Leveraging reforecasts for flood estimation with long continuous simulation: a proof-of-concept study

Daniel Viviroli[1], Martin Jury[2], Maria Staudinger[1], Martina Kauzlaric[3,4], Heimo Truhetz[2], and Douglas Maraun[2]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
[2]Wegener Center for Climate and Global Change, University of Graz, Graz, Austria
[3]Institute of Geography, University of Bern, Bern, Switzerland
[4]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Correspondence: Daniel Viviroli (daniel.viviroli@geo.uzh.ch)

Abstract. Flood estimation is critical for risk assessment, but traditional ~~methods~~ approaches are often constrained by the limited length of observation data. This study explores the potential of reforecasts (RFs) ~~to enhance flood estimation through use in~~ for flood estimation using long continuous simulation (CS) with a hydrological model at high (hourly) temporal resolution. As a proof of concept, we first processed individual RFs from the ~~vast database~~ extensive archive of the European ~~Center~~ Centre for Medium-Range Weather Forecasts (ECMWF) with bias correction, stochastic downscaling and disaggregation with analogs to ~~finally~~ obtain mean areal precipitation and mean areal temperature for a set of test catchments in Switzerland. We subsequently concatenated these RFs into a time series of ~~close to~~ nearly 10 000 years ~~length~~ and used them in long CS to derive flood return levels. ~~Results demonstrate the potential of RFs as a complementary tool~~

Results show hydrological consistency of the concatenated RFs and demonstrate their potential in flood estimation, providing ~~insights into extreme event magnitudes and frequencies. Moreover, RFs can provide a relevant alternative view on exceptionally high extremes when compared to flood~~ information on the magnitude and frequency of extreme events. In addition, RFs offer a relevant complementary perspective on exceptionally large floods when compared with estimates derived from using other inputs to long CS, such as ~~those generated by~~ from a stochastic weather generator. ~~Limitations apply to~~ However, structural uncertainties – particularly related to the underlying numerical weather prediction system – must be considered, along with the reliance on a single model framework, non-stationarity and internal climate variability. Further limitations arise for catchments smaller than approximately 500 km^2, ~~where the~~ for which stochastic downscaling becomes increasingly inadequate~~, especially for resolving convective events. There~~. For resolving the relevant convective events in such catchments, dynamical downscaling would be more appropriate~~, but~~ ; however, this was not feasible with the ~~data currently available~~ currently available data.

1

Rare to very rare floods, associated with return periods of 1000 to 10 000 years, can cause severe human and economic damage. However, their estimation is ~~limited~~ constrained by the comparatively short ~~streamflow records available and the unknown representativity of~~ length of available streamflow records and by uncertainty about whether disaster-rich and disaster-poor

periods within these records ~~for~~ are representative of present conditions (**????**). At the upper end of the flood magnitude ~~scale~~spectrum, the estimation of very rare or ~~even~~ unprecedented floods – and of weather hazards more ~~broadly, see ?~~ generally (see **?**) – is ~~explored through various~~ addressed using a variety of approaches. These include ~~transforming estimates of possible~~ converting estimates of probable maximum precipitation (PMP) into ~~estimates of possible~~ probable maximum floods (PMF) (**?**), ~~employing weather generators for simulations with hydrological models (?), using large~~ applying weather generators in hydrological simulations (**?**), pooling large regional data sets of observed floods ~~and pooling them into regions~~ (**?**), and developing storylines based on extremely rare observed events (e.g., **?**).

~~Regarding~~ With respect to meteorological extremes more ~~generally~~broadly, reforecasts (RF) have been used to explore plausible yet unobserved extremes that ~~would be difficult to anticipate based on historical records . Even though~~ are difficult to infer from historical records alone. Although the primary purpose of ~~reforecasts is~~ RFs is forecast skill assessment and calibration ~~of forecasts~~ (**?**), the ~~large~~ extensive archive of model realisations under current atmospheric conditions ~~that these~~ provided by ensemble prediction systems ~~provide (?) is also an interesting possibility~~ (**?**) offers an opportunity to explore meteorologically plausible extreme events beyond the ~~range of observations. RFs often have lead times of several weeks. That is, while starting from observed weather, they simulate the atmospheric state beyond the time horizonover which the atmosphere is well predictable and, therefore, substantially deviate from the weather which has actually manifested. Thus, RFs simulate plausiblebut non-realized weather states,~~ observed range. RFs are typically initialised from observed atmospheric states but simulate conditions beyond the predictability horizon, diverging from the observed meteorological evolution. Consequently, RFs represent plausible, yet unrealized, atmospheric evolutions, including extreme events that have not ~~yet been observed (?). In this sense,~~ been observed to date (**?**).

RFs have been ~~used to estimate , for example, extremes of~~ applied to estimate extreme sea surges (**?**), precipitation (**??**), temperature (**??**) and wind (**?**), ~~and to map~~ as well as to characterize compound events (**?**). In ~~connection with~~ relation to estimation of extreme floods, **?** pooled RF ensemble members of river discharge from the European Flood Awareness System (EFAS) to examine the frequency and intensity of extreme floods~~(~~, with return periods of up to 200 years ~~return period) at the daily scale across numerous catchments in Central Europe. They demonstrated that, compared to observation-derived flood estimates~~ across Central European catchments. Compared to flood estimates derived from observations, this approach ~~yields estimates with~~ yielded considerably narrower uncertainty bounds for extreme events occurring less than twice a century. Similarly, **?** investigated the robustness of frequency estimates for ~~both~~ daily precipitation and streamflow extremes ~~(~~with return periods of up to 500 years ~~return period),~~ using synthetic time series derived from pooled ~~meteorological seasonal reforecasts~~ SEAS5 seasonal RFs (**?**). These ~~time series were then~~ series were used to generate hydrological ~~reforecasts with the E-HYPE~~ RFs with the process-based ~~model (?) for the pan-European domain~~E-HYPE model (**?**) across Europe. By generating 100 samples with replacement for ~~several time series~~ different time-series length, they showed that the relative interquartile range of ~~the~~ estimated 100-year return levels – ~~used to quantify~~ a measure of sampling uncertainty – decreased markedly ~~as the time series~~length increased, dropping from over with longer time series, from above 100% ~~with~~ for 50-year time series to below 10% ~~with~~ for 500-year ~~time series. Extending on these methods~~series. Building on these approaches, **?** introduced an additional step ~~in the workflow~~ to account for different flood type clusters ~~, modifying the approach~~ by applying a mixing distribution

after pooling. ~~In their study, they used~~ Using RF data from the Global Flood Awareness System (GLOFAS) and ~~applied the UNprecendented~~ the UNprecedented Simulated Extreme ENsemble (UNSEEN) method (**??**)~~for pooling, focusing on~~ , they analysed two catchments in Germany ~~,~~ with return periods of up to 200 years. Their results indicate that explicitly ~~considering different flood types considerably alters the~~ representing flood types alters the estimated 100-year return level ~~estimate (≥~~by at least 15% ~~)~~ and reduces the Root Mean Square Error (RMSE) compared ~~to the conventional approach, which assumes a single distribution. The application~~ with conventional single-distribution approaches. Applications of the UNSEEN ~~method using RFs extends~~ framework using RFs therefore extend beyond flood frequency estimation ~~; it has also been shown to provide plausible unprecedented events, supporting the development of flood~~ and also support the construction of plausible unprecedented-event storylines under varying initial conditions (**?**).

Here, we ~~explore the feasibility of utilizing extensive RF data on~~ examine the feasibility and added value of using extensive RF precipitation and temperature ~~within a~~ data within an established long continuous simulation (CS) framework ~~, leveraging RFs with a total length of close to 10 000 years obtained from the vast database of the European Center for Medium-Range Weather Forecasts (ECMWF). Our proof of concept combines these data with key advantages of CS, enabling the estimation of return periods considerably larger than previous RF-related studies. In particular, this approach~~ at hourly resolution. This framework avoids assumptions about antecedent conditions and their spatial patterns while enabling a physically coherent representation of the full ~~spatial-temporal development~~ spatio-temporal evolution of floods over larger areas (**?**). ~~For~~ To ensure comparability with an existing long CS ~~approach in the application domain that utilizes~~ framework that relies on weather generator (WGEN) input, we ~~employed~~ adopted the hydrometeorological simulation chain introduced by **?**~~that serves as~~ , which forms the foundation for the ~~projects~~ EXAR (hazard information for extreme flood events on the rivers Aare and Rhine) (**?**) and EXCH (Extreme Floods in Switzerland) (**?**) ~~. Both EXAR and EXCH aim at providing~~ projects. Both initiatives aim to deliver methodologically and spatially ~~coherent~~ consistent flood hazard assessments, ~~and the project results have been examined comprehensively as far as feasible from the data~~ evaluated as comprehensively as possible given the limited information available on rare events (**???**).

~~To arrive at the spatial and temporal scales necessary for hydrological modelling, bias adjustment, downscaling and disaggregation of the relatively coarse RFs were necessary. To ensure~~ For consistency and comparability with the EXAR/EXCH framework, ~~some methodological choices had to be carried over, most notably~~ several methodological choices were retained, including the use of a lumped catchment model ~~combined~~ with hydrological routing where ~~necessary,~~ required and the interpolation of mean catchment meteorological inputs from point values. ~~While specifics~~ Although specific elements of this procedure could ~~certainly be altered or fine-tuned, this framework offers the possibility to juxtapose~~ be refined, the adopted framework enables direct juxtaposition of RF-based ~~(~~flood estimates, derived from physically consistent and plausible weather conditions~~)~~ , and WGEN-based ~~(from~~ estimates generated within a stochastic multi-site framework~~) flood estimates and~~ . This juxtaposition provides valuable context for assessing very rare floods.

~~Departing from a broad scale range of approximately 20 to 3000 km^2, we explored the feasibility of both a statistical and a dynamical downscaling approach, and discovered that only the statistical approach – suitable for larger scales – was feasible with the data publicly available. The main parts of this paper therefore focus on this approach, while dynamical downscaling~~

~~– which would be more suitable for small scales than the statistical approach– is briefly discussed in Sect. ??.~~ Against this background, this study is designed as a proof of concept to address the following research questions:

1. Can raw RF data be processed to hourly resolution and concatenated into very long time series suitable for continuous hydrological simulation in a challenging Alpine setting?
2. How do flood frequency estimates derived from such RF-based continuous simulations compare with, and complement, estimates obtained using an established stochastic weather generator framework?
3. What are the key limitations of the RF-based approach within a long CS framework?

To address these questions, we processed RFs from the European Centre for Medium-Range Weather Forecasts (ECMWF) using bias correction, stochastic downscaling, and temporal disaggregation, and concatenated the RFs into a 10 000-year time series. These were then used as input for hydrological simulations, which were compared with simulations driven by a stochastic weather generator to evaluate feasibility, added value and limitations of the RF-based approach.

2 Data

2.1 Test catchments

For our proof of concept, we examined a total of 20 test catchments (Table 1, Fig. 1) from across Switzerland. These catchments represent important climate regions and runoff regime characteristic of a region with complex topography. Note that three catchments in each of the Aare river basin (Kander at Hondrich, Aare at Thun, Aare at Bern), the Thur river basin (Thur at Alt St. Johann, Thur at Jonschwil, Thur at Andelfingen) and the Maggia river basin (Riale di Calneggia at Cavergno, Isorno at Mosogno, Maggia at Locarno) are nested. The sites on the Aare, Drance de Bagnes, Maggia, Saltina and Sarine rivers are impacted by hydropower, and the Thun and Bern sites on the Aare River are influenced by the regulated lakes Brienz and Thun.

2.2 Reforecast data

Our work is based on the ensemble RFs provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) that allows considering plausible yet not actually occurred weather extremes (**?**). For the statistical modelling, we used precipitation and ~~near surface~~ near-surface air temperature as input. The data were available on a 0.25° regular grid (~19×28 km^2 in the study domain) until forecast day 15, and on a 0.5° regular grid (~38×56 km^2) thereafter. The temporal resolution of the data was 6 hours.

ECMWF's IFS has undergone several changes since its inception in 2008 (**???**). As detailed in Tab. 2, the ECMWF RFs (control and perturbed) are

- initialised weekly 2008/03–2015/05, and bi-weekly from 2015/05 onwards,
- initialised for the leading 18 years 2008/03–2012/06, and for the leading 20 years from 2012/06 onwards,

Table 1. Test catchments examined in this proof-of-concept study, with area (A), mean catchment elevation (E) and runoff regime type (**?**). The catchments are grouped by climate region (**?**) and listed in order of decreasing area.

Climate region	River and site	ID	A [km^2]	E [m a.s.l.]	Runoff regime type
Northern Alps	Aare @ Bern *†	AarBrn	2941	1584	nivo-glaciaire
	Aare @ Thun *†	AarThu	2459	1739	nivo-glaciaire
	Sarine @ Broc *	SarBro	636	1501	nival de transition
	Thur @ Jonschwil	ThuJon	493	1026	nivo-pluvial préalpin
	Kander @ Hondrich	KanHon	491	1846	nivo-glaciaire
	Kleine Emme @ Emmen	KEmEmm	478	1058	nivo-pluvial préalpin
	Minster @ Euthal, Rüti	MinEut	59.1	1352	nival de transition
	Thur @ Alt St. Johann	ThuASJ	42.3	1489	nival alpin
	Allenbach @ Adelboden	AllAde	28.8	1858	nival alpin
Western Central Alps	Drance de Bagnes @ Le Châble *	DdBLeC	254	2601	b-glaciaire
	Lonza @ Blatten	LonBla	77.4	2615	a-glaciaire
	Saltina @ Brig *	SalBrg	76.5	2014	b-glacio-nival
	Goneri @ Oberwald	GonOwd	38.5	2375	b-glacio-nival
Eastern Plateau	Thur @ Andelfingen	ThuAnd	1702	773	pluvial supérieur
Eastern Central Alps	Rein da Sumvitg @ Sumvitg	RdSSum	21.8	2445	b-glacio-nival
Engadin	Inn @ S-Chanf	InnSCh	616	2460	b-glacio-nival
Southern Alps and Ticino	Maggia @ Locarno *	MagLcn	927	1534	nivo-pluvial méridional
	Isorno @ Mosogno	IsoMsg	125	1570	nivo-pluvial méridional
	Riale di Calneggia @ Cavergno	RdCCav	23.9	1986	nival méridional
	Krummbach @ Klusmatten	KruKlu	19.4	2265	nival méridional

* Flows at these sites are impacted by hydropower operations (see **?**)

† Flows at these sites are impacted by regulated natural lakes

- provide 4 ensemble members from 2008/03–2015/05, and 10 ensemble members from 2015/05 onwards, and
- provide forecasts for a period covering 31 days (2008/03–2015/05), and 46 days from 2015/05 onwards.

In addition, the IFS underwent regular updates over the years, often multiple times per year (Tab. 2). To benefit from as much data as possible and derive a large collection of yearly time series for hydrological modelling, data from the different IFS cycles initialised within the same year were ~~not treated differently~~treated uniformly.

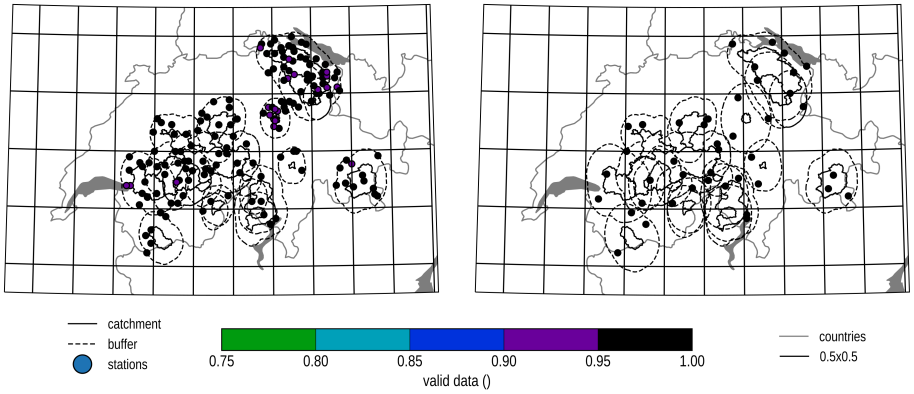


Figure 1. Study region with catchments (solid black lines) and meteorological stations (dots), also showing temporal coverage of station data for precipitation 1961–2019 (left) and temperature 1991–2019 (right). The 0.5° grid (~38×56 km^2) shown is equivalent to that of the reforecasts used. Dashed black lines indicate the buffer zone around the catchment used in the statistical downscaling.

However, a ~~detailed analysis regarding inconsistencies and drifts revealed two issues~~visual examination of the RF precipitation and temperature by forecast lead time and by initialisation year revealed two systematic inconsistencies (Fig. **??** in the Supplement):

1. The precipitation data show a spike on the transition day between medium-range ensemble (ENS) RFs (days 1–15) and the extended range RFs (days 16–46)~~(see Fig. ?? in the Supplement, left).~~ .
2. The temperature data for RFs initialised in 2015 have a strong drift over the forecast period~~(see Fig. ?? in the Supplement, right).~~ .

To address the first issue, we discarded the first 15 days and used only data from day 16 onward, corresponding to the extended range RFs. This also ensures a high degree of independence among individual ensemble members, as discarding the first 10 days alone would have been ~~already~~ sufficient for our application domain (**?**). To resolve the second issue, we ~~excluded~~ discarded the year 2015 entirely~~from the analysis. Moreover, we did not include~~ . We also excluded the year 2008, as it does not provide a full year of data. Residual inhomogeneities arising from the evolution of the IFS model across the archive period were further addressed in the bias adjustment (Sect. 3.2.1) and concatenation steps (Sect. 3.2.4). As a result, data with a total length of more than 10 000 years were available for analysis and processing.

2.3 Meteorological data

We employ a statistical model to relate the gridded RF data to the point scale. As a meteorological reference for grid-scale precipitation, we used RhiresD, a spatial precipitation analysis ~~available from~~ provided by **?**, which provides daily precipitation data from 1961 onwards at a 1 km raster resolution. Additionally, we evaluated the hourly MeteoSwiss product CombiPrecip (**?**), which combines weather radar fields with point precipitation measurements from 2005 onward, also at a 1 km resolution.

Table 2. Reforecasts by initialisation year, showing number of initialisation per week, number of years reforecasted, number of members, the available forecast length (from day 16 onwards) and the respective IGS IFS cycles for each year (**???**). Note that years 2008 and 2015 were not used in this study.

year	initialisation frequency	n years before	members	forecast length	IFS cycles
2008 *	weekly	18	1+4	16	32r3V; 33r1; 35r1/33r2; 35r2
2009	weekly	18	1+4	16	35r2; 35r3; 36r1
2010	weekly	18	1+4	16	36r1; 36r2; 36r4; 37r2
2011	weekly	18	1+4	16	37r2; 37r3; 38r2
2012	weekly	18 / 20	1+4	16	38r2
2013	weekly	20	1+4	16	38r1; 38r2; 40r1
2014	weekly	20	1+4	16	40r1
2015 *	weekly / twice a week	20	1+4 / 1+10	16 / 30	40r1; 41r1; 41r2
2016	twice a week	20	1+10	30	41r2; 43r1
2017	twice a week	20	1+10	30	43r1; 43r3
2018	twice a week	20	1+10	30	45r1
2019	twice a week	20	1+10	30	45r1; 46r1
2020	twice a week	20	1+10	30	46r1; 47r1

* Not used in this study

However, ~~we used CombiPrecip~~ CombiPrecip was used only for disaggregating daily precipitation (Sect. 3.2.3) because the data, despite MeteoSwiss' elaborate processing (**?**), are noticeably affected by terrain shielding of the radar beam~~by terrain~~. Moreover, CombiPrecip data are only available from 2005 onward, which is too short for robust bias adjustment. We also examined two further gridded observation products but discarded them due to their limitations in the present context: E-OBS (**?**), which has a comparatively low station density; and WFDE5 (**?**), which matches the grid point resolution of the coarse RFs (0.5°) but is ~~misaligned~~ offset by half a grid point.

As a meteorological reference for grid-scale temperature, we considered a spatial analysis of mean, minimum and maximum temperature ~~available from~~ provided by **?**, which provides daily data from 1961 onwards at a 1 km raster resolution. However, we ~~opted to directly correlate the~~ instead directly bias-adjusted RF temperature data ~~to the station scale~~ against station observations and did not use this gridded analysis. Unlike precipitation, which exhibits strong stochastic sub-grid variability that direct quantile mapping cannot reproduce without introducing artefacts (**??**), temperature varies smoothly in space, making station-based bias adjustment appropriate.

As a meteorological reference at the point scale for both precipitation and temperature, we used station data obtained from **?**, retrieved 01.05.2021. The temporal and spatial coverage varied between the available daily and the hourly gauging networks, and for consistency with the approach used in the projects EXAR/EXCH, we used hourly station ~~data~~ observations aggregated to the daily scale for both the stochastic downscaling of precipitation and the quantile mapping of temperature (see Sect. 3.2).

2.4 Hydrological data

The continuous hydrological data encompassed hourly discharge records for the test catchments listed in Tab. 1. These data were available for a maximum period from 1974–2021. However, most time series were shorter, with a median length of 30 years and a range of 6–47 years. In addition, records of annual maximum floods (AMFs) were available for most of these stations. These records refer to instantaneous peak flow and date back even further, with a median length of 63 years and a range of 23–118 years. Extrapolations of the observed AMFs via the generalized extreme value (GEV) distribution were also available for these stations (**?**). Maximum flood data from all over Switzerland were available from **?**, covering measurements and historical reconstructions at 740 sites in total.

2.5 Hydrometeorological scenarios based on weather generator

For comparison, 300 000 years of hourly simulation results from a hydrometeorological model chain using the Generator of Weather EXtremes (GWEX) WGEN were available. GWEX is a multi-site, two-part stochastic weather generator for precipitation and temperature, based on the structure proposed by **?**. It is designed to reproduce the statistical behaviour of weather events at various temporal and spatial resolutions, with a focus on extremes. Since comparatively long events are relevant in the main EXAR/EXCH framework, GWEX first generates 3-day precipitation amounts. These amounts are then disaggregated into daily and ultimately hourly values using meteorological ~~analogues~~analogs. Details on GWEX are found in **??**, while its application to flood estimation is discussed in **?** and ~~**?**~~**?**.

3 Methods

~~As mentioned in the introduction, some methodological choices from the EXAR/EXCH framework (??) were retained to facilitate comparisons. A lumped catchment model – combined with hydrological routing where necessary – was run using mean catchment~~

3.1 Experimental design

The methodology was designed to address the research questions outlined in the Introduction by processing ECMWF RF data into spatially coherent hourly series, concatenating these into multi-millennial continuous sequences and evaluating their suitability for long continuous hydrological simulations to estimate flood frequency at the catchment scale.

The workflow starts with bias adjustment of RF precipitation and temperature ~~inputs interpolated from point values. The raw RF datawere bias adjusted, stochastically downscaled~~ data, stochastic downscaling to multiple meteorological stations, ~~and disaggregated to hourly temporal resolution . Subsequently, mean catchment precipitation and temperature were interpolated~~ disaggregation to hourly resolution and concatenation into long series (Sect. 3.2). Mean catchment values were then derived via interpolation and adjusted to mean catchment elevation (Sect. 3.3). The hydrological simulations employed a lumped catchment model with hydrological routing where necessary. As noted in the Introduction, we retained the main components

and methodological choices of an established long CS framework (**??**) to enable direct comparison with simulations based on a stochastic weather generator. The only modification in the present study is that the meteorological forcing is derived from RFs rather than from a stochastic weather generator.

3.2 Statistical postprocessing of reforecasts

~~The processing of the RF data involved multiple steps~~ Statistical postprocessing was applied to transform the coarse-resolution RF data into spatially coherent, hourly time series of precipitation and temperature in Alpine catchment conditions, while preserving extreme events beyond the observed record (Fig. 2)~~, which~~ . All steps were applied to ~~all test catchments~~ each test catchment (Sect. 2.1).

As a model-derived product, the RF data are subject to biases, particularly at longer lead times~~. The statistical modelling chain served three main purposes: adjustment of model biases~~, and they are available at coarse spatial and temporal resolution. The postprocessing therefore comprised three components: bias adjustment, downscaling to ~~the point scale to multiple meteorological stations, and performing~~ station locations, and temporal disaggregation.

~~Bias adjustment is often calibrated between climate model output and station data as a simple downscaling method that both adjusts model biases and downscales to the point-scale station data. This approach has been shown to cause substantial artefacts when applied to precipitation(?). We therefore~~ For precipitation, we followed the conceptual approach by **?** and first bias-adjusted the gridded RF precipitation data at ~~its native grid and subsequently statistically downscaled the adjusted data. The bias adjustment was carried out~~ their native resolution using quantile mapping against the gridded observations of RhiresD (Sect. 3.2.1). ~~Subsequently, a~~ This separation of bias adjustment and downscaling avoids artefacts associated with direct calibration to station data (**?**). The adjusted precipitation data were subsequently downscaled to multiple locations using a multi-site ~~stochastic downscaling method was applied to the adjusted RFs to generate spatially coherent precipitation time series at multiple station locations (?). This model is~~ stochastic downscaling model based on a truncated and transformed multivariate Gaussian distribution (**?**), preserving spatial coherence among stations (Sect. 3.2.2). ~~To derive the hourly precipitation time series, daily precipitation values were disaggregated via~~ Daily precipitation totals were then disaggregated to hourly resolution using analogs (see Sect. 3.2.3).

Two-meter air temperature (TAS) was bias adjusted directly to the station locations using daily data, ~~and applying the correction value~~ with correction factors applied to the 6-hourly RF data. ~~Hourly values were subsequently built, after~~ After concatenating single 15/30 day long RFs~~via a~~ , hourly series were built using cubic spline interpolation~~of 6-hourly temperatures. Time series of~~ . Finally, mean catchment precipitation and temperature were ~~finally derived via~~ derived by Thiessen interpolation, using a constant adjustment factor for precipitation and a prescribed lapse rate for temperature ~~as described in~~ following **?**.

3.2.1 Bias adjustment

In a first step, the RFs were ~~corrected~~ bias-corrected separately for each month via basic quantile mapping (QM) (**?**), a common state-of-the-art bias adjustment method. ~~We applied quantile mapping to the RF data separately for each month.~~ For precipitation, we used RhiresD as a reference, with daily data from the period 1991–2020. For temperature, we used station

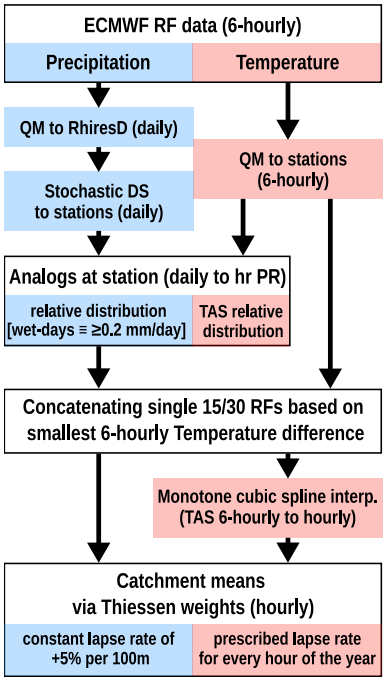


Figure 2. Workflow of statistical postprocessing, beginning with ECMWF raw reforecast (RF) data and proceeding through bias adjustment and stochastic downscaling (DS), application of analogs, concatenation, and interpolation to mean catchment values. PR is precipitation, TAS is temperature.

data located within a ~10–20 km buffer zone around the catchment as a reference, with data from the period 1991–2020. Temperature ~~has been~~ was bias corrected using daily data, and applying the correction value to the 6-hourly data. Since the bias structure for temperature and precipitation in the RFs varied by year of initialisation (see Sect. 2.2), the periods 2009–2014 and 2016–2020 were adjusted separately ~~.~~(excluding 2015, which was discarded, see Sect. 2.2).

~~Due to the large sample of ensembles, RFs include extremes not yet observed, and a~~ A key objective was to retain the expected upper tail of the RF precipitation distribution, as the added value of RFs for flood estimation lies precisely in their ability to sample extremes beyond those observed. Since direct bias correction would ~~result in the elimination of~~ eliminate such heavy precipitation events~~when fitting to the distribution of the observational data. This would negate the key advantage of the RFs, which is their ability to provide insight into unseen extremes. Therefore,~~ , a transfer function was ~~constructed by estimating the RF data distribution~~ applied: the RF distribution was estimated using non-overlapping samples of the same length as the observational data~~. Simply put, the raw RF data were corrected by the~~ , and corrections were applied based on quantile differences between ~~the~~ observations and the ~~sample mean to derive the corrected RF data~~ RF sample mean (Fig. 3).

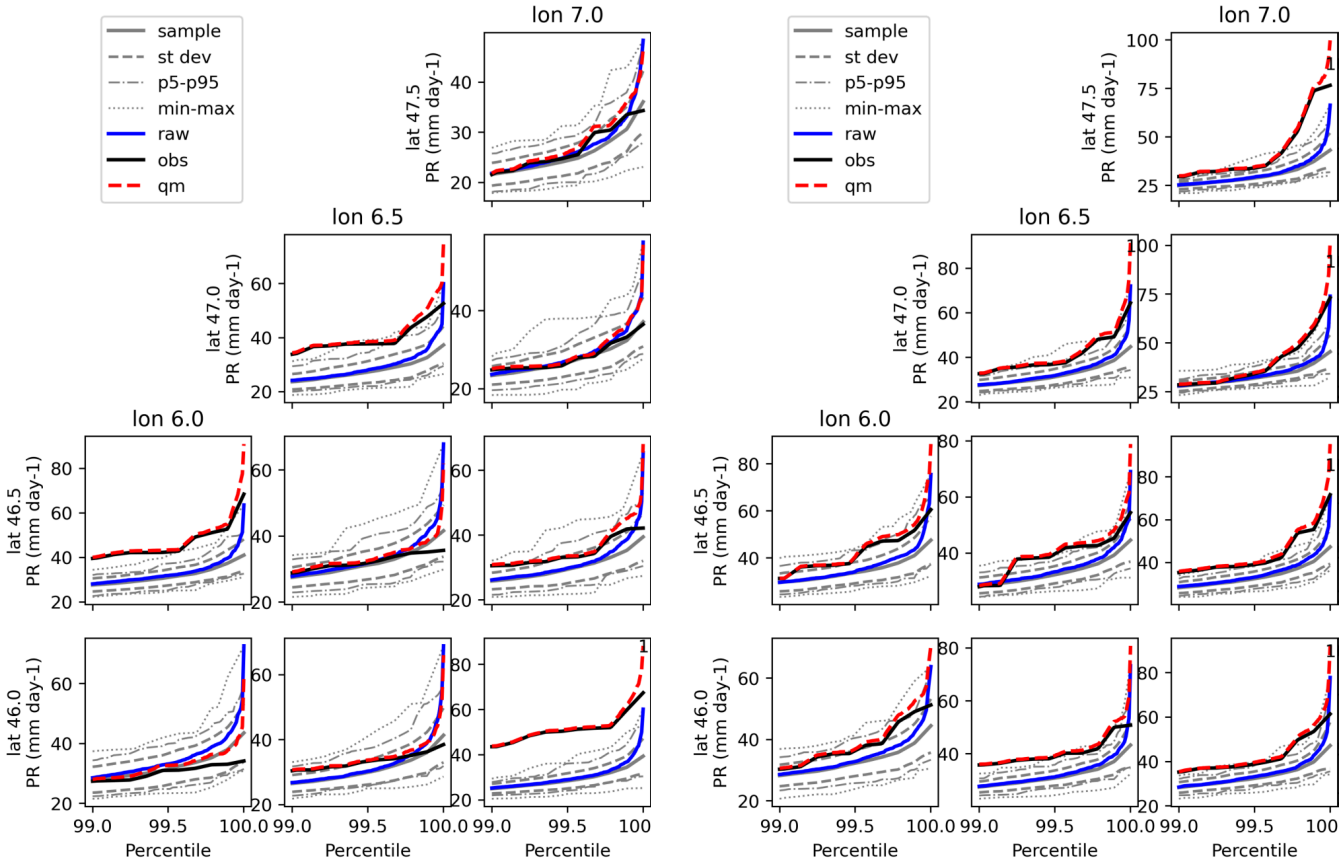


Figure 3. Daily precipitation (mm day^{-1}) starting from above the 99th percentile for January (left) and August (right). Shown are RhiresD ('obs', dashed black line), ECMWF raw reforecast data ('raw', blue line), bias adjusted reforecast data ('qm', red dashed line), and the transfer function on which the correction is based ('sample', gray solid line) for selected grid points over Switzerland.

3.2.2 Stochastic downscaling

~~For stochastic downscaling , we~~ The downscaling ensures that small-scale spatial variability characteristic of Alpine precipitation is represented consistently across multiple stations, which is essential for realistic flood simulations at the catchment scale. We used a stochastic downscaling approach and adapted code of **?** to handle missing station data. ~~We then~~

We applied this stochastic downscaling for each month and for each test catchment individually, using RhiresD and daily station data for calibration over the period ~~1961—2019 for each month and for each of the test catchments separately. We did not use stations that had~~ 1961–2019. Stations with more than 25% ~~of~~ missing data during this period ~~. Using~~ were excluded. Calibration with 6-hourly instead of daily station data ~~for calibration yielded~~ produced unrealistic values for some sites and was therefore discarded. In addition to stations within ~~a catchment, all stations that were~~ each catchment, we included all

stations located within a ~10 km buffer zone around the catchmentwere used. We applied the adjustment of the mean bias to the simulated data; however, the .

We adjusted the simulated data for mean bias but omitted the quantile-based multiplicative bias adjustment was omitted as because it led to poorer results in a split sample performance in a split-sample test (see **?**). We also tested the impact of extending the range of grid points included as predictorspredictor grid points, but found no discernible difference between including grid points of up to 0.75° or 1.25° distance from the nearest stations. Therefore, we We therefore used grid points with a distance of up to within 0.75° for the final downscaling.

3.2.3 Analogs and disaggregation

Sub-daily precipitation dynamics are decisive for flood peaks in Alpine catchments. Therefore, realistic hourly disaggregation is critical for evaluating the suitability of RFs for flood frequency estimation. The bias-corrected and downscaled daily precipitation data had to be were therefore disaggregated to the hourly scale necessary resolution required for hydrological modelling. For this, We applied the analog method was used as because it is relatively easy to implement , straightforward to implement and generally produces spatially and temporally coherent results. To have provide a large pool of reference data to draw the analogs from, we used daily and hourly station data (point valuesmeasurements) as well as hourly CombiPrecip data (values on 1×1 km^2 raster).

First, stations within each single catchment and its vicinity were selected. Then, for each station, days with both daily and hourly data available over the period during 1981–2019 were identified. For days where with only daily station datawere available, disaggregation was performed using daily fractions of hourly CombiPrecip values from the 2005–2019 period based on daily fractions derived from hourly CombiPrecip data for the period 2005–2019. Analogs were finally drawn for every month separately drawn separately for each month using a moving window of ± 1 month. Further, the data were separated into dry and wet days (using a threshold of 0.2 mm day^{-1} for daily catchment precipitation interpolated via Thiessen weights)with Thiessen weights. For dry days (<0.2 mm day^{-1}), the hourly precipitation data were hourly precipitation was set to zero. For wet days, the daily precipitation data daily precipitation totals were disaggregated to hourly values by using relative hourly contributions of the selected analogs at the station scale. As reference, all days within plus or minus one All days within ± 1 month of the respective modelled day were used. The best analog day was determined by calculating the RMSE at the station locations of both identified by computing the RMSE of precipitation and mean daily temperature to all reference days , at the station locations for all candidate days and selecting the analog day that showed the smallest mean precipitation and temperature day with the smallest combined RMSE.

3.2.4 Concatenation

The single RFs, which depending on their initialisation yearhave a length of 15 To enable long continuous simulation, the short (15- or 30 days, had to be concatenated into yearly time series for use in continuous hydrological simulations. A 30-day, depending on the initialisation year) RF segments were concatenated into synthetic annual sequences while minimizing artificial discontinuities at the stitching points. For consistency across years, a length of 360 days was assumedfor each year,

consisting of 24 ~~and~~ or 12 single RFs, respectively. ~~For each year, drawing of individual RFs was restricted to~~ RF segments were drawn exclusively from the same initialisation year ~~, minimizing the mixing of RFs from~~ to avoid mixing different IFS cycles~~. To reproduce~~ , which could introduce inconsistencies (cf. Sect. 2.2).

To approximate the annual cycle, ~~we selected single RFs based on the dates that roughly represent a~~ RFs were selected such that their initialisation dates collectively span one calendar year. For instance, for forecasts before 2015 (initialised weekly ~~,~~ ~~length~~ with a length of 15 days), ~~we built years~~ synthetic years were constructed using every second initialisation week, ~~which resulted~~ resulting in two sets of years with ~~the same~~ identical initialisation dates. ~~This approach introduces a slight offset with the ongoing year, due to the difference of~~ Because the 15-day forecast length ~~and~~ differs from the 14-day ~~gap~~ interval between initialisation dates~~. Therefore,~~ , a small temporal offset accumulated over the year. This was addressed by discarding single initialisation dates~~had to be discarded to keep this offset reasonably small. With this approach, a total~~ . This procedure yielded a total of 9920 ~~single~~ synthetic 360-day years~~were obtained.~~ .

The selection of ~~the single RFs building~~ individual RFs used to construct a year was based on minimizing ~~the~~ temperature difference at the ~~intersection~~ stitching points at the catchment scale. ~~Using raw~~ Raw 6-hourly temperature ~~enabled us to minimize the temperature difference~~ data were used to evaluate differences at the same time step. Grid points were weighted according to ~~the~~ station locations and their Thiessen weights. At the ~~start of building years for each~~ beginning of the assembly for a given initialisation year, ~~succeeding~~ subsequent RFs can be drawn from a large pool ~~(reforecasted years ensemble members). With each built year, this pool diminishes resulting in~~ of reforecast-year ensemble members, allowing selection of RF segments with minimal temperature discontinuities. As more years are constructed, the pool decreases, leading to larger temperature differences ~~to succeeding RFs~~between consecutive RF segments. Mean differences ~~over~~ across all date combinations ~~stayed~~ remained well below 1 °C for about ~~three-quarters of the selected reforecasted years ensemble members selection steps ,~~ three quarters of the ensemble-member selection steps and increased to about 5 °C for the ~~last built year (see Fig. ?? in the Supplement, left)~~final year assembled in the sequence. The mean overall maximum temperature differences ~~of the single~~ across individual years and selection steps ~~increases~~ increased from approximately 0.1 °C to about 14 °C for the last ~~built year(see Fig. ?? in the Supplement, right). The differences shown are for temperature at 6-hourly resolution.~~ assembled year. In the final hourly time series, derived from 6-hourly ~~temperature station data via a~~ station temperature data using monotone cubic spline interpolation, ~~differences are around one-sixth of the valuesshown.~~ these differences were around one sixth of those values. A detailed view of the temperature differences at the stitching points is provided in Fig. ~~??~~ **??** in the Supplement~~shows one of the extreme gap years with an extreme~~ , and Fig. **??** illustrates a year with a pronounced temperature jump. The ~~behaviour of the hydrological simulations in relation to the stitching points is examined~~ influence of these stitching artefacts on simulated AMFs is evaluated in Sect. 4.2.

3.3 Continuous hydrological simulation and routing

The following hydrological modelling framework was used to evaluate whether RF-based meteorological inputs produce internally consistent discharge simulations suitable for flood frequency analysis within a long continuous simulation approach. The complete concatenated RF precipitation and temperature data were finally used as input for the bucket-type catchment

model HBV (Hydrologiska Byråns Vattenbalansavdelning model, see **?**). HBV was calibrated on hourly data spanning 1983–2019, and run at an hourly ~~time-step~~time step, employing the non-linear response function of **?** to ~~enhance the representation of flood behaviour~~ better represent flood behaviour (see **?**). To achieve ~~the time series of~~ mean areal precipitation and mean areal temperature time series required for hydrological modelling in the given long CS framework, the RF values downscaled to ~~locations of meteorological stations were interpolated via~~ meteorological station locations were interpolated using Thiessen polygons. ~~Based on the difference of Thiessen weighted~~ Precipitation was adjusted for differences between the Thiessen-weighted mean station elevation ~~to~~ and the mean catchment elevation, ~~a constant adjustment~~ applying a constant factor with a linear increase of 5% for every 100 m ~~was used for precipitation~~ (see e.g., **????**). For temperature, ~~calendar lapse rates estimated from data~~ calendar-based lapse rates were estimated from 1981–2019 data for each hour and each day of the year~~(and differentiated by~~, differentiated for six large river basins ~~) were used (see ?)~~(see **?**).

~~Observations of~~ Observed precipitation and temperature data from 1931–2019 were interpolated similarly to mean areal values and used ~~as a basis~~ for a control run with HBV.

In catchments strongly affected by hydropower operations, lake retention, lake regulation, bank overflow or floodplain retention, hydrological routing was applied using RS Minerve (**?**) to account for these effects.

~~For further~~ Further details on the hydrological simulation set-up ~~we refer to~~ are provided in **?**.

4 Results

4.1 Stochastic downscaling

Fig. 4 shows simulated and observed mean daily catchment precipitation (left) and mean wet day frequency (right). Note that there is a mismatch between the station observations used in processing the RFs and the data shown at the point scale: The station observations are based on all available observed days over 1991–2019, which roughly coincides with the period used for the quantile mapping (1991–2020), while the stochastic downscaling was calibrated using grid data and station observations over the period 1961–2019. The stochastic downscaling reproduced these mean characteristics well. Fig. 5 shows simulations and observations at single station locations (circles) and in the Thiessen weighted mean (crosses) for the catchment of the Aare River at Bern. Also for the single stations, the mean characteristics are reproduced well. There is a tendency for a small underestimation of mean precipitation for single stations and months, and some deviations from observed wet-day frequency, which might be explained by the different data periods used for the bias adjustment and for calibrating the stochastic downscaling, as well as associated sampling differences of ~~large scale~~ large-scale modes of variability such as the Atlantic Multidecadal Oscillation (AMO) (see also Sect. 5.4).

The following evaluation of the downscaling results refers to the catchment scale. A clear and concise evaluation is hindered by two aspects. First, the bias adjustment of temperature and precipitation was done using reference data spanning the periods 1991–2019 and 1991–2020, respectively, while the stochastic downscaling was trained using grid and station data over the period 1961–2019, and the analogs were built using station data over 1981–2019. We here use the periods 1991–2019 and

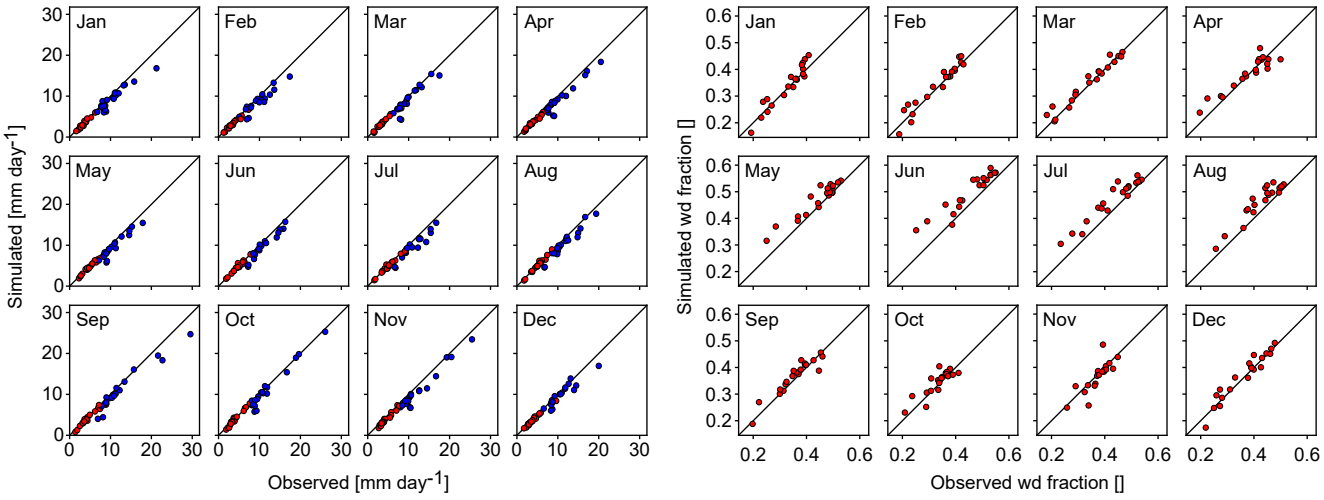


Figure 4. Left: Mean simulated and observed daily precipitation per month over all catchments for all day mean (red circles) and wet day mean (>1.0 mm day^{-1}, blue circles). Right: Mean simulated and observed wet day frequency per month (expressed as fraction of days per month) over all catchments (red circles). Station observations are for the period 1991–2019.

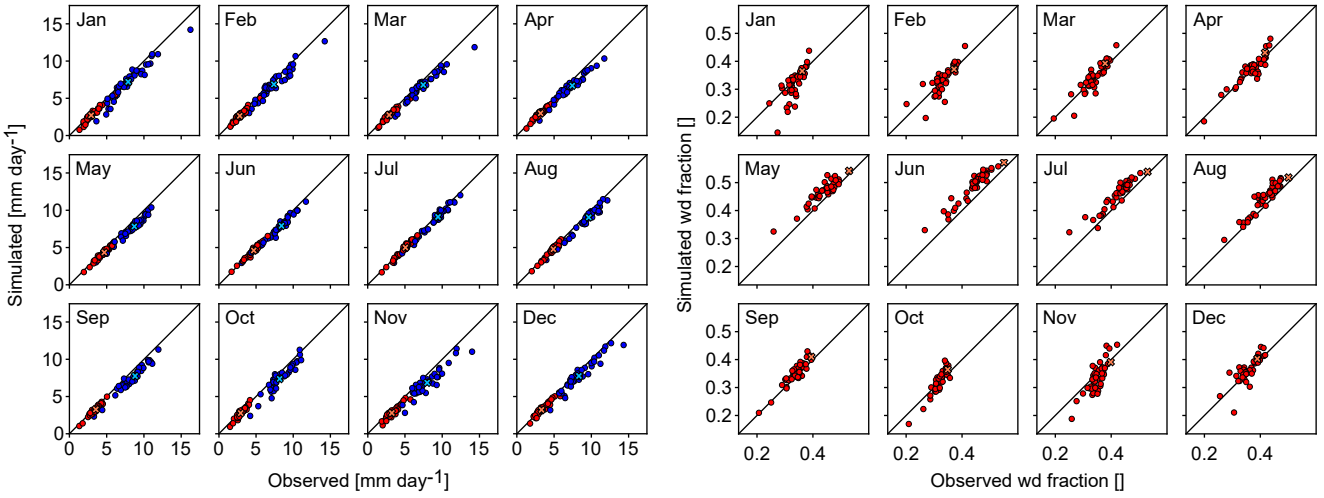


Figure 5. Same as Fig. 4 but for single stations within the Aare River catchment upstream of Bern (AarBrn). The crosses denote mean catchment all day and wet day precipitation and wet day frequency.

345 1981–2019 as reference. Second, the resulting 360-day calendar was built from the single RFs, which ~~leads~~ led to small shifts in some of the seasonal statistics presented.

For monthly mean precipitation (Fig. 6, first column~~; for all results see Figs. ?? ?? in the Supplement~~), good agreement was found between observed and downscaled RF values for large and medium-sized catchments. Good agreement was also

found for several small and very small catchments, while others – in particular Lonza, Saltina and Krummbach – showed notable differences for single months, with the mean observed annual cycle being outside the interquartile range (25th to 75th percentile) of the RFs.

In terms of annual 90th percentile daily precipitation (Fig. 6, second column), good agreement was found between the observations and the RFs for large and medium-sized catchments. Specifically, the values for the single observed years lie within the distribution of the RFs. While the same is true for many of the small and very small catchments, observed values for single years are larger than all downscaled values for the Lonza, Saltina and Krummbach river catchments, and smaller than all downscaled values for the Rein da Sumvitg river catchment.

Annual maximum 1-day and 3-day precipitation (Fig. 6, third and fourth column) show good agreement for the 29 years of observations used for large and medium-sized catchments, with the RFs falling within the confidence intervals of the observations. For 1-day maxima, the RFs show higher return levels for return periods greater than 10 years over the Thur river catchment at Jonschwil and at Andelfingen. For 3-day maxima, no such disagreement is noted. For 3-day maxima in the Inn and Kleine Emme river catchments, RFs indicate a smaller return value than observed above the 10-year return period. There is also a good agreement for several of the small catchments, however, over the Isorno river catchment, both 1- and 3-day maxima deviate strongly from the observed return levels. The largest differences are found over the very small catchments starting already at a return period of 1 year.

The analog method was used to disaggregate daily precipitation. Results over the catchments are mixed, also within catchments of the same scale range. For many of the large and medium-sized catchments, results indicate a relatively good overlap with the intensity-duration-frequency (IDF) curves of the observations, in particular for the 2-year return level, with the exception of the Sarine, Inn and Minster river catchments~~(see Figs. ?? and ?? in the Supplement).~~. For small and very small catchments, the downscaled RFs show overly strong precipitation intensities in particular for short durations~~(see~~. A detailed comparison of the full IDF curves is provided in Figs. ~~?? and~~ **??**–**??** in the Supplement~~)~~.

IDF curves for 1- to 7-day precipitation (Figs. **??**–**??** in the Supplement) show good agreement between the downscaled RFs and the observations for the 2-year return level for all large and medium-sized catchments, and the small catchments Allenbach, Drance de Bagnes and Thur at Alt St. Johann. The remaining catchments show larger differences. For some of the catchments showing a good performance, there are indications that higher return level precipitation episodes spanning several days are underrepresented (e.g., the 25-year return level over the Aare at Thun), whereas they are represented well for other catchments (e.g., the Sarine river catchment).

Fig. 7 shows density plots based on annual and seasonal mean as well as daily mean temperature and precipitation for the large Aare River at Bern and the very small Krummbach River catchments. The observed seasonal and annual mean values (red crosses, rows 1, 2, 4 and 5) are within the density cloud of the RFs, except for single extreme years and seasons. For very small catchments, precipitation means tend to be closer towards the edge of the point cloud. Seasons in rows 1 and 4 were built using the RF forecasted days, while seasons in rows 2 and 5 are based on the 360-day calendar of the final time series. Not surprisingly, the concatenation of the single RFs to a 360-day calendar artificially increases seasonal variability, in particular for the transitional seasons. At the same time, observed extreme years indicated in the seasonal means are not depicted by

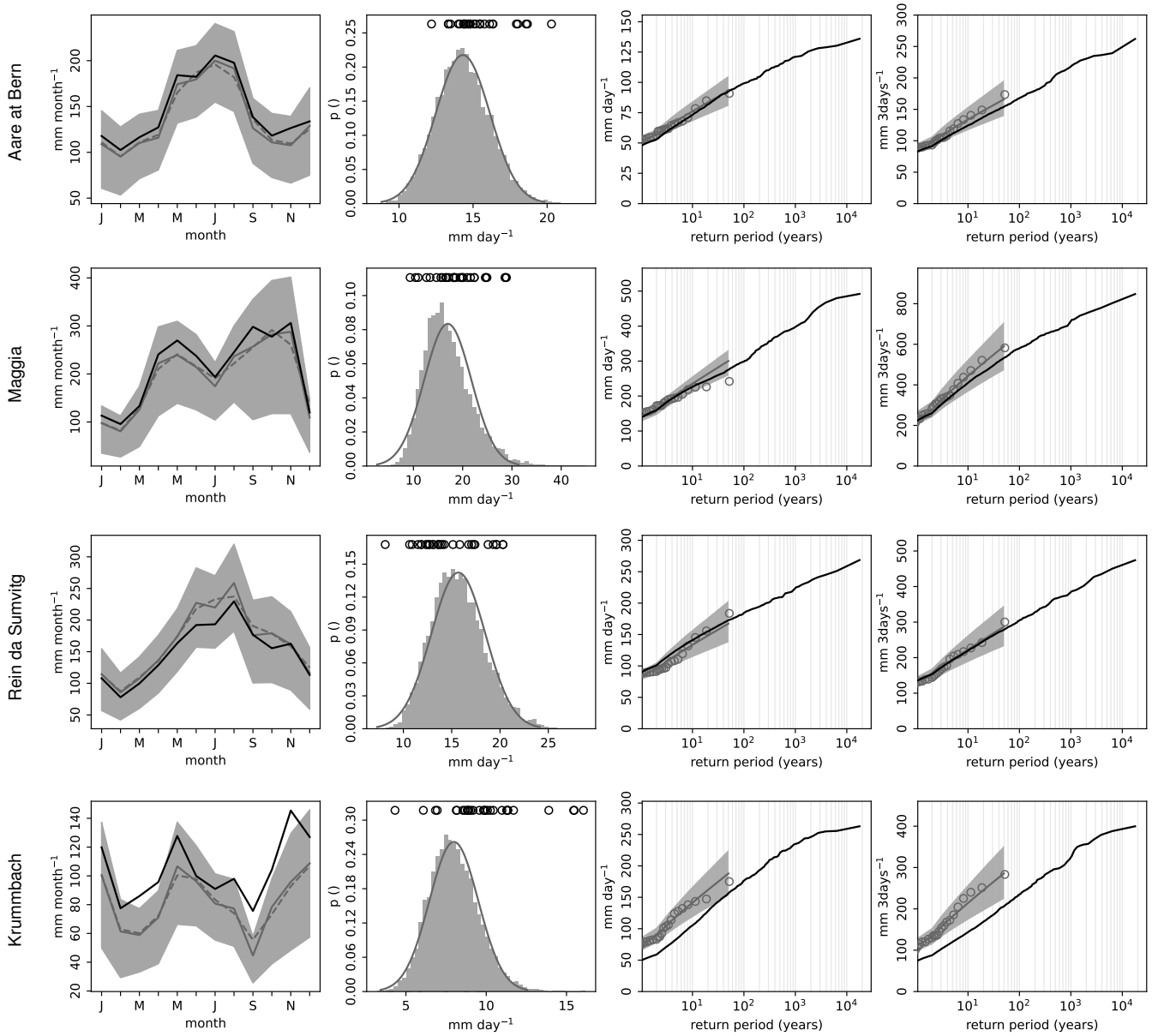


Figure 6. Evaluation of reforecast (RF) precipitation for selected catchments across different scale ranges: Aare at Bern (large), Maggia (medium), Rein da Sumvitg (small) and Krummbach (very small). From left to right: Annual cycle of mean monthly precipitation for observations (1991–2019, black) and RFs (grey; interquartile range (25^{th}–75^{th} percentile) in shading; mean based on 360-day calendar in dashed line); histogram of annual daily 90^{th} percentile precipitation for RFs, and observations (black circles); return levels for annual maximum daily and maximum 3-day precipitation for RFs (black) and observation (grey circles) using the Gringorten plotting position. For observations, return levels have additionally been fitted using a Gumbel distribution (using a GEV distribution resulted in overly broad confidence intervals) and maximum likelihood estimates, confidence intervals are based on 5000 bootstrap samples (mean in dark grey, 2.5^{th}–97.5^{th} percentile in grey shading). Results for all test catchments are shown in Figs. **??**–**??** in the Supplement.

the RFs, which is likely due to the opportunistic concatenation of the single RFs to yearly time series. Also for the remaining catchments there is overall a good agreement between RFs and observations~~(see Figs. ??–?? in the Supplement).~~. However, there are some indications of the RFs being slightly warmer than the observations in spring in the catchments of Thur (at Jonschwil and at Alt St. Johann), Maggia and Isorno. Density plots for all catchments examined are available in Figs. **??–??** in the Supplement.

4.2 Hydrological validation

To ~~validate~~ assess the suitability of the bias-corrected and downscaled RFs for hydrological simulations, we ~~calculated~~ derived flow duration curves (FDCs) from the RF-based simulations. ~~To achieve these, we sampled daily values~~ Daily values were sampled over the length of the ~~observations (or, if shorter, the length of the control run)~~ observational record, or the control run if shorter, from each block of 1000 years~~and calculated the exceedance probabilities~~ , and exceedance probabilities were computed separately. For ~~plotting~~visualisation, we calculated the 95% confidence intervals for each exceedance probability over the flow values and show these together with the observations and the control run. Comparing the FDCs from RF, control run and observations for selected sites (Fig. 8~~in the Supplement~~), we find very similar behaviour for the Aare River at Bern, the Thur River at Andelfingen, and the Minster River. ~~However~~In contrast, for the Maggia River and~~especially~~ , more pronouncedly, the Inn River, both RFs and control run ~~show higher values than observations~~ yield higher flows than observed in the upper flow ranges. ~~In the case of the~~ For the Maggia River, control run and RF-based simulation agree well to very well. The discrepancy to observations is partly or entirely due to the absence of the large flood events of 1981 and 1982 ~~in~~ from the records, as the station was destroyed ~~by a massive~~ during a severe flood in 1978~~and did not resume operation until~~ , and measurements resumed only in 1995 (**?**). ~~In~~ For the Saltina River, control run and observations agree well, ~~but~~ whereas RFs show lower values overall, particularly in the upper flow ranges. This suggests limitations in the RF input for reproducing large floods in this small catchment.

We also examined the seasonality of AMF occurrence in the RF-based hydrological simulations and compared it with the seasonality in control run and observed AMFs at selected sites (Fig. 9~~in the Supplement~~). The months with frequent AMF occurrence are generally consistent, with differences typically no greater than one month. However, discrepancies are noted in the strength of AMF seasonality in some cases. For example, in the Inn River, RF-based AMFs are more evidently distributed across the months of frequent occurrence and do not show the pronounced peak in observed AMF occurrence centered in June, which is well captured in the control run. For the interpretation of all sites, it should be noted that the time periods and their durations for AMF observations, control run and RFs do not fully align.

As individual RF segments were concatenated to obtain a continuous hourly time series (see Sect. 3.2.4), we verified whether the resulting AMFs do not show exceptional behaviour by comparing the time of occurrence and magnitude of the AMF with the stitching dates. For the smaller catchments, the densest occurrence of the AMFs is within a relatively narrow time window, occurring markedly after the stitching point, typically at least 100 days later. For the larger catchments, the AMFs are more widely distributed throughout the entire year following the stitching. ~~Fig. ?? in the Supplement shows results for a small and a~~

~~larger sub-catchment from each large river basin.~~ In conclusion, the simulated AMFs appear to be unrelated in their magnitude and timing to the stitching points ~~.~~ (see Fig. **??** in the Supplement for details).

4.3 Flood estimation results

Exceedance curves ~~based on RF inputs are shown~~ for selected examples are shown in Fig. 10. In the following, we compare RF-derived AMFs with observed AMFs, control-run AMFs simulated with observed weather for 1931–2019, and AMFs simulated using weather generator inputs (GWEX).

Looking at large catchments (Fig. 10, top row), very good agreement between RF-based AMFs and observed AMFs is found for the Aare River at Thun. Note that the observed time series ~~is~~ are not stationary at this site (**?**) because a flood relief tunnel was ~~taken~~ put into operation in 2009, and regulation rules for the upstream lakes Thun and Brienz were altered. The CS modelling chain used in this study depicts the current state, and comparison to observations of the period 2009–2021 is most appropriate, even though the statistics of observed floods from this period show an exceptionally wide confidence interval due to the short record length. Further downstream at Bern (not shown), ~~the return levels based on~~ return levels derived from RF simulations are higher ~~in comparison to those of observations and GWEX,~~ than those based on observations and on the control run, whereas GWEX-derived AMFs show better agreement with observations. The discrepancy in the RF-based results is presumably due to ~~overestimation of floods~~ an overestimation of flood contributions from smaller tributaries ~~that join~~ joining the Aare River. For the Thur River at Andelfingen, return levels ~~are higher than expected from observations for control run,~~ from the RF-based simulation ~~and GWEX-based simulation~~ exceed those inferred from observations for return periods greater than ~~5–10 years. It is difficult to decide whether this points at floods that could be systematically higher than expected, or whether the model tends to overestimate AMFs in this case~~approximately 5–10 years. The same is true for the control run and the GWEX-based simulation. However, at least three AMF measurements prior to 1998 were strongly influenced by bank overflow, before the river channel was modified to convey higher flows (**?**). An adjusted dataset incorporating reconstructions of unattenuated AMFs (**?**) and the resulting flood statistics shows substantially better agreement with all simulations, which consistently represent the current, modified river state. For the large Maggia River catchment, which exhibits very challenging meteorological conditions, the RF-based flood exceedance curve agrees remarkably well with that of the observations. The highest AMFs simulated from RFs, with estimated return periods between 1000 and 10 000 years, reach values of approximately 5300–8400 m^3s^{-1}. The corresponding return levels estimated from GWEX are approximately 5500–9150 m^3s^{-1}, which is higher, but does not indicate a fundamental disagreement.

For medium-sized catchments (Fig. 10, middle row), we note curves lower (Kander River) or higher (Inn River, Thur River at Jonschwil) than observed exceedance curves.

For small catchments (Fig. 10, bottom row), results matching well with observations are possible, such as for the Riale di Calneggia and Lonza rivers. But as expected, marked underestimation can also occur, like for the Saltina River, where the highest RF-based AMF is only slightly higher than the highest observed AMF in the ~~55 year~~ 55-year long records.

Table 3. Precipitation gauges used for juxtaposition with reforecasts and weather generator (GWEX), with coordinates (x, y) and elevation (z).

Code	Name	x [m]	y [m]	z [m a.s.l.]
BOS	Bosco-Gurin	680 879	130 027	1486
CEV	Cevio	689 688	130 565	417
BEP	Belp	605 140	193 805	515
FRF	Frauenfeld	709 480	270 170	393
GRH	Grimsel Hospiz	668 583	158 215	1980
INT	Interlaken	633 019	169 093	577
LTB	Lauterbrunnen	635 851	160 291	815
MSG	Mosogno	692 803	117 050	771
SAE	Säntis	744 200	234 920	2502
URN	Urnäsch	739 205	241 765	825

5 Discussion

5.1 ~~Comparison~~ Juxtaposition of reforecasts and weather generator precipitation

In the following, RF precipitation maxima are juxtaposed to the fundamentally different approach of constructing precipitation scenarios with the stochastic multi-site weather generator GWEX (see Sect. 2.5). Three catchments were selected for comparison as they well represent different conditions regarding climatology, station density and station representativeness: The Aare River at Bern, the Thur River at Andelfingen, and the Maggia River. These were evaluated in terms of annual peaks of areal precipitation, and three to four meteorological stations were selected ~~in addition~~ per river basin (Tab. 3) with focus on different station elevations.

To achieve a consistent ~~comparison~~juxtaposition, a time series of 9900 years was taken each from RF and GWEX, and divided into 99 blocks with a length of 100 years each for computing confidence intervals. Observations (OBS) with a maximum length of 90 years in the period 1930–2019 were furthermore available for juxtaposition, with confidence intervals computed from a parametric bootstrap General Extreme Value (GEV) distribution fitted with L-Moments. The following discussion focuses on return periods that can be reasonably estimated from OBS, approximately 150 years ~~as per~~ according to the **?** plotting position.

~~It~~ As context for the juxtaposition, it should be noted that ~~GWEX was parameterized using OBS data from the same period 1930–2019, whereas the~~ the RF data cover the ~~period~~ years 1991–2020 and were bias corrected using RhiresD over this period. ~~Instead, the~~ The stochastic downscaling was calibrated using the full available RhiresD ~~data~~ dataset 1961–2019 and station ~~data 1961–2019. Caution is therefore~~ observations from the same period. In contrast, GWEX was parameterized using OBS data from 1930–2019. Therefore, caution is warranted when comparing OBS, RF and GWEX, ~~especially considering~~ particularly

given that the Atlantic Multidecadal Oscillation (AMO) index went through different phases between 1930 and 2019, indicating a relevant impact of internal climate variability (see also Sect. 5.4). The latest heavy precipitation statistics by MeteoSwiss (**?**, retrieved 27.03.2023) are shown for further reference. These use the entire measurement series available up to 2022 at each station, roughly compatible with the block length chosen for RF and GWEX.

At-station maximum 1-day and 3-day precipitation sums (Fig. 11, columns 1–4) generally show good to very good agreement between OBS, RF and GWEX for all three catchments considered. The largest disagreement is noted at the station Mosogno (MSG) in the Ticino River basin, where RF values are lower than OBS and GWEX. At the level of mean catchment precipitation, agreement is also high in all three river basins (Fig. 11, right). GWEX confidence intervals tend to exceed those of the RFs in the upper range of maximum 1-day and 3-day precipitation sums for return periods larger than 50 to 100 years at both station and river basin level. This also means that GWEX in general reaches higher precipitation extremes in the 9900 years of data analysed as compared to the RFs. Given the differences between how the RF and GWEX time series are composed, it ~~appears not~~ does not appear possible to draw firm conclusions beyond this statement. Also, ~~it does not appear justified to make comparisons to the range found for observations since their~~ comparisons with the observational range are not warranted, as the corresponding confidence intervals are derived from ~~a maximum of~~ records of at most 90 years~~of record only~~. Both for 1-day and 3-day maxima, it is not possible to determine whether RFs or GWEX is closer to reality as regards high extremes.

5.2 Envelope curves for floods

Fig. 12 shows an evaluation of the ten highest RF-based flood estimates in the context of maximum flood (~~Qmax~~Q_{max}) data from across Switzerland (**?**). Two envelope curves ~~based on~~ derived from these data are shown. Their slope was derived from a ~~log-log~~ log–log linear regression, after which the ~~intercept was adjusted to either encompass~~ intercepts were adjusted to encompass either all data or 95% of the data, respectively. ~~Additionally,~~ In addition, two relevant envelope curves for European ~~(?) and global (?) maximum floods~~ data (**?**), based on a regionally broader dataset, are shown. The global envelope curve by **?** provides further context, as it includes data from substantially wetter climates and different hydro-climatic regimes.

Overall, the magnitude ~~and estimated frequency~~ of RF-based ~~floods appear~~ maximum floods appears plausible, although they tend to be lower at smaller scales than expected from ~~Qmax~~ Q_{max} data and envelope curves. This is evident in catchments where storms increasingly dominate the generation of AMFs. Such catchments would require dynamical downscaling of RF, rather than stochastic downscaling, which was not feasible with the available data (see Sect. ~~??~~**??**). In contrast, RF-based ~~values~~ maximum floods for the Ticino region – specifically for the Isorno River (IsoMsg, 124.7 km^2) and the Maggia River (MagLcn, 926.9 km^2) – are notably higher than in other regions. This is expected given the region's climatological characteristics, which favour particularly heavy precipitation events, as well as the crystalline bedrock, which promotes ~~a~~ rapid and strong flood ~~response.~~ responses. In this context, it is interesting to note that the two highest RF-based maximum floods for the Maggia River exceed the Swiss envelope curve, and are comparatively close to the world record curve (note the logarithmic y-axis), which contains data from wetter climates but is limited by record lengths and number of sites. The comparatively low values for the Aare River at Thun (AarThu, 2459 km^2) and at Bern (AarBrn, ~~2965~~2941 km^2) can be attributed to the marked attenuation

of floods by Lake Brienz and Lake Thun further upstream, which is represented in the simulation chain but not in the envelope curve.

The highest ten GWEX-derived values from the full 300 000-year simulations often reach considerably higher magnitudes than those derived from RFs (again note the logarithmic y-axis). This is not surprising given the substantially longer duration of the weather generator scenarios. However, when considering only 9920 years of GWEX weather scenarios – matching the length of the RFs – the magnitudes align more closely. Notably, the exceptionally high RF results for the ~~Ticino region~~ Maggia River are consistent with GWEX results and remain plausible.

~~A comparison within~~ In the context of large river basins~~(Fig. ?? in the Supplement) shows that the~~ , the RF-based flood estimates align well with the regional patterns and are often positioned near the regional 95th percentile envelope. Values that are noticeably lower can be explained by lake retention in the case of the Aare River at Thun and Bern, as noted above, and by a combination of slightly different climatological regime and small catchment area in the case of the Krummbach (KruKlu), which is the smallest test catchment considered. Compared to other large river basins, flood estimates for sites not influenced by lake attenuation fall somewhat more noticeably below the regional 95th percentile envelope in the Aare River basin. Full regional envelope curves are provided in Fig. **??** in the Supplement.

5.3 ~~Feasibility of dynamical downscaling~~

~~Statistical and dynamical downscaling each have their own pros and cons, and should not be understood as competing approaches. Statistical downscaling is comparatively cheap but involves a trade-off between model complexity and realism, which among other things has an impact on spatio-temporal consistency, replication of extremes and temporal disaggregation. In contrast, dynamical downscaling is based on models that use fundamental laws of physics, providing physically and spatio-temporally consistent results. However, its high computational demand limits its applicability. Specifically, within the framework presented in this paper, dynamical downscaling of the entire set of nearly 10 000 years of RF data is not feasible, requiring the selection of extreme events for dynamic downscaling.~~

~~We tested the feasibility of dynamically downscaling the ECMWF RFs to the convection-permitting scale (~3 km grid size), which would improve the representation of short local extreme precipitation events and be valuable for small catchments. To model rapid regional weather changes, as occurring during small-scale extreme events, the raw data (in our case RFs) should be available at high temporal resolution. Typically, these data (e.g., reanalyses or data from global climate models) are stored every 6 or 3 hours and have to be interpolated to the required temporal resolution. The ECMWF RFs we used were stored at 6-hour resolution for the first two forecast weeks, but only at 12-hour resolution for the time beyond. This poses the risk that short-term extreme events, lasting no more than a few hours, will not be captured. This risk emerges due to the methodology of dynamical downscaling: The limited-area regional model needs atmospheric driving data at each model time step (20 seconds) along its lateral boundaries. These driving data are interpolated linearly in time from the available RF data.Hence, if the RF data are available only on a 12-hour resolution, processes beyond this resolution are hardly represented, although the regional model is to a certain extent capable of reintroducing sub-daily processes.~~

~~We therefore compared whether downscaling applied to 12-hourly input data is capable of simulating local extreme events that essentially correspond to those from downscaled hourly data. Since the RF data are only available at a coarse temporal resolution, ERA5 reanalysis data (?) with a grid width of 25 km and a temporal resolution of 1 hour were used as alternative input data. The ECMWF forecast model underlying ERA5 is very similar to that used for the RFs. The CCLM model (??) was used for the dynamic downscaling.~~

~~The two extreme events of 8 July 1996 and 23 June 2021 were considered. While the simulation results looked realistic at intervals of 1 hour, there was a marked reduction in the simulated extreme precipitation as well as spatial distortions in the precipitation field at 12-hour intervals. Due to the reduced simulated precipitation, there is a risk that extreme events (which are not based exclusively on slow processes) will no longer be represented as such.~~

~~In summary, it was not possible to identify the relevant extreme events for dynamical downscaling because the link between the scales was not pronounced enough, especially with the limited 12-hour resolution available. For successful selection, the ECMWF RF data would need to be available at a higher temporal and spatial resolution to simulate relatively small-scale extreme precipitation events (~500 km^2 or smaller). While dynamic downscaling would work for large-scale extreme events, no additional benefit is expected compared to stochastic downscaling. In the future, dynamic downscaling would be highly relevant if RF data became available at a higher temporal resolution.~~

5.3 Flood return levels

Some caution is warranted when interpreting the estimated return periods of AMFs derived from the RF-based long CS: ~~Although~~ although the individual RF ensemble members are assumed independent after discarding days 1 to 15, their ability to accurately represent the flood-relevant meteorological behaviour across different regions cannot be examined systematically with the test catchments used in this proof of concept. The exceedance curves for the larger test catchments~~studied~~, however, indicate a generally plausible range when compared to observed AMFs and their extrapolation. However, it should be borne in mind that the observed AMFs are also subject to uncertainty (**?**), and that the ~~observed AMFs available~~ available observed AMFs are instantaneous peak values, as opposed to the hourly simulated values.

An important question ~~to which~~ that RF-based model runs can ~~contribute~~ help address is how the exceedance curve behaves as the return period increases. Notably, the observed AMFs sometimes appear to level off, whereas GWEX-derived AMFs typically do not. Interestingly, the RF-derived AMFs also generally show no levelling off or a tendency towards a plateau value. Moreover, the agreement between RF and GWEX-based flood estimates in the tail is remarkably high where comparison can be made (Fig. 10), particularly also in the case of the Maggia River, which shows an exceptionally heavy tail. This general agreement is noteworthy because the RF offer a more physically based approach to precipitation scenarios, whereas the GWEX weather generator employs a stochastic approach without explicit physical boundary conditions. This suggests that indeed the AMF observations – paired with extrapolation using a GEV distribution – might underestimate the potential of catchments to generate increasingly higher floods within the return periods considered here. This interpretation is preliminary, as the range of test catchments examined in this study is limited, and the return periods covered extend only to 1000 to 10 000 years.

It should be noted that the RF-based exceedance curves primarily reflect the internal consistency of the chosen modelling chain. Structural and epistemic uncertainties inherent to the numerical weather prediction system, such as systematic biases in extreme precipitation or storm dynamics, remain largely unresolved (see Sect. 5.4.8). Consequently, the RF-derived return levels should be interpreted as conditional on the applied RF ensemble and postprocessing framework.

5.4 Limitations

Several limitations of the current approach should be considered when interpreting the results.

5.4.1 Interpolation

~~Some methods,~~ Methods such as Thiessen interpolation of point values, precipitation adjustment factor and temperature lapse rates ~~, were adopted to enable long CSs~~ were applied to enable exceedingly long CS within the EXAR and EXCH project frameworks (**?**) (Sect. 3.3). Along with statistical RF downscaling, these methods have limitations for smaller catchments, which however were not the focus of EXAR and EXCH but here serve to explore the RF-CS approach across a broad range of scales from approximately 20 to 3000 km^2. Addressing some of the associated complexities – such as through dynamical downscaling, targeting mean areal precipitation and temperature, or even a spatially distributed modelling approach – will be time-consuming and require further research. In particular, dynamical downscaling could improve representation of short-duration, small-scale precipitation extremes. However, due to the coarse temporal resolution of the available RF data and the high computational demand, it was not feasible for the long multi-millennial series considered here. For large-scale extremes, stochastic downscaling remains adequate, while dynamical approaches could be applied in the future if higher-resolution data become available.

5.4.2 RF resolution

Further limitations are related to the resolution of the RF data. Initially, we planned to use both the ENS Model Climate (day 1–15) and the Extended Range Model Climate (day 16–45) for the statistical downscaling. However, due to inconsistencies in the data and concerns about the independence of short forecasting times from observed weather, only the Extended Range Model Climate data were used (see Sect. 2.2), effectively discarding days 1–15. In consequence, the statistically downscaled time series were only 9920 years long. On the other hand, ENS Model climate data are available at 0.25° resolution, whereas the longer forecasts are available on a coarser 0.5° grid, limiting the applied approach in terms of catchment size.

5.4.3 RF evolution

Since its introduction, the ECMWF IFS and hence the RFs have undergone continuous updates (see Sect. 2.2). When building the yearly time series required for the hydrological modelling, we were drawing from RFs initialised in the same year. For many initialisation years, this resulted in concatenation of RFs from different IFS cycles. Certainly, over the 12 initialisation years of RFs, considerable model evolution has occurred. However, while the extent of updates between successive IFS cycles

varies, the updates are gradual, often introducing only minor changes. The resulting yearly time series are hence, at most, based on slightly different IFS model versions.

5.4.4 Small and very small catchments

While results of the statistical downscaling approach match well with observations for medium and large catchments, the stochastic downscaling is not well suited for very small catchments, where only a few stations are scattered over few or even only one grid-point. The Extended Range Model Climate RFs used have a spatial resolution of 0.5° (~38×56 = 2128 km^2 in the study domain), whereas some of the catchments examined cover an area of less than 50 km^2, and only 2–3 stations were used in the calibration of the stochastic downscaling. We noted a systematically poor performance for these catchments. The smaller number of stations, together with the coarse resolution of the RFs, impedes ~~estimating~~ the estimation of reliable statistical relationships between the large and small scale during the calibration of the stochastic downscaling.

5.4.5 Temporal disaggregation

For the temporal disaggregation, we have experimented with several procedures to draw the best analogs. Additionally, we initially planned to disaggregate temperature using the analog method, alike precipitation. ~~We tried using~~ Approaches tested included mean daily temperatures~~and ,~~ absolute values of temperature ~~as well as~~ and differences between minimum and maximum temperatures at the station level~~, all of which~~ . All of these approaches led to very strong temperature jumps when ~~concatenating~~ the single RFs were concatenated to the yearly time series. To smooth those steep changes from one hour to the next, we pragmatically interpolated between the bias corrected 6-hourly temperature values. For precipitation, the applied analogs performed satisfactorily for large and ~~medium sized~~ medium-sized catchments, but showed excessive precipitation intensities in particular for short durations and in small to very small catchments.

5.4.6 Sampling of interannual variability

The concatenation of individual RFs to yearly time series furthermore disregarded other potentially more important aspects. Future efforts could seek to additionally incorporate variables accounting for the dynamic state of the atmosphere such as geopotential height, and aim at sampling inter-annual variability which might require the reuse of single RFs when building the annual time series. However, comparison of observations and RFs on a seasonal and annual scale indicated that the time series produced do not contain years with values overly above or below observed precipitation and temperature.

5.4.7 Sampling of long-term internal variability

The RFs ~~have been~~ were introduced in March 2008, reforecasting weather ~~occurring~~ up to 18 years prior. ~~Counting~~ Considering full years only, the RFs used ~~here~~ in this study span the period 1991–2019 (excluding 2015, see Sect. 2.2), with fewer data ~~in~~ at the beginning and end of ~~that~~ this period and more in the middle. Counting the number of RF days per month and year ~~revealed~~ indicated that the Atlantic Multidecadal Oscillation (AMO) ~~has been~~ was predominantly in a positive phase ~~over~~ during the

RF period (see **?**), ~~in particular for~~ particularly in the middle of the period ~~for which most RF were available(see Fig. ?? in the Supplement).~~ when the largest number of RFs were available. This results in ~~an~~ oversampling of positive AMO phases in the RFs processed here. A detailed figure of AMO phases and RF sampling is provided in Fig. **??** in the Supplement.

5.4.8 Structural and model uncertainty of reforecasts

Although the RF-based time series of nearly 10 000 years provides a robust estimate of sampling variability, it is important to note that all realizations share the same model physics, parameterizations, and structural assumptions. Increasing the number of RF realizations can therefore not reduce epistemic or structural uncertainty inherent in the underlying numerical weather prediction system. Similarly, minor inhomogeneities arising from the continuous evolution of the underlying forecast model cannot be fully eliminated, despite the mitigation steps described in Sects. 2.2, 3.2.1 and 3.2.4. Systematic biases in precipitation extremes, storm persistence, or co-variability of temperature and precipitation are propagated into the hydrological simulations. Bias correction and stochastic downscaling adjust marginal statistics but do not fully address errors in event dynamics, spatial coherence, or representation of processes relevant for extreme floods. Consequently, the narrower confidence intervals of RF-based precipitation extremes compared with GWEX (see Sect. 5.1) primarily reflect internal consistency of weather model and postprocessing. Interpretations of RF-based flood return levels should therefore be considered conditional on the modelling framework and postprocessing applied, with the relative importance of uncertainty sources depending on catchment characteristics such as size, elevation, geology, climatology and human disturbance (see **?**).

5.4.9 Non-stationarity

With ongoing climate change, mean precipitation over the Alps is projected to increase during winter and decrease during summer. Virtually all ~~state-of-the-art~~ state-of-the-art global and regional climate model ensembles indicate an increase in extreme precipitation over the Alpine region, except for summer, where only ~~high resolution~~ high-resolution global climate models and regional climate models indicate an increase (**?**). The RF simulation period spans the period 1991–2019, which was also used for the bias adjustment. The RFs are thus representative for the current state of the climate, whereas aspects of non-stationarity are not accounted for. Given the dominant role of precipitation for the subsequent hydrological simulations (**?**), any resultant flood estimates should be re-assessed within a time-frame of approximately 10 years.

6

~~We have demonstrated~~ This study evaluated whether ECMWF reforecasts (RFs) can be transformed into meteorological inputs suitable for long continuous simulation and derived flood estimation in a challenging Alpine environment, addressing three research questions.

Regarding the feasibility of processing ~~extensive RF data to generate long continuous inputs for a hydrological model, with statistical downscaling proving to be a suitable approach. Good meteorological and hydrological results were achieved for~~ and concatenation, we demonstrated that bias adjustment, stochastic downscaling and temporal disaggregation allow

the construction of consistent hourly forcing data from RF data. Individual RF segments can be concatenated into multi-millennial sequences, and the stitching points do not systematically influence simulated AMFs. Hydrological validation further showed that the simulated discharge series reproduce the frequency of daily discharge magnitudes and the distribution of Annual Maximum Floods (AMFs) with sufficient realism for flood frequency analysis. For large and medium-sized catchments (larger than approximately 500 km^2), performance ranged from satisfactory to good, while limitations were identified for small and very small catchments. ~~In addition to contributing information on rare floods, the resulting~~ These scale-dependent limitations primarily reflect constraints of the statistical downscaling approach applied. Processing for the largest catchments was completed within a few hours, confirming that the approach is feasible also in terms of computational cost.

Regarding comparison with and complementarity to an established stochastic weather generator approach, the RF-based flood ~~estimates provide valuable insights for verifying the magnitude of flood estimates derived from other continuous simulation studies, such as those using weather generator inputs. In our study, we found that the stochastic weather generatoremployed as an alternative is not likely to underestimate possible but unobserved floods. Feasibility is also ensured in terms of computational cost, as processing for the largest catchmentswas run within a few hours in the current setting using daily data as a reference~~frequency estimates are comparable to those of the weather generator approach and do not suggest systematic underestimation of rare yet unobserved floods. The RFs thus provide an independent and physically based complement to the purely statistical approach of the weather generator, as they are initialised with observed states and evolve according to model physics.

Regarding limitations specific to the RF-based approach, three issues should be acknowledged. First, RFs inherit structural uncertainties in the underlying numerical weather prediction system, which may affect the realism of simulated extremes and cannot be removed through statistical postprocessing. Second, the RF archive does not explicitly account for long-term climate variability or non-stationarity, introducing uncertainty when constructing multi-millennial synthetic sequences. Third, the spatial and temporal resolution of the RFs, together with the applied stochastic downscaling approach, constrains their applicability in small and very small catchments, where small-scale precipitation variability strongly controls flood generation.

Snowmelt-related processes were not analysed explicitly, as AMFs in the study catchments predominantly occur during summer convective events when rainfall dominates flood generation (see **?**). Extending the framework to explicitly assess temperature- and snowmelt-related return periods represents a relevant topic for future research (see, e.g., **?**, in the context of continuous simulation and extreme floods).

~~In the future, ways to apply~~ Several methodological developments could further improve the applicability of the framework. In particular, extending the stochastic downscaling to higher resolved temporal data should be explored. In ~~our~~ this proof-of-concept study, ~~one obstacle in this regard~~ a key limitation was data availability, ~~in particular that of~~ particularly with respect to gridded observations at high temporal resolution. ~~Although the~~ The hourly CombiPrecip dataset~~currently provides data~~, for example, currently extends only from 2005 onward, ~~the gap to create a sufficiently long data record~~ although the record will eventually reach the length required for use as a predictor dataset ~~within the stochastic downscaling framework is expected to close over time. Also regarding station data, ongoing measurements will improve the data basis.~~ for stochastic downscaling calibration. In the meantime, ~~disaggregated data could be generated using other sophisticated approaches over the applied~~

analogs, such as stochastic disaggregation (**?**). As a way forward for achieving better alternative approaches for generating sub-daily variability could be explored in place of the analog-based method applied here, for example through stochastic disaggregation methods (**?**).

To improve results in small and very small catchments, the possibilities potential of dynamical downscaling should also be further explored, on the condition provided that RF data become available at higher temporal and spatial resolution.

Other possible avenues to derive local information from the RFs include the use of emulators (see **?**, for an overview). **??** proposed a regional climate model (RCM) emulator , combining the strengths of both that combines empirical statistical downscaling methods and dynamical downscaling via neural networks. The emulator learns the complex spatial structure and daily variability simulated by the RCMs, particularly how the RCM refines the low-resolution climate patterns, and with dynamical downscaling through a neural network architecture and could be applied to the RFs datadirectly. Paired with recent efforts in convection permitting simulations directly to the RF data. In combination with recent convection-permitting modelling efforts such as the CORDEX FPS on convective phenomena over Europe and the Mediterranean (**?**)this approach could proof very useful, such emulators could improve the representation of local-scale extremes.

Overall, the processed ECMWF RFs provide a viable and computationally efficient meteorological forcing for long continuous simulation in Alpine catchments at medium to large spatial scales. Their added value lies in being physically based rather than statistically derived, which allows RF-based flood estimates to serve as a process-based counterpart to approaches based on weather generators.

. Code and data generated in this study can be obtained from the first author upon reasonable request.

. Conceptualization: DM; funding acquisition: DV; methodology: DM, MJ, DV; investigation: MJ, DV, MS, MK, HT; visualization: MJ, DV; original draft preparation: DV, MJ; review and editing: all authors.

. The contact author has declared that none of the authors has any competing interests.

. We thank the ECMWF for producing and making available the vast database of RFs that made this study possible. We also thank MeteoSwiss, the Federal Office for the Environment FOEN, as well as the cantons of St. Gallen and Ticino for providing hydrometeorological data. Heimo Truhetz gratefully acknowledges the computational resources granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JURECA at the Julich Jülich Supercomputing Centre (JSC) through grant JJSC39 and by the Vienna Scientific Cluster (VSC) through grant 71193. We also thank the two anonymous referees for their comments, which improved the manuscript.

720 . This research was funded by the Federal Office for the Environment FOEN and the Swiss Federal Office of Energy SFOE as a part of the project “Extreme Floods in Switzerland” (EXCH).

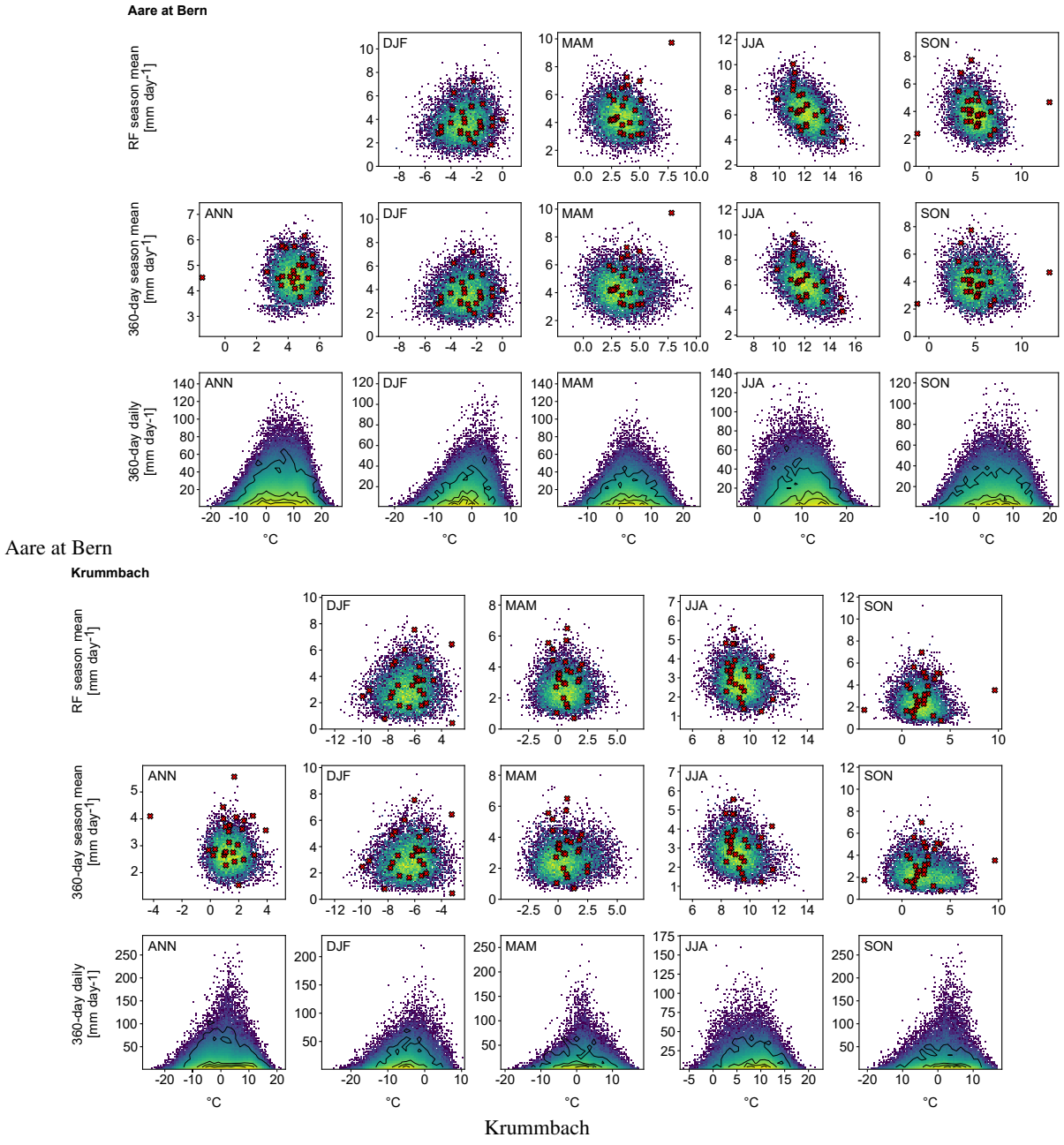


Figure 7. Density plots for the Aare River at Bern (top) and the Krummbach River (bottom), showing annual and seasonal precipitation and temperature based on seasonal means and the RF calendar dates (first row each), on seasonal means and the final 360-day calendar (second row each), and as based on daily means and the final 360-day calendar (third row each, days with a precipitation sum of $\geq$1 mm). Colours depict RF values, ranging from blue for low density to yellow for high density. Observations over the period 1991–2019 are denoted by red crosses (first and second row each) and contour lines (from low to high density in the innermost contours, third row each).

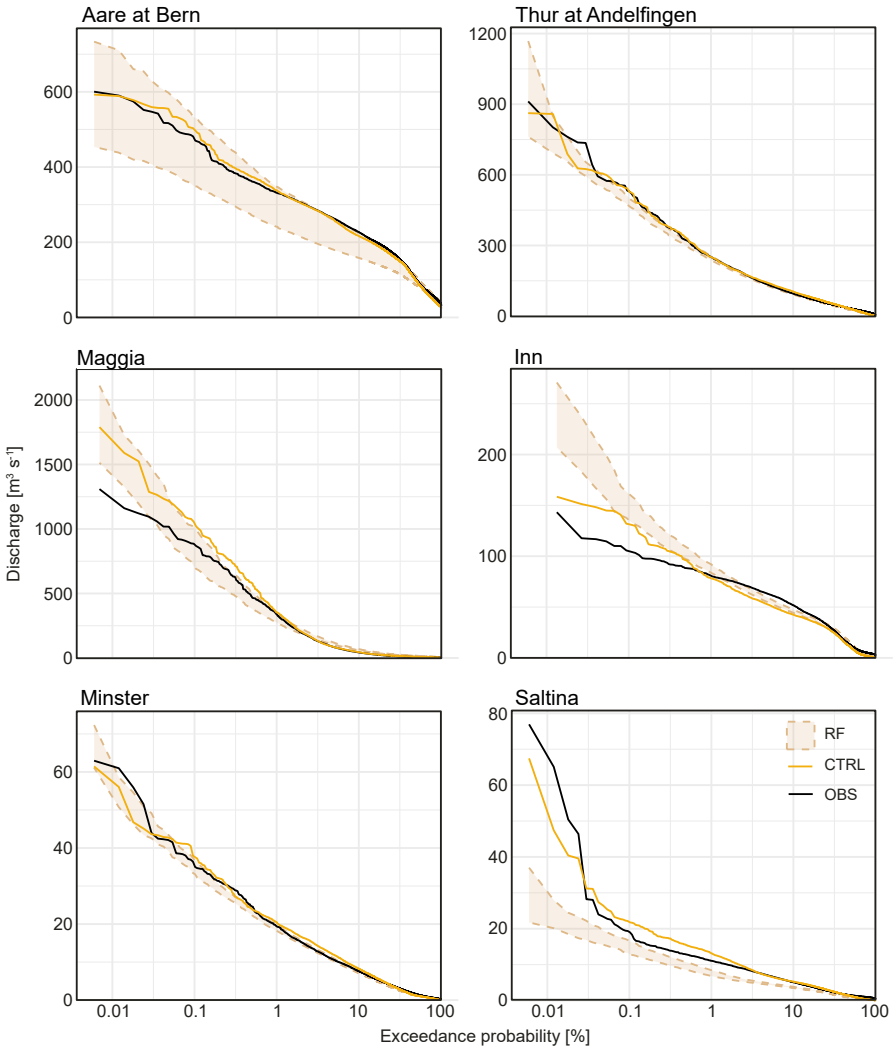


Figure 8. Flow Duration Curves (FDCs) for selected sites, comparing RF-based simulations (RF), control simulation (CTRL) and observations (OBS). Note that the x axis is scaled logarithmically. For RF-based simulations, samples matching the length of the observations (or, if shorter, the control simulation) were used to compute exceedance probabilities and corresponding 95% confidence intervals.

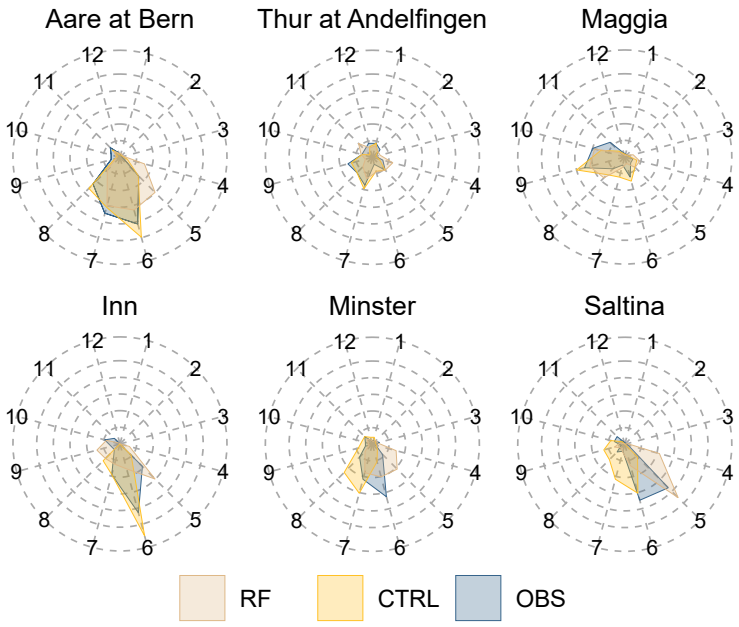


Figure 9. Seasonality of annual maximum floods for selected sites, comparing reforecast-based simulations (RF), control simulation (CTRL) and observations (OBS).

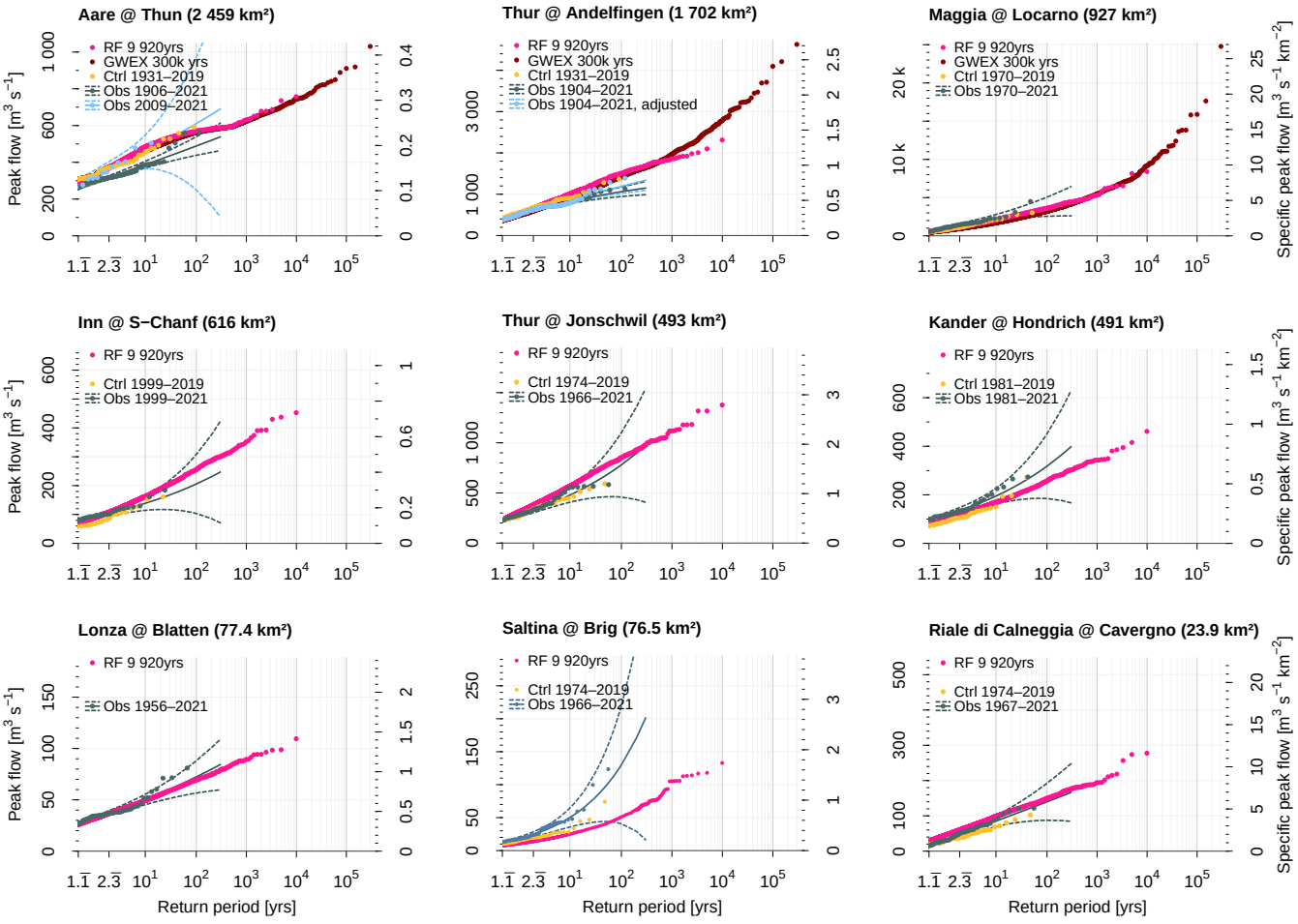


Figure 10. Exceedance curves from hydrological simulations using reforecast (RF) input. The top row shows three large catchments, where for which long continuous simulations with input from the weather generator GWEX are available for comparison. The middle row shows three medium-sized catchments, the bottom row three small catchments. 'RF' refers to annual maximum floods (AMFs) from the long continuous simulation (CS) based on reforecasts, 'GWEX' to AMFs from the long CS based on weather generator input (only top row), 'Ctrl' to AMFs from a control run with observed weather, and 'Obs' are observed AMFs and a corresponding extrapolation.

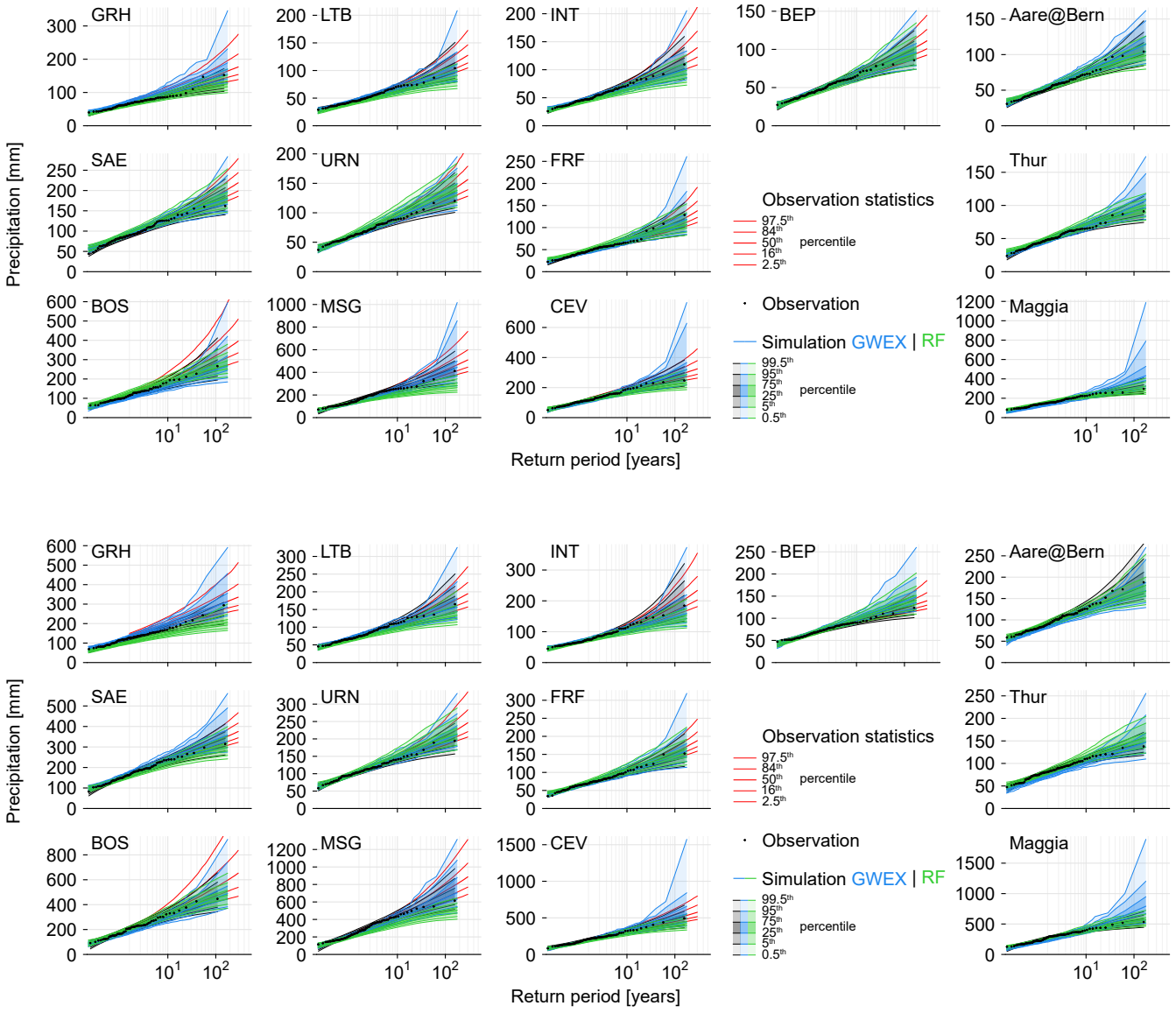
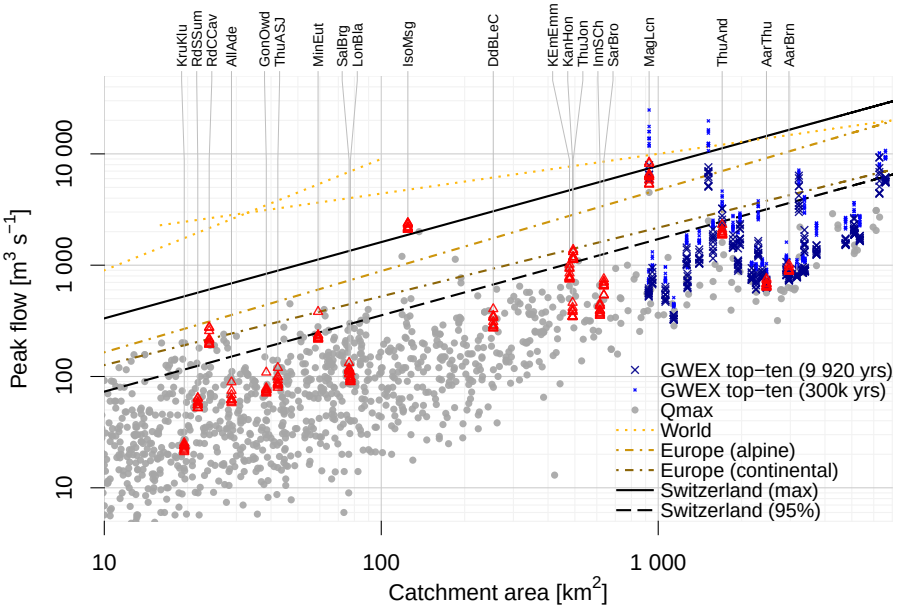


Figure 11. Comparison of maximum 1-day (top) and 3-day (bottom) precipitation sums from reforecasts (RF), weather generator (GWEX) and observations (OBS). Columns 1–4 show selected precipitation gauging stations (see Tab. 3) within the full catchment shown in column 5.



Highest ten annual maximum floods simulated using reforecast (RF) and weather generator (GWEX) input. Maximum observed floods in Switzerland as per **?** (gray dots) are shown for context, along with corresponding envelope curves (the solid line refers to all data, the dashed line to the 95th percentile). Also shown are envelope curves for maximum floods in Europe (**?**) (two regions are relevant: alpine, applies to AarBrn, AarThu, KEmEmm, MinEut, SarBro, ThuAnd, ThuJon; continental, applies to the remaining catchments) and worldwide (**?**). For catchment IDs see Table 1, for analyses separately for large river basins see Fig. **??**.

Figure 12. Highest ten annual maximum floods simulated using reforecast (RF) and weather generator (GWEX) input. Maximum observed floods in Switzerland as per **?** (gray dots) are shown for context, along with corresponding envelope curves (solid line: all data; dashed line: 95th percentile). Also shown are envelope curves for maximum floods in Europe (**?**) (two regions are relevant: alpine, applies to AarBrn, AarThu, KEmEmm, MinEut, SarBro, ThuAnd, ThuJon; and continental, applies to the remaining catchments) and worldwide (**?**). For catchment IDs see Table 1, for analyses separately for large river basins see Fig. **??**.

Supplementary material for "Leveraging reforecasts for flood estimation with long continuous simulation: a proof-of-concept study"

Daniel Viviroli[1], Martin Jury[2], Maria Staudinger[1], Martina Kauzlaric[3,4], Heimo Truhetz[2], and Douglas Maraun[2]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
[2]Wegener Center for Climate and Global Change, University of Graz, Graz, Austria
[3]Institute of Geography, University of Bern, Bern, Switzerland
[4]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

S1 Feasibility of dynamical downscaling

Statistical and dynamical downscaling provide complementary approaches to generating high-resolution precipitation data. Statistical downscaling, as applied in this study, is computationally efficient but has limitations in representing small-scale convective extremes. Dynamical downscaling, in contrast, explicitly resolves physical processes and provides spatio-temporally consistent results, making it particularly relevant for small-scale convective precipitation events. However, its high computational cost restricts its applicability for exceedingly long datasets such as the nearly 10 000 years of RF data considered here, which would require pre-selection of extreme events for dynamical downscaling.

We tested the feasibility of dynamically downscaling the ECMWF RFs to the convection-permitting scale (~3 km grid spacing), which would improve the representation of short local extreme precipitation events and be valuable for small catchments. To model rapid regional weather changes such as those associated with small-scale extreme events, the raw data (in our case RFs) should be available at high temporal resolution. Typically, these data (e.g., reanalyses or data from global climate models) are stored every 6 or 3 hours and must be interpolated to the required temporal resolution. The ECMWF RFs we used were stored at 6-hour resolution for the first two forecast weeks, but only at 12-hour resolution for the extended range beyond day 15. This raised the concern that short-term extreme events lasting only a few hours would not be captured. This risk arises due to the methodology of dynamical downscaling: the limited-area regional model requires atmospheric driving data at each model time step (20 seconds) along its lateral boundaries. These driving data are linearly interpolated in time from the available RF data. Hence, if the RF data are available only at 12-hour resolution, processes beyond this resolution are poorly resolved, although the regional model is partially capable of reintroducing sub-daily processes.

We therefore tested whether dynamical downscaling driven by 12-hourly input data could reproduce the local extreme events produced when hourly input data were used. Since the RF data are only available at a coarse temporal resolution, ERA5 reanalysis data (**?**) with a grid spacing of 25 km and a temporal resolution of 1 hour were used as a surrogate to test the sensitivity to input temporal resolution. The ECMWF forecast model underlying ERA5 is very similar to that used for the RFs. Dynamical downscaling was performed with the CCLM model (**??**).

The two extreme events of 8 July 1996 and 23 June 2021 were considered. While the simulations driven by 1-hourly input produced realistic results, simulations driven at 12-hour intervals showed a marked reduction in simulated extreme precipitation and spatial distortions in the precipitation field. As a result, extreme events driven by fast convective processes may be substantially underestimated or lost entirely.

In summary, it was not possible to identify the relevant extreme events for dynamical downscaling because the link between the scales was not pronounced enough, especially with the limited 12-hour resolution available. Higher temporal and spatial resolution of ECMWF RF data would be required to reliably identify relatively small-scale extreme precipitation events for catchments of approximately 500 km^2 or smaller. While dynamical downscaling remains suitable for large-scale extreme events, it offers no clear advantage over stochastic downscaling at these scales. In the future, dynamical downscaling could substantially improve the representation of short-duration convective events, provided that RF data become available at a higher temporal resolution.

S2 Supplementary figures

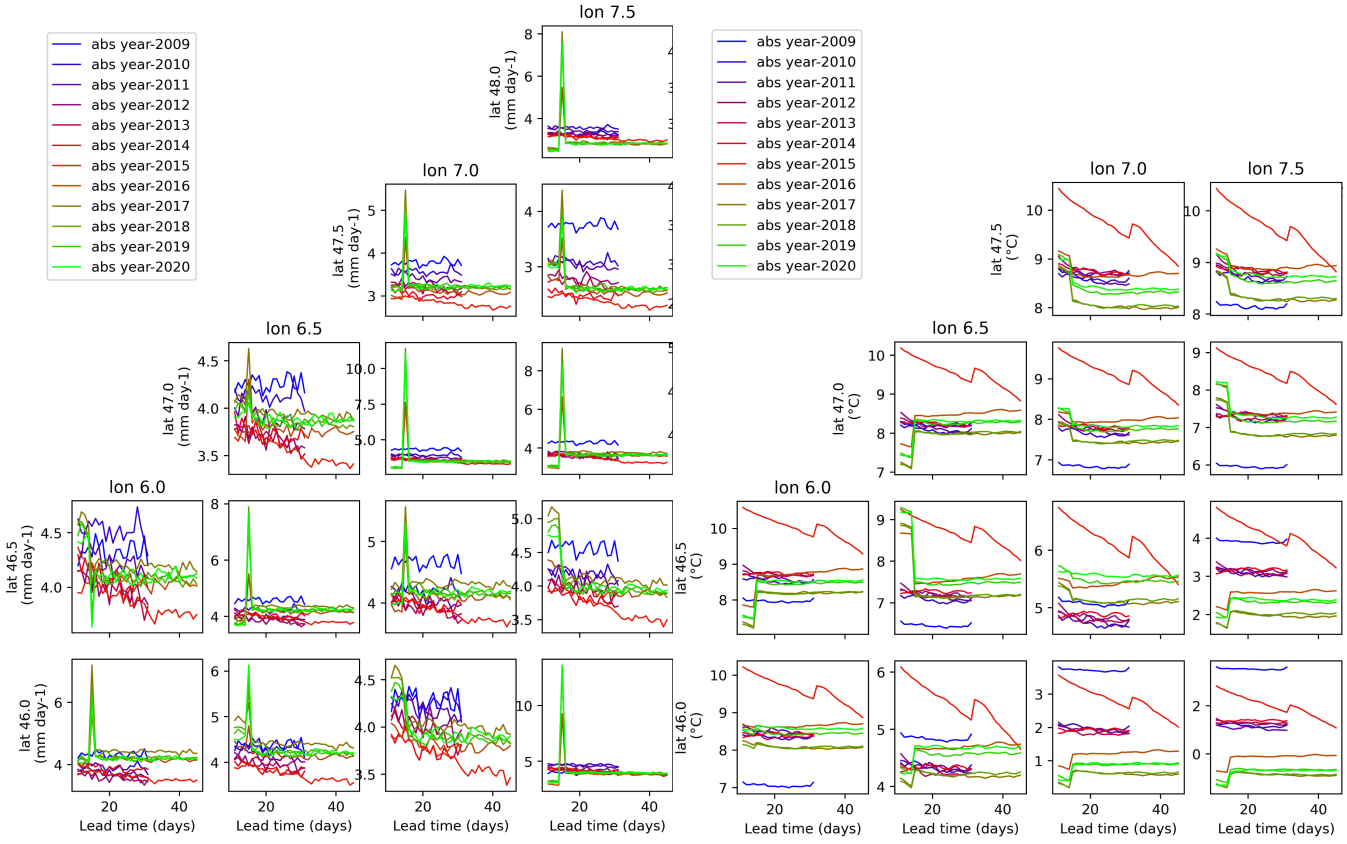


Figure S1. Annual mean ECMWF reforecast precipitation (left) and temperature (right) for selected grid points over Switzerland, plotted by forecast time (days 11 to 45) and by the year of initialization year.

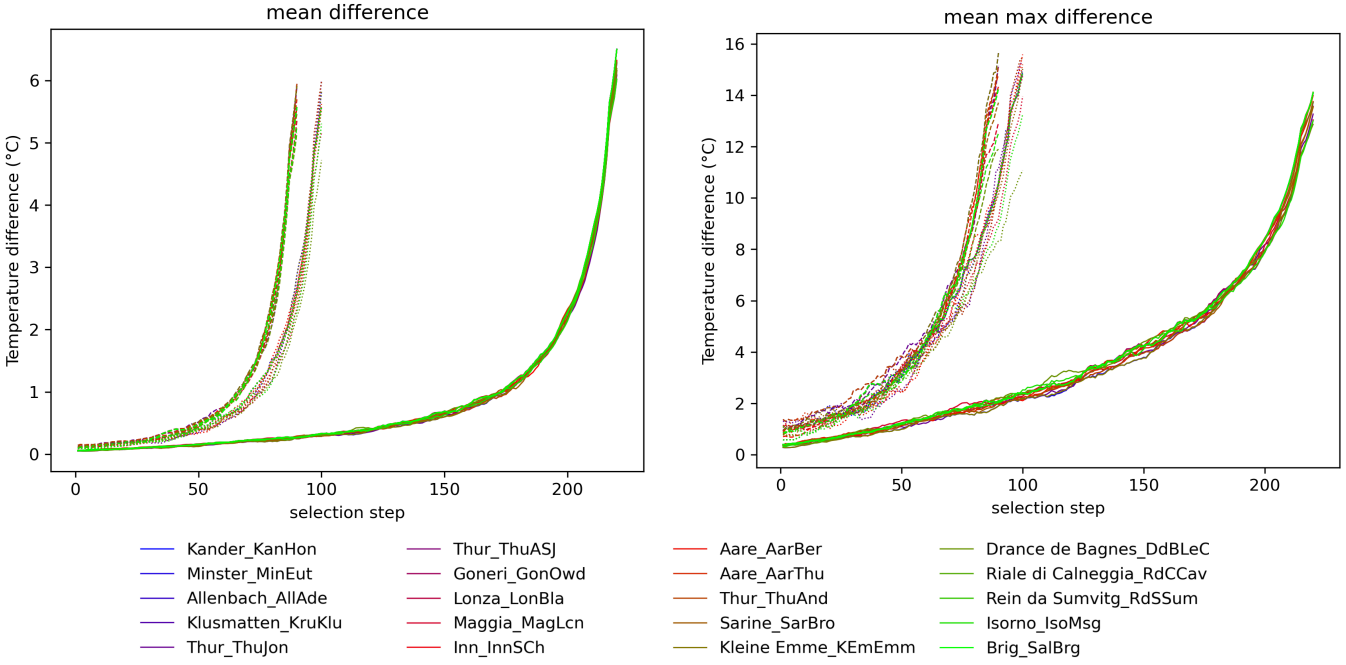


Figure S2. Mean (left) and maximum (right) 6-hourly raw temperature difference at the intersection time of concatenated single ECMWF RFs. Dashed lines are for RFs initialized between 2009–2012, dotted lines for RFs initialized between 2013–2014, and solid lines for RFs initialized between 2016–2020. A running mean of 11 selection steps was applied.

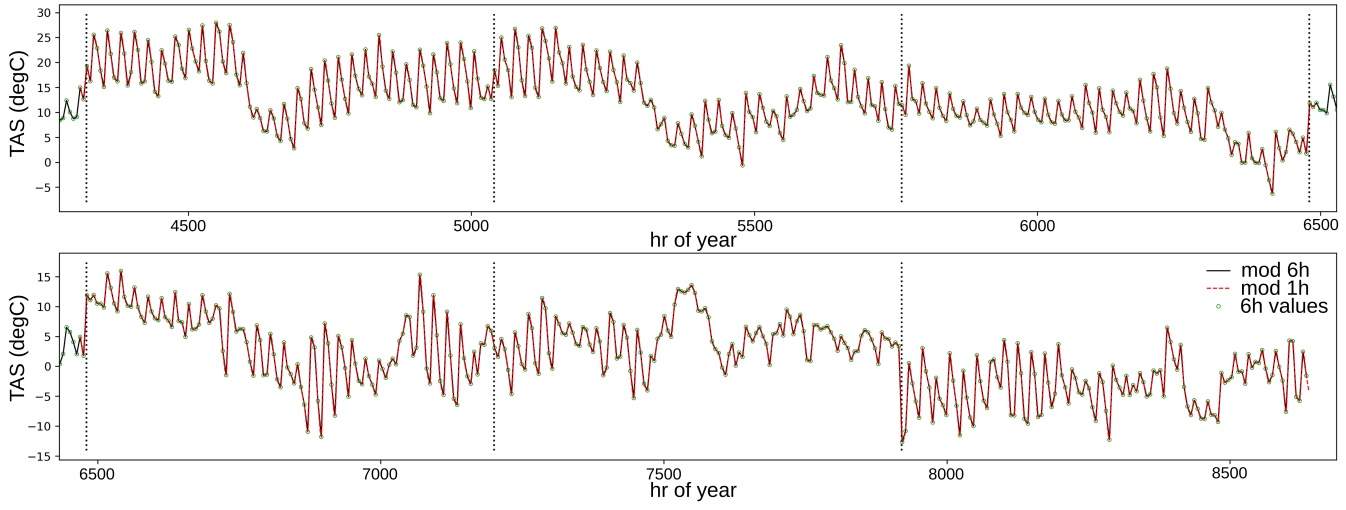


Figure S3. Time series of temperature (TAS) for the second half-year of one extreme 'gap'-year an example year (year 9920) with an extreme temperature jump at the station Adelboden (ABO) in the Kander River catchment (at hr ~approximately hour 7900). The intercept stitching points between single reforecasts is are indicated by the vertical dotted linelines.

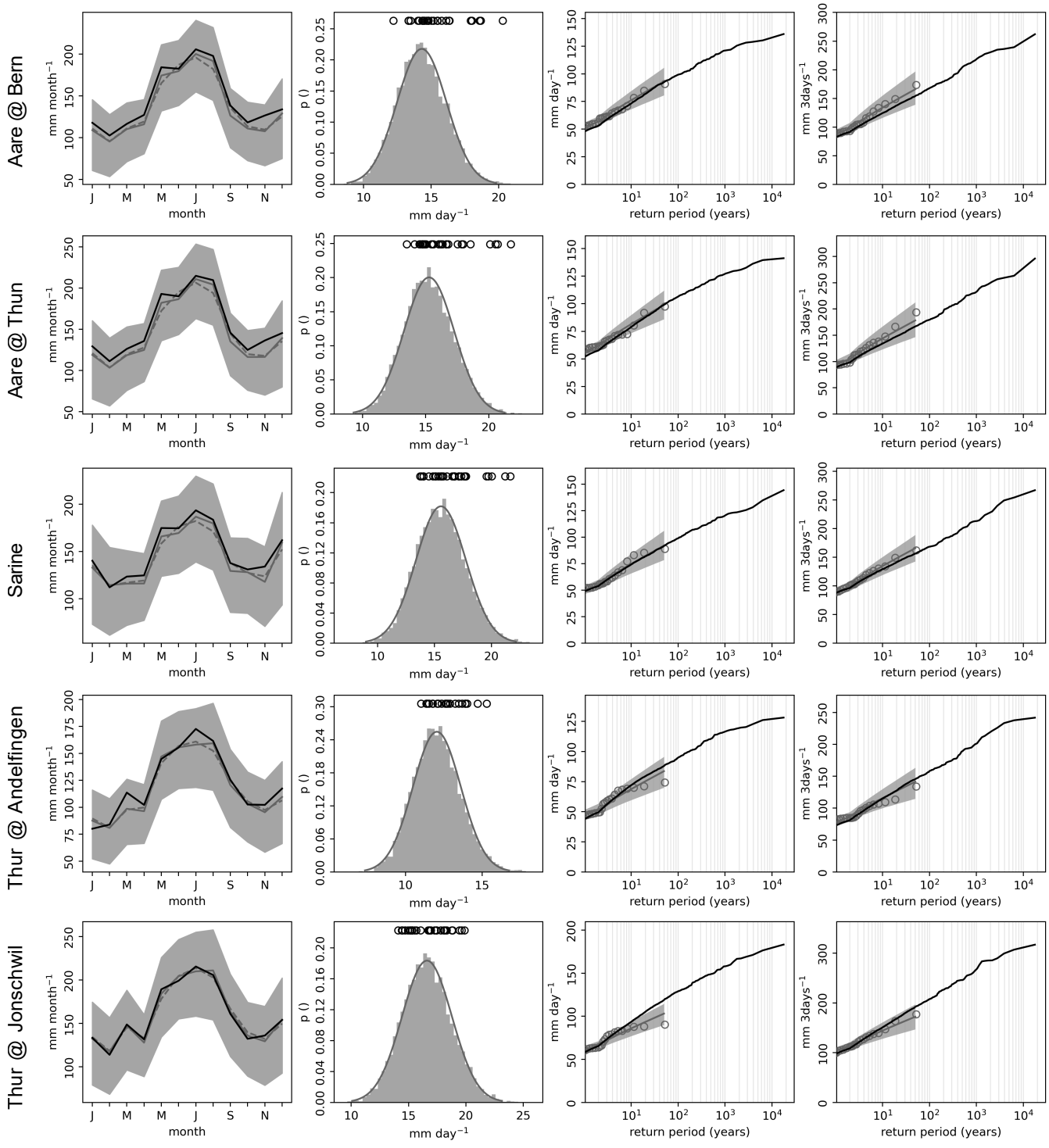


Figure S4. Evaluation of reforecast (RF) precipitation for large catchments: Aare at Bern, Aare at Thun, Sarine, Thur at Andelfingen, Thur at Jonschwil. From left to right: Annual cycle of mean monthly precipitation for observations (1991–2019, black) and RF (grey, shading p25–p75; dashed line based on 360-day calendar); histogram of annual daily p90 precipitation for RF, and observations (black circles); return levels for annual maximum daily and maximum 3-day precipitation for RF (black) and observation (grey circles) using the Gringorten plotting position. For observations, return values have additionally been fitted using a Gumbel distribution (using a GEV distribution resulted in overly broad confidence intervals) and maximum likelihood estimates, confidence intervals are based on 5000 bootstrap samples (mean in dark grey, p2.5–97.5 grey shading).

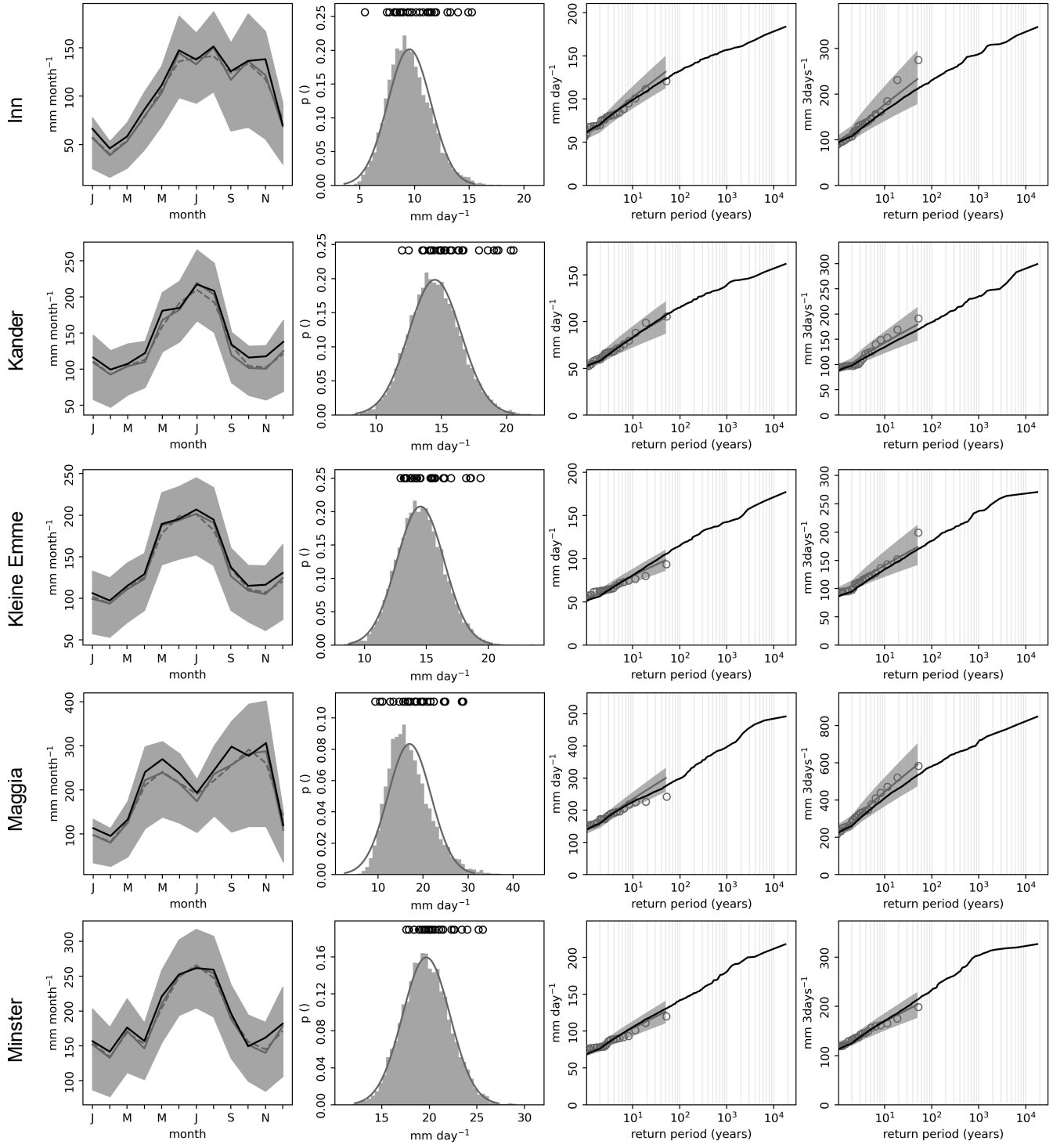


Figure S5. Same as Fig. S4, but for medium-sized catchments: Inn, Kander, Kleine Emme, Maggia, and Minster.

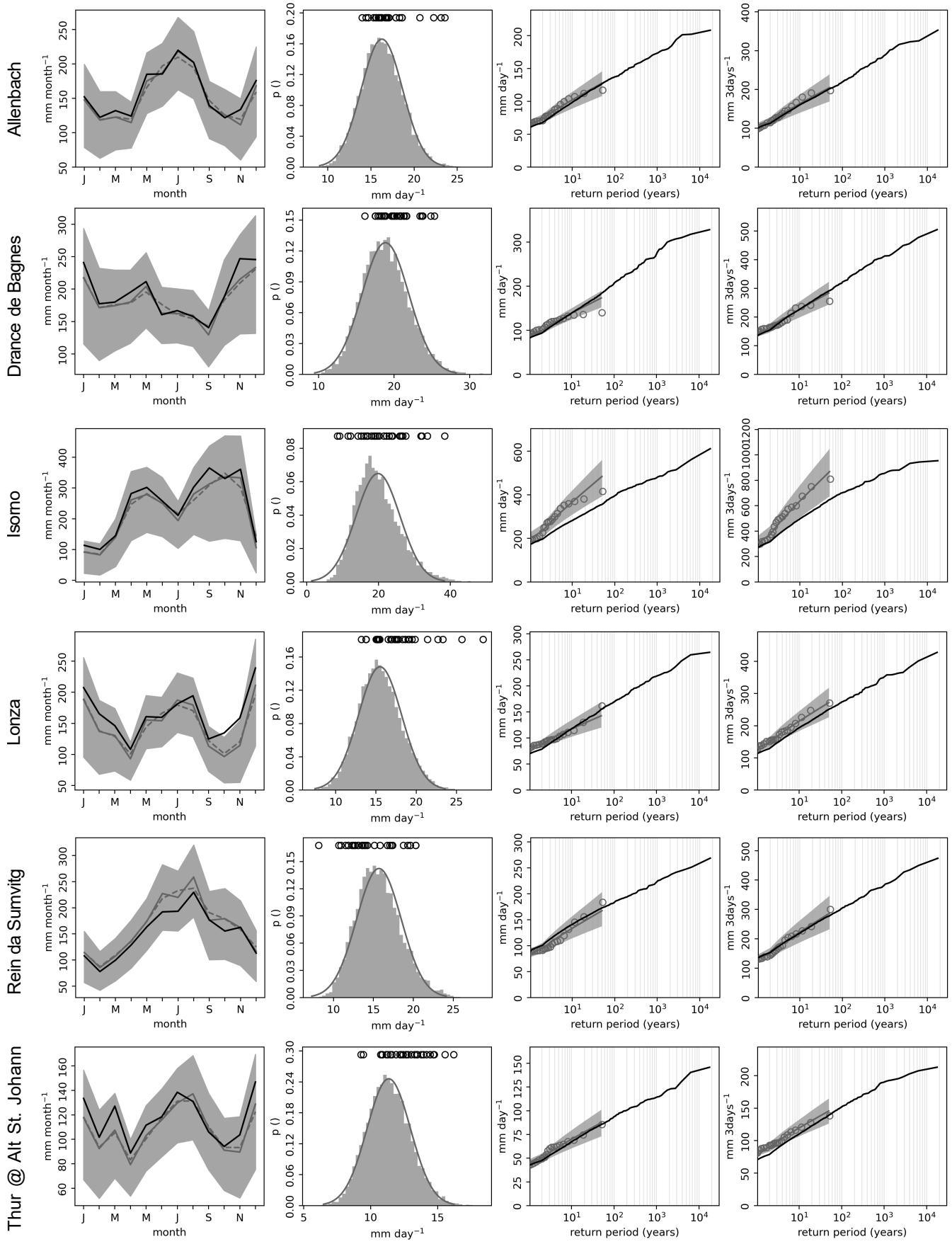


Figure S6. Same as Fig. S4, but for small catchments: Allenbach, Drance de Bagnes, Isorno, Lonza, Rein da Sumvitg and Thur at Alt St. Johann.

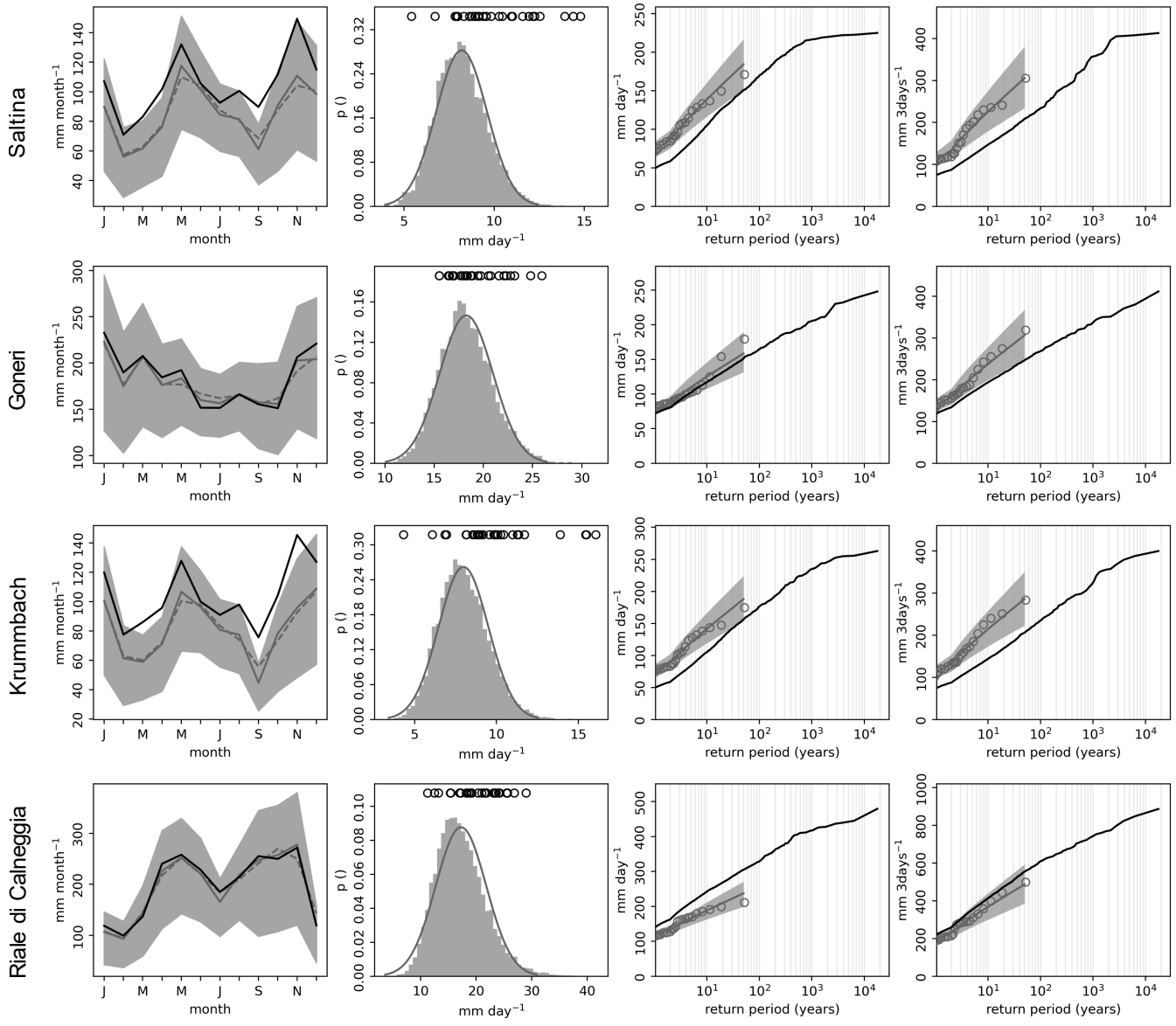


Figure S7. Same as Fig. S4, but for very small catchments: Saltina, Goneri, Krummbach and Riale di Calneggia.

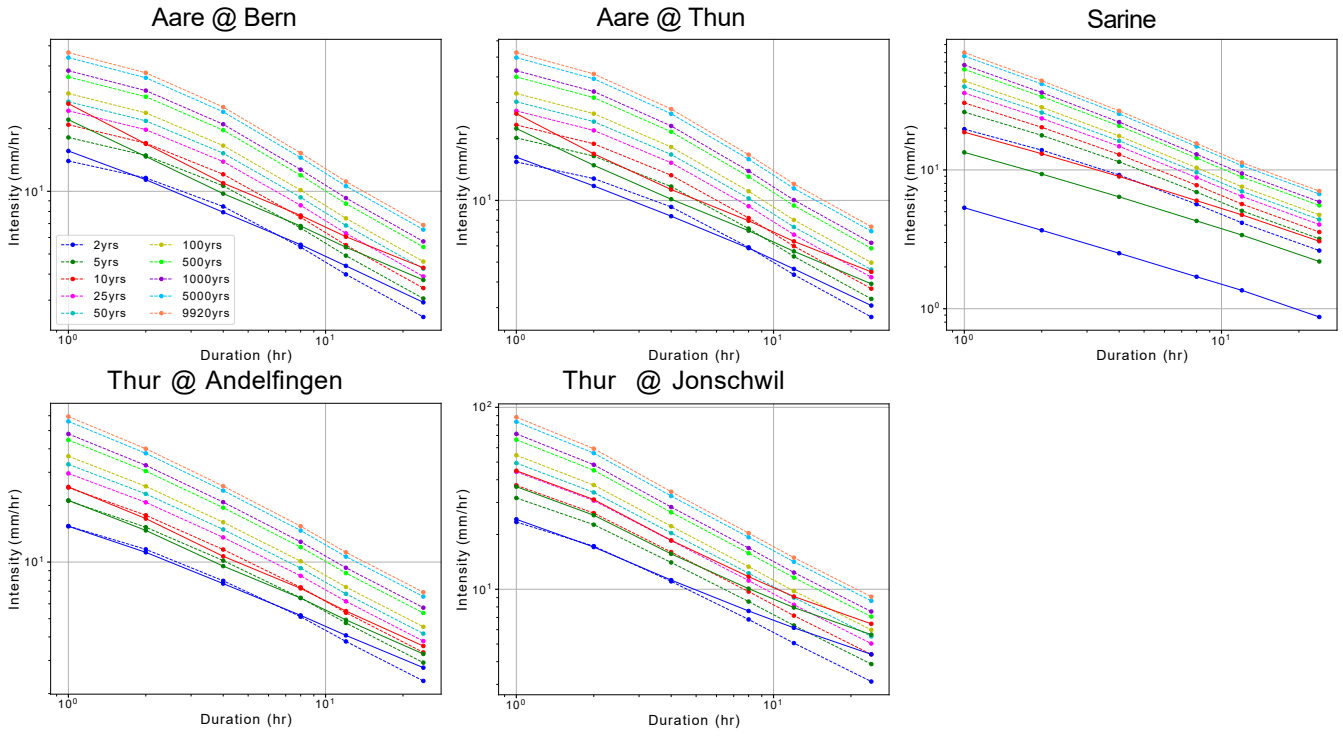


Figure S8. Average intensity-duration-frequency (IDF) curves (1h to 24h) for different return periods (colors) in large catchments: Aare at Bern, Aare at Thun, Sarine, Thur at Andelfingen and Thur at Jonschwil. Reforecasts results are shown with dashed lines, observations spanning 1981–2019 and used for the analog method are shown with solid lines. The IDFs have been derived using the Gumbel method (**?**), and a 360-day calendar was used.

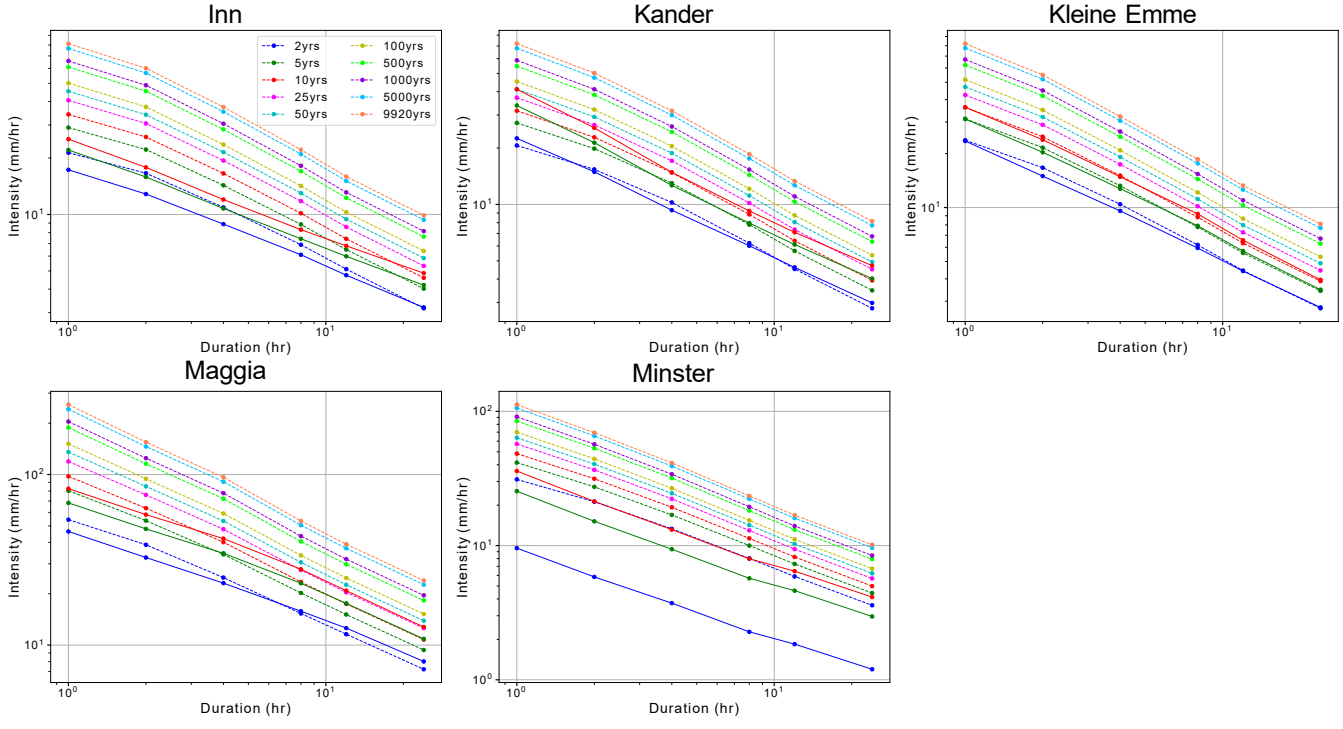


Figure S9. Same as Fig. S8 but for medium catchments: Inn, Kander, Kleine Emme, Maggia and Minster.

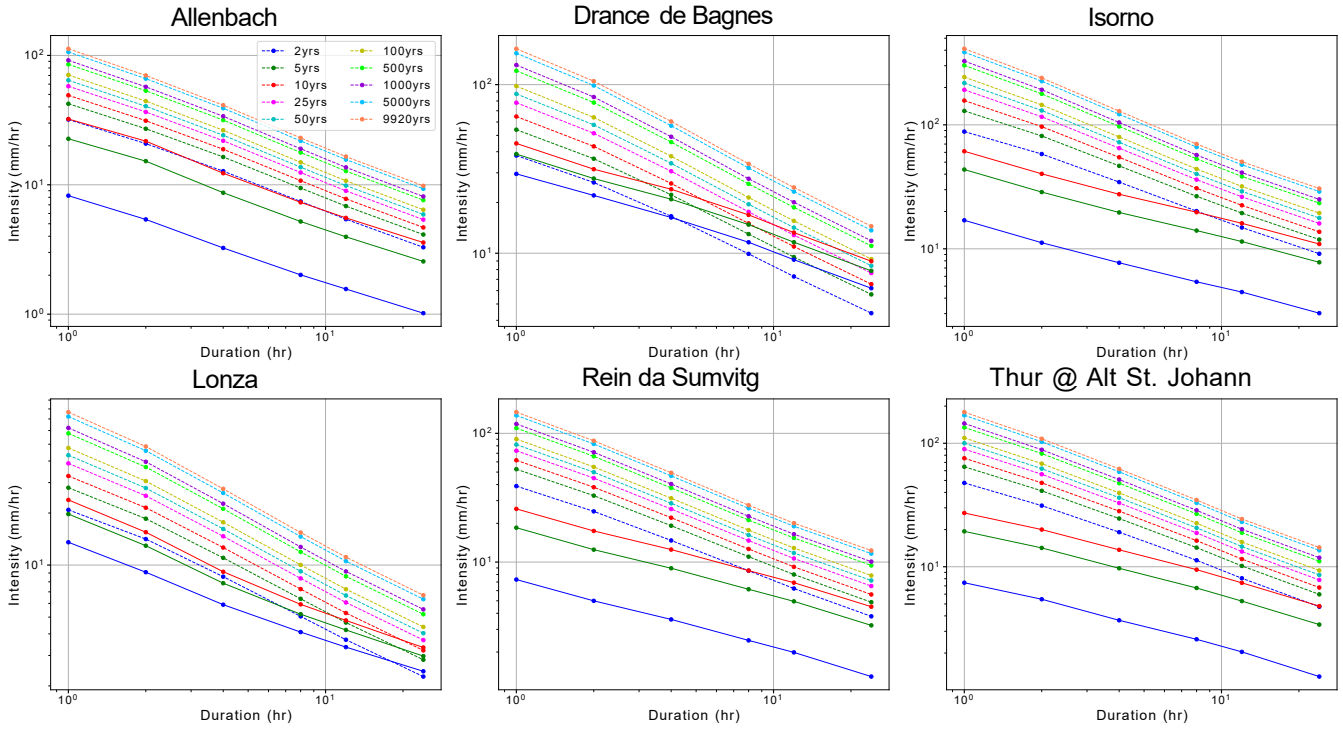


Figure S10. Same as Fig. S8 but for small catchments: Allenbach, Drance de Bagnes, Isorno, Lonza, Rein da Sumvitg and Thur at Alt St. Johann.

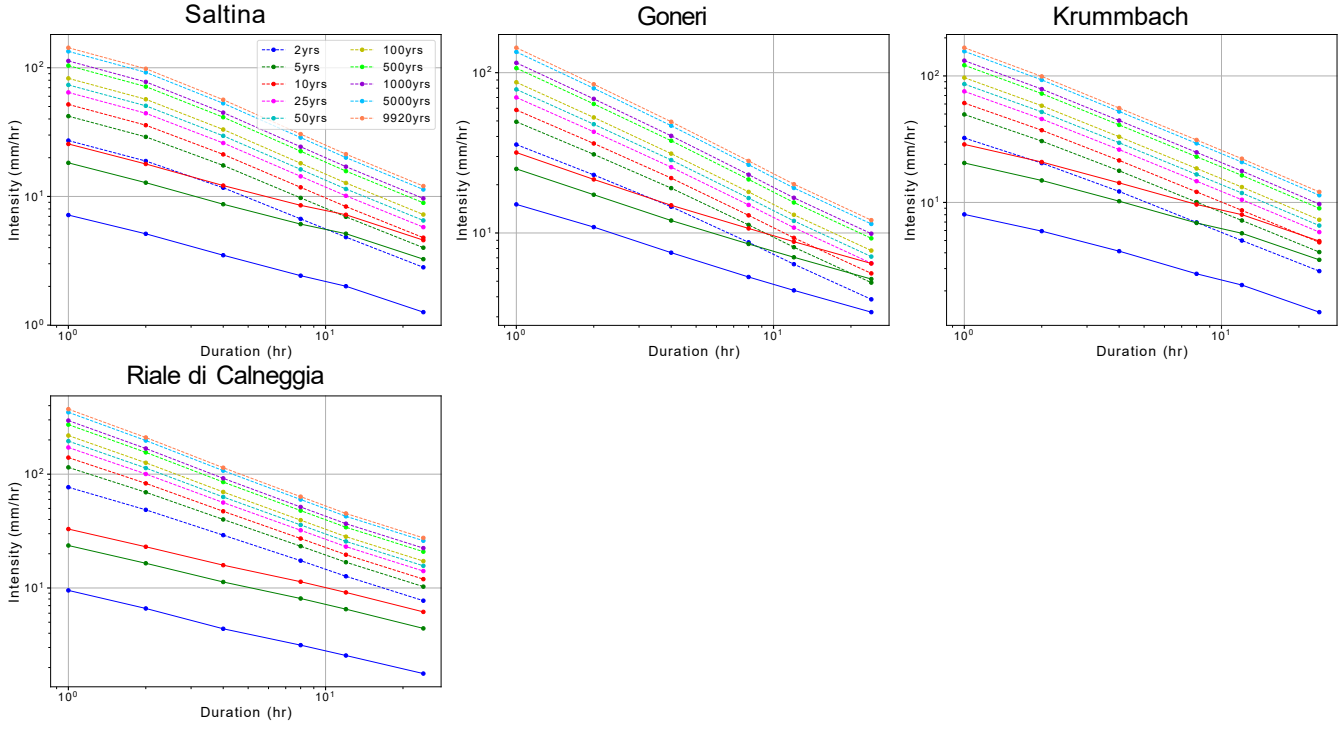


Figure S11. Same as Fig. S8 but for very small catchments: Saltina, Goneri, Krummbach and Riale di Calneggia.

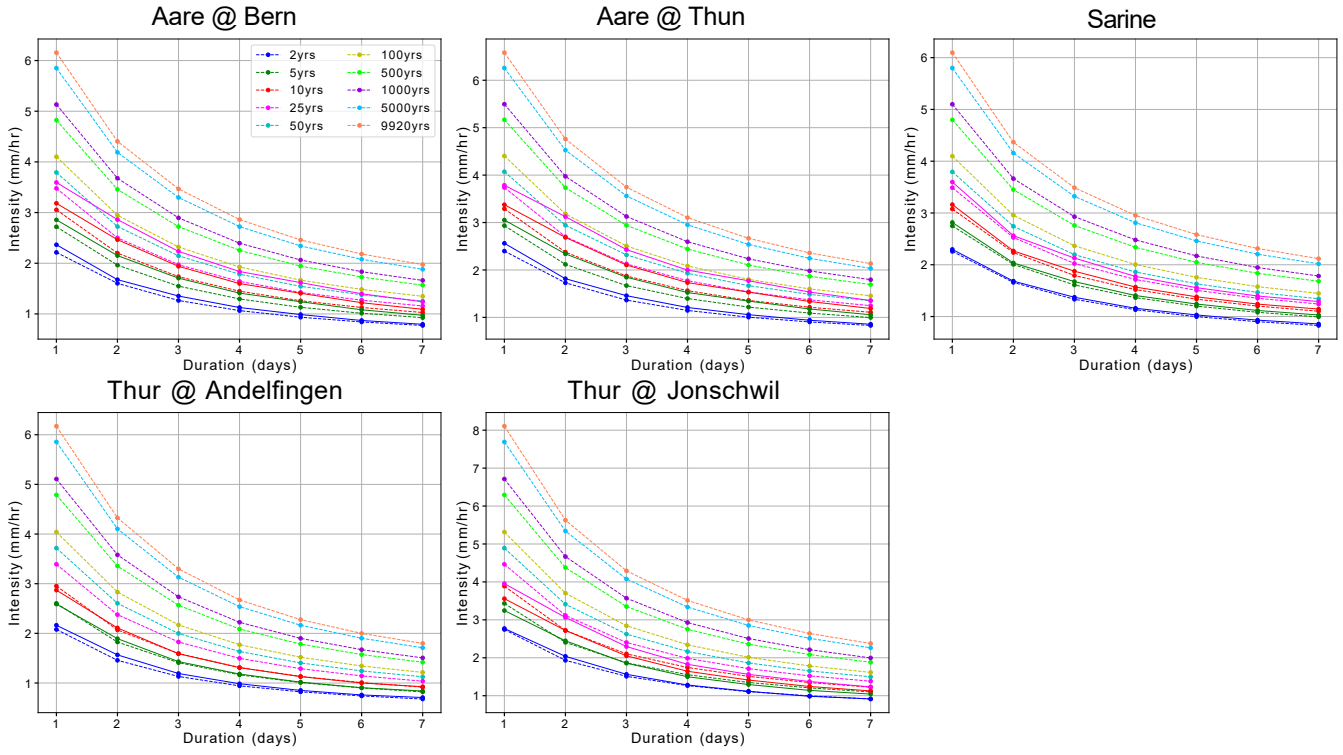


Figure S12. Average intensity-duration-frequency (IDF) curves (1-day to 7-days) for different return periods (colors) in large catchments: Aare at Bern, Aare at Thun, Sarine, Thur at Andelfingen and Thur at Jonschwil. Reforecasts results are shown with dashed lines, observations spanning 1981–2019 and used for the bias adjustment analog method are shown with solid lines. The IDFs have been derived using the Gumbel method (**?**), and a 360-day calendar was used.

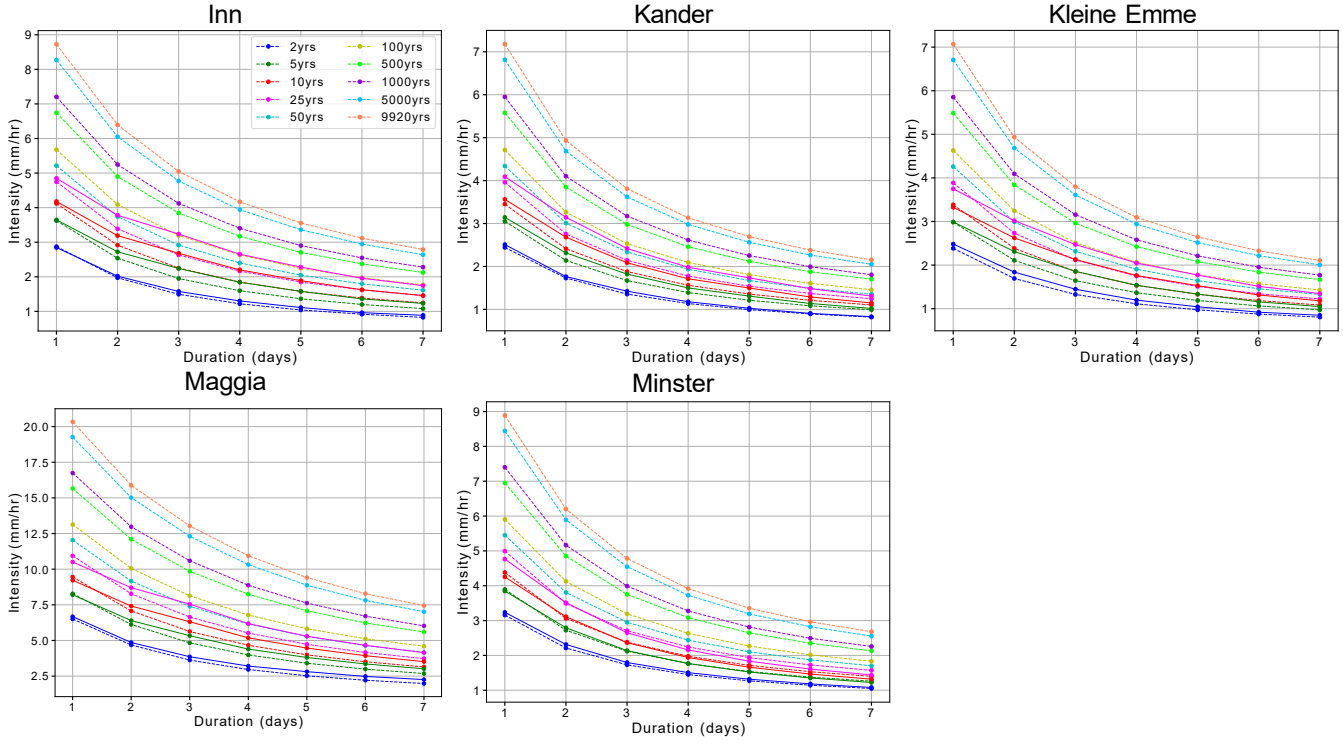


Figure S13. Same as Fig. S12 but for medium catchments: Inn, Kander, Kleine Emme, Maggia and Minster.

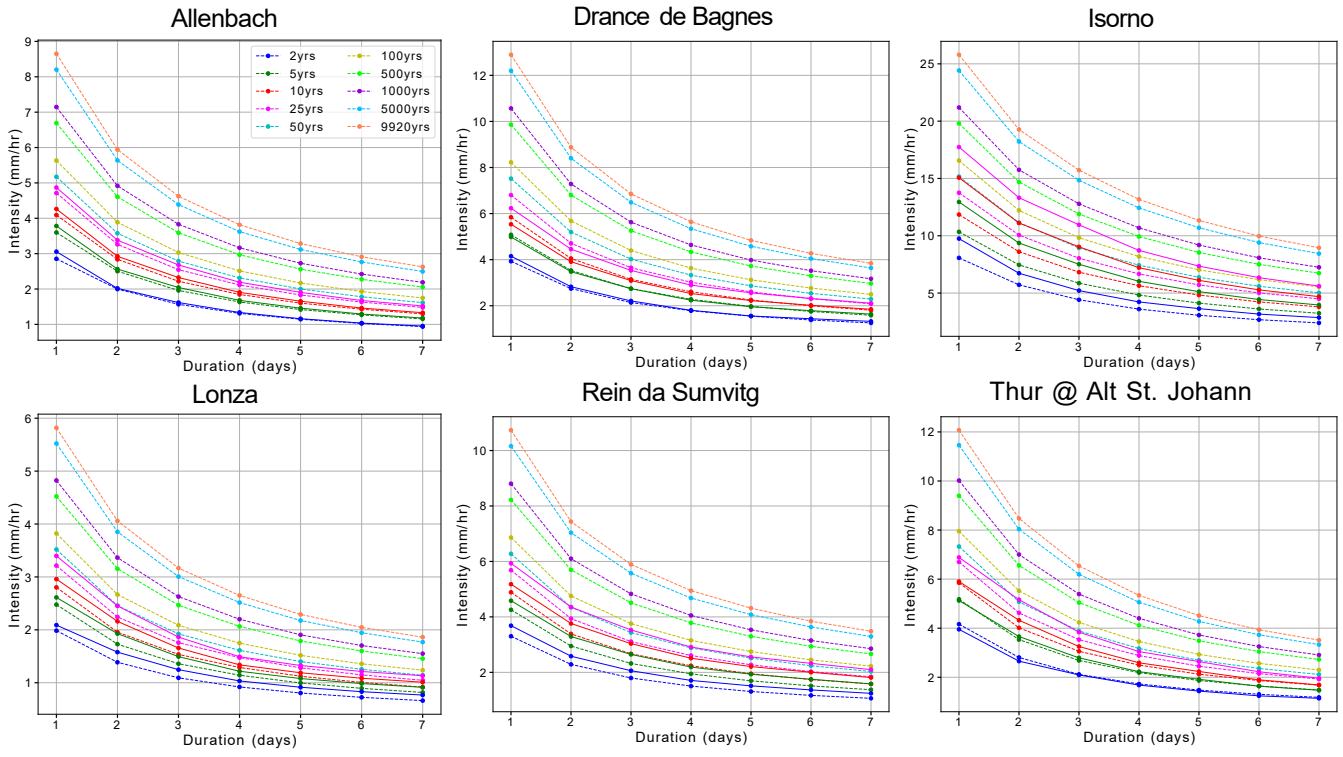


Figure S14. Same as Fig. S12 but for small catchments: Allenbach, Drance de Bagnes, Isorno, Lonza, Rein da Sumvitg and Thur at Alt St. Johann.

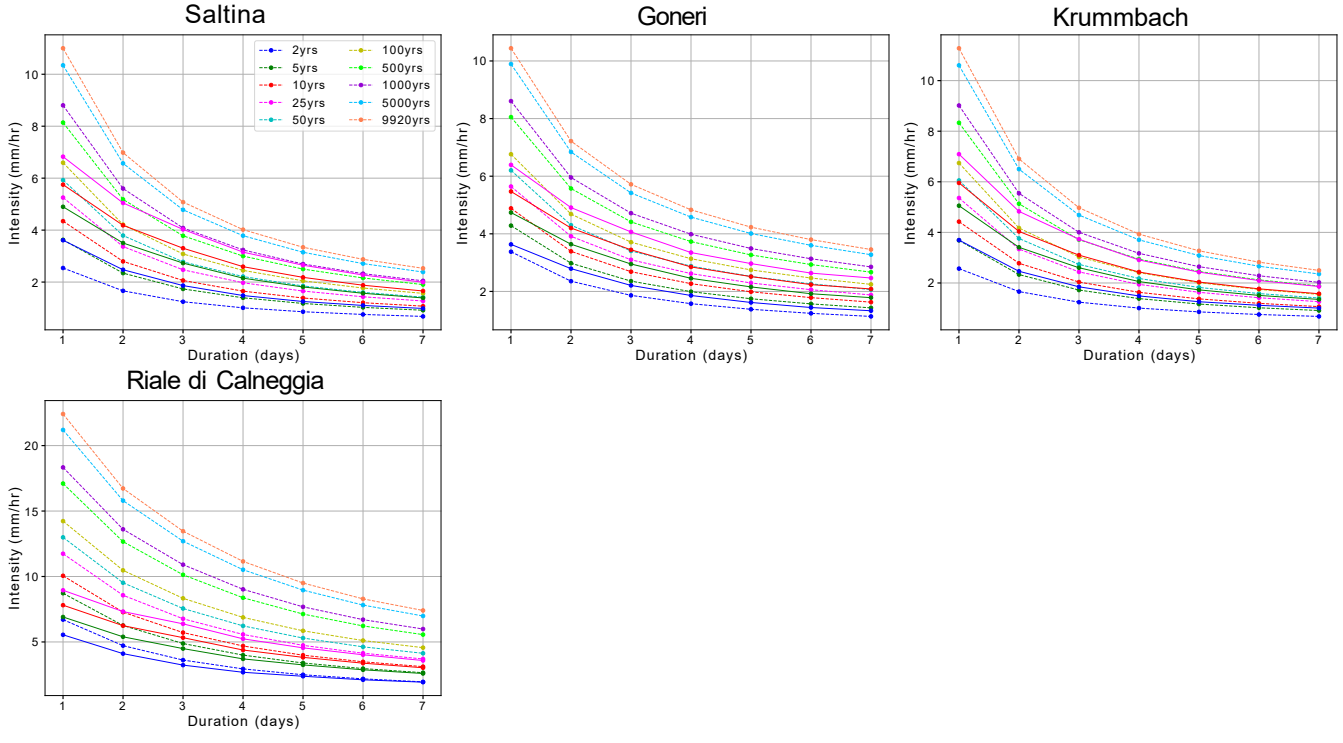


Figure S15. Same as Fig. S12 but for very small catchments: Saltina, Goneri, Krummbach and Riale di Calneggia.

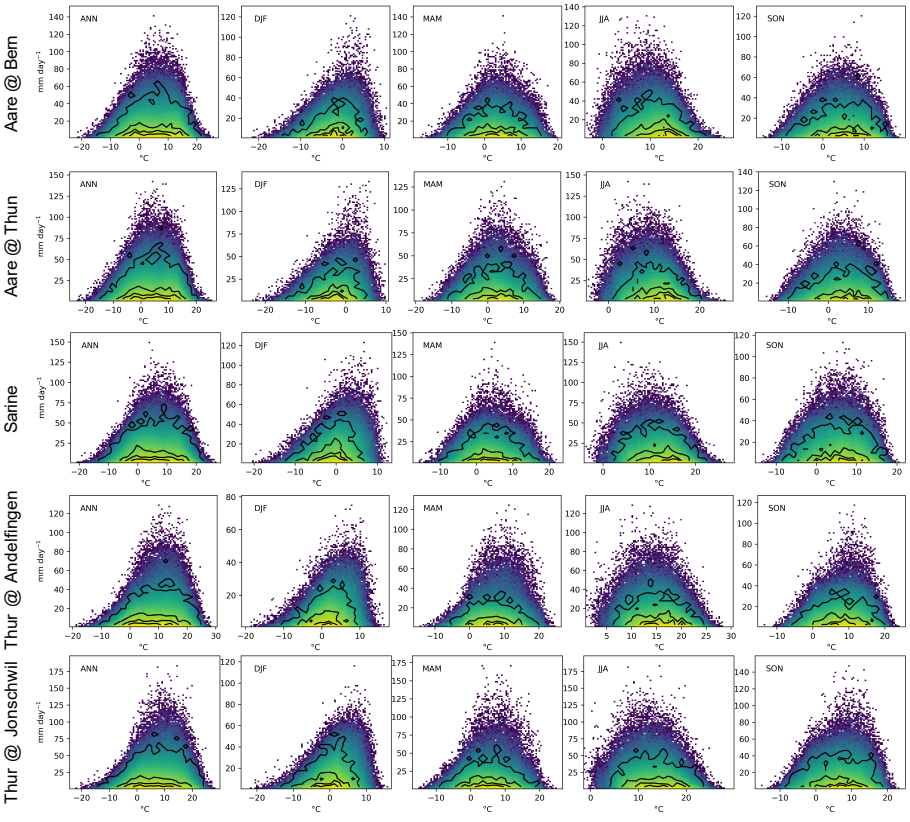


Figure S16. Annual and seasonal density plots based on daily mean temperature and precipitation ($\geq$1 mm day^{-1}) and the final 360 day calendar for large catchments. Colours range from blue for low density to yellow for high density. Observations over the 1991–2019 period are denoted by contour lines.

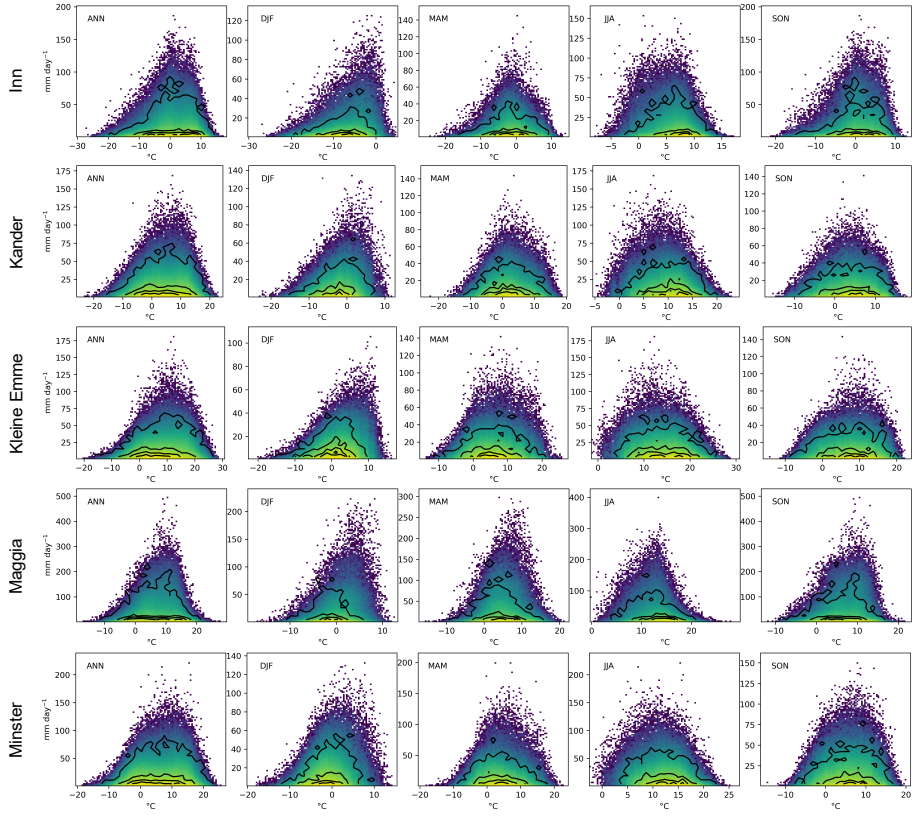


Figure S17. Same as Fig. S16, but for medium-sized catchments.

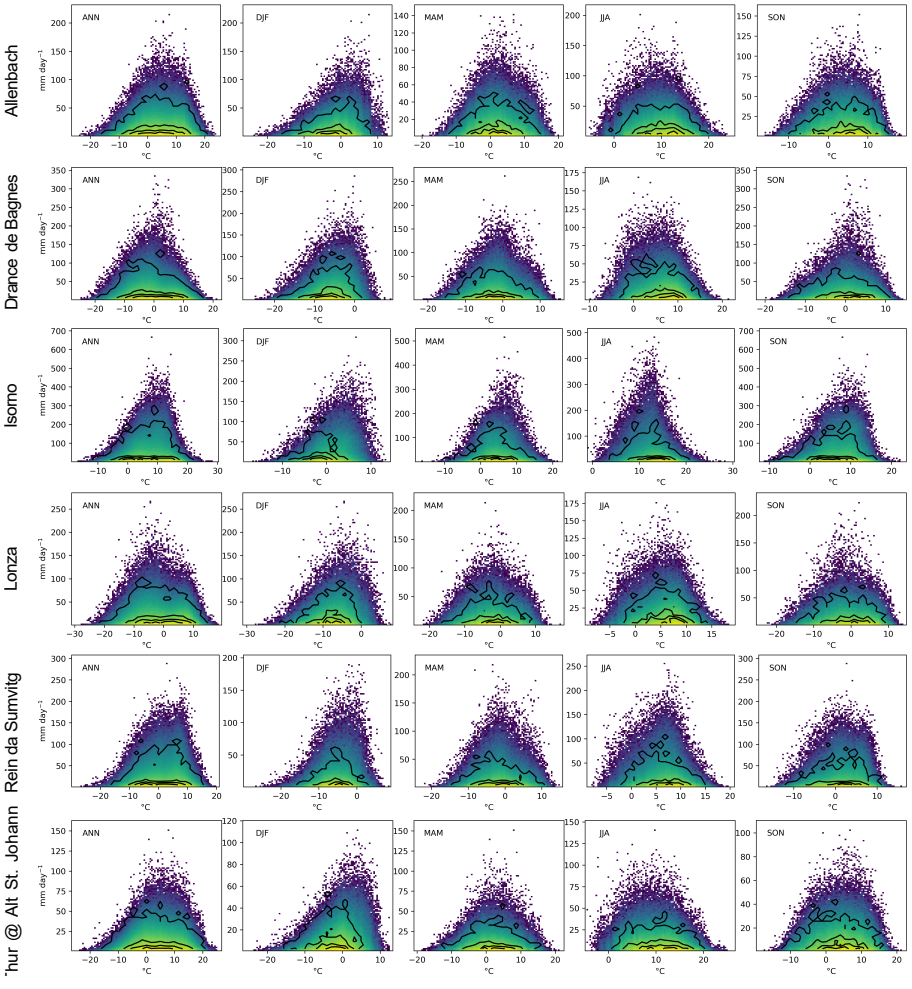


Figure S18. Same as Fig. S16, but for small catchments.

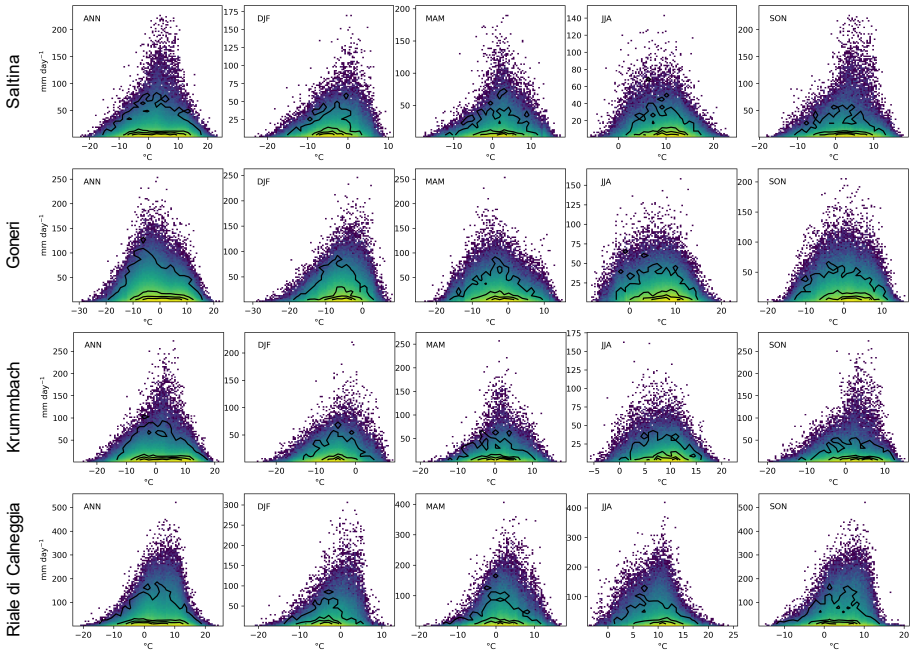


Figure S19. Same as Fig. S16, but for very small catchments.

Flow Duration Curves (FDCs) for selected sites, comparing RF-based simulations (RF), control simulation (CTRL) and observations (OBS). Note that the x axis is scaled logarithmically. For RF-based simulations, samples matching the length of the observations (or, if shorter, the length of the control simulation) were taken, exceedance probabilities calculated, and 95% confidence intervals were computed.
Seasonality of Annual Maximum Floods for selected sites, comparing reforecast-based simulations (RF), control simulation (CTRL) and observations (OBS).

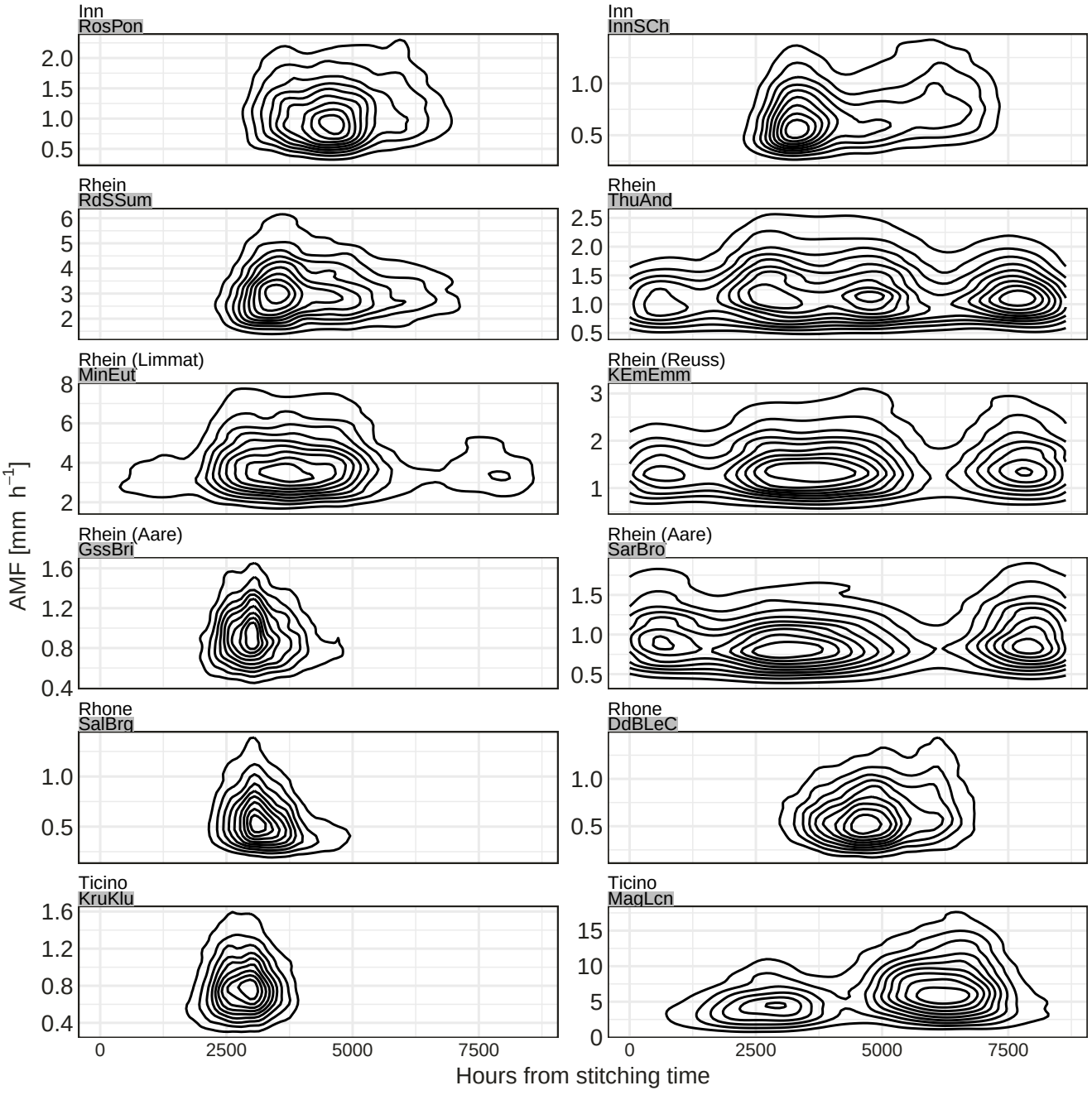


Figure S20. Contour density plots showing the magnitude of simulated AMFs annual maximum floods as a function of their occurrence time relative to the stitching points of individual RFs. Where available, both a smaller and a larger catchment was chosen were selected for each river basin.

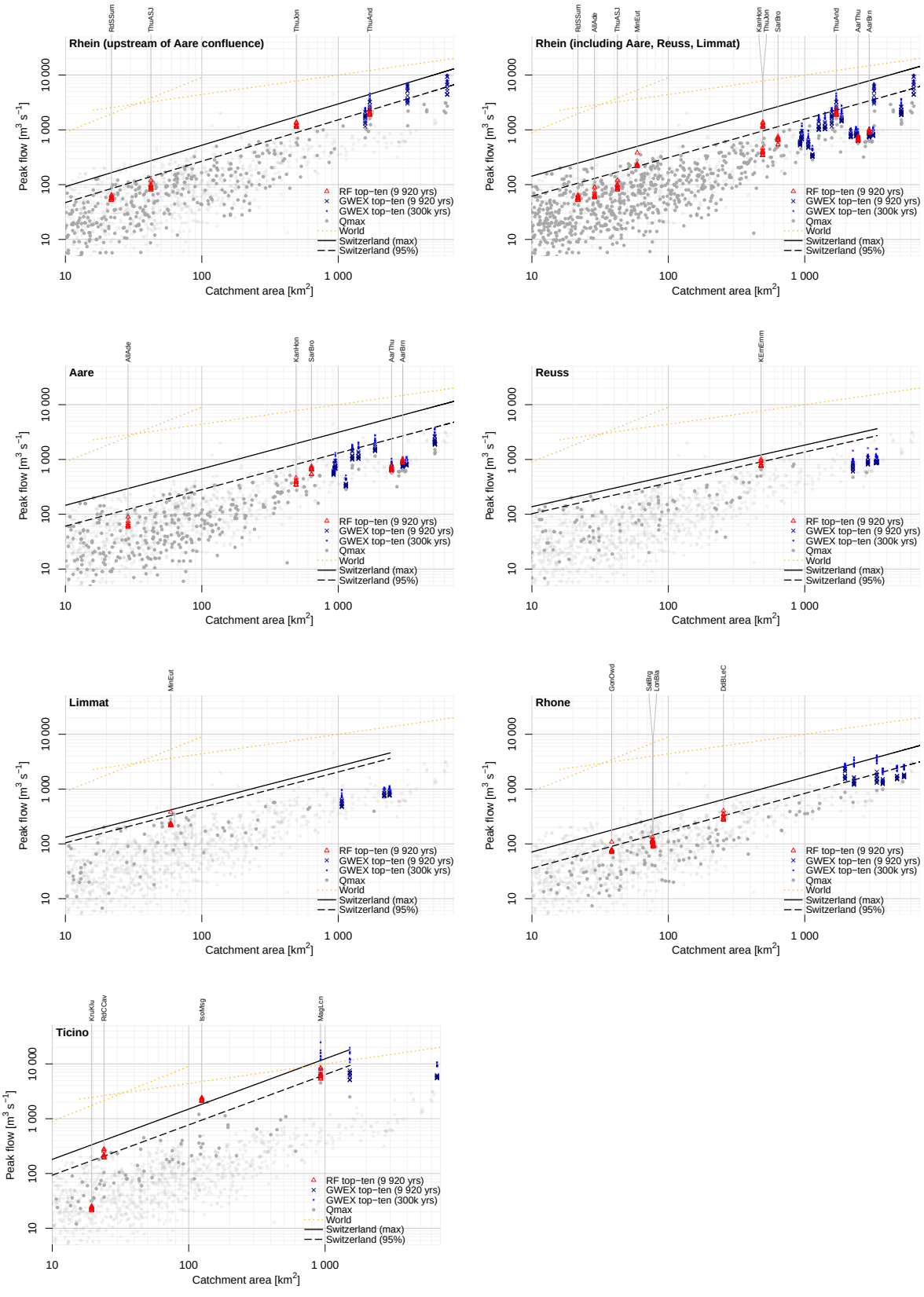


Figure S21. Highest ten annual maximum floods simulated using reforecast (RF) and weather generator (GWEX) input, presented separately for large river basins. Maximum observed floods in the corresponding large river basin (dark grey) and Switzerland (light grey) as per **?** (gray dots) are shown for context, along with an envelope curve for the corresponding large river basin (solid line: all data; dashed line: 95th percentile). Also shown is the envelope curve for maximum floods worldwide (**?**). For catchment IDs see Table **??**.

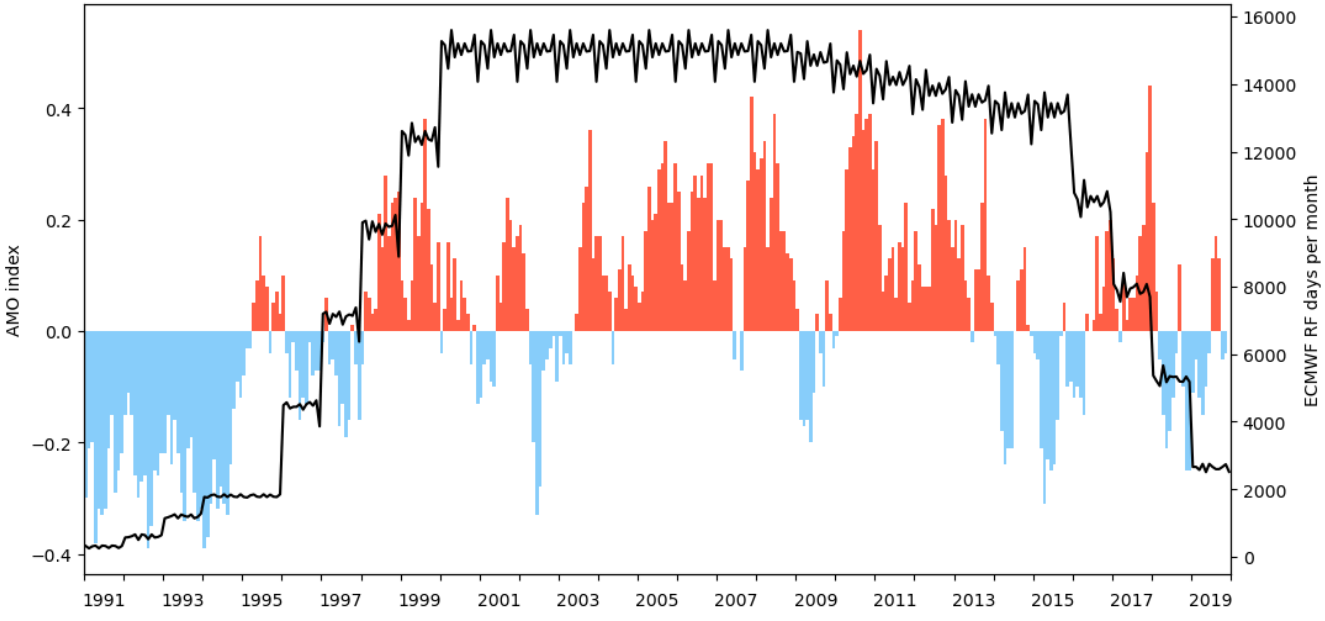


Figure S22. The Atlantic Multidecadal Oscillation (AMO) index in comparison to compared with the sampled reofrecasts as total days of the sampled reforecasts within the 9920 RF years processed. The AMO data stems were obtained from NCAR and is are based on **?** and HadISST1 (**?**).