



**HAL**  
open science

## Robustness of hydrometeorological extremes in surrogated seasonal forecasts

Katharina Klehmet, Peter Berg, Denica Bozhinova, Louise Crochemore,  
Yiheng Du, Ilias Pechlivanidis, Christiana Photiadou, Wei Yang

► **To cite this version:**

Katharina Klehmet, Peter Berg, Denica Bozhinova, Louise Crochemore, Yiheng Du, et al.. Robustness of hydrometeorological extremes in surrogated seasonal forecasts. *International Journal of Climatology*, 2024, 44 (5), pp.1725-1738. 10.1002/joc.8407 . hal-04591849

**HAL Id: hal-04591849**

**<https://hal.inrae.fr/hal-04591849v1>**

Submitted on 29 May 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Robustness of hydrometeorological extremes in surrogated seasonal forecasts

Katharina Klehmet<sup>1</sup>  | Peter Berg<sup>1</sup>  | Denica Bozhinova<sup>1</sup>  |  
 Louise Crochemore<sup>2</sup>  | Yiheng Du<sup>1</sup>  | Ilias Pechlivanidis<sup>1</sup>  |  
 Christiana Photiadou<sup>3</sup> | Wei Yang<sup>1</sup> 

<sup>1</sup>Hydrology Research Unit, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

<sup>2</sup>Institute of Environmental Geosciences, Université Grenoble Alpes, CNRS, INRAE, IRD, Grenoble INP, Grenoble, France

<sup>3</sup>Climate Change Impacts and Adaptation, European Environment Agency, Copenhagen, Denmark

## Correspondence

Peter Berg, Hydrology Research Unit, Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, Norrköping 60176, Sweden.  
 Email: [peter.berg@smhi.se](mailto:peter.berg@smhi.se)

## Funding information

EU Horizon Europe Project MedEWSa, Grant/Award Number: 101121192; EU Horizon 2020 Framework Programme, Grant/Award Numbers: 101003876, 776787

[Correction added on 15 March 2024, after first online publication: The first author name and surname were transposed, and have been corrected in this version.]

## Abstract

Water and disaster risk management require accurate information about hydrometeorological extremes. However, estimation of rare events using extreme value analysis is hampered by short observational records, with large resulting uncertainties. Here, we present a surrogate world setup that makes use of data samples from meteorological and hydrological seasonal re-forecasts to explore extremes for long return periods. The surrogate timeseries allow us to pool the re-forecasts into 1000-year-long timeseries. We can then calculate return values of extremes and explore how they are affected by the size of sub-samples as method for estimating the uncertainty. The approach relies on the fact that probabilistic seasonal re-forecasts, initialized with perturbed initial conditions, have limited predictive skill with increasing lead time. At long lead times re-forecasts will diverge into independent samples. The meteorological seasonal re-forecasts are taken from the SEAS5 system, and hydrological re-forecasts are generated with the E-HYPE process-based model for the pan-European domain. Extreme value analysis is applied to annual maxima of precipitation and streamflow for return periods of 100 years. The analysis clearly demonstrates the large uncertainty in long return period estimates with typical available samples of only few decades. The uncertainty is somewhat reduced for 100-year samples, but several 100 years seem to be necessary to have robust estimates. The bootstrap with replacement approach is applied to shorter timeseries, and is shown to well reproduce the uncertainty range of the longer samples. However, the main estimate of the return value can be significantly offset. Although the method is model based, with the associated uncertainties and bias compared to the real world, the surrogate approach is likely useful to explore rare and compounding extremes.

## KEYWORDS

extremes, precipitation, robustness analysis, seasonal forecasts, statistical uncertainty, streamflow

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *International Journal of Climatology* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

## 1 | INTRODUCTION

Extremes of precipitation and floods are among the most damaging water-related hazards for society and ecosystems (van Loenhout et al., 2020). Still, the fraction of people living in flood-prone areas has increased in the last decades (Tellman et al., 2021). The hazard is in many places expected to become worse with a warming climate, with a growing probability for changes in the frequency, intensity and duration of many types of extreme events including heavy precipitation, floods, droughts or heat waves (Belusic et al., 2019; Pachauri et al., 2014; Reyer et al., 2012; Seneviratne et al., 2012, 2021; Stott, 2016). Although societies adapt to the more frequent extremes, with return periods of a few decades, the largest consequences remain for higher return periods. Further, the damages may even be worsened by adaptation measures to frequent extremes, such as bursting levees, if they were built using too low design levels. Reliable and robust information, that is, reduced uncertainty in the estimation of hydrometeorological extremes is therefore essential for effective water and disaster risk management.

Disciplines working with risk analysis or engineering often use return periods to define extreme thresholds (Poschlod et al., 2020), which can be calculated by fitting an extreme value distribution to the data (Coles, 2001). However, many meteorological and hydrological estimates of, for example, 100- or 1000-year return levels are limited by relatively short observational data records of only few decades, with resulting large uncertainties (van den Brink et al., 2004, 2005). In cases where the variable's distribution is homogeneous in space, one can increase the sample by appending several spatially separated timeseries into a single longer timeseries. The independence must of course be ensured between the spatial samples, and this approach is sometimes called the *station-year method* (Olsson et al., 2019). For precipitation, this may be applicable for cases where translational invariance is fulfilled, such as over relatively flat land regions (Olsson et al., 2019; Overeem et al., 2008). Where precipitation is affected by, for example, strong orographic lifting, land-sea contrast and so forth, the method is not applicable. Similarly, the station-year method is far from straightforward to apply to hydrological extremes, as they are strongly affected by the characteristics of the catchment, such as its location, size, surface properties, response time and so forth. However, a detailed assessment of the characteristics for each catchment can be worthwhile to allow regionalization of the samples (Hosking & Wallis, 1997). The issue is further exacerbated by observation issues with streamflow timeseries, as sedimentation and other changes to the river may affect the rating curve

over time, causing inhomogeneity or discontinuities in the streamflow records.

Several methods have been proposed to solve the sample issue by using model simulations to increase the length of the timeseries. van den Brink et al. (2004) proposed to use probabilistic seasonal re-forecasts to increase the sample from a few decades to over 1000 years, when investigating extreme storm surge levels in the Netherlands. Similarly, Brunner and Slater (2022) made use of extended range re-forecasts of streamflow using the European Flood Awareness System (EFAS), and concluded that the increase in sample size allowed reducing the uncertainty bounds by on average 80% for the over 200 catchments studied. Global climate models have also been used to produce single model large ensembles by running several so-called realizations of the model, often starting from different natural oscillations of a pre-industrial quasi-equilibrium simulation. Such simulations produce a large number of timeseries that can be appended to feed the extreme value analysis with more data (van der Wiel et al., 2019). For example, Poschlod et al. (2020) used a single-model initial condition ensemble with 50 members in order to assess the frequency of heavy precipitation events over Europe, thus extending the historical timeseries from 30 to 1500 years. In a series of studies, pooling of single model ensembles was performed using the UNSEEN approach (UNprecedented Simulated Extreme ENsemble) (Thompson et al., 2017). The different studies explored extreme rainfall and summer heat waves in the United Kingdom (Thompson et al., 2017; Thompson et al., 2019) and drought hazards in China (Kent et al., 2019), based on global climate models. (Kelder et al., 2020) applied the UNSEEN approach to seasonal forecasts to assess trends in extreme precipitation.

In this study, we investigate the uncertainty in extreme value estimation of daily precipitation and streamflow across Europe using the ECMWF SEAS5 seasonal re-forecasts (Johnson et al., 2019) as well as re-forecasts from the hydrological model E-HYPE (Pechlivanidis et al., 2020), respectively. We pose the following scientific questions: (1) How does the sample size of constructed timeseries influence the robustness of precipitation and streamflow extremes? (2) Can the uncertainty bounds be estimated by common bootstrap methods? and (3) How do the spatial characteristics of return values differ depending on the sample size? To address these questions we construct synthetic timeseries by pooling single seasonal forecasts and forecast members into timeseries of several 1000 years; suitable for extreme value theory analysis. We employ the forecast skill as an indicator of independence of the members. Initial months of the forecasts with positive skill are excluded from the pooling. We then fit a Generalized Extreme Value

(GEV) distribution to annual maxima to derive return values for daily precipitation and streamflow using different sample sizes of the constructed surrogate timeseries.

## 2 | DATA AND METHODS

### 2.1 | Seasonal meteorological re-forecasts

We use outputs of daily mean temperature and precipitation for Europe provided by the ECMWF SEAS5 system (Johnson et al., 2019). Our focus is on the re-forecast period of 1993–2015, where SEAS5 provides 25 members with perturbed initial conditions. Each re-forecast covers 215 days and are initialized from the 1st of each month, thus always covering seven complete months. The data are extracted from the ECMWF MARS archive directly to a 0.5° regular grid. Analysis of precipitation extremes are based on these original SEAS5 data, without any bias adjustment applied.

### 2.2 | Hydrological seasonal re-forecasts

Hydrological seasonal forecasts are performed with version 3.0 of the E-HYPE model (Hundecha et al., 2016), with over 35,000 catchments across Europe. E-HYPE is forced by daily precipitation and temperature, and was set up and calibrated with the version 2.0 of the HydroGFD reference data (Berg et al., 2021). The same meteorological data is used to initialize the model until the start of each re-forecast, starting from a continuous historical simulation.

SEAS5 deviates from HydroGFD to an extent that urges bias adjustment to fit with the calibration of E-HYPE. Bias adjustment was performed with a modified version of the Distribution Based Scaling (DBS) method (Yang et al., 2010). The original version of DBS performs bias adjustment by fitting a distribution (double gamma for precipitation and normal for temperature) to the reference and the model data, and then producing a transfer function that maps the model distribution to that of the reference. The modification of DBS is to include all the 25 members of SEAS5 in the model distribution fit by first appending all members to a long timeseries, rather than performing the calibration of DBS separately for each member. The reason is to retain variability between the members. Further, the transfer function is derived separately for each lead month and start month for the forecasts. A detailed description of the seasonal hydrological forecasts and the forecasting skill is available in

Pechlivanidis et al. (2020), and aspects of this are further explored in the current paper.

The E-HYPE model performance has been explored in several studies (Donnelly et al., 2016; Hundecha et al., 2016, 2020). The main philosophy behind the E-HYPE calibration is to regionalize parameters across cluster regions based on catchment characteristics, to allow a reliable performance also in ungauged basins. Hundecha et al. (2016) present the current calibration of E-HYPE, which is based on catchment physiographic and climate descriptors to set parameters in a final clustering of eight groups with similar character. Evaluation using a subset of 538 stations which were divided into calibration stations and independent validation stations across Europe resulted in a median NSE of over 0.5 for both calibration and validation data. Hundecha et al. (2016) also explored the Q95 flows, and found generally good performance across the stations, and no systematic bias in relation to flow magnitude.

### 2.3 | Extreme value analysis

By applying extreme value theory, it is possible to calculate return values of different return periods of precipitation and streamflow by fitting extreme value distributions to a sample of independent and identically distributed events. In this study, we use the “block maxima” method by selecting the maximum value of 1-year blocks (Coles, 2001), for example, using calendar years. We note that in some catchments the 1-year block is not always sufficient—likely due to the extremes occurring around the break point at the 1st of January, or due to long-term memory effects that make extremes dependent. Sample independence can be assured in several ways, such as using multi-year blocks, a more flexible definition of the dates of the 1-year block, or using a peak-over-threshold approach. For the current analysis, the first option was investigated with 2 and 5-year blocks, without significant impact on the overall results presented here.

According to the Fisher-Tippett theorem, the distribution of the block maxima sample,  $\chi$ , can be described with the Generalized Extreme Value (GEV) distribution,  $G$ , which is defined as:

$$G(x; \xi) = \begin{cases} \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{1}{\xi}}\right), & \xi \neq 0 \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), & \xi = 0 \end{cases} \quad (1)$$

with  $\mu$ ,  $\sigma$  and  $\xi$  representing the location, scale and shape parameters of the distribution. Note that  $G$  will be undefined



unless  $1 + \xi ((x - \mu) / \sigma) > 0$  when  $\xi \neq 0$ . The parameter estimation was explored with both maximum likelihood (Jenkinson, 1955) and L-moments (Hosking, 1990). It was found that the maximum likelihood estimation often gave huge discrepancies for larger rivers where the estimation did not converge. This issue is not apparent with the L-moment estimator, while the results were similar for other catchments and grid points. The analysis shows the benefits of using L-moments, and as the results are very similar for all timeseries where the maximum likelihood did converge, we opted for using L-moments for all calculations presented in the paper. Return values are calculated by inverting Equation (1).

## 2.4 | Estimation of sample independence

The method of using seasonal forecasts as surrogate data rests on the assumption that the data are physically and statistically independent (van den Brink et al., 2004). The chaotic development at weather scales means that precipitation typically becomes independent after around 5–10 days, although large scale features may sometimes have longer predictability with some impact on precipitation (Kelder et al., 2020) and thus, to some extent, streamflow. Streamflow forecasting is to a large extent determined by the initial hydrological states (Musuza et al., 2023; Shukla & Lettenmaier, 2011), and can for some catchments and seasons retain traces of the initial state for days, weeks or even months. This implies that the different forecast members will retain a dependence with each other even though the meteorological drivers have become independent.

Previous studies have explored the forecast member independence of annual maxima using Spearman's rank correlation with member-pairs of annual maxima time series. Kelder et al. (2020) found independence after short lead times of less than a few weeks for precipitation in the SEAS5 re-forecasts. Brunner and Slater (2022) explored streamflow re-forecasts in a medium range forecast system and found sometimes much longer lead times with some dependence, in particular for catchments with snow-fed summer flood regimes in the Alps and Scandinavia. They used a general limit where the first 22 days' lead time is considered dependent and removed from each forecast member.

Forecast skill is a measure of the added value of a forecast, typically compared to climatology. This can serve as a proxy to estimate the independence between the members of a forecast ensemble. For example, a skilful forecast can be interpreted as a dependence between several of the forecast members on the initial states and the evolution of the hydrology (Girons Lopez et al., 2021; Pechlivanidis et al., 2020; Sutanto & Lanen, 2022).

Therefore, as a first step, the seasonal forecast skill for precipitation and streamflow extremes is assessed with the hypothesis that when the skill drops below a certain level, we can assume independence between the forecasts. The skill of precipitation and streamflow forecasts is assessed based on the Continuous Ranked Probability Skill Score (CRPSS). Additionally, the skill of extreme streamflow forecasts is assessed based on the Brier Score Skill score (BSS). The skill assessment was applied to the 25 members for 1993–2015.

CRPSS offers an overall picture of forecast skill by accounting for reliability, resolution and uncertainties (Hersbach, 2010). The reference used in the evaluation was HydroGFD for precipitation and for streamflow the perfect forecast, that is, a historical simulation with E-HYPE forced with HydroGFD. A benchmark is constructed for each forecast date by sampling 25 years from the historical timeseries for the same dates as the reforecast excluding the forecasted year. The CRPSS was thus assessed for each forecast initialisation month (January to December) and each forecast horizon (weeks ahead), as well as for the corresponding benchmark.

In addition, to confirm the skill assessment for extreme streamflow forecasts, we applied the Brier Score (Brier, 1950; Wilks, 1995) for high flows (defined as weekly streamflows above the 90th non-exceedance streamflow percentile). BS follows a strictly proper scoring rule to measure the accuracy of the probabilistic forecasts, and is given by:

$$BS = \frac{1}{T} \sum_{t=1}^T (P(X(t)) - \text{sgn}(\text{ref}))^2 \quad (2)$$

Here,  $\text{sgn}(\text{ref})$  gives a binary value of 0 and 1, indicating whether the reference exceeds the event threshold (0.9 in this study),  $P(X(t))$  provides the probability of exceeding the threshold from every model forecast at time  $t$ .

The Skill Score (both BSS and CRPSS) is computed to determine the added performance of the forecasts ( $S_{\text{fst}}$ ) with respect to a benchmark ( $S_{\text{ben}}$ ). In this study, the precipitation climatology serves as benchmark in the CRPSS in precipitation, and the simulated streamflow climatology serves as benchmark in the BSS in streamflow. The skill is computed with the following formula:

$$SS = 1 - \frac{S_{\text{fst}}}{S_{\text{ben}}} \quad (3)$$

The skill scores (BSS or CRPSS) range from  $-\infty$  to 1, with 1 indicating perfect skill and negative values indicating superiority of the benchmark.

BSS90, that is BSS for an event threshold of 0.9 of high streamflow extremes, is examined only for time steps within high streamflow periods to account for the intra-annual variability of the hydrological response between sub-basins, given Europe's strong hydro-climatic gradient (Du et al., 2023). The high streamflow period is defined by flows exceeding the upper tercile (66th percentile) as terciles are one of the standard measures to classify streamflow conditions as above-normal, near-normal or below-normal in operational applications. Detailed steps are as follows:

1. Define high streamflow periods for each sub-basin based on the upper tercile (66th percentile) from the streamflow climatology in the reference simulation (i.e., the historical simulation with E-HYPE forced with HydroGFD). Weeks with streamflow higher than the 66th percentile are considered as high streamflow weeks.
2. Define the high streamflow event threshold as the 90th percentile of weekly mean streamflow from all weeks in the reference simulation, and calculate  $BS_{fst}$  based on the threshold for each sub-basin, target week and lead time.
3. Construct the benchmark system for each forecast date by sampling 20 years from the historical time series for the same dates as the re-forecast excluding the current, previous and following year, to minimize artefacts related to similar initial conditions; calculate  $BS_{ben}$  accordingly.
4. Calculate BSS90 based on results from step 2 and 3. BSS within the high streamflow periods identified in step 1 are pooled regardless of initialization month and analysed as a function of lead weeks.

The skill of the forecasts deteriorates with time. We argue that the lack of forecast skill, that is, after CRPSS or BSS90 score become lower than 0, indicates independent ensemble members—suitable for the present analysis. We note that large scale features may lead to member dependence even if the skill drops below zero. Here, the use of randomly selected historical years should allow for a benchmark whose members evenly represent large scale configurations, and whose spread should thus ensure independence. The CRPSS is influenced by the forecast spread: for an unbiased forecast, a forecast ensemble sharper (wider) than climatology will have a positive (negative) skill. This condition of no bias is ensured here by using model climatology for streamflow and bias adjusted precipitations. In the case of the BSS90, any deviation from the probability of being above the 90th percentile indicates that the ensemble spread

defers from the climatological spread, indicating some dependence.

Moreover, observational uncertainty could partially impact the results from forecast verification (Jolliffe, 2017), particularly in regions where the reanalysis products lack of accuracy, the monitoring weather networks lack of high density and/or due to the representativeness error, that is, the mismatch between gridded forecasts and the interpolated observations (Bouallegue et al., 2020). Here, we are driven by operational limitations where a single product is used to represent the meteorological information and hence our investigation does not consider the aspect of observational uncertainty assuming that the reference dataset (HydroGFD) is “accurate”.

## 2.5 | Construction of surrogate timeseries

Because the seasonal forecast form incomplete years, we combine two or more individual seasonal forecast members based on their start times into a sequence of complete surrogate years. This is a valid approach since the continuity of the timeseries is not required when sampling daily extremes with block maxima. However, if multi-day extreme events are analysed, care would need to be taken to not split the days of maximum values when pooling different forecasts to a one-year timeseries. In practice, one would like to work with samples of 1, 2, 3, 4 or 6 months forecasts lengths, which can easily be aggregated to yearly data needed for the block maxima analysis. The pooling of individual seasonal forecasts to surrogate years is done after assessing the independence of the samples (see Section 2.4, Section 3.1). In this step we determine how many lead months we need to exclude from the forecasts before they are appended to yearly timeseries. For example, if independence is achieved after 1 month or less, we remove the first month from the forecasts and pair it with forecasts starting half a year later to form complete years (see Figure 1). When 4 months are retained, three forecasts will form a complete year, and so on.

The first step of merging forecasts to one surrogate timeseries allows the record to be extended by a multiplier equal to the number of available months. This is already a large increase from the original 23 years and 25 members, resulting, for example, in 3450 surrogate years when working with samples of 6 months forecasts lengths (23 years  $\times$  25 members  $\times$  6 months). In addition, one could repeat the construction of the surrogate years with different combinations, which can provide information for further analysis of robustness. For example, the

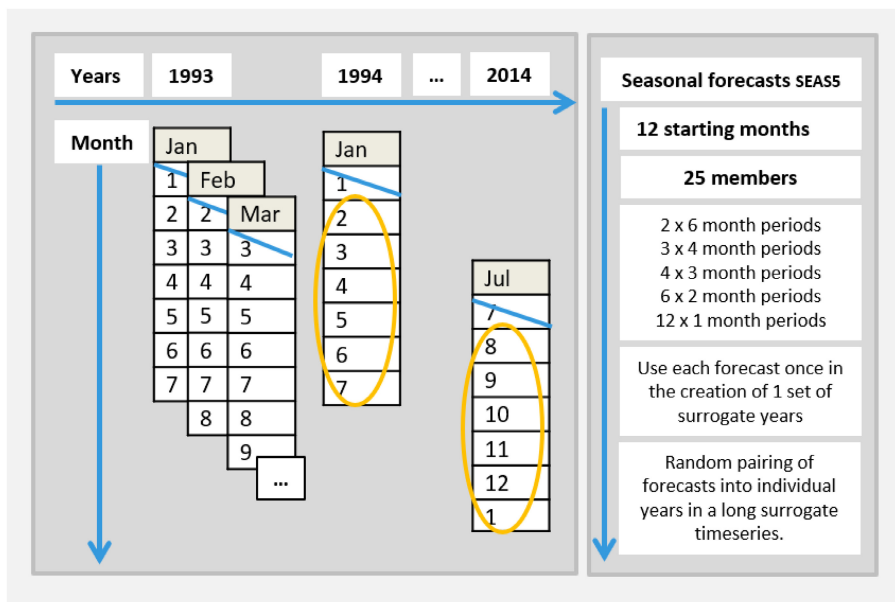


FIGURE 1 Diagram of the procedure to append re-forecasts to construct the surrogate years.

annual maxima might be different depending on which two timeseries were appended, which would affect a block maxima, but not a peak-over-threshold approach to extreme value analysis. Note that the re-combinations were initially explored, but from the perspective of having over 1000 years of data at disposal the results and conclusions made in the paper did not change, and the analysis was therefore left out.

## 2.6 | Robustness of the extreme value estimation

To evaluate the robustness of return values for different sample sizes, that is, the number of years in the timeseries, a random selection of years of 50, 100, and 500 years is made from the complete timeseries, for example, from a timeseries of 3450 surrogate years (see Section 2.5). The selection was repeated 100 times, which is found sufficient to inform on the uncertainty intervals and the median value of the selections. GEV fits are made based on each selection and used for the analysis, which consists of inspections of box plots for single cases, and the inter-quartile range for spatial analysis.

In addition, the robustness is assessed from single shorter samples, to simulate the case with a limited observed timeseries. The robustness analysis is then performed using bootstrap with replacement (Gilleland, 2020), that is, to select a set of  $N$  annual maxima from an  $N$  year timeseries, while allowing the same year to be selected several times. The random selection is repeated until a stable statistic is found. Three sets of bootstrap experiments are performed starting from the original long

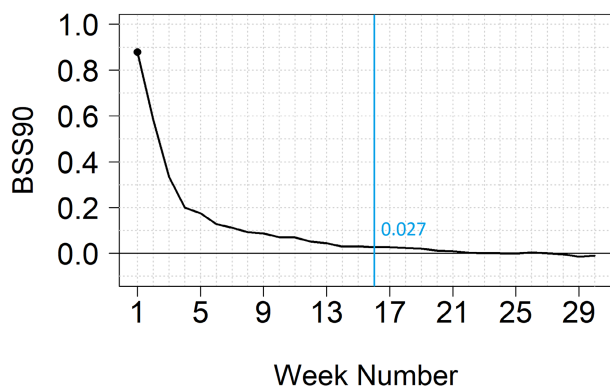
timeseries, where the respective shorter timeseries are selected to avoid overlap within the same-length periods: BS1 uses years 1–50, 1–100 and 1–500 of the main time series, BS2 uses years 51–100, 101–200, and 501–1000, and BS3 uses years 101–150, 201–300, and 1001–1500. A set of 100 bootstraps are found sufficient for robust statistics of the GEV fit and return value calculations.

## 3 | RESULTS

### 3.1 | Forecast skill and sample independence

The forecast skill analysis for precipitation shows a very sharp decline in the skill with time (not shown), which is well known (Kelder et al., 2020). With some regional variations and a slight dependence on the start month, independence of precipitation forecasts can safely be assumed from 2 to 4 weeks into the forecasts for most regions across Europe. Such relatively homogeneous patterns were expected and are likely due to the chaotic nature of the atmosphere which leads to uncertainties within the range of climatological records for horizons further than 10 days ahead. Therefore, we decided to exclude the first month from each of the precipitation forecasts and work with 6 month forecast lengths (instead of the full 7-month forecasts) when pooling to surrogate years.

Streamflow is more challenging, due to the longer predictability scales. Figure 2 shows a summary of the median BSS90 for all catchments, with fast declining skill over the first 4 weeks, and then a gradually slower rate of decline for longer lead times. The skill generally drops



**FIGURE 2** Median BSS90 of streamflow for all catchments as a function of the lead week into the forecast. The skill level at week 16 is indicated, see main text.

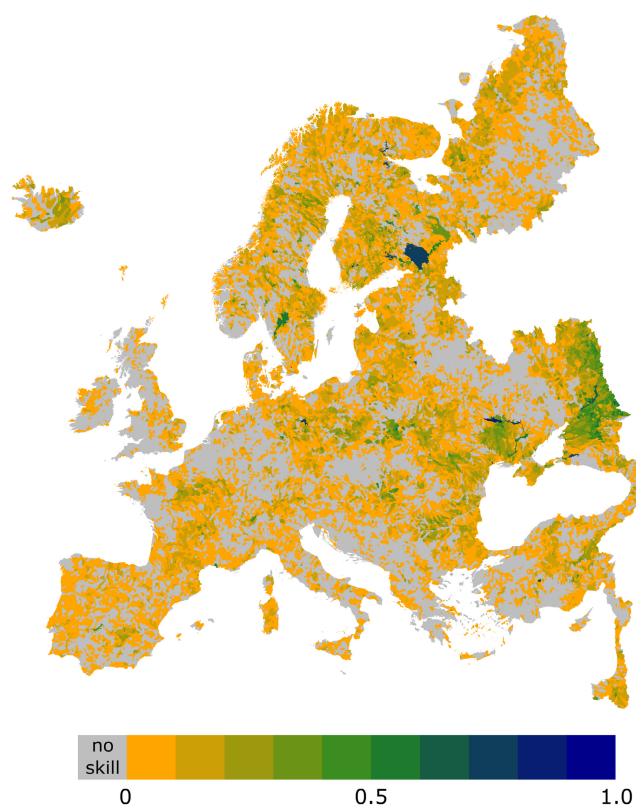
below 0.1 from lead week 8 (Figure 2), but this is more uniformly observed across all catchments from lead week 16 (Figure 3).

Larger regions have zero or negative skill, and most other regions have skill below 0.1. There are, however, regions with higher skill, which roughly correspond with regions of low runoff coefficient and a larger spring flood or large base flow (Pechlivanidis et al., 2020). A few large lakes have much higher remaining skill, which is likely due to the large reservoirs and flow management. The results are similar for the CRPSS (not shown).

Based on this result, we set a general lead time of 4 months after which we assume independence of the individual members. This is larger than the limit of 22 days set by Brunner and Slater (2022), and can be considered a conservative choice for this pan-European study. A more detailed assessment of single catchments might lead to a less conservative choice and allow for a larger sample. Removing the four lead months in each forecast period for streamflow allows us to construct a timeline with 1725 surrogate years (23 years  $\times$  25 members  $\times$  3 months), while removing only a single month from the precipitation data results in 3450 surrogate years.

### 3.2 | Estimation of return periods for different sample sizes

To investigate the robustness of the extreme value estimation, we perform GEV fits to data of different record lengths (see Section 2.6). Figure 4 presents selected results from the experiment based on a set of 100 surrogates and sample sizes from 50, 100 and 500 years. We focus first on the blue boxplots that show the effect of different sub-samples from the main timeseries. The median

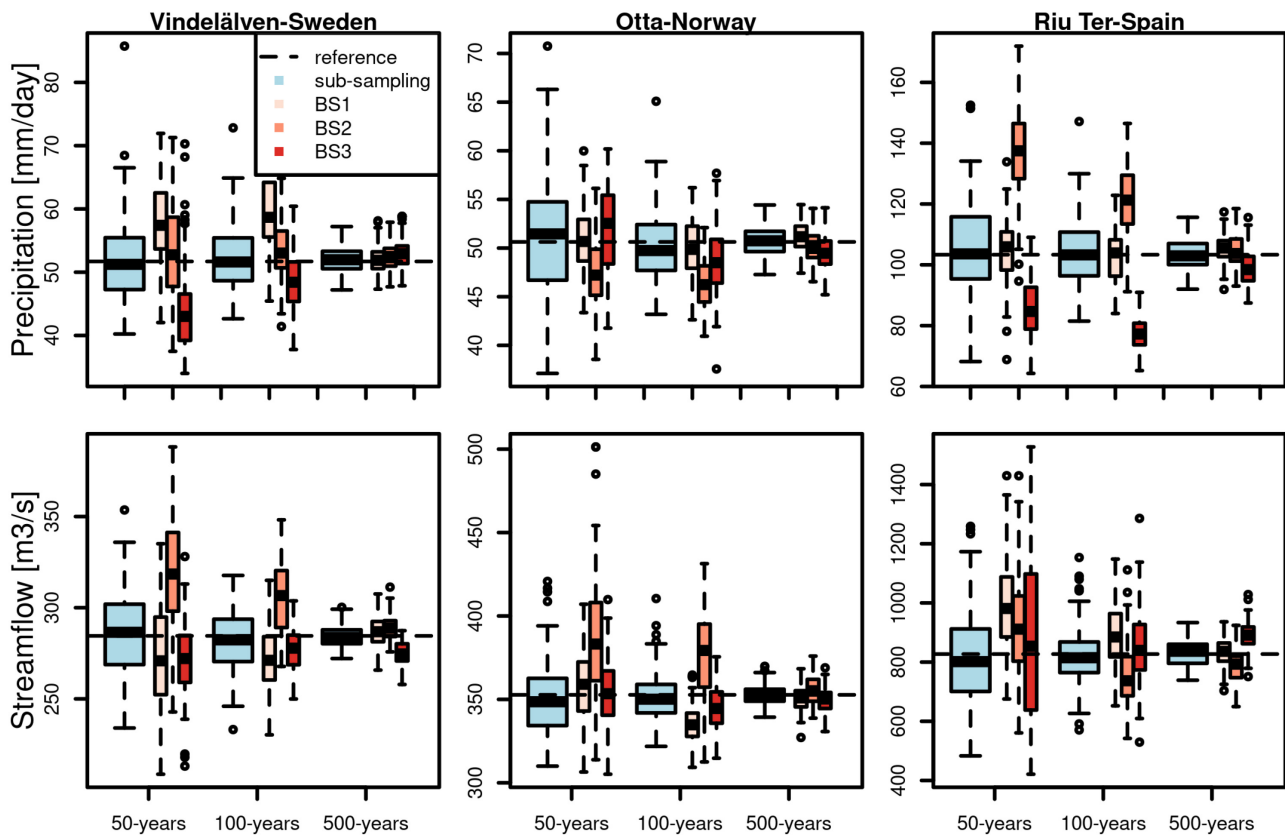


**FIGURE 3** A map of the skill level for streamflow in each catchment for lead week 16.

of the return values for all sub-samples converges on that calculated from the complete timeseries, which is a sign that the number of random samples is sufficient. The uncertainty is very large for the 50 year samples, which would be considered a long timeseries had it been observations. Taking Vindelälven as an example, the range around the median (about  $285 \text{ m}^3 \text{ s}^{-1}$ ) is about  $\pm 55 \text{ m}^3 \text{ s}^{-1}$ , which is about 20%—a substantial uncertainty. A similar uncertainty is present for all cases. With 100 year samples, the spread is somewhat reduced, but still substantial. However, at 500 years the uncertainty is much better constrained with a range of about  $\pm 15 \text{ m}^3 \text{ s}^{-1}$ , which is about 5%.

We simulate the access to only a single shorter record using a bootstrap with replacement strategy, which is shown as red boxplots for each of the record lengths, with three completely separate selections of sub-periods. Clearly, the sub-period strongly affects the median result of the bootstrap, which is offset from the target value of the complete timeseries. This is expected from the resulting large uncertainty shown in the blue boxplots. However, the uncertainty range of the boxplots is still similar in magnitude to that of the blue boxes, although it varies quite a lot between samples and case studies. That is, the bootstrap based on a single shorter timeseries represents





**FIGURE 4** 100-year return values for precipitation (top row) and streamflow (bottom row) for three representative catchments in Europe. The return values are presented as a function of the sample size, where each  $x$ -year long timeseries was sampled 100 times to derive the confidence intervals shown as boxplots. The dashed lines mark the return value calculated from the complete timeseries (3450 years for precipitation and 1725 years for streamflow). The boxplots are grouped by the length of the timeseries used and show the effect of random sub-sampling from the complete timeseries (blue) and bootstrap from a fixed timeseries from three independent samples (BS; red colours). The locations are at (latitude, longitude): NW-Sweden (65.95, 16.33), SW-Norway (62.10, 8.86), and NE-Spain (42.05, 2.19).

the full range of uncertainty from a much longer record. The issue of the offset from the target value remains, which gives an error in the estimate of the return values. There is further a large random component in the selection of the sub-samples, which is seen as the large offsets between the different red boxplots for each single record length.

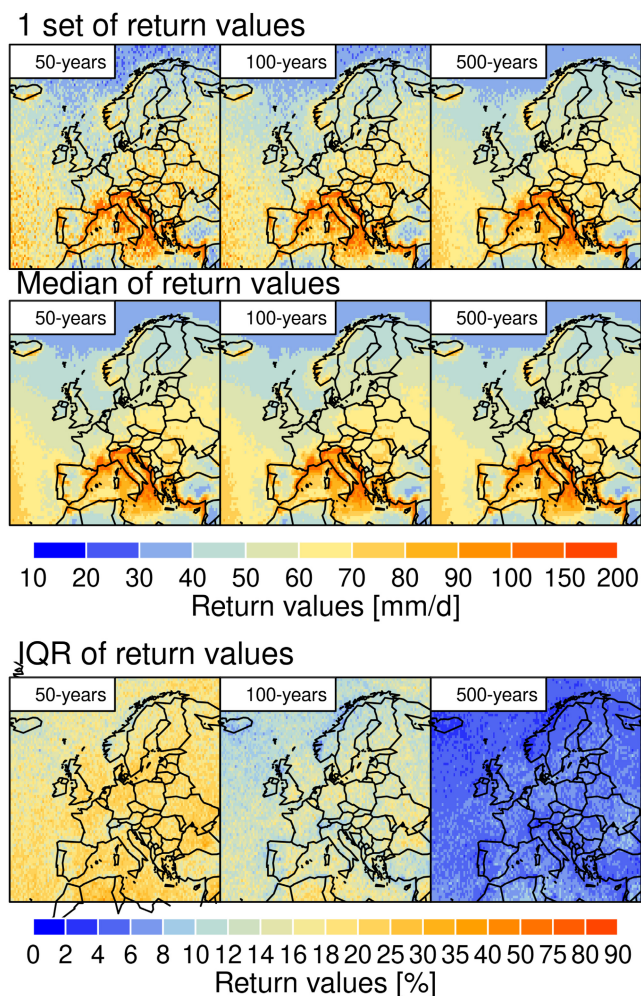
### 3.3 | Spatial patterns of 100-year return value

We now extend the analysis to all grid points and catchments in Europe. For precipitation, Figure 5 shows 100-year return values in three rows for (i) randomly selected single timeseries of different lengths, (ii) the median of 100 samples for each timeseries, and (iii) the interquartile range (IQR) of the samples relative to the median and expressed as a percentage. In comparison with Figure 4, the rows correspond to (i) any random

point along the boxplot whiskers, (ii) the median line in the boxplot, (iii) the difference between the top and bottom of the box divided by the median and multiplied by 100.

Focusing first on the 50 year timeseries (left column of Figure 5), the single return value estimates show a generally consistent pattern with that seen for the median of all samples. This means that there are spatial features of the precipitation extremes that are more prominent than the “noise” of the data. However, at any given grid point, the return value can differ by tens of mm/d, depending on the character of the single timeseries. This can lead to both over- and underestimations in different regions. We note that the station-year method is taking advantage of this fact, but requires careful selection of stations to remain in a spatial region with coherent extremes. The relative IQR in the bottom row shows that the range of values is rather homogeneous across Europe at between 15% and 30% for the 50-year long timeseries. However, some regions with strong orography, such

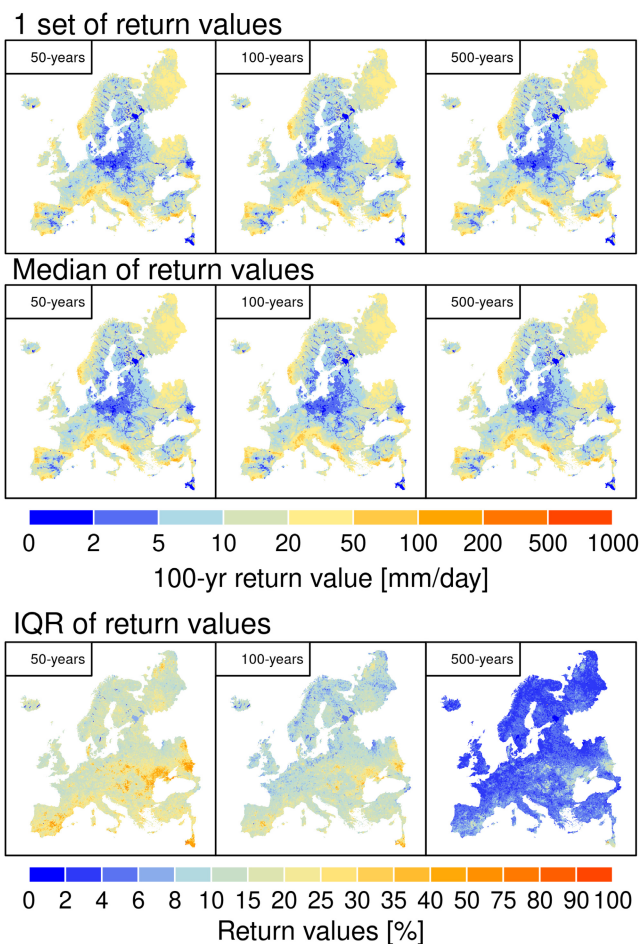




**FIGURE 5** 100-year return values for precipitation obtained from timeseries of 50, 100 and 500 years. Upper row: the 100-year return values of a single randomly selected surrogate timeseries. Middle row: the median of the 100-year return values of all 100 surrogate timeseries. Lower row: the interquartile range (IQR) of the 100-year return values of all 100 surrogate timeseries relative to the median.

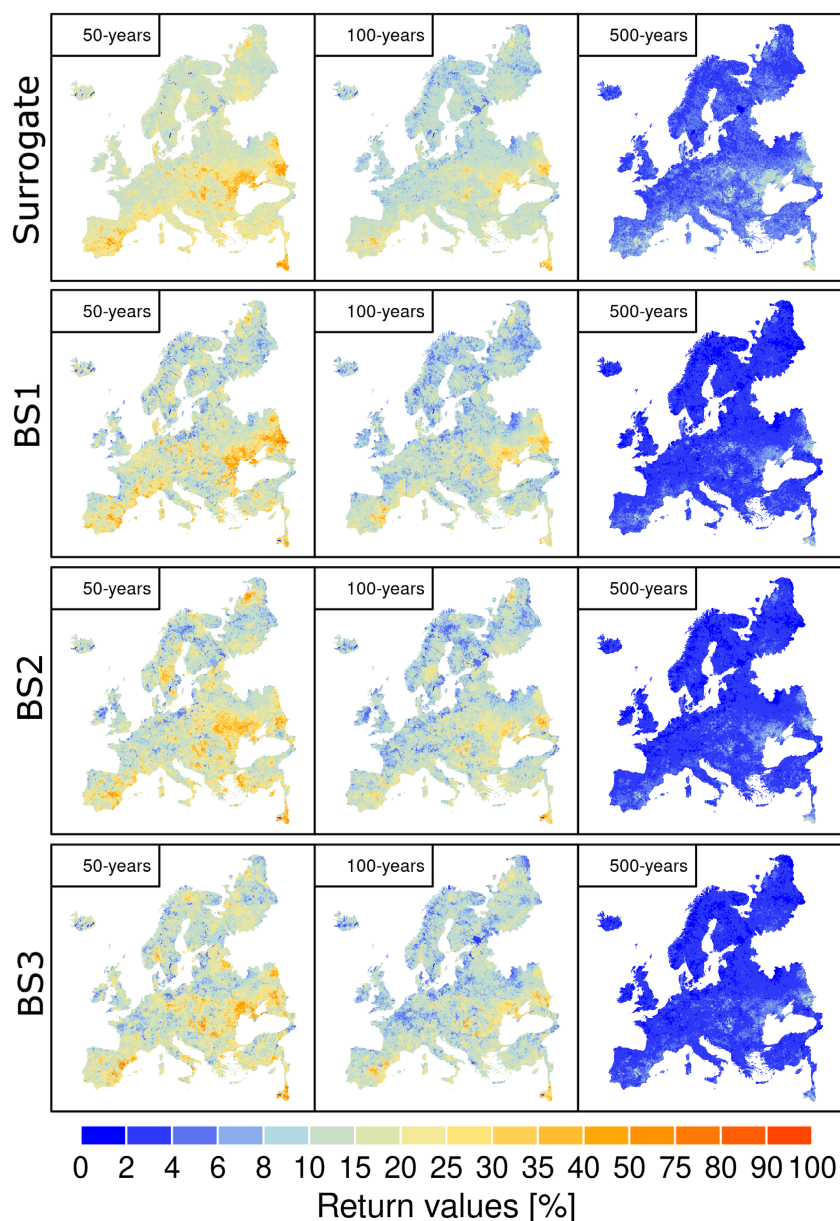
as the coast of Norway and the Alps, have a systematically lower relative IQR.

When timeseries of 100 years are sampled (middle column of Figure 5), the single sample shows a reduced spatial variability compared to the 50 year sample, and looks more similar to the median of all samples. The relative IQR has been reduced by about half compared to the 50 year long timeseries. For the 500 year timeseries (right column of Figure 5), the relative IQR is close to zero, and there are only few visible differences between a single sample and the median of all 100 samples. The 500-year timeseries have yet less difference between the single sample and the median, and the IQR is now reduced to less than 10%.



**FIGURE 6** 100-year return values for streamflow normalized by the upstream area obtained from the experiment of 100 surrogate timeseries selecting 50, 100 and 500 years. Upper row: the 100-year return values of 1 surrogate timeseries is shown. Middle row: the median of the 100-year return values of all 100 surrogate timeseries is shown. Lower row: the interquartile range (IQR) of the 100-year return values of all 100 surrogate timeseries is shown relative to the median.

Streamflow follows a very similar pattern as precipitation, but with overall larger variability across the domain, although the streamflow has been normalized by the upstream catchment area to give units of mm/d instead of  $\text{m}^3/\text{s}$  (Figure 6). This larger spatial variability is due to the vastly different volumes of water passing through the different river basins and the various hydrological regimes and catchment responses, and the figures convey only the larger differences and not much nuances. The relative IQR best conveys the uncertainty for each sample, with 10 to several 100% for the 50-year timeseries, which are reduced to 0%–100% for the 100-year timeseries. At 500 years, the IQR is again reduced to less than 10%, as for precipitation. A region in eastern Europe, just north of the Black Sea, stands out with markedly higher



**FIGURE 7** 100-year return values interquartile ranges (IQR) for streamflow normalized by the upstream area obtained from the experiment of 100 surrogate timeseries selecting 50, 100 and 500 years. The top row shows the results for the surrogate dataset, and is identical to the bottom row of Figure 6. The following three rows show results for the three bootstrap experiments that use 100 bootstrap samples from different timeseries.

return values and IQR for streamflow. This is a region with substantial anthropogenic influences, such as irrigation, and thus poor performances of the E-HYPE model (Hundecha et al., 2016), resulting in highly uncertain streamflow and return periods.

Finally, we conclude again on the robustness that can be achieved from bootstrapping within a single timeseries. As shown in Figure 4, the difference between the median result can be substantial depending on the chosen timeseries. There is no way around that, as one cannot easily estimate this value from a short timeseries. We focus instead on how well the bootstrap method describes the uncertainty of the return value, in the form of the IQR. Figure 7 shows the IQR from the surrogate method, as before, and for the three bootstrap experiments. The general pattern is similar between BS1–3 and the

surrogate method already for the shortest 50-year timeseries. However, the BS experiments are all both under- and overestimating the IQR, depending on the region and on the chosen timeseries. The differences are reduced for the 100-year timeseries, and for 500-years BS1–3 are generally underestimating the IQR by a few percentage units. In a broad perspective, the bootstrap method gives a good overview of the uncertainty range, but the local deviations can be substantial.

## 4 | DISCUSSION

When using models as a replacement for observations to estimate extremes, one is essentially substituting the uncertainty in the observations to that of the model's

performance in realistically simulating extremes and their frequency. The model needs to achieve at least a basic accuracy in simulating extremes to be a plausible replacement of observations. At the same time, the presented results indicate the substantial noise levels in the extreme value estimations, which makes such evaluation very challenging when only short timeseries of a few decades are available.

The gain of using model surrogate timeseries is in the long timeseries that can be constructed with the models. This constitutes a huge advantage over observations, considering the rapidly widening confidence intervals of any extreme value theory estimation outside the range of the data records. With increasing attention to extremes, often in the context of adapting society to current and a changing climate, there is a great need for this kind of background information about uncertainty.

The similarities between single small sample timeseries to the more robust large samples in the overall pattern across large areas (on the order of several decades of percent of the land mass of Europe) means that one can expect noise levels to cancel out. A path forward in assessing model performance in extremes is therefore to perform analysis over large regions with accumulated statistics. This is similar to the station-year method (Olsson et al., 2019) mentioned above, or continent-scale analysis performed by, for example, Guerreiro et al. (2018).

Independence between the ensemble members is explored using skill metrics of the forecast system, assuming independence when there is no skill.

This approach could be refined to pair different forecast lengths together subject to the skill score of each grid point and catchment. Furthermore, the CRPSS and BSS90 skill metrics are applied to weekly averages. This procedure certainly smooths the high extremes to some extent.

The question whether large-scale atmospheric drivers have a higher predictive skill in the SEAS5 forecasts than precipitation or streamflow extremes itself and how this affects the assumption of independent ensemble members deserves further research. We have explored the weather regimes of the SEAS5 re-forecasts in earlier work, with emphasis on Sweden. In the study 12 weather regimes were classified based on daily anomalies of mean sea level pressure from ERA-Interim and optimized by historical observed precipitation in Sweden and then predict the occurrence of those patterns from ensemble seasonal forecast. We found essentially no correlations between the ensemble members of the large scales on day-to-day basis. However, some studies have indicated predictability on certain weather regimes over the North Atlantic and Europe, such as NAO+ and ENSO (Falkena et al., 2021). This could affect the ensemble member

independence of precipitation and streamflow extremes over Europe. Additional analysis along this reference and with respect to the understanding of the connections between hydrometeorological extremes and large-scale weather patterns (Lavers et al., 2013) would be needed. However, this analysis is beyond the scope of this paper.

A somewhat different approach for the independence testing is applied in the UNSEEN studies following the idea of the potential predictability (Kelder et al., 2020; Lavers et al., 2014). Kelder et al. (2020) instead assessed the dependency of data by applying a pairwise correlation test between all ensemble members. The data were resampled and data points were randomly selected from all members, years and lead times to remove potential correlations. Our independence testing of ensemble members provides an alternative to the one in the UNSEEN approach, with data that can generally be derived from forecast evaluations.

The 23 years of the meteorological and hydrological re-forecasts used in this study represent only part of the present-day climate variability. For a full spectrum of extreme value estimation under general climate conditions one would need to include data of several decades, if available (van den Brink et al., 2005). A viable option is to make use of single model multiple realizations from climate models, for example, CMIP6 (Coupled Model Intercomparison Project Phase 6), where centennial climate simulations are performed with varying initial conditions. Although of coarser spatial resolution, these models allow for large ensembles across long historical and future projections to assess extremes under different climatic states, as well as natural variability thereof. However, unless such ensembles include hundreds of members, they do not reach the same number of samples for a given historical period as the currently presented use of seasonal forecasts. The optional use of a multi-model ensemble from CMIP6 requires additional checks to ensure that the included models share a sufficiently similar distribution of extremes.

In this study, precipitation and streamflow extremes are estimated separately. In a future development this method can be extended to investigate multiple extremes or compound events to account for the increasing risk of interconnected extremes.

Further, the analysis is based on the GEV distribution to determine extreme precipitation and streamflow estimates. This might not be ideal for all regions. Various processes affect especially streamflow such as regulations included in the model. A naturalized hydrological model could be used, but is generally not of interest for the users of the forecasts. If specific regions and catchments are investigated, one way forward could be to allow the method to find best fits for alternative or extended



distributions (Nascimento et al., 2016) or even mixed distributions (Gruss et al., 2020). However, here we wanted to focus our analysis on the European scale, we therefore had to sacrifice some detail in the distribution fits.

## 5 | CONCLUSIONS

We present a method to pool seasonal re-forecasts to generate synthetic timeseries with several 1000 surrogate years suitable for extreme value analysis. The method is evaluated on a pan-European scale using both meteorological and hydrological seasonal re-forecasts. The pairing of seasonal forecasts ensembles used in this study enables the record to be extended from the original 23 years (1993–2015) to 3450 and 1725 surrogate years for precipitation and streamflow, respectively. The different record lengths arise from the exclusion of 1 and 4 month (s) from the full 7-month forecasts after assessing the sample independence. We investigate the robustness of return value estimates for precipitation and streamflow using a 1-year block-maxima fitted to the GEV distribution. We do this by applying two approaches. In a first step we assess the robustness of the GEV fits to data of different record lengths. This investigation provides the opportunity to identify the timeseries length needed to obtain more robust extreme value estimates. In a second step we assess the robustness to only a single shorter record using a bootstrap with replacement to assess the uncertainty. The main conclusions are:

- The forecast skill and testing of independence between ensemble members for precipitation indicates some regional variations and a slight dependence on the start month. Independence can be assumed from 2 to 4 weeks into the forecasts for most regions across Europe. The skill and independence analysis for streamflow indicates a more complex pattern with topographic and hydrological heterogeneities varying with the month of forecast initialisation.
- Constructing long surrogate timeseries offers a clear statistical advantage in the extreme value theory estimation, particularly over short observed data records. It provides a great potential in reducing the uncertainty of return period estimates for both precipitation and streamflow. The improvement is particularly visible for precipitation all over Europe from a sample size of about 500 years. This increase in robustness is also evident for streamflow.
- The more common approach of assessing uncertainty, or confidence levels, with a bootstrap method shows overall similar assessments of the range of possible values, although it might severely under- or overestimate

at the local scale. Still, the best estimate of the return value is still subject to the large uncertainty in the timeseries sample, and the uncertainty range will be equally offset.

With increasing use of seasonal forecasts in impact modelling, the presented method may gain attention and be useful to more robustly assess the current state of extremes as background information to increase resilience to floods or other natural hazards.

The larger sample sizes allow better assessments of multiple extremes. The more robust return period estimates of unprecedented extreme streamflow events may be helpful in improving the risk estimation of hazards associated with flooding and in designing risk management options for decision-makers.

## AUTHOR CONTRIBUTIONS

**Katharina Klehmet:** Investigation; writing – original draft; methodology; visualization; writing – review and editing. **Peter Berg:** Conceptualization; investigation; writing – original draft; funding acquisition; methodology; visualization; writing – review and editing. **Denica Bozhinova:** Software; data curation; methodology. **Louise Crochemore:** Investigation; methodology; formal analysis. **Yiheng Du:** Investigation; methodology; formal analysis. **Ilias Pechlivanidis:** Writing – review and editing. **Christiana Photiadou:** Writing – review and editing. **Wei Yang:** Writing – review and editing; software; methodology; investigation.

## ACKNOWLEDGEMENTS

This study was partially funded by the EU Horizon Europe Project MedEWSa (Mediterranean and pan-European forecast and Early Warning System against natural hazards) under Grant Agreement No. 101121192 and the EU Horizon 2020 Framework Programme project CLINT (Climate Intelligence: Extreme events detection, attribution and adaptation design using machine learning) under Grant Agreement No. 101003876. Funding was also received from the EU Horizon 2020 project S2S4E (Subseasonal to seasonal forecasting for the energy sector) under Grant Agreement No. 776787.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The seasonal meteorological forecasts SEAS5 of ECMWF in 1° spatial resolution are freely accessible from the Copernicus Climate Data Store (<https://cds.climate.copernicus.eu/cdsapp#!/dataset/seasonal-original-single-levels>). This dataset

is coarser than the one used in this study (0.5° resolution) but is essentially the same. The HYPE model code is available from the HYPEweb portal (<http://hypeweb.smhi.se/model-water>). The created surrogate timeseries based on the seasonal meteorological and hydrological re-forecasts, return values of precipitation and streamflow and scripts used to perform the analysis can be accessed from the corresponding author upon request.

## ORCID


Katharina Klehmet  <https://orcid.org/0000-0002-1503-2908>

Peter Berg  <https://orcid.org/0000-0002-1469-2568>

Denica Bozhinova  <https://orcid.org/0000-0003-0611-321X>

Louise Crochemore  <https://orcid.org/0000-0001-5776-6275>

Yiheng Du  <https://orcid.org/0000-0002-5176-8111>

Ilias Pechlivanidis  <https://orcid.org/0000-0002-3416-317X>

Wei Yang  <https://orcid.org/0000-0002-6803-5563>

## REFERENCES

- Belusic, D., Berg, P., Bozhinova, D., Bärring, L., Doescher, R., Eronn, A. et al. (2019) Climate extremes for Sweden. SMHI.
- Berg, P., Almén, F. & Bozhinova, D. (2021) Hydrogfd3.0 (hydrological global forcing data): a 25 km global precipitation and temperature data set updated in near-real time. *Earth System Science Data*, 13, 1531–1545.
- Bouallegue, Z.B., Haiden, T., Weber, J.N., Hamill, T.M. & Richardson, D.S. (2020) Accounting for representativeness in the verification of ensemble precipitation forecasts. *Monthly Weather Review*, 148, 2049–2062.
- Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.
- Brunner, M.I. & Slater, L.J. (2022) Extreme floods in Europe: going beyond observations using reforecast ensemble pooling. *Hydrology and Earth System Sciences*, 26, 469–482.
- Coles, S. (2001) An introduction to statistical modeling of extreme values. In: *Statistics*. London, England: Springer Series.
- Donnelly, C., Andersson, J.C.M. & Arheimer, B. (2016) Using flow signatures and catchment similarities to evaluate the e-hype multi-basin model across Europe. *Hydrological Sciences Journal*, 61, 255–273.
- Du, Y., Clemenzi, I. & Pechlivanidis, I.G. (2023) Hydrological regimes explain the seasonal predictability of streamflow extremes. *Environmental Research Letters*, 18, 094060. Available from: <https://doi.org/10.1088/1748-9326/acf678>
- Falkena, S.K.J., de Wiljes, J., Weisheimer, A. & Shepherd, T.G. (2021) Detection of interannual ensemble forecast signals over the north Atlantic and Europe using atmospheric circulation regimes. *Quarterly Journal of the Royal Meteorological Society*, 148, 434–453. Available from: <https://doi.org/10.1002/qj.4213>
- Gilleland, E. (2020) Bootstrap methods for statistical inference. Part II: extreme-value analysis. *Journal of Atmospheric and Oceanic Technology*, 37, 2135–2144.
- Gruss, Ł., Pollert, J., Jr., Pollert, J., Sr., Wiatkowski, M. & Czaban, S. (2020) The application of new distribution in determining extreme hydrologic events such as floods. *Hydrology and Earth System Sciences Discussions*, 2020, 1–31.
- Guerreiro, S.B., Fowler, H.J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S. et al. (2018) Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, 8, 803–807.
- Hersbach, H. (2010) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.
- Hosking, J.R.M. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B: Methodological*, 52, 105–124.
- Hosking, J.R.M. & Wallis, J.R. (1997) *Regional frequency analysis: an approach based on L-moments*. Cambridge, England: Cambridge University Press.
- Hundecha, Y., Arheimer, B., Berg, P., Capell, R., Musuuza, J., Pechlivanidis, I. et al. (2020) Effect of model calibration strategy on climate projections of hydrological indicators at a continental scale. *Climatic Change*, 163, 1287–1306.
- Hundecha, Y., Arheimer, B., Donnelly, C. & Pechlivanidis, I. (2016) A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, 6, 90–111.
- Jenkinson, A.F. (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Royal Meteorological Society*, 81, 158–171.
- Johnson, S., Stockdale, T., Ferranti, L., Balmaseda, M., Molteni, F., Magnusson, L. et al. (2019) Seas5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12, 1087–1117.
- Jolliffe, I.T. (2017) Probability forecasts with observation error: what should be forecast? *Meteorological Applications*, 24, 276–278. Available from: <https://doi.org/10.1002/met.1626>
- Kelder, T., Müller, M., Slater, L.J., Marjoribanks, T.I., Wilby, R.L., Prudhomme, C. et al. (2020) Using unseen trends to detect decadal changes in 100-year precipitation extremes. *NPJ Climate and Atmospheric Science*, 3, 47.
- Kent, C., Pope, E., Dunstone, N., Scaife, A.A., Tian, Z., Clark, R. et al. (2019) Maize drought hazard in the northeast farming region of China: unprecedented events in the current climate. *Journal of Applied Meteorology and Climatology*, 58, 2247–2258.
- Lavers, D., Prudhomme, C. & Hannah, D.M. (2013) European precipitation connections with large-scale mean sea-level pressure (MSLP) fields. *Hydrological Sciences Journal*, 58, 310–327. Available from: <https://doi.org/10.1080/02626667.2012.754545>
- Lavers, D.A., Pappenberger, F. & Zsoter, E. (2014) Extending medium-range predictability of extreme hydrological events in Europe. *Nature Communications*, 5, 5382.
- Girons Lopez, M., Crochemore, L. & Pechlivanidis, I.G. (2021) Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden. *Hydrology and Earth System Sciences*, 25, 1189–1209. <https://doi.org/10.5194/hess-25-1189-2021>
- Musuuza, J.L., Crochemore, L. & Pechlivanidis, I.G. (2023) Evaluation of earth observations and in situ data assimilation for seasonal hydrological forecasting. *Water Research*, 59, e2022WR033655. Available from: <https://doi.org/10.1029/2022WR033655>



- Nascimento, F.F., Bourguignon, M. & Leao, J.S. (2016) Extended generalized extreme value distribution with applications in environmental data. *Hacetatepe Journal of Mathematics and Statistics*, 45, 1847–1864.
- Olsson, J., Södling, J., Berg, P., Wern, L. & Eronn, A. (2019) Short-duration rainfall extremes in Sweden: a regional analysis. *Hydrology Research*, 50, 945–960.
- Overeem, A., Buishand, A. & Holleman, I. (2008) Rainfall depth-duration-frequency curves and their uncertainties. *Journal of Hydrology*, 348, 124–134.
- Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R. et al. (2014) Climate change 2014: synthesis report. In: *Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Geneva: IPCC.
- Pechlivanidis, I.G., Crochemore, L., Rosberg, J. & Bosshard, T. (2020) What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resources Research*, 56, e2019WR026987.
- Poschlod, B., Ludwig, R. & Sillmann, J. (2020) Return levels of sub-daily extreme precipitation over Europe. *Earth System Science Data*, 2020, 1–35.
- Reyer, C.P.O., Leuzinger, S., Rammig, A., Wolf, A., Bartholomeus, R.P., Bonfante, A. et al. (2012) A plant's perspective of extremes: terrestrial plant responses to changing climatic variability. *Global Change Biology*, 19, 75–89.
- Seneviratne, S., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J. et al. (2012) Changes in climate extremes and their impacts on the natural physical environment. In: Field, C.B., Barros, V., Stocker, T.F., Qin, D., Dokken, D., Ebi, K.L. et al. (Eds.) *A special report of working groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*. New York, NY: Cambridge University Press, pp. 109–230.
- Seneviratne, S., Zhang, X., Badi, M.A.W., Dereczynski, C., Luca, A.D., Ghosh, S. et al. (2021) Weather and climate extreme events in a changing climate. In: *Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
- Shukla, S. & Lettenmaier, D.P. (2011) Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, 15, 3529–3538.
- Stott, P. (2016) How climate change affects extreme weather events. *Science*, 352, 1517–1518.
- Sutanto, S.J. & Lanen, H.A.J.V. (2022) Catchment memory explains hydrological drought forecast performance. *Scientific Reports*, 12, 2689.
- Tellman, B., Sullivan, J.A., Kuhn, C., Kettner, A.J., Doyle, C.S., Brakenridge, G.R. et al. (2021) Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, 596, 80–86.
- Thompson, V., Dunstone, N.J., Scaife, A.A., Smith, D.M., Hardiman, S.C., Ren, H.-L. et al. (2019) Risk and dynamics of unprecedented hot months in south East China. *Climate Dynamics*, 52, 2585–2596.
- Thompson, V., Dunstone, N.J., Scaife, A.A., Smith, D.M., Slingo, J.M., Brown, S. et al. (2017) High risk of unprecedented UK rainfall in the current climate. *Nature Communications*, 8, 107.
- van den Brink, H.W., Können, G.P., Opsteegh, J.D., van Oldenborgh, G.J. & Burgers, G. (2004) Improving 104-year surge level estimates using data of the ecmwf seasonal prediction system. *Geophysical Research Letters*, 31(17), L17210. <https://doi.org/10.1029/2004GL020610>
- van den Brink, H.W., Können, G.P., Opsteegh, J.D., van Oldenborgh, G.J. & Burgers, G. (2005) Estimating return periods of extreme events from ecmwf seasonal forecast ensembles. *International Journal of Climatology*, 25, 1345–1354.
- van der Wiel, K., Wanders, N., Selten, F.M. & Bierkens, M.F.P. (2019) Added value of large ensemble simulations for assessing extreme river discharge in a 2°C warmer world. *Geophysical Research Letters*, 46, 2093–2102.
- van Loenhout, J., Below, R. & McClean, D.C. (2020) Human costs of disasters. An overview of the last 20 years, 2000–2019. Technical Report, UNDRR and CRED.
- Wilks, D. (1995) *Statistical methods in the atmospheric sciences: an introduction*. International Geophysics Series. San Diego: Academic Press. 464. [https://books.google.se/books?id=sJ\\_ZCddUW6oC](https://books.google.se/books?id=sJ_ZCddUW6oC)
- Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J. & Wetterhall, F. (2010) Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrology Research*, 41, 211–229.

**How to cite this article:** Klehmet, K., Berg, P., Bozhinova, D., Crochemore, L., Du, Y., Pechlivanidis, I., Photiadou, C., & Yang, W. (2024). Robustness of hydrometeorological extremes in surrogated seasonal forecasts. *International Journal of Climatology*, 44(5), 1725–1738. <https://doi.org/10.1002/joc.8407>