is the most important predictor for likelihood of a station-month to be intermittent, yet we cannot assume that inclusion of this predictor in the RF model development is adequately representing the effect of upstream area on the likelihood of intermittence given the existing distribution of our target data. To represent upstream area appropriately in the target data, we would need a data set of streamflow gauging stations that show the same distribution of upstream areas as the stream reaches; however, the distributions are extremely different (Table 4). While 77.7% of the reaches have an upstream area of up to 50 km², this is only the case for 12.3% of the gauging stations. The largest size class for the reaches is, with 31%, the class 2-5 km², but only 0.6% of the station-months are in this class. As an illustration, if we would like to have the same size distribution with the 22 stations in the class 2-5 km² that were available, then we would have to consider only 70 stations in total, instead of 3706. As the fraction of perennial months is higher for reaches than for the stations, especially for drainage areas below 50 km² (except for the smallest size class due to the definition of the smallest reaches, Table 5), a further decrease of the average fraction of intermittent months for the gauging stations by the extension of the data set might have led to an even stronger underestimation of intermittence in these headwater reaches. However, by our extension, we more than doubled the number of stations in the class 2-5 km² by raising the number of stations from 10 to 22 (Table 5) which increased the information base upon which the RF models were trained.

The performance of our model certainly suffers from the general problem of imbalanced target data, with 97.2% of the station-months being perennial. The most important approach to handle this problem was the two-step approach whereby the prediction of perennial months in step 1 was followed by the prediction of the number of no-flow days per month only for those 2.8% of all station-months for which at least one no-flow day was observed. In addition, various alternative methods for handling imbalanced data were tested for the step 2 RF. Oversampling resulted in slightly better BACC values than undersampling and the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002).

6. Conclusions and Outlook

For the first time, streamflow intermittence dynamics could be quantified at the continental scale at a high spatial resolution, i.e., for stream reaches with an upstream area down to only 2 km² (or even smaller in wet regions). We simulated monthly time series of streamflow intermittence in
five classes (0, 1-5, 6-15, 16-29 and 30-31 no-flow days per month) in the period 1981-2019 for more than 1.5 million stream reaches in Europe. This was achieved by 1) downscaling the 0.5 arc-deg output of the global hydrological model WaterGAP to obtain time series of monthly streamflow at about 73 million 15 arc-sec grid cells and 2) combining this information with daily data of streamflow as observed at 3706 gauging stations and a number of static hydro-environmental characteristics of the upstream basins (plus two WaterGAP-related datasets) in an RF modeling approach. The model captures the interannual variability of the number of intermittent months satisfactorily, and the monthly time series of the predicted five streamflow intermittence states is highly correlated with observations. The spatial prevalence of weakly intermittent conditions appears to be underestimated, while the number of intermittent months is overestimated in the dry regions of Europe where reservoirs tend to perennialize streamflow. While the generated streamflow intermittence data set does diverge from reality for many reach-months, it is nevertheless a valuable basis for macro-scale studies of biodiversity, ecosystem functions and ecosystem services under conditions of potential streamflow intermittence.

The presented modeling approach was designed to enable the computation of intermittence changes due to climate change. For this purpose, the LR output of a WaterGAP run that is driven by the bias-corrected output of global climate models, instead of observed historic climate, can be downscaled to calculate monthly time series of HR streamflow in, for example, a 30-yr reference period and a 30-year period in the future. These time series, together with the LR WaterGAP time series of monthly diffuse groundwater recharge, runoff from land and the number of wet days under climate change, can then serve to compute the dynamic predictor values that are, in addition to the unchanged static predictor values, the input for the two developed RF models. In addition, the developed modeling approach can be used to analyze the occurrence of drought in intermittent streams (Sarremejane et al., 2021).

**Data Availability Statement**

WaterGAP 2.2e input and output used for deriving HR streamflow and as LR predictors in the RF model is available from Müller Schmied et al. (2023b). The code for deriving HR streamflow (Trautmann, 2023) is available at https://doi.org/10.5281/zenodo.10301003, and the code and workflow for the RF modeling (Abbasi & Messager, 2023) at https://github.com/mahabbasi/europeanIRmap.git. Due to the very large file sizes, the HR monthly streamflow time series are only available on request from first authors. The following data are available at https://doi.org/10.6084/m9.figshare.24591807: 1) Input files for deriving HR streamflow (Text S1), 2) the monthly time series of streamflow at the 3706 gauging stations, 3) shapefiles of locations of streamflow gauging stations and European reaches, 4) all predictors and target variables for the 3706 gauging stations used to generate the step 1 and step 2 RFs and 5) shapefiles with the five intermittence classes for each reach-month in the period 1981-2019 as well as the shapefiles for generating all figures (Döll et al., 2023b). The original data used for deriving the HR static predictors are available as described in Section 2.3.2 and Table 1.

**Author Contributions**

- Conceptualization: Petra Döll, Bernhard Lehner
- Methodology: Mahdi Abbasi, Tim Trautmann, Petra Döll, Bernhard Lehner, Mathis L. Messager, Nicolas Lamouroux
Acknowledgments

The authors thank their collaborators in the DRYvER project, in particular Jean-Philippe Vidal and Thibault Datry, for their contributions. MA was funded by the EU Horizon 2020 DRYvER project (Project ID 869226). MLM was supported by a Vanier Canada Graduate Scholarship (Natural Sciences and Engineering Research Council of Canada) and an H2O'LYON Doctoral Fellowship (Université de Lyon, ANR-17-EURE-0018).

References


Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., & Gerdener, H., et al. (2023b). The global water resources and use model WaterGAP v2.2e - model output driven by gswp3-w5e5 and historical setup of direct human impacts. Goethe-Universität Frankfurt. https://doi.org/10.25716/GUDE.0TNY-K JPG [Dataset]


ONDE (Observatoire national des étiages) (2020). French Biodiversity Office https://onde.eaufrance.fr/content/t%C3%A9l%C3%A9charger-les-donn%C3%A9es-campagnes-par-ann%C3%A9e (accessed 21 June 2020)


Streamflow intermittence in Europe: Estimating high-resolution monthly time series by downscaling of simulated runoff and Random Forest modeling

Petra Döll1,2*, Mahdi Abbasi1*, Mathis Loïc Messager3,4, Tim Trautmann1, Bernhard Lehner4, Nicolas Lamouroux3

1Institute of Physical Geography, Goethe University Frankfurt, Frankfurt/Main, Germany
2Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F) Frankfurt, Frankfurt/Main, Germany
3INRAE, UR RiverLy, Lyon-Villeurbanne, France
4Department of Geography, McGill University, Montreal, Canada

Corresponding author: Petra Döll (p.doell@em.uni-frankfurt.de)

*Equal contribution

Key Points:

- Streamflow intermittence at more than 1.5 million European reaches was estimated for every month during 1981-2019.
- 18.7% of the European river network length and 3.8% of all reach-months are intermittent, predominantly with 30-31 no-flow days.
- 15 arc-sec monthly streamflow obtained by downscaling the output of a global hydrological model serves as input to Random Forest modeling.
Abstract

Knowing where and when rivers cease to flow provides an important basis for evaluating riverine biodiversity, biogeochemistry and ecosystem services. We present a novel modeling approach to estimate monthly time series of streamflow intermittence at high spatial resolution at the continental scale. Streamflow intermittence is quantified at more than 1.5 million river reaches in Europe as the number of no-flow days grouped into five classes (0, 1-5, 6-15, 16-29, 30-31 no-flow days) for each month from 1981 to 2019. Daily time series of observed streamflow at 3706 gauging stations were used to train and validate a two-step Random Forest modeling approach. Important predictors were derived from time series of monthly streamflow at 73 million 15 arc-sec (~500 m) grid cells that were computed by downsampling the 0.5 arc-deg (~55 km) output of the global hydrological model WaterGAP, which accounts for human water use. Of the observed perennial and intermittent station-months, 97.8% and 86.4%, respectively, are correctly predicted. Interannual variations of the number of intermittent months at intermittent reaches are satisfactorily simulated, with a median Pearson correlation of 0.5. While the spatial prevalence of intermittent reaches is underestimated, the number of intermittent months is overestimated in dry regions of Europe where artificial storage abounds. Our model estimates that 3.8% of all European reach-months and 17.2% of all reaches were intermittent during 1981-2019, predominantly with 30-31 no-flow days. Although estimation uncertainty is high, our study provides, for the first time, information on the continent-wide dynamics of intermittent rivers and streams.

Plain Language Summary

Even in wet climates, small streams can seasonally dry up. In drier areas, large rivers might not carry water for weeks or months. However, as streamflow observations are lacking for most drying rivers, we know little about when, where, and how long rivers experience such a streamflow intermittence that is crucial for both river life and human water supply. We developed and applied a novel approach to estimate, for the first time, the temporal dynamics of streamflow intermittence across European rivers and streams, including small ones. This approach combines the output of a global hydrological model with streamflow observations and other data. We refined the global model output available for 50 km cells to monthly streamflow in 500 m cells. We then applied a machine learning model to predict the number of days without water flow in each month during the period 1981-2019 for over 1.5 million river segments. We found that 17% of all European segments and 4% of all months at all segments experienced at least one day without flow. In the future, the model will be used to estimate the impact of climate change on streamflow intermittence.

1. Introduction

It has recently been estimated that most rivers and streams on Earth have reaches that naturally cease to flow or dry at least one day per year (Messager et al., 2021). Natural streamflow intermittence is most prevalent in semi-arid and arid regions, where it may occur even in large rivers, but it is also widespread in smaller headwater streams across humid regions. For example, 25–40% of the total length of streams and rivers in France are estimated to be intermittent (Snelder et al., 2013). In most basins, the likelihood and degree of streamflow intermittence, i.e., the fraction of no-flow days, increases with decreasing mean streamflow or upstream area (Datry et al., 2014;
Messager et al., 2021). Waterways can also cease to flow without being dry due to freezing conditions, so intermittent streams are a significant feature of cold landscapes as well (Buttle et al., 2012; Shanafield et al., 2021). Anthropogenic alterations of the natural flow regime resulting, for example, from human water abstractions or the operation of artificial reservoirs, can increase or decrease the number of no-flow days (Richter, 1997).

While streamflow intermittence can be monitored by measuring streamflow at gauging stations, these measurements come with numerous limitations (Zimmer et al., 2020) and only cover a very small part of all reaches, being particularly sparse where intermittent conditions prevail (Sauquet et al., 2021b; Krabbenhoft et al., 2022). In addition, streamflow observations are insufficient to derive projections of future changes in intermittence due to anthropogenic drivers, including climate change as well as artificial reservoirs and land and water use (Sauquet et al., 2021a; Döll & Müller Schmied, 2012). Therefore, comprehensive analyses of streamflow intermittence and its effects on water resources for humans and other biota require a modeling approach.

Large-scale modeling of streamflow intermittence is necessary for assessments of biodiversity, ecosystem functions and ecosystem services of rivers and streams at national to global scales. Until now, however, continental- or global-scale modeling studies on streamflow intermittence have either provided a static classification of river reaches into intermittent or perennial at high spatial resolution (15 arc-sec, ca. 500 m; Messager et al., 2021) or time series of intermittent streamflow conditions at a low spatial resolution (0.5 arc-deg, ca. 50 km; Döll & Müller Schmied, 2012). Messager et al. (2021) used Random Forest modeling to estimate which river reaches cease to flow at least one day per year or for at least thirty days per year; this was achieved for 23.3 million km of mapped rivers and streams across the globe (except Antarctica) whose long-term average naturalized discharge exceeds 0.1 m$^3$/s. Despite its fine resolution, such a static classification of reaches as either perennial or intermittent fails to characterize the temporal structure of flow intermittence (e.g., the number of no-flow days or seasonality of intermittence) which is required for analyzing the biodiversity and ecosystem functions of intermittent streams and rivers (Datry et al., 2018). By contrast, daily streamflow time series simulated by global hydrological models such as the WaterGAP model used in Döll and Müller Schmied (2012) do represent the temporal dynamics of streamflow intermittence. However, these coarser models overlook headwater stream reaches with small drainage basins, which are more prone to intermittence than larger downstream reaches and comprise the majority of global river length (Messager et al., 2021).

Simulating daily streamflow in small headwater streams requires small computational grid cell sizes (e.g., 500 m or less). Such small grid cells can easily be implemented in hydrological models if the drainage basin of study is small (Mahoney et al., 2023). However, this is not feasible across large geographic extents like entire continents or the world due to the lack of high-resolution climate data at these scales and computational constraints resulting from the large number of small (high-resolution) grid cells (Bierkens et al., 2015, Döll et al., 2016). A 0.5 arc-deg grid cell, typical for global hydrological models, contains 14,400 individual 15 arc-sec grid cells; in Europe alone (without Russia and Turkey), for example, there are about 73 million 15 arc-sec cells. Furthermore, hydrological models are often less successful in simulating low flows than mean flows (Zaherpour et al., 2018). Most hydrological models are process-based, i.e., they attempt to estimate water storage and fluxes across the different compartments of the terrestrial part of the hydrological cycle with sets of mathematical equations (Telteu et al., 2021). However, a satisfactory process-based simulation of low-flow, and particularly no-flow conditions, is very difficult even at small scales,
in part because the simulation of two-way exchange flows between surface water bodies and groundwater bodies requires coupling of a hydrological model with a gradient-based groundwater model (Döll et al., 2016). To help advance the science and management of freshwater ecosystems globally, new approaches are thus needed to produce large-scale high-resolution models of streamflow intermittence that provide information on the frequency, duration and timing of flow cessation across the entire river network, from the headwaters to river mouths.

Machine learning methods such as Random Forest (RF) have the advantage over process-based models that they do not require detailed knowledge of the processes underlying the phenomenon of interest and are thus a promising tool to produce large-scale high-resolution predictions of no-flow conditions. However, to achieve temporally explicit predictions, these models require temporally explicit predictors. The respective strengths of global hydrological models and machine learning methods can hence be combined by using the dynamic output of the former as an input predictor for the latter to achieve large-scale high-resolution modeling of the temporal structure of streamflow intermittence.

Here, we present such a combined modeling approach for computing monthly time series of streamflow intermittence conditions at the continental scale for river reaches that can be defined with a spatial resolution of 15 arc-sec. Our RF modeling approach combines temporally explicit predictor variables derived from the low resolution (LR, 0.5 arc-deg) state-of-the-art global hydrological model WaterGAP 2.2e (Müller Schmied et al., 2021) with several high-resolution (HR, 15 arc-sec) static predictor variables (e.g., drainage area and irrigated area). As part of this approach, WaterGAP LR output is spatially downscaled to derive HR monthly time series of streamflow. While all predictors used in the model are based on globally available data, the approach was developed using time series of daily streamflow observed at 3706 gauging stations throughout Europe (resulting in more than 1 million station-month with information on the number of no-flow days). It was then applied to estimate streamflow intermittence in Europe.

Section 2 presents the data and methods of this study. In section 3, the downscaled HR monthly streamflow time series are compared to observations at all gauging stations that were used to set up and calibrate the RF model. RF model performance and results of the RF application are presented in section 4. Section 5 provides validation and discussion of the streamflow intermittence modeling approach, while conclusions are drawn in section 6.

2. Methods and data

Below, we first explain the downscaling method applied to derive HR time series of streamflow from the LR output of the global hydrological model WaterGAP (Section 2.1). We then describe the compilation of the dataset of observed daily streamflow in Europe that was used for both validating the HR streamflow and for deriving the target data of the RF modeling approach (Section 2.2). This is followed by the description of the RF modeling approach, which consists of two sequential RF models (Section 2.3) and the definition of European river reaches for which model predictions are made (Section 2.4). In Section 2.5, performance metrics are explained.

The hydrographic dataset applied throughout this study is the global HR drainage direction map of HydroSHEDS v1 (Lehner et al., 2008; www.hydrosheds.org). This dataset represents, for each 15 arc-sec grid cell on land, the direction in which water would flow from that cell to its neighboring cells given topography. It serves to downscale LR outputs from WaterGAP, to co-
register streamflow gauging stations, to delineate river reaches for which the RF model produces predictions and to quantify predictors that are aggregated over the upstream areas of stations and reaches. In this study for Europe, HydroSHEDS was modified in three drainage basins (each about 200 km²) in Finland, Hungary and Croatia due to their use as case study basins in the related DRYvER project (see Döll et al., 2023a and https://www.dryver.eu/about/case-studies.

### 2.1. Downscaling of LR WaterGAP output to obtain time series of monthly HR streamflow

#### 2.1.1. WaterGAP

WaterGAP is a global water resources and use model that covers all continents except Antarctica (see Müller Schmied et al., 2021, for details). It computes time series of water use for irrigation, livestock, manufacturing, cooling of thermal power plants and households, distinguishing groundwater and surface water sources. It also simulates water flows (e.g., evapotranspiration and runoff) and water storages (e.g., in soil and groundwater), taking into account the impact of net abstractions from groundwater and surface water bodies as well as of artificial reservoirs. However, only operation of the globally largest 1109 artificial reservoirs (including regulated lakes) is simulated explicitly in WaterGAP, while smaller reservoirs only add to the fraction of each LR cell that is made up by of so-called local lakes, thus affecting evapotranspiration and flow dynamics in a very coarse way only. In WaterGAP, daily water flows and storages of 10 storage compartments are simulated in each LR grid cell. Total runoff from land is partitioned into fast (surface) runoff and groundwater recharge. Surface runoff from within a grid cell reaches surface water bodies (wetlands, lakes, reservoirs and rivers) on the same day, while groundwater recharge flows from the soil into the groundwater, which then releases groundwater discharge to surface water bodies as a function of groundwater storage. Only one river is assumed to exist within each LR grid cell, and the streamflow computed by WaterGAP refers to the outflow from the LR grid cell to the next downstream grid cell, which is prescribed by the LR drainage direction map DDM30 (Döll & Lehner, 2002). Groundwater discharge to surface water bodies may become zero in case of groundwater depletion, but loss of streamflow to the groundwater cannot be simulated. The LR WaterGAP output used in this study was computed by forcing version 2.2e of WaterGAP with the climate data set GSWP3-W5E5 (Müller Schmied et al., 2023a) for the time period 1901-2019. The model was calibrated against long-term mean annual streamflow observations observed at 1509 gauging stations globally (with a drainage area of at least 9,000 km²) by adjusting 1-3 model parameters.

#### 2.1.2. Downscaling approach

A number of approaches for generating time series of high-resolution streamflow from the output of global hydrological models were recently developed (Lin et al.; 2019, Kallio et al., 2021; Chuphal and Mishra, 2023). Our approach for downscaling the LR output of a global hydrological model to HR streamflow is based on the conceptual framework developed by Lehner and Grill (2013) which was globally applied, for example, in Linke et al. (2019). In this study, we generalized and adapted the approach, including some simplifications, to enable a computationally efficient generation of HR time series of monthly streamflow. As a distinct feature, the downscaling approach does not simply disaggregate and then route the sum of LR surface runoff and groundwater recharge (i.e., total runoff from land) along the HR river network, as this would disregard water retention in the groundwater and in surface water bodies, evaporation from surface water bodies and as well as human water use. Instead, our approach uses both surface runoff and
groundwater discharge estimates from the LR model and projects the results onto the HR river network using geospatial interpolation methods. Considering the original LR groundwater discharge estimates allows for better representation of HR streamflow variability because it takes into account the storage capacity of surface water bodies and groundwater aquifers that smoothen and delay the streamflow signal. Further corrections take into account the LR net cell runoff of WaterGAP, which, in addition to surface runoff and groundwater discharge, takes into account the dynamics of surface water bodies and human water use.

Here, we only describe the core elements of the downscaling method; for details see Supplementary Information Text S1. The sum of LR monthly surface runoff and groundwater discharge (expressed as specific volume flow per unit area, i.e., m³ s⁻¹ km⁻²) is first interpolated from 0.5 arc-deg to an intermediate resolution of 0.1 arc-deg to avoid abrupt changes in streamflow at the edges between LR cells. This is performed using an inverse distance interpolation with a power of 2 and taking into account the nearest 9 LR data points. A maximum interpolation radius of 1.8 arc-deg is allowed to extend data into areas where land cells are represented in the HR hydrography but not in the LR river network. This is the case in coastal regions and in missing cells within large lakes of the LR model. In the next step, the 0.1 arc-deg values are disaggregated to the 15 arc-sec HR grid cells by assigning the same 0.1 arc-deg value to all respective 15 arc-sec cells, assigning null values to HR cells outside of the continental boundaries of HydroSHEDS.

This disaggregated “runoff” $dR_{Li,Hj}$ of HR cell $j$ located within LR cell $i$, expressed as volume flow, is then corrected to integrate information from the routing routine of the LR model, in particular about the impact of surface water bodies and human water use on streamflow. Finally, HR streamflow $Q_{Li,Hj}$ in each month is computed by accumulating the sum of $dR_{Li,Hj}$ and a weighted correction term along the HR drainage direction, with

\begin{equation}
Q_{Li,Hj} = \text{flowacc}(dR_{Li',Hj'} + C_{Li'} * W_{Li',Hj'})
\end{equation}

\begin{equation}
C_{Li} = n_{cR_{Li}} - \sum_{j=1}^{14400} dR_{Li,Hj}
\end{equation}

\begin{equation}
W_{Li,Hj} = \frac{\text{flowacc}(dR_{Li,Hj})}{\sum_{j=1}^{14400} \text{flowacc}(dR_{Li,Hj})}
\end{equation}

where $C_{Li}$ is correction term for each LR cell $i$ (Equation 2), $W_{Li,Hj}$ is the correction weight to apply the correction term for each HR cell $j$ within LR cell $i$ (Equation 3), $\text{flowacc}()$ represents the flow-accumulated variable computed by summing the values of all HR grid cells ($L_i;H_j$) upstream of and including cell $L_i,H_j$. The net cell runoff of LR grid cell $i$, $n_{cR_{Li}}$, is calculated as the streamflow of LR grid cell $i$ minus the streamflow of all upstream LR grid cells, and is corrected for the different land masks (i.e., ocean-continent boundaries) of WaterGAP and HydroSHEDS.

The GHM WaterGAP computes streamflow not only by routing surface runoff and groundwater discharge. It additionally considers the impacts of reservoirs, lakes and wetlands as well as human abstractions of groundwater and surface water within each LR grid cell. These impacts are included in $n_{cR_{Li}}$. This is why the disaggregated runoff is corrected by $C_{Li}$, i.e., the difference between the LR $n_{cR_{Li}}$ and the sum of HR $dR_{Li,Hj}$ within the LR cell (Equation 2). For example, if human water use leads to a streamflow reduction in the LR cell, net cell runoff is lower than the sum of surface runoff and groundwater discharge ($n_{cR_{Li}} < \sum_{j=1}^{14400} dR_{Li,Hj}$), such that $C_{Li}$ becomes negative and HR streamflow is reduced, too.
Further correction terms are added to $C_{Li}$ as computed by Equation 2 in specific grid cells, such as HR endorheic sinks, outflow cells of lakes and reservoirs, or cells containing large rivers with a drainage area of more than 50,000 km² (see Supporting Information Text S1). The final correction term is applied in a spatially weighted way to the HR grid cells. With the help of the correction weight $W_{Li,Hi}$, a greater share of the total correction occurs in the downstream HR grid cells within each LR cell, which reflects the assumption that downstream HR cells are more affected by surface water bodies and human water use than upstream cells within the LR. The downscaling approach was implemented in Python.

2.2. Compilation and processing of measured streamflow for the computation of target observations and for the validation of simulated HR streamflow

Long-term historical information on the number of no-flow days per month in waterways, the target for the RF modeling, can only be derived consistently from continuous daily observations of streamflow at gauging stations. We collected most of these observations from the Global Runoff Data Centre (GRDC; https://www.bafg.de/GRDC) and the Global Streamflow Indices and Metadata archive (GSIM; Do et al., 2018; Gudmundsson et al., 2018), the largest existing global repositories of streamflow gauging station data. Altogether, daily streamflow records for 2930 GRDC and GSIM stations are available through these datasets for Europe. However, most of the GRDC and GSIM stations are on perennial streams, without any no-flow days in their record, which reflects the global underrepresentation of streamflow gauging stations on intermittent river reaches (Krabbenhoft et al., 2022). Therefore, we used metadata on gauging stations with flow intermittence in 19 European countries from the SMIRES meta-database (Sauquet, 2020) to obtain daily streamflow time series directly from national streamflow data providers for 375 additional gauging stations listed in the database. As flow intermittence in Europe is most prevalent in Mediterranean regions, we additionally retrieved daily streamflow data for a total of 55, 648 and 1031 gauging stations from governmental websites for Corsica (https://www.sandre.eaufrance.fr/), Italy (http://meteoniardo.altervista.org/) and Spain (https://ceh.cedex.es/anuarioafos/demarcaciones.asp), respectively.

From this compiled streamflow dataset, records suitable for deriving target observations were selected for subsequent analyses. We first checked whether each gauging station was correctly located on the updated 15 arc-sec HydroSHEDS drainage direction map by comparing the upstream area given in the metadata with the upstream area of the HR cell where the station was located. Confirmatory checks also involved inspecting high-resolution satellite imagery and comparing the river and station names provided in the metadata to topographic maps (ESRI ArcGIS basemaps). If the drainage areas deviated by more than 10%, the stations were manually relocated to a suitable HR grid cell with a deviation of less than 10% and/or associated to a river or stream with the correct name in topographic maps (if provided in the metadata). If this was not possible, the station was excluded from the RF modeling. For the remaining stations, we excluded all station-months with any missing or suspicious daily flow values following the approach of Gudmundsson et al. (2018). We then excluded all stations that had less than 36 station-months of daily streamflow data. Finally, we labeled all days with a mean streamflow of 0.001 m³ s⁻¹ or less as no-flow days and computed, as the target of the RF modeling, the number of no-flow days per month and station (i.e., per station-month). The maximum period with observed no-flow days and streamflow per station-month is 1981-2019 (468 months).

In total, data on streamflow at 3706 stations during 1981-2019 were used for calibrating and
validating the RF models, corresponding to 1,166,944 station-months (26 years of useable data per station on average). While 2.8% of the station-months were intermittent, 24% of the stations had at least one no-flow day. In addition, daily streamflow values were aggregated to monthly values for the same station-months to serve for the validation of the simulated HR streamflow (Section 2.1).

2.3. Random Forest modeling approach

2.3.1. Overview

The supervised machine learning method RF is well suited for both classification and regression tasks (Breiman et al., 2001). RF modeling has already been used for hydrological classification problems, i.e., for predicting classes of hydrological characteristics including intermittence (global: Messager et al., 2021; Australia: Bond and Kennard, 2017; France: Snelder et al., 2013). Tyralis et al. (2019) provide a review of RF methods with a focus on hydrological applications.

With less than 3% of all observed station-months in our European streamflow dataset being intermittent (Section 2.2), the dataset of target observations used for training the model is highly imbalanced, which can severely bias the resulting predictions (Japkowicz and Stephen, 2002). To mitigate this problem, two RFs are set up sequentially in our modeling approach. The first RF is developed to predict months with and without no-flow days (intermittent station-month and perennial station-months, respectively) in a binary way. The second RF is trained only with data for intermittent station-months to predict the number of no-flow days in four classes. The two calibrated RFs were then successively applied to predict the occurrence of five intermittence classes (0, 1-5, 6-15, 16-29 and 30-31 no-flow days per month) for each of the 468 months from 1981 to 2019 at more than 1.5 million river reaches in Europe (without Russia and Turkey, see Section 2.4 for the definition of reaches).

In this study, we used a derivative of the standard RF algorithm for making probabilistic predictions of class membership (Malley et al., 2012), which is included in the ‘ranger’ R package (Wright & Ziegler, 2017) that we used for the RF modeling. The ‘ranger’ R package is a fast implementation of RF suited for high-dimensional data (Tyralis et al., 2019). The two consecutive RF models were trained and optimized by cross-validation, i.e., calibrated, by relating observations of the number of no-flow days per station-month at streamflow gauging stations (target of RFs) to 23 predictors, 9 of them temporally-explicit (i.e., dynamic). For the RF training, each streamflow station is assigned to a HR grid cell. For RF predictions, each European river reach is assigned to the HR cell containing its downstream end. In the following two sections, the compilation and processing of the predictor variables and the 2-step RF modeling approach are explained.

2.3.2. Predictors

A total of 23 predictor variables were used in both RFs (Table 1). We selected predictors based on their potential causal influence on streamflow intermittence. In contrast to multiple regression analysis, RF can leverage information from highly correlated predictors while producing unbiased predictions (Tyralis et al., 2019). Five HR dynamic predictors were derived from the WaterGAP HR monthly streamflow time series and indicate the streamflow conditions in each HR stream reach (represented by its respective most downstream HR cell) in the current month and past 3 and 12 months. Prior to computing these predictors, streamflow was converted into area-specific streamflow by dividing it by the drainage area of the HR cell (i.e., the area of the upstream drainage basin). This is done because the spatial variation of streamflow is mainly due to the area that drains...
to the point of the streamflow observation and streamflow intermittence is expected to depend on area-specific streamflow. In addition, drainage area of the HR cell was included as a static predictor (Table 1). Of the 2 LR dynamic predictors, one is the ratio of diffuse groundwater recharge to total runoff from land for each month as computed by WaterGAP; a higher ratio of delayed groundwater discharge is expected to decrease the likelihood of no-flow days. The other LR dynamic predictor is the average number of days with substantial precipitation (i.e., >2.5 mm/d) per month according to the WaterGAP climate forcing; a low number of days with substantial precipitation in a month may lead to an increased likelihood of no-flow days. For both LR dynamic predictors, the average value over the upstream basin of each HR grid cell was computed assuming that the values in all upstream HR cells are identical within a given LR cell. The 5 HR and 2 LR dynamic predictors vary between the 468 months of the study period. Three of the 16 static HR predictors vary with the calendar month, the two predictors that quantify the interannual variability of monthly streamflow and the aridity index, which is included as the long-term mean values for the 12 calendar months. Interannual variability was computed from the HR monthly time series of area-specific streamflow as either the standard deviation or the coefficient of variation of all streamflow values of each of the 12 calendar months for the period 1981-2019 (Table 1).

In addition to drainage area, the other 12 HR static predictors include the dominant potential natural and actual land cover class in the upstream basins and the average slope, glacier fraction and lake fraction in the upstream basin. These static HR predictors are selected from the set of globally important predictors from Messager et al. (2021). Additional static predictors include five suspected anthropogenic drivers of streamflow intermittence and two newly developed karst-related predictors derived from the World Karst Aquifer Map (WOKAM) of karstifiable areas (Table 1). In the case of the static predictors of karst status and degree of regulation, the value for the HR grid cell for which the number of no-flow days is to be predicted (target cell) is used as a predictor. In the case of the anthropogenic drivers irrigated area fraction and human population, two sets of predictor values are taken into for each: one set of values computed by aggregating over the (total) upstream basin and the other computed by aggregating over the immediate upstream basin, which only encompasses upstream HR grid cells that drain directly into the respective stream reach (see Linke et al., 2019, for additional descriptions of these spatial units).

To train the RF models, the values of these predictors were assembled for each of the 1,166,944 station-months for which daily streamflow observations are available, i.e., for the 3706 HR grid cells that contain a gauging station. For model application, the predictor values for each reach (i.e., for the most downstream HR grid cell of each reach) were computed to predict the occurrence of one of the five intermittence classes for each reach-month.
Table 1

Predictors Used in RF Modeling, with their Abbreviations, Units and Data Sources

<table>
<thead>
<tr>
<th>Category</th>
<th>Predictor type</th>
<th>Predictor</th>
<th>Abbreviation (unit)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrology</td>
<td></td>
<td>Monthly area-specific streamflow</td>
<td>Q ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td></td>
<td>Minimum monthly area-specific streamflow of the past 12 months</td>
<td>Q_min_p12 ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Monthly time series</td>
<td>Mean monthly area-specific streamflow of the past 12 months</td>
<td>Q_mean_p12 ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Monthly time series</td>
<td>Minimum monthly area-specific streamflow of the past 3 months</td>
<td>Q_min_p3 ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td>Monthly time series</td>
<td>Mean monthly area-specific streamflow of the past 3 months</td>
<td>Q_mean_p3 ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td></td>
<td>Ratio of diffuse groundwater recharge to runoff from land, mean over ub^a</td>
<td>gwr_to_runoff_ratio (-)</td>
<td>WaterGAP 2.2e</td>
</tr>
<tr>
<td>Climate</td>
<td></td>
<td>Number of wet days, mean over ub^a</td>
<td>wet_days (days mon^{-1}/100)</td>
<td>WaterGAP 2.2e</td>
</tr>
<tr>
<td>Hydrology</td>
<td></td>
<td>Interannual variability of monthly area-specific streamflow, per</td>
<td>Q_iav_sd ($m^3 s^{-1} km^{-2}$)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calendar month, in terms of standard deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydrology</td>
<td></td>
<td>Interannual variability of monthly area-specific streamflow, per</td>
<td>Q_iav_cv (-)</td>
<td>Downscaled WaterGAP 2.2e</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calendar month, in terms of coefficient of variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td>Static HR</td>
<td>Aridity index (long-term average P/PET), per calendar month, mean over</td>
<td>P_to_PET_ratio (1/10000)</td>
<td>Global-AL_PET_v3^c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ub^a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Land cover</td>
<td></td>
<td>Potential natural vegetation classes (ranges: 1-15), spatial</td>
<td>pot_nat_vegetation (-)</td>
<td>EarthStat^d</td>
</tr>
<tr>
<td>Category</td>
<td>Description</td>
<td>Code</td>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>Land cover</td>
<td>Land cover classes (ranges: 1-22), spatial majority in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>land_cover (-)</td>
<td>GLC2000&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Land cover</td>
<td>Glacier area fraction in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>glacier_frac (%)</td>
<td>GLIMS&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Physiography</td>
<td>Drainage area</td>
<td>drainage_area (km&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>HydroSHEDS&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Physiography</td>
<td>Terrain slope, mean over ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>slope (deg/100)</td>
<td>EarthEnv-DEM90&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Geology</td>
<td>Fraction of karst area in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>karst_frac (%)</td>
<td>WOKAM&lt;sup&gt;i&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Geology</td>
<td>Occurrence of karst (1 if karst, 0 if not) at HR grid cell</td>
<td>karst_status (-)</td>
<td>WOKAM&lt;sup&gt;i&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Anthropogenic</td>
<td>Fraction of area equipped for irrigation in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>irri_frac (% /100)</td>
<td>HID v1.0&lt;sup&gt;j&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Anthropogenic</td>
<td>Fraction of area equipped for irrigation in iub&lt;sup&gt;b&lt;/sup&gt;</td>
<td>irri_frac_im (% /100)</td>
<td>HID v1.0&lt;sup&gt;j&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Geology</td>
<td>Population density in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>pop_dens (people km&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>WorldPop&lt;sup&gt;k&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Geology</td>
<td>Population density in iub&lt;sup&gt;b&lt;/sup&gt;</td>
<td>pop_dens_im (people km&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>WorldPop&lt;sup&gt;k&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Anthropogenic</td>
<td>Degree of regulation (total upstream artificial reservoir storage volume / annual streamflow volume) at HR grid cell</td>
<td>dor (% /10)</td>
<td>HydroSHEDS&lt;sup&gt;g&lt;/sup&gt; &amp; GranD&lt;sup&gt;i&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Lakes</td>
<td>Fraction of lake area in ub&lt;sup&gt;a&lt;/sup&gt;</td>
<td>lake_frac_(-(% /100)</td>
<td>HydroLAKES&lt;sup&gt;m&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

Note. Area-specific streamflow is streamflow at the HR grid cell divided by upstream drainage area. The units are those for the data sets used as input to the RF modeling, in which the integer values were partly multiplied by 10, 100 or 1000 to increase the precision. <sup>a</sup>ub: HR upstream basin. <sup>b</sup>iub: HR immediate upstream drainage basin, refers to all the HR grid cells that drain directly into the respective stream reach. <sup>c</sup>Zomer et al. (2022). <sup>d</sup>Ramankutty and Foley (1999). <sup>e</sup>Bartholomé and Belward (2005). <sup>f</sup>GLIMS & NSIDC (2012). <sup>g</sup>Lehner et al. (2008). <sup>h</sup>Robinson et al. (2014). <sup>i</sup>Chen et al. (2017). <sup>j</sup>Siebert et al. (2015). <sup>k</sup>Bondarenko et al. (2020). <sup>l</sup>Lehner et al. (2011). <sup>m</sup>Messager et al. (2016).
2.3.3. Two-step RF modeling approach

The first RF model in our approach (Figure 1, step 1) results in a binary classification of station-months as either intermittent or perennial, whereas the second RF model (step 2) was only applied to intermittent station-months and classified them into four ordinal intermittent classes: 1-5, 6-15, 16-29 and 30-31 no-flow days per month. We performed a classification into four classes based on a previous study with a two-step RF model (with less target observations and slightly different predictors) where the performance for six classes was not satisfactory (Döll et al., 2023a). The four classes were defined such that they are informative for biodiversity and ecosystem function studies while keeping the number of observations per class approximately balanced. Following model training and validation for each of the RFs (Figure 1, left-hand side), we sequentially applied the calibrated models (right-hand side) to predict monthly streamflow intermittence for all reaches in Europe derived from the HR drainage direction map (see Section 2.4).

Despite implementing the two-step approach, class imbalance persists in each step of the modeling process, with many more perennial station-months than intermittent ones in step 1 and a relatively
large number of station-months with 30-31 no-flow days in step 2. Therefore, we applied standard
oversampling of the minor class (intermittent) in step 1 by a factor of 34.68, the ratio of perennial
to intermittent months. In step 2, the three minor classes were oversampled such that, for each
minority class, the number of training observations in that class was equal to the number of
observations in the majority class (30-31 no flow days).

Unlike hydrological models, RF models, which are ensembles of classification trees, do not have
parameters that represent properties of the modeled system. However, they have hyperparameters
that determine how the RF algorithm exactly works, which can be tuned to maximize predictive
accuracy while minimizing overfitting. RF hyperparameters are 1) the sample fraction, i.e., the
fraction of the training data that is randomly sampled without replacement for generating each
individual tree, 2) the number of predictors that are sampled from the full set of predictors and
considered by each tree when splitting each node (MTRY) and 3) the minimum number of
observations that a terminal node can contain, which influences the depth of the trees (i.e., when
tree construction stops). Model performance increases asymptotically with the number of decision
trees. In this study, the number of decision trees was set to 800 to limit run times.

In each step, the RF was tuned and evaluated by twice-repeated three-fold nested cross-validation.
Nested cross-validation, a resampling method that combines two levels of cross-validation loops
(outter and inner loops) separates hyperparameter tuning in the inner loop from model performance
evaluation in the outer loop (Bischl et al. 2012). In each loop, cross-validation uses different
portions of the data to iteratively test and train a model on the different subsets of the data. A three-
fold cross-validation means that the RF is trained with a random selection of two-thirds of the
samples (training data), each sample consisting of the predictors and the target for one station-
month. The predictive accuracy of the model is then evaluated with the remaining third of the
samples (testing data). In a twice-repeated three-fold nested cross-validation, there are six rounds
of cross-validation in total with different training and test data. Hyperparameter tuning for RF
step 1 and step 2 was performed through 15 and 55 unique combinations of hyperparameter,
respectively. For each round, hyperparameters were tuned by evaluating the performance of 15
and 55 unique combinations of hyperparameters in the case of the step 1 RF and step 2 RF,
respectively.

Model validation (Sections 4.1 and 4.2) was done using the results of the six rounds of cross-
validation, i.e. the results of the six RF models with an optimal combination of hyperparameters
as determined by the inner loop. For each station-month, the two predicted probabilities of
pertaining to a certain class were averaged and the class was assigned. The threshold for assigning
the perennial or intermittent class was set to a probability of 50%, consistent with our efforts to
balance the training data set.

The RF showing the highest balanced accuracy (BACC; Section 2.5) across all six rounds was
used for model application (Section 4.3). This resulted in a calibrated RF model consisting of the
best-performing step 1 RF and the best-performing step 2 RF. For step 1, the optimal values for
sample fraction, MTRY and minimum number of observations for the terminal node were 0.25, 4
and 2, respectively; the corresponding values for step 2 were 0.75, 6 and 10, respectively.

We computed the relative contribution of predictors to the predictive ability of the model, in the
form of the Actual Impurity Reduction (AIR) predictor importance metric. The higher the AIR,
the more important the predictor. The role of predictor variables was also evaluated with partial
dependence plots, which depict the marginal relationship between each predictor variable and the
probability of a predicted class while holding the rest of the predictors at their respective mean values. Using 20 processors (Intel Xeon silver 4114 2.2 GHz) in parallel, the run time for setting up the step 1 RF was about 14 days, and 14 hours for setting up the step 2 RF.

2.4. Definition of stream reaches for model application

It would be computationally too expensive to estimate the streamflow intermittence status for all HR grid cells in Europe, regarding both computation time and data storage. With 73 million HR grid cells across Europe and 468 months (1981-2019), more than 34 billion predictions would have to be computed. Therefore, we applied the two RF models sequentially to predict the streamflow intermittence status of river reaches rather than individual grid cells. Predictions are made for the most downstream HR grid cell of each river reach and are assumed to represent the mean conditions over the whole river reach.

River reaches at the HR resolution of 15 arc-sec are available in HydroSHEDS (HydroRIVERS, Lehner and Grill, 2013, https://www.hydrosheds.org/products/hydrorivers) but they insufficiently cover headwater streams for the purpose of our study (Döll et al., 2023a); in addition, we had slightly modified the HydroSHEDs drainage direction map. Therefore, river reaches were newly generated from the modified HydroSHEDS HR drainage direction map by applying the following delineation thresholds: streams were defined to start at all HR grid cells with an upstream drainage area of more than 2 km² (instead of 10 km² in 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 m³/s (instead of 0.1 m³/s in HydroRIVERS). Decreasing the threshold for streamflow to 0.02 m³/s would lead to potential "aggregates" of multiple streams in one grid cell in wet areas. Using these delineation thresholds, the resulting number of reaches in Europe is 1,533,471, with an average reach length of 2.0 km (standard deviation 1.7 km), representing a total stream network length of 3.06 million km. Accordingly, the European data set of monthly streamflow intermittence status contains a total of 717,664,428 reach-months covering the period 1981-2019.

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 delineated inside the boundaries of lakes and artificial reservoirs. Users of the streamflow intermittence dataset may therefore need to mask out simulated reaches as appropriate.

2.5. Performance metrics

As the observation data were strongly imbalanced, we evaluated model performance through the cross-validation of the two RFs based on the BAlanced ACCuracy (BACC). BACC provides a better indication of the classification performance of imbalanced models than raw accuracy (the percentage of correctly classified observations). In the binary case of step 1, BACC is the mean of sensitivity and specificity, with

\[
sensitivity = \frac{TP}{(TP+FN)}
\]

\[
specificity = \frac{TN}{(TN+FP)}
\]

where TP: true positive, FN: false negative, TN: true negative and FP: false positive, resulting from the confusion matrix (Figure S1). In the multiclass case of step 2, we follow the definition of Urbanowicz & Moore (2015) whereby the mean of sensitivity and specificity is calculated for each
of the four classes and then averaged over the classes.

Model performance was also evaluated with the Nash-Sutcliffe efficiency (or model efficiency), a traditional performance metric in hydrological modeling. It provides an integrated measure of model performance concerning mean values and variability and is computed as

\[
\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(\text{sim}(t) - \text{obs}(t))^2}{\sum_{t=1}^{n}(\text{obs}(t) - \mu_{\text{obs}})^2}
\]

where \(\mu_{\text{obs}}\) is the mean of observations across all time steps; \(\text{sim}(t)\) and \(\text{obs}(t)\) refer to the simulated and observed values respectively, at time-step \(t\) of a total number of time steps \(n\). NSE can range from \(-\infty\) to 1; a value of 0 indicates that the model performs no better than simply using the mean of the observed data to predict the values, and a value of 1 indicates perfect agreement between the observed and modeled values.

### 3. Validation of HR time series of monthly streamflow in Europe

Comparing the downscaled HR monthly streamflow time series to the monthly time series of observed streamflow at the 3706 gauging stations across Europe yielded a median NSE value of 0.41; NSE exceeds 0 for 69% of the stations, and 25% of stations exceed the value of 0.64 which indicates a relatively good performance. When NSE is computed with the logarithm of streamflow, which puts a larger weight on low-flow months of interest for intermittence, NSE exceeds 0 for 63% of stations and 0.57 for 25% of stations. This shows that streamflow during the low-flow months is also estimated reasonably well. However, the performance of simulated HR streamflow is very poor in most of Spain, where human activities strongly impact streamflow (Figure 2). Although the impact of artificial reservoirs as well as groundwater and surface water abstractions are simulated by WaterGAP, the coarse resolution of the original model calculations (at LR grid cells) prevents the identification of the specific locations of these impacts in the downscaling procedures. Also, the HR location of natural surface water bodies, i.e., lakes and wetlands, is not explicitly taken into account in the downscaling method, causing potential misallocation of their attenuating effects on HR streamflow. Furthermore, other anthropogenic disturbances such as weirs are not accounted for in the original WaterGAP estimates. A poorer performance of HR streamflow in strongly altered streams is therefore due to both downscaling constraints and the difficulty of simulating human impacts at the LR resolution.
The continental area considered to belong to Europe in this study is shown in grey. The performance of the downscaling algorithms can be assessed by comparing the NSE values at gauging stations with different upstream areas (Figure 3). LR streamflow as computed by WaterGAP is generally only compared to streamflow observed at gauging stations with upstream areas of more than 10,000 km², as a single LR grid cell can cover more than 2,500 km² (Müller Schmied et al. 2021). The high uncertainty of the global climate datasets used as the input of WaterGAP also inherently limits model performance for smaller basins. The performance of simulated streamflow does not decrease much with decreasing upstream area of the gauging stations (Figure 3a). For example, the median NSE for drainage basins larger than 10,000 km² is 0.51, while it is only slightly lower at 0.38 for the smallest drainage basins with areas below 2 km². The median NSE of logarithmic streamflow decreases from 0.40 for the basins larger than 10,000 km² to 0.14 for basins smaller than 2 km² (Figure 3b). When interpreting the NSE values, it should be noted that the stations are not equally distributed between the different catchment area classes; for instance, there are less than 100 stations with an upstream area of less than 10 km². Furthermore, this relationship between predictive performance and catchment area is not consistent among stations on intermittent and perennial waterways. In the case of intermittent stations (n=885), there is a decline in NSE values from basins with upstream areas of less than 50 km² to basins larger than 10,000 km² (Figure S2), whereas the opposite is true for perennial stations.
While small intermittent basins smaller than 2 km$^2$ are characterized by a median NSE of 0.49 (median NSE for log streamflow: 0.21), the large basins over 10,000 km$^2$ show a very poor performance with a median NSE of less than 0 (Figure S3). This might be due to the difficulty of simulating the impact of reservoir operations on intermittence. Considering the size class of 50-500 km$^2$, which include the most stations of both intermittent (>100 stations) and perennial types (>1000 stations), the median NSE is 0.23 for intermittent stations and 0.43 for perennial stations (Figures S2 and S3).

**Figure 3.** NSE of monthly streamflow time series (left) and of the logarithm of monthly streamflow time series (right) for all 3706 streamflow stations with observations, grouped in size classes of the upstream area of the streamflow gauging stations. The boxes indicate the 25$^{th}$, 50$^{th}$ (median) and 75$^{th}$ percentiles, the whiskers indicate the 5$^{th}$ and 95$^{th}$ percentiles of the samples. The blue lines of the violin plot show the smoothed distribution of the data points. The “number of stations not shown” indicates the number of stations with an NSE of less than -1.

**4. RF modeling results**

**4.1. Model validation**

**4.1.1. Step 1 RF**

The cross-validation of the calibrated step 1 RF resulted in a BACC of 0.92. Of all perennial station-months, 98% were correctly identified as perennial, i.e., without any no-flow day (Table 2). Consequently, 25,496 (2%) of all perennial station-months were erroneously identified as intermittent. Of the intermittent station-months, 86% were correctly identified as intermittent, i.e.,...
4,463 intermittent months were wrongly identified as perennial. Thus, the step 1 RF tends to overestimate the occurrence of intermittent months in absolute terms. In Europe, streamflow intermittence is more prevalent in the summer (JJA) and in the fall (SON) than in winter (DJF) and spring (MAM), and this is also the case for the number of predicted intermittent months (Table 2). A higher percentage of intermittent station-months, about 88%, was correctly identified as intermittent in JJA and SON than in the other two seasons (Table 2).

### Table 2

**Number of Observed and Correctly Simulated Perennial and Intermittent Stations and Station-months**

<table>
<thead>
<tr>
<th></th>
<th>Number of stations</th>
<th>Number of station-months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>DJF</td>
</tr>
<tr>
<td>Correctly simulated as perennial</td>
<td>2806</td>
<td>1,108,741</td>
</tr>
<tr>
<td>Observed as perennial</td>
<td>2821</td>
<td>1,134,237</td>
</tr>
<tr>
<td></td>
<td>99.5%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Correctly simulated as intermittent</td>
<td>551</td>
<td>28,244</td>
</tr>
<tr>
<td>Observed as intermittent</td>
<td>885</td>
<td>32,707</td>
</tr>
<tr>
<td></td>
<td>62.3%</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

Note. Observed (bottom numbers) and correctly simulated (top numbers). Information on station-months is provided for all months and the four seasons December to February (DJF), March to May (MAM), June to August (JJA) and September to November (SON).

The overestimation of intermittent months mainly occurs at stations that are both observed and simulated to be intermittent, i.e., stations that have at least one no-flow day in the whole period 1981-2019, as only 15 perennial gauging stations, scattered throughout Europe, were erroneously predicted to be intermittent (dark red symbol in Figure 4b). Thus, 99.5% of all 2,821 stations observed to be perennial were correctly simulated to be perennial (Table 2, grey symbols in Figure 4b). The 885 gauging stations with at least one intermittent month, i.e., 24% of all stations considered in this study, are particularly concentrated on the Iberian Peninsula, Sardinia and Cyprus (Figure 4a), where gauging stations commonly recorded more than 20% of intermittent months. Elsewhere, almost all intermittent stations have less than 20%, and mostly less than 10%, of intermittent months. No intermittence is observed in winter months in the northern parts of Scandinavia, even though no-flow conditions are commonly reported in these climates because of dry conditions, the storage of precipitation as snow, and freezing (Buttle et al., 2012). Intermittence was not even observed at a station on a northern Norwegian island with a small drainage area of 19 km². Intermittence at the stations in western Finland occurs only in the summer and only at
stations with small upstream areas. The two intermittent stations in northern Sweden are located downstream of large artificial reservoirs.

Over a third of intermittent stations (334 out of 885) were wrongly simulated to be perennial by the step 1 RF (dark blue dots in Figure 4b); these stations are distributed across Europe with no clear spatial clustering. Many of these stations are located on streams that normally flow year-round but that exceptionally dried, for example, during a severe drought. Indeed, these intermittent stations that were wrongly classified as perennial have a median of only 2 intermittent months across their entire record (range: 1-19 months), while the 551 correctly classified intermittent stations have a median of 35 months (range 2-431 intermittent months).

When considering only the 885 intermittent stations, the median and mean percent of observed intermittent months are 5.6% and 15.8%, respectively. Whereas 86% of all observed intermittent stations months (28,244 out of 32,707) are correctly predicted to be intermittent, 11% of observed perennial station-months at intermittent stations (25,398 out of 233,195) are wrongly predicted to be intermittent. This resulted in a general overestimation of the total share of station-months at intermittent stations. While 13% of all station-months at intermittent stations are observed to be intermittent (and 11% correctly predicted as such), 21% are predicted to be intermittent. The overestimation is concentrated in regions with a relatively high prevalence of intermittence, i.e., large parts of the Iberian Peninsula, Sardinia and Cyprus (compare Figures 4a and 4b), where intermittent months are often overestimated by a factor of more than 2 (Figure 4b). The main suspected reasons for this overestimation are the poor ability of the downscaled streamflow estimates (Figure 2) and the RF model to capture the strong human impacts on streamflow dynamic in large parts of Spain as well as Cyprus (not Sardinia). In these semi-arid regions, a multitude of small and large dams as well as water transfers by canals often make naturally intermittent streamflow perennial (Chiu et al., 2017). Even though some large reservoirs are considered when computing LR net cell runoff used to estimate HR streamflow, simulation of reservoir outflow is very uncertain already at LR. In addition, information on reservoirs, weirs or canals in the individual HR cells within each LR is not taken into account in the streamflow downscaling approach. The reservoirs included in the computation of the static HR predictor by the predictor dor (degree of regulation by upstream dams; Lehner et al., 2011) (Table 1) only takes into account a subset of the actual reservoirs and misses small ones.

Interannual variability of the number of intermittent months per year is simulated quite satisfactorily, in particular for gauging stations in southern Spain (Figure 4c). Considering all 885 intermittent stations, the median Pearson correlation coefficient between the observed and predicted annual time series of the number of intermittent months is 0.50. Thus, the step 1 RF is able to capture the interannual variability of climatic conditions. That said, the corresponding NSE values (i.e., based on the annual time series; not shown) are below zero at almost all stations due to the strong overall overestimation of intermittent months.
Figure 4. Percentage of observed intermittent months (with at least one no-flow day) per gauging station for all observations during 1981-2019 (a), ratio of the number of predicted months to the number of observed intermittent months (P: perennial, I: intermittent) (b) and Pearson correlation of the annual time series of the number of intermittent months (c), as simulated by the step 1 RF model.

As expected, gauged streams in smaller drainage basins are both observed and simulated to be more strongly intermittent than larger drainage basins, especially in the two smallest drainage basin classes 0-2 km² and 2-5 km² (Figure S4). However, intermittent months are also most overestimated in these size classes; the predicted median proportion of intermittent months for these stations is twice the observed median of about 13%. For drainage basins larger than 2500 km², on the contrary, the step 1 RF tends to underestimate the already low percentage of intermittent months (though it strongly overestimates intermittence for a few basins, too; Figure S4).
4.1.2. Step 2 RF

The target of the step 2 RF are the observations of the number of no-flow days, in four classes (1: 1-5, 2: 6-15, 3: 16-29, 4: 30-31 no-flow days) in observed intermittent months. At most intermittent gauging stations, class 1 (1-5 no-flow days) dominates, whereas class 4 (30-31 no-flow days) dominates in many stations with more than 10% of intermittent months, in particular in the central and southern part of the Iberian Peninsula and in Cyprus (Figure S5). With a BACC of 0.67 (averaged over the four classes) in the cross-validation of the calibrated step 2 RF, the classification performs satisfactorily. More than three quarters of the station-months with observed class 4 (30-31 no-flow days) are correctly classified, and almost half of the station-months with 1-5 and 16-29 no-flow days are correctly classified (Figure 5). Although the model exhibits weaker performance for station-months with 6-15 observed no-flow days, these observations are still more likely to be correctly classified than pertaining to any of the three other classes. Classification performance is highest for the class with most observations, 30-31 no-flow days, as can be expected in RF modeling. In total, 54% of the 32,707 station-months are classified into the correct observed class, and of the wrongly classified observations, 70% are predicted to belong to neighboring classes (Figure 5).

Figure 5. Confusion matrix of predicting four classes of no-flow days per station-month. The top number in each box shows total number of station-months belonging to the observed and simulated intermittence class, the bottom number the percent of the total number of station-months that are observed to be in the intermittence class (step 2 RF model).
The percentage of intermittent months that are correctly classified into the four classes shows no spatial pattern across Europe (Figure 6a), although the overestimation of no-flow days is most pervasive in Spain where the number of observed no-flow days is already high (red in Figure 6b). The step 2 RF tends to overestimate the number of no-flow days in the intermittent station-months where the step 1 RF also overestimates the number of intermittent months (e.g., in many stations on the Iberian Peninsula). The bias shown in Figure 6b correlates weakly with the ratio of predicted to observed intermittent months shown in Figure 4b, with a Pearson correlation coefficient of 0.11. The correlation between the monthly time series of observed and simulated intermittence classes, as measured by the Spearman rank correlation coefficient, is positive for most gauging stations, and larger than 0.3 for 38% of stations (Figure 6c). This correlation analysis does not include the perennial months at a station. The overall performance of the monthly time series of five classes, with class 0 for perennial months, reflects the combined performance of the step 1 and step 2 RFs and thus the overall RF modeling approach used for estimating streamflow intermittence for all reach-months in Europe. These correlation values, shown in Figure 6d, are much higher than the correlation for just the intermittent months; values larger than 0.9 dominate. The median Spearman rank correlation coefficient for the monthly time series of the five intermittence classes is 0.81, with 90% of the stations exceeding a value of 0.58 and 14% of the stations exceeding a value of 0.99.
Figure 6. Percentage of intermittent months that are correctly classified into the four classes (1: 1-5 no-flow days per month, 2: 6-15, 3: 16-29, 4: 30-31) by the step 2 RF at each of 885 gauging stations with at least 1 no-flow day in their record (a), bias expressed as simulated mean class number (1 through 4) minus observed mean class number (green: correct average classification, red: overestimation of no-flow days, blue: underestimation of no-flow days) (b), and Pearson correlation coefficient for the monthly time series of simulated and observed intermittence classes, for four classes 1-4 (c) and five classes 0-5, with class 0: 0 no-flow day (d). All correctly classified perennial stations were omitted from the maps and would show a correlation coefficient of 1.

4.2. Importance of predictors and dependence of predicted class on predictor values

All 23 predictors were found to be significant at the p-value = 0.05 level. The relative importance of the 23 predictors differs strongly between step 1 RF (identifying whether a station-month is
intermittent) and step 2 RF (identifying the number of no-flow days in intermittent months, in four
classes) (Figure 7). However, two predictors computed from the downscaled HR monthly
streamflow, namely the monthly area-specific streamflow (Q) and the mean of the area-specific
streamflow of the previous three months (Q_mean_p3), are among the five most important
predictors in both RFs. Both are predicted to be negatively correlated to the probability of
intermittence, as was expected (Figure S6).

The most important predictor in the step 1 RF is the size of the drainage basin of the streamflow
gauging station (Figure 7), with the probability of intermittence decreasing with increasing size up
to a drainage area of about 20,000 km² (Figure S6). Terrain slope (slope) and the precipitation to
potential evapotranspiration ratio (P_to_PET_ratio) show a similar importance in step 1 and take
up ranks 4 and 5, respectively. The partial dependence plots for the step 1 RF show, for all but 2
of the 23 predictors, correlations between the predictor and the likelihood of intermittence that are
expected by hydrologists. For example, 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 for Q_iav_cv > 0.4. Exceptions to
this correspondence between model predictions and hydrological understanding include the terrain
slope (slope) and, albeit less conclusively, the degree of regulation (dor) (Figure S6). Steeper
slopes across the upstream drainage area are expected to make intermittence more likely
(Šarauškienė et al., 2020) due to a decrease in the fraction of runoff that recharges groundwater
and thus a decrease in baseflow, but the RF predicted the opposite correlation. This 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, i.e., wet regions with large runoff. As for the degree of regulation, artificial reservoirs
can make streams either more perennial or more intermittent, depending on reservoir management
(e.g., for hydroelectricity, irrigation, flood control) and river type (Datry et al., 2023). Here, the
step 1 RF showed that increased regulation was associated with greater levels of intermittence
(Figure S6). A likely reason for this correlation is that many stations downstream of large dams in
our training dataset were located in dry areas like Spain, where intermittence is common and flow
regulation by reservoirs is associated with extensive water withdrawal (Sabater & Tockner, 2009).
This predictor’s importance in RF 1 is very low (Figure 7), so the impact of this counterintuitive
relationship on model predictions is minor.

In the step 2 RF, all of the five most important predictors are dynamic predictors. They include
four HR predictors derived from the downscaled WaterGAP output (Figure 7). In addition to the
highest ranking Q and Q_mean_p3, the minimum area-specific streamflow over the previous three
months (Q_min_p3) and the mean area-specific streamflow over the previous 12 months
(Q_min_p12) are among the five most important predictors. The LR predictor of the number of
wet days per month is ranked second in importance.
Figure 7. Predictor importance for step 1 RF (left) and step 2 RF (right). The higher the impurity reduction, the larger the relative importance of a predictor. The higher absolute values for the step 1 RF are due to the larger number of station months available as target. Error bars show the standard deviation across the six cross-validation training sets calculated for both the step 1 and step 2 RFs. The relatively larger error bars for the step 2 RF are due to considering four classes instead of only two in the step 1 RF. Dynamic HR predictors are indicated by * and dynamic LR predictors by +.

4.3. Predicted time series of monthly streamflow intermittence status of stream reaches in Europe

In total, 96.2% of the approximately 718 million reach-months at more than 1.5 million stream reaches in Europe are simulated as perennial in the period 1981-2019 (Table 3). 82.2% of the stream reaches and 81.3% of the European network length of 3.06 million km are simulated to never have experienced a no-flow day during this period. Reaches with intermittent months are simulated to exist in almost all European countries, but high percentages of intermittent months are prevalent on the Iberian Peninsula, Sardinia and Cyprus and also occur in southern Italy and
Greece (Figure 8). Large regions with low fractions of intermittent months exist in France but also in Finland, Belarus and Ukraine.

Table 3

Occurrence of the Five Intermittence Classes at the Gauging Stations and All Reach-months in Europe

<table>
<thead>
<tr>
<th>Class</th>
<th>Station-months</th>
<th>Reach-months</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Predicted</td>
<td>Predicted</td>
</tr>
<tr>
<td>0: Perennial</td>
<td>1,134,237</td>
<td>1,113,204</td>
<td>690,269,534</td>
</tr>
<tr>
<td></td>
<td>97.20%</td>
<td>95.39%</td>
<td>96.20%</td>
</tr>
<tr>
<td>1: 1-5 no-flow days</td>
<td>5,643</td>
<td>5,248</td>
<td>413,786</td>
</tr>
<tr>
<td></td>
<td>0.48%</td>
<td>0.45%</td>
<td>0.06%</td>
</tr>
<tr>
<td>2: 6-15 no-flow days</td>
<td>6,030</td>
<td>5,338</td>
<td>549,107</td>
</tr>
<tr>
<td></td>
<td>0.52%</td>
<td>0.45%</td>
<td>0.08%</td>
</tr>
<tr>
<td>3: 16-29 no-flow days</td>
<td>8,634</td>
<td>8,484</td>
<td>1,742,476</td>
</tr>
<tr>
<td></td>
<td>0.74%</td>
<td>0.73%</td>
<td>0.24%</td>
</tr>
<tr>
<td>4: 30-31 no-flow days</td>
<td>12,400</td>
<td>13,637</td>
<td>24,689,525</td>
</tr>
<tr>
<td></td>
<td>1.06%</td>
<td>1.17%</td>
<td>3.43%</td>
</tr>
<tr>
<td>Total</td>
<td>1,166,944</td>
<td>1,145,911</td>
<td>717,664,428</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>98.20%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Note. In this study, Europe does not include Russia and Turkey. The gauging stations represent those which were used to set up the RF model, where the fraction of all station-months with observed and simulated classes is provided. In each column, the first row shows the total number of stations or reaches in the class and the second row shows the percentage in the class. The percentage values for the reach-months relate to the total number of reach months (468) during 1981-2019; and for the station-months, to the number of station-months with observations. As the step 2 RF model predicting the four classes of no-flow days was set up only for the station-months that are observed to be intermittent, the predicted class percentages do not add up to 100%.
Figure 8. Percentage of months with at least one no-flow day for European stream reaches during the period 1981-2019.

The predicted prevalence of perennial conditions across reaches is similar to the observed prevalence in streamflow gauging stations where 97% of the observed station-months and 76% of the stations are perennial. As drainage area is the most important predictor for a station-month being perennial or intermittent, with small basin size leading to a higher probability of intermittence, it is surprising that a higher percentage of reaches is simulated to be perennial as compared to the gauging stations. Reaches with small upstream basins of less the 50 km$^2$ comprise 78% of all reaches, whereas only 12% of gauging stations have such small basins (Table 4). This highlights the importance of the interplay of all predictors of the step 1 RF and may be affected by our voluntary addition of intermittent data in observations.
Table 4  
Mean Streamflow per Station-month and Reach-months Averaged for Drainage Basin Area Classes

<table>
<thead>
<tr>
<th>Upstream area [km$^2$]</th>
<th>Gauging stations</th>
<th>Reaches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Observed (m$^3$ s$^{-1}$)</td>
<td>Mean Predicted (DSS$^a$) (m$^3$ s$^{-1}$)</td>
</tr>
<tr>
<td></td>
<td>[0-2]</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>[2-5]</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>[5-10]</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>[10-50]</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>[50-500]</td>
<td>4.55</td>
</tr>
<tr>
<td></td>
<td>[500-2500]</td>
<td>17.29</td>
</tr>
<tr>
<td></td>
<td>[2500-10000]</td>
<td>56.91</td>
</tr>
<tr>
<td></td>
<td>&gt;10000</td>
<td>512.31</td>
</tr>
</tbody>
</table>

$^a$Downscaled streamflow.

The fraction of reach-months with 30-31 no-flow days (3.4%) is much higher than the corresponding fraction of the stream-months that are observed and predicted to occur at the streamflow gauges (1.1%; Table 3). This is not due to the much higher prevalence of reaches with small upstream basins than of stations with such small basins (Table 4) as also in each drainage area size class, the fraction of months with 30-31 no-flow days is larger for the reach-months than for the station-months (Table 5). Both station observations and reach predictions agree that the likelihood of perennial months increases and the likelihood of 30-31 no-flow days decreases with increasing size of the drainage basin (Table 5). The exception are the smallest reaches with an upstream area of 2 km$^2$ or smaller, because we only generated such small reaches from the 15 arc-sec drainage direction map where the mean annual downscaled HR streamflow during the period 1981-2019 exceeds 0.03 m$^3$/s (section 2.4); this explains the high fraction of perennial months in the smallest size class. One reason for the higher prevalence of the class 30-31 no-flow days for the reach-months as compared to the station-months in all size classes between 2 km$^2$ and 500 km$^2$ may be that the average streamflow for all reach-months of a certain size class is smaller than for the gauges (both observed and predicted) (Table 4). This discrepancy likely led to more dry months, because streamflow is the most important predictor in the step 2 RF (Figure 7). At the same time, the fraction of perennial reach-months, which is determined by the step 1 RF, is also
higher than the fraction of perennial station-months in each size class, such that the other three intermittence classes are predicted to be very rare among the reach-months. The reason for this is unknown, but one aspect to consider may be that streamflow is not the most important predictor in the step 1 RF (Figure 7).

**Table 5**

*Percent of Observed Station-months and Predicted Reach-months (1981-2019) in the Five Intermittence Classes*

<table>
<thead>
<tr>
<th>Upstream area [km$^2$]</th>
<th>Observed station-months in classes 0-4 (%)</th>
<th>Predicted reach-months in classes 0-4 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(0-2]</td>
<td>87.30</td>
<td>2.57</td>
</tr>
<tr>
<td>(2-5]</td>
<td>88.48</td>
<td>1.93</td>
</tr>
<tr>
<td>(5-10]</td>
<td>94.18</td>
<td>1.11</td>
</tr>
<tr>
<td>(10-50]</td>
<td>95.85</td>
<td>0.59</td>
</tr>
<tr>
<td>(50-500]</td>
<td>96.83</td>
<td>0.34</td>
</tr>
<tr>
<td>(500-2500]</td>
<td>98.33</td>
<td>0.24</td>
</tr>
<tr>
<td>(2500-10000]</td>
<td>99.32</td>
<td>0.10</td>
</tr>
<tr>
<td>&gt;10000</td>
<td>98.85</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Note. These values are represented as a function of upstream drainage area [km$^2$] of the streamflow gauging stations or the reach. Classes: 0: perennial, 1: 1-5 no-flow days, 2: 6-15 no-flow days, 3: 16-29 no-flow days, 4: 30-31 no-flow day. In total, 1,166,944 station-months and 717,664,428 reach-months are considered.

The actual number of perennial months in reaches with upstream areas of 2-50 km$^2$, the dominant upstream area classes listed in Table 4, may even be higher as the step 1 RF tends to underestimate the fraction of perennial station-months (Figure S4). However, the number of streamflow gauging stations for that class, in particular in the size class under 10 km$^2$, is rather small (Table 4).

The prevalence of intermittence across the European river network shows a clear seasonal and interannual variability. While 97.6-99.8% of the European reaches are perennial in January and February, this is the case for only 89.6-93.4% in August and September (Figure 9). There is no overall trend over the whole 39-year period, but seasonal minima and maxima of the fraction of perennial reaches decreased from 2013 to 2019, while the opposite is true for the fraction of months with 30-31 no-flow days (Figure 9). The southern European countries of Portugal, Spain, Italy, Greece and Cyprus have a much higher seasonal range of the fraction of perennial reaches; in July
to August, only about 70% of the reaches are perennial, while in winter, it is close to 90-99%, depending on the year (Figure S8a). In the Scandinavian countries Norway, Sweden and Finland, the (very low) occurrence of intermittent conditions is larger in the second half of the study period, but the highest level of intermittence occurred in 1996, related to unusually low precipitation (Figure S7b). As an illustration of the spatial distribution and seasonality of streamflow intermittence, the European maps for streamflow intermittence in January and August 2019 are shown in Figure S8.

![Figure 9. Monthly time series of the percent of all European stream reaches in the five intermittence classes for the period 1981-2019.](image)

5. Discussion

In this study, we chose to use all daily streamflow observations available for the study period to set up the RF model, to obtain a robust model based on the maximum amount of information. A
temporal validation of the models with independent data was not conducted. The RF model of Döll et al. (2023a), with less streamflow observations and slightly different predictor variables and intermittence class definitions, was trained with data for a calibration period that encompassed, for each gauging station, the first two-thirds of the available observed months, while the rest was left for independent validation; 99% and 95% of the perennial station-months were predicted correctly for the calibration and the validation period, respectively. Considering only the intermittent station-months (i.e., predictions from the step 2 RF), the frequency of predicting the correct class decreased from 56% in the calibration period to 47% in the validation period. In this study, we validate our predictions with a data set of visually observed intermittence for France. Then, we discuss the challenges of deriving continental-scale high-resolution estimates of monthly streamflow and streamflow intermittence.

5.1. Validation of streamflow intermittence predictions using ONDE observations for France

We used observations from the French national river drying observatory (ONDE, 2020; Beaufort et al., 2018) to validate our predictions for 2,865 reaches and 148,004 reach-months in France of whether each reach-month was intermittent (with at least one no-flow day) or perennial, i.e., the step 1 RF. The ONDE network consists of a stable set of approximately 3,300 sites on river and stream reaches of Strahler orders under five, which have been inspected since 2012 by trained public staff from the French Biodiversity Office (OFB in French), at least monthly between May and September with the objective of identifying all drying events. If either the status “no visible flow” or “dried out” was assigned in any month, we considered the reach-month to be observed as intermittent. Considering that its objective is to track intermittence in mostly headwater streams, the ONDE dataset has a much higher percentage of intermittent reaches and reach-months than the European streamflow gauging station dataset used to set up the RF model. While 61% of the reaches and 15% of the reach-months are intermittent in the ONDE dataset, only 24% of the European gauging stations and 2.8% of the station-months are intermittent. Considering only French gauging stations, the respective values are 38% and 3.5%. About 73% of ONDE reaches have a drainage area of less than 50 km², which is similar to the fraction of European reaches in the size class (Table 4), whereas this is the case for only 12% in the European data set of gauging stations.

Compared to the ONDE data, the step 1 RF model underestimates the number of intermittent reach-months (Figure 10d), whereas it tends to overestimate the number of intermittent station-months relative to the 3706 European streamflow stations. With a BACC of 0.53, only 8% of the intermittent reach-months in ONDE are correctly identified (Figure 10c). Underestimation occurs in all size classes, increasing from an underestimation of, on average, 4 months for upstream areas of less than 10 km² to an underestimation of 6-7 months for basins between 10 km² and 2,500 km².

Considering whether reaches are intermittent or perennial, only 23% of the intermittent reaches were correctly predicted as such, compared to 62% for the European stations used to set up the step 1 RF (Table 2). Our RF model achieves a balanced accuracy of only 0.54 (Figure 10a) in its binary classification of ONDE reaches, while the global static RF model of naturally intermittent reaches of Messager et al. (2021) yielded a slightly higher value of 0.59 (Figure 10b). The spatial pattern of agreement of the static global model is less patchy than that of our model. The global model predicts intermittence to occur in large contiguous areas, as it is mainly driven by larger-scale climatic predictors whereas our dynamic European model is strongly driven by small-scale streamflow characteristics. In addition, our model is based on more streamflow gauging stations.
For unknown reasons, our RF model cannot predict intermittent reaches along the Mediterranean coast, which differs from the static global model. If the threshold for perennial conditions is increased to a probability of 75%, which does not lead to a decrease of BACC, 91% of the intermittent reaches would be correctly identified, but then 85% of the perennial reaches would be incorrectly predicted as intermittent.

**Figure 10.** Comparison of simulated intermittence of reaches and reach-months with the ONDE data set of visually observed intermittence. Correspondence between the simulated and observed intermittence state of reaches for our RF model (a) and the RF model of Messager et al. (2021) (b), percent of correctly classified reach-months in our model (c) and ratio of predicted to observed intermittent months in our model (d).
5.2. Challenges of continental-scale high-resolution estimation of streamflow and streamflow intermittence

Given the amount, spatial resolution and uncertainty of available data, it is very challenging to achieve a good prediction of HR streamflow intermittence for all of Europe. One reason is that continental and global-scale streamflow simulations for relatively large rivers represented by LR models often strongly differ from observations. This mismatch stems from large-scale models not being calibrated in a basin-specific manner against observed daily streamflow, as is done with small-scale models. In addition, their input data are coarser and usually less accurate than those of small-scale models. Here, using an advanced downscaling algorithm, the output of a LR hydrological model (Figure 11a, b) was downscaled, by a factor of 14,400, to generate monthly time series of streamflow at 15 arc-sec resolution (Figure 11c). It is encouraging that these simulated HR streamflow time series show skill for most streamflow stations with upstream areas smaller than an LR cell, even down to upstream areas of less the 5 km$^2$ (Figures 2 and 3). However, the number of evaluated gauging stations with such small upstream areas was very small (Figure 3). The estimated HR streamflow time series enabled, together with other predictors, the estimation of HR streamflow intermittence. A comparison of Figures 11b and 11d, which shows LR and HR intermittence, respectively, for a part of France illustrates the strongly increased information content of the European HR streamflow intermittence dataset as compared to an LR intermittence estimation.
Figure 11. Illustration of downscaling of LR WaterGAP output to the HR stream network of HydroSHEDS and the resulting resolution-dependent characterization of intermittence. Panels show LR (0.5 arc-deg) grid cells with the sum of surface runoff and groundwater discharge (the main input to the downscaling algorithm) (a), LR reaches with their intermittence status (b), HR (15 arc-sec) grid cells with downscaled streamflow (c) and HR reaches with intermittence status in 5 classes (d). The figure shows the situation for the example of August 2003. In c and d, the locations of the streamflow gauging stations used for validation of downscaled streamflow and as target for the RF model are added.
It is very difficult to judge the realism or plausibility of the predicted reach intermittence. The model validation against an independent data set for France (section 5.1) indicates a severe underestimation of observed intermittent reach-months, while the comparison of predicted intermittence to streamflow observations used for setting up the RF model indicates that the RF model overestimates intermittence, particularly for the relatively dry regions of Europe such as large parts of the Iberian Peninsula. The latter may be explained by the suspected anthropogenic perennialization of streamflow by many small and large dams that have been constructed to make water available even in periods of low or no flow. Still, the BACC for predicting intermittent station-months (0.92) was very good. We found that the RF model can simulate well the interannual variability of the number of intermittent months at the streamflow gauging stations (Figure 4c), which is an important positive characteristic if the model is to be used for assessing the impact of drought conditions or climate change. The partial dependence plots for the step 1 RF show that the model identifies tendencies in the probability of a station-month being intermittent that agree with expert expectations (except for terrain slope), which increases our trust in the derived RF. Moreover, the correlation between the observed and predicted monthly time series of the five intermittence classes is high at most intermittent stations (Figure 6d), which indicates a good representation of the seasonality of streamflow intermittence. Averaged over all station-months with available intermittence observations, there is no bias in the prediction of the five intermittence classes per size class of upstream area as compared to observations (Table 3). However, the prediction of the number of no-flow days per reach-month in four classes must be considered to be less reliable than the prediction of a reach-month as either intermittent or perennial, as indicated by the lower BACC of 0.67 for the step 2 RF.

In this study, we estimate that 18.7% of the European stream network length were intermittent in the period 1981-2019, while the global model of Messager et al. (2021) predicts a value of 17.1% for the our European study area. However, these values cannot be compared directly for various reasons. Our river network includes smaller headwater streams than the global model (representing 12.4% of the European river network used in this study; see Section 2.4) and the definition of intermittent reach is slightly different (global model: 1 no-flow day per year, our model: 1 no-flow day during the period 1981-2019). In addition, the global model aimed to predict natural intermittence by excluding heavily influenced gauging stations, relying on naturalized hydrology for the period 1971-2000. Still, our model predicts a similar prevalence of intermittent reaches in Europe.

The prevalence of intermittence across European rivers and streams by this study, with 17.8% of intermittent reaches and 3.8% of the reach-months, is much lower than in the study of Döll et al. (2023a), with values of 39.6% and 9.1%, respectively, even though the same HR streamflow estimates were used in the RF modeling. Even though some predictors (related to irrigation, population and the degree of regulation by reservoirs) were added and one (daily streamflow variability) removed in this study, we attribute this strong discrepancy to the different observations of the RF target variables, which were derived from daily time series of streamflow observations. We explicitly tried to obtain streamflow from dry areas and small streams, with a higher likelihood of intermittence, and added data from Cyprus and Italy (for Sardinia and the Po basin, but time series for the rest of Italy were shorter than our inclusion threshold of 36 months), but we could not obtain in time any data for, e.g., Greece, Albania and Bulgaria. The data set was rather extended by stations for more humid regions such as Scandinavia, the three Baltic states, Poland and Belarus, most of which are perennial (comp. Figure 4a). When the streamflow observations data set was
extended from the one used by Döll et al. (2023a), i.e., from 1915 gauging stations to the 3706 stations in this study (see Table 4 for station numbers per drainage area class), the additional stations had a smaller fraction of intermittent months than the original data set. In this way, we have further biased the target data set and therefore may have caused a biased streamflow intermittence prediction for the reaches. Still, we expect that almost doubling the amount of target observations as compared to the study of Döll et al. (2023a) increased the reliability of the RF models. The additional streamflow data and predictors are informative because the BACC for the step 1 RF increased from 0.85 in Döll et al. (2023a) to 0.92 in this study (while the BACC for the step 2 RFs are the same). However, the fit to the ONDE data as measured by the BACC for the identification of intermittent reaches remained the same as that of the step 1 RF model by Döll et al. (2023a). The comparison of the European streamflow intermittence estimation by Döll et al. (2023a) and in this study shows the major impact of available target observations on RF modeling results.

Upstream area is the most important predictor for likelihood of a station-month to be intermittent, yet we cannot assume that inclusion of this predictor in the RF model development is adequately representing the effect of upstream area on the likelihood of intermittence given the existing distribution of our target data. To represent upstream area appropriately in the target data, we would need a data set of streamflow gauging stations that show the same distribution of upstream areas as the stream reaches; however, the distributions are extremely different (Table 4). While 77.7% of the reaches have an upstream area of up to 50 km², this is only the case for 12.3% of the gauging stations. The largest size class for the reaches is, with 31%, the class 2-5 km², but only 0.6% of the station-months are in this class. As an illustration, if we would like to have the same size distribution with the 22 stations in the class 2-5 km² that were available, then we would have to consider only 70 stations in total, instead of 3706. As the fraction of perennial months is higher for reaches than for the stations, especially for drainage areas below 50 km² (except for the smallest size class due to the definition of the smallest reaches, Table 5), a further decrease of the average fraction of intermittent months for the gauging stations by the extension of the data set might have led to an even stronger underestimation of intermittence in these headwater reaches. However, by our extension, we more than doubled the number of stations in the class 2-5 km² by raising the number of stations from 10 to 22 (Table 5) which increased the information base upon which the RF models were trained.

The performance of our model certainly suffers from the general problem of imbalanced target data, with 97.2% of the station-months being perennial. The most important approach to handle this problem was the two-step approach whereby the prediction of perennial months in step 1 was followed by the prediction of the number of no-flow days per month only for those 2.8% of all station-months for which at least one no-flow day was observed. In addition, various alternative methods for handling imbalanced data were tested for the step 2 RF. Oversampling resulted in slightly better BACC values than undersampling and the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002).

6. Conclusions and Outlook

For the first time, streamflow intermittence dynamics could be quantified at the continental scale at a high spatial resolution, i.e., for stream reaches with an upstream area down to only 2 km² (or even smaller in wet regions). We simulated monthly time series of streamflow intermittence in
five classes (0, 1-5, 6-15, 16-29 and 30-31 no-flow days per month) in the period 1981-2019 for
more than 1.5 million stream reaches in Europe. This was achieved by 1) downscaling the 0.5 arc-
deg output of the global hydrological model WaterGAP to obtain time series of monthly
streamflow at about 73 million 15 arc-sec grid cells and 2) combining this information with daily
data of streamflow as observed at 3706 gauging stations and a number of static hydro-
environmental characteristics of the upstream basins (plus two WaterGAP-related datasets) in an
RF modeling approach. The model captures the interannual variability of the number of
intermittent months satisfactorily, and the monthly time series of the predicted five streamflow
intermittence states is highly correlated with observations. The spatial prevalence of weakly
intermittent conditions appears to be underestimated, while the number of intermittent months is
overestimated in the dry regions of Europe where reservoirs tend to perennialize streamflow.
While the generated streamflow intermittence data set does diverge from reality for many reach-
months, it is nevertheless a valuable basis for macro-scale studies of biodiversity, ecosystem
functions and ecosystem services under conditions of potential streamflow intermittence.

The presented modeling approach was designed to enable the computation of intermittence
changes due to climate change. For this purpose, the LR output of a WaterGAP run that is driven
by the bias-corrected output of global climate models, instead of observed historic climate, can be
downscaled to calculate monthly time series of HR streamflow in, for example, a 30-yr reference
period and a 30-year period in the future. These time series, together with the LR WaterGAP time
series of monthly diffuse groundwater recharge, runoff from land and the number of wet days
under climate change, can then serve to compute the dynamic predictor values that are, in addition
to the unchanged static predictor values, the input for the two developed RF models. In addition,
developed modeling approach can be used to analyze the occurrence of drought in intermittent
streams (Sarremejane et al., 2021).

### Data Availability Statement

WaterGAP 2.2e input and output used for deriving HR streamflow and as LR predictors in the RF
model is available from Müller Schmied et al. (2023b). The code for deriving HR streamflow
(Trautmann, 2023) is available at https://doi.org/10.5281/zenodo.10301003, and the code and
workflow for the RF modeling (Abbasi & Messager, 2023) at https://github.com/mahabbasi/europeanIRmap.git. Due to the very large file sizes, the HR monthly
streamflow time series are only available on request from first authors. The following data are
available at https://doi.org/10.6084/m9.figshare.24591807: 1) Input files for deriving HR
streamflow (Text S1), 2) the monthly time series of streamflow at the 3706 gauging stations, 3)
shapefiles of locations of streamflow gauging stations and European reaches, 4) all predictors and
target variables for the 3706 gauging stations used to generate the step 1 and step 2 RFs and 5)
shapefiles with the five intermittence classes for each reach-month in the period 1981-2019 as well
as the shapefiles for generating all figures (Döll et al., 2023b). The original data used for deriving
the HR static predictors are available as described in Section 2.3.2 and Table 1.

### Author Contributions

**Conceptualization**  
Petra Döll, Bernhard Lehner

**Methodology**  
Mahdi Abbasi, Tim Trautmann, Petra Döll, Bernhard Lehner, Mathis L. Messager, Nicolas Lamouroux
Acknowledgments

The authors thank their collaborators in the DRYvER project, in particular Jean-Philippe Vidal and Thibault Datry, for their contributions. MA was funded by the EU Horizon 2020 DRYvER project (Project ID 869226). MLM was supported by a Vanier Canada Graduate Scholarship (Natural Sciences and Engineering Research Council of Canada) and an H2O'Lyon Doctoral Fellowship (Université de Lyon, ANR-17-EURE-0018).

References


Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., & Gerdener, H., et al. (2023b). The global water resources and use model WaterGAP v2.2e - model output driven by gswp3-w5e5 and historical setup of direct human impacts. Goethe-Universität Frankfurt. https://doi.org/10.25716/GUDE.0TNY-KJPG [Dataset]


ONDE (Observatoire national des étiages) (2020). French Biodiversity Office https://onde.eaufrance.fr/content/t%C3%A9l%C3%A9charger-les-donn%C3%A9es-campagnes-par-ann%C3%A9e (accessed 21 June 2020)


Supporting Information for

Streamflow intermittence in Europe: Estimating high-resolution monthly time series by downscaling of simulated streamflow and Random Forest modeling

Petra Döll¹,²*, Mahdi Abbasi¹*, Mathis Loïc Messager³,⁴, Tim Trautmann¹, Bernhard Lehner⁴, Nicolas Lamouroux³

¹Institute of Physical Geography, Goethe University Frankfurt, Frankfurt/Main, Germany, ²Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F) Frankfurt, Frankfurt/Main, Germany, ³INRAE, UR RiverLy, Lyon-Villeurbanne, France, ⁴Department of Geography, McGill University, Montreal, Canada

Corresponding author: Petra Döll (p.doell@em.uni-frankfurt.de)

*Equal contribution

Contents of this file

Text S1

Figures S1 to S8

Text S1. Further adjustments of the correction factor used in the downscaling algorithm

The correction factor $C_{Li}$ (Equation 2) is augmented for certain grid cells as explained below.

S1.1 Redistribution of water storage modifications in large lakes and reservoirs

In WaterGAP, reservoirs with a maximum storage capacity of at least 0.5 km³, regulated lakes with 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 considered as so-called ‘global surface water bodies’ (Müller Schmied et al., 2021) that receive water not only from the surface runoff and groundwater discharge generated within the LR cell but also from upstream streamflow. Global surface water bodies may spread over more than one LR grid cell and their overall water balance is calculated in their assigned outflow cell. Thus, the initial net cell runoff ($ncR_{Li}$) of this outflow grid cell includes the runoff generated by the global surface water bodies (a single grid cell may represent the outflow of multiple global surface water bodies), which needs to be redistributed to all LR grid cells that intersect with these global surface water bodies and their respective HR cells. This is done by
calculating the change of water storage in the global surface water body for each month compared to the previous month. This amount is subtracted from the net cell runoff of the outflow cell and redistributed in an area-weighted way to all upstream LR cells intersecting one of the global surface water bodies. Every LR cell has its net cell runoff from global surface water bodies assigned based on the area of the cell that intersects the global surface water body. Then these LR values are applied to those HR cells that are covered by polygons of global surface water bodies. As HR grid cells have different grid cell areas, the distribution of runoff from global surface water bodies is area-weighted.

S1.2 Additional correction for remaining discrepancies in large rivers

Routing in WaterGAP is performed along the 0.5 arc-deg DDM30 river network (Döll & Lehner, 2002), but HR streamflow is computed based on a slightly modified version of the 15 arc-sec river network of HydroSHEDS (Lehner et al., 2008). Given their different spatial resolutions and generation processes, these two river networks differ locally in their representation of river courses and related characteristics. This, in turn, may cause the correction term $C_{Li}$ (Eq. 2) to not take effect in the desired way. One major issue is that the HydroSHEDS river network contains additional endorheic sinks, typically smaller ones, that are not covered by the DDM30 river network. Endorheic sinks (or depressions) are basins without an outlet to the ocean, represented topographically by one or multiple grid cells that are surrounded by higher elevation values. Those local endorheic sinks are not covered by the LR DDM30 because they occur at a smaller geographic scale that cannot be represented by the LR (0.5 arc-deg) grid cells. In such cases, for example if a subgrid endorheic sink covers half of an LR cell, the initial correction term $C_{Li}$ would be applied to all HR cells in the LR cell. But in the subsequent routing of discharge along HR grid cells, the discharge within the endorheic sink would not contribute to the discharge of the mainstem river, and thus the original $C_{Li}$ term alone would not be capable to correct the mainstem’s flow quantities.

An additional correction mechanism, already included in the original method of Lehner and Grill (2013), aims at correcting for such HR endorheic sinks but also covers other remaining artefacts that cause deviations between LR and HR streamflow estimates. Importantly, this additional correction mechanism is only applied to relatively large rivers, i.e., those with an upstream area of at least 50,000 km², and for locations with a reasonable accordance in drainage areas between DDM30 and HydroSHEDS: for rivers with catchment areas between 50,000 and 100,000 km², they are allowed to differ by up to 20%, and for rivers with catchment areas of >100,000 km², they are allowed to differ by up to 50%. These criteria are necessary because the two river networks can diverge strongly at local scale, especially in headwater areas and at confluences. For example, an HR grid cell may represent only a tributary to a mainstem, whereas the corresponding LR grid cell from DDM30 may represent the (much larger) mainstem. Therefore, if the above conditions are not fulfilled, additional corrections could cause major deteriorating effects on the results.

For those LR cells that fulfill the above criteria, the initial correction term $C_{Li,init}$ (Eq. 2) is extended by an additional correction term. This modification of the correction term $C_{Li}$ is calculated by comparing the net cell runoff of the LR cell with the net cell runoff of the HR grid cell with the maximum upstream area in that LR cell, with

$$C_{Li} = C_{Li,init} + (ncR_{Li}^{LR} - ncR_{Li}^{HR})$$ (S1)
The HR net cell runoff representation of LR grid cell $L_i$ ($ncR_{Li}^{HR}$) is calculated as the streamflow of the HR grid cell with the maximum upstream area in $L_i$ ($Q_{Li,Hi}$, Eq. 1) minus the corresponding streamflow values of direct upstream LR grid cells.

**S1.3 Equalizing correction terms by partially shifting them to the next downstream LR grid cell**

The additional correction term (see S1.2) can introduce correction gaps caused by discrepancies between the DDM30 and HydroSHEDS river networks. These gaps can then lead to oscillating upward and downward corrections in neighboring LR grid cells. To smooth such oscillating corrections, the correction terms are partially propagated to the next downstream LR grid cell and are thus balanced with the correction term in that cell. The partial shifts of the correction terms along the LR river network are only applied if the maximum HR upstream area in the downstream LR cell is at least 90% of that in the evaluated LR cell. This criterion guarantees that the correction values are solely shifted to larger streams (within a 10% tolerance to consider minor discrepancies such as endorheic sinks) and that shifting between LR cells with mismatching river networks is avoided. The fraction of the correction term that is shifted downstream depends on the difference between upstream and downstream basin area such that 50% of the correction term is shifted downstream if the two neighboring cells represent equal basin size (i.e., along the same river), and an increasingly higher fraction is shifted downstream if a smaller river flows into a larger one (as applying a correction in a larger river leads to less potential distortion). The fractional shift is computed as

$$f_{Li}^{shift} = \begin{cases} 
\frac{(2 * \text{up}A_{Li,down}^{Max} - \text{up}A_{Li}^{Max})}{2 * \text{up}A_{Li,down}^{Max}}, & \text{if } \text{up}A_{Li,down}^{Max} > 0.9 \times \text{up}A_{Li}^{Max} \\
0, & \text{otherwise} 
\end{cases} \quad (S2)$$

with $\text{up}A_{Li,down}^{Max}$ representing the maximum HR upstream area in the downstream LR grid cell and $\text{up}A_{Li}^{Max}$ representing the maximum HR upstream area in the evaluated LR grid cell $L_i$. Following this approach, the modified correction term for a given LR grid cell $L_i$ consists of the part which is not shifted downstream and the parts which originate from the shifted correction terms from direct upstream cells $C_{Li,upj}$.

$$C_{Li} = C_{Li,init} \times (1 - f_{Li}^{shift}) + \sum_{j=1}^{n}(C_{Li,upj} \times f_{Li,upj}^{shift}) \quad (S3)$$

**S1.4 Negative and extreme correction values**

Despite the various correction and balancing algorithms described above, it is possible that in singular cases negative values of streamflow or extreme correction values are calculated. This can happen, for example, in places where there are major discrepancies between the LR and HR river network alignments or where their upstream areas differ substantially. Three measures are sequentially implemented to limit potential artefacts caused by applying the final correction term. First, the final correction term is limited to a maximum threshold value of 0.001 m³ s⁻¹ per km² of upstream area. Second, any negative HR streamflow values, which may originate from side effects of the correction mechanisms, are not accumulated along the river network. A negative streamflow
value can turn positive during flow accumulation if streamflow is added from upstream cells, but negative correction values are not propagated along the river network. Third, all remaining negative streamflow values are set to zero in the final step.

S1.5 Technical implementation

The software implementation of the downscaling algorithm was developed in Python. A set of Python scripts (with ArcPy dependency) was developed to preprocess necessary static data. The static data listed below are necessary to run the downscaling algorithm.

<table>
<thead>
<tr>
<th>Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>flow_dir_15s_by_continent.gdb</td>
<td>HydroSHEDS flow directions [ESRI flow direction codes]</td>
</tr>
<tr>
<td>pixel_area_skm_15s.gdb</td>
<td>HydroSHEDS area of HR grid cells [km²]</td>
</tr>
<tr>
<td>flowdir_30min.tif</td>
<td>DDM30 flow directions [ESRI flow direction codes]</td>
</tr>
<tr>
<td>landratio_correction.tif</td>
<td>Ratio between percent of LR cell covered by HydroSHEDS landmask and percent of cell covered by WaterGAP landmask [-]</td>
</tr>
<tr>
<td>orgDDM30area.tif</td>
<td>Area of LR cells of WaterGAP [km²]</td>
</tr>
<tr>
<td>pixareafraction_glolakres_15s.tif</td>
<td>Ratio of global surface water bodies that is covered by the HR grid cell [-]</td>
</tr>
</tbody>
</table>
Figure S1. Binary confusion matrix in case of two classes (perennial and non-perennial) only.

Figure S2. NSE of monthly streamflow time series (left) and of the logarithm of monthly streamflow time series (right) for all 885 intermittent streamflow stations with observations, grouped in size classes of the upstream area of the streamflow gauging stations. The boxes indicate the 25th, 50th (median) and 75th percentiles, the whiskers indicate the 5th and 95th percentiles of the samples. The blue lines of the violin plot show the smoothed distribution of the data points. The “number of stations not shown” indicates the number of stations with an NSE of less than -1.
Figure S3. NSE of monthly streamflow time series (left) and of the logarithm of monthly streamflow time series (right) for all 2,821 perennial streamflow stations with observations, grouped in size classes of the upstream area of the streamflow gauging stations. The boxes indicate the 25th, 50th (median) and 75th percentiles, the whiskers indicate the 5th and 95th percentiles of the samples. The blue lines of the violin plot show the smoothed distribution of the data points. The “number of stations not shown” indicates the number of stations with an NSE of less than -1.
**Figure S4.** Performance of the step 1 RF as a function of upstream drainage area [km$^2$] of the streamflow gauging stations. The box plot shows the percent 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. The boxes indicate P25 (25th percentile), P50 (median) and P75, the whiskers P5 and P95 of the samples.
Figure S5. Percentage of intermittent months with observations of the four intermittence classes (1-5, 6-15, 16-29, 30-31 no-flow days per month) at gauging stations in the complete streamflow dataset.
Figure S6. Partial dependence plots for the 23 predictors of the step 1 RF.
Figure S7. Monthly time series of the percent of reaches in the five intermittence classes in southern Europe (Portugal Spain, Italy, Greece and Cyprus) (left) and in Scandinavia (Norway, Sweden and Finland (right)).

Figure S8. Number of no-flow days, in five classes, in January 2019 (left) and August 2019 (right).