This is an original manuscript of an article published by Wiley in Journal of Geophysical Research: Atmosp

Floods and Heavy Precipitation at the Global Scale: 100-year Analysis and 180-year Reconstruction

B. Renard[1,2,3], D. McInerney[2], S. Westra[2], M. Leonard[2], D. Kavetski[2], M. Thyer[2], J.-P. Vidal[1]

[1]INRAE, RiverLy, Lyon, France

[2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia

[3]INRAE, Aix Marseille Univ, RECOVER, Aix-En-Provence, France

Key Points:

- We perform a joint analysis of station-based flood and heavy precipitation data, at the global scale and over a long 100-year period
- Results highlight wide-ranging increasing trends affecting heavy precipitation, whereas flood trends appear weaker and less consistent
- A 180-year reconstruction of flood and heavy precipitation probabilities is proposed, using atmospheric predictors from the 20CR reanalysis

Abstract

Floods and heavy precipitation have disruptive impacts worldwide, but their historical variability remains only partially understood at the global scale. This article aims at reducing this knowledge gap by jointly analyzing seasonal maxima of streamflow and precipitation at more than 3,000 stations over a 100-year period.

The analysis is based on Hidden Climate Indices (HCIs). Like standard climate indices (e.g. Nino 3.4, NAO), HCIs are used as covariates explaining the temporal variability of data, but unlike them, HCIs are estimated from the data. In this work, a distinction is made between common HCIs, that affect both heavy precipitation and floods, and specific HCIs, that exclusively affect one or the other. Overall, HCIs do not show noticeable autocorrelation, but some are affected by noticeable trends. In particular, strong and wide-ranging trends are identified in precipitation-specific HCIs, while trends affecting flood-specific HCIs are weaker and have more localized effects.

A probabilistic model is then derived to link HCIs and large-scale atmospheric variables (pressure, wind, temperature) and to reconstruct HCIs since 1836 using the 20CRv3 reanalysis. In turn this allows estimating the probability of occurrence of floods and heavy precipitation at the global scale. This 180-year reconstruction highlights flood hot-spots and hot-moments in the distant past, well before the establishment of perennial monitoring networks. The approach presented in this study is generic and paves the way for an improved characterization of historical variability by making a better use of long but highly irregular station datasets.

Plain Language Summary

Floods and heavy precipitation events still hold some mystery despite their disruptive impacts. As an illustration, the latest IPCC report (recently released in 2021) indicates that "the frequency and intensity of heavy precipitation events have increased since the 1950s", but that at the same time "confidence about peak flow trends over past decades on the global scale is low". Why this apparent disconnect between floods and heavy precipitation? Beyond trends, do floods and heavy precipitation vary together at the global scale? How are they related to atmospheric variables such as winds, temperature, atmospheric pressure?

This article describes a 100-year analysis of floods and heavy precipitation data at the global scale. This analysis is made possible by an original probabilistic model adapted to station datasets with highly variable data availability (`https://vimeo.com/802751683`). The analysis first highlights wide-ranging increasing trends affecting heavy precipitation, whereas flood trends appeared weaker and less consistent. It is then used to identify climate configurations associated with the occurrence of floods and heavy precipitation, and to build a 180-year (1836-2015) reconstruction of floods and heavy precipitation probabilities at the global scale. This contributes to a better understanding of the historical variability of hydrologic extremes in the distant past.

1 Introduction

Understanding the historical variability of floods and heavy precipitation in the context of a changing climate is an important endeavor (Sharma et al., 2018). At a global scale, this understanding is hampered by the spatial sparsity of station data and the scarcity of long series spanning more than 50 years. Yet some long series do exist and may be highly informative when analyzed with adapted methods. The first aim of this work is hence to provide a 100-year global analysis of the joint historical variability of floods and heavy precipitation, and to compare the outcome with literature results mostly based on shorter 50-to-60-year analysis periods. The second aim is to infer relations between hydrologic extremes and large-scale climate variables from this long analysis, and to use these relations to estimate probabilities of occurrence of extremes since 1836 at the global scale.

Many studies have analyzed historical changes in floods and heavy precipitation, as summarized in the latest IPCC report (IPCC, 2021, chapters 8 and 11). Focusing on large-scale studies, there is now growing evidence that heavy precipitation has increased over land since the 1950's (e.g. Westra et al., 2012; Papalexiou & Montanari, 2019; Dunn et al., 2020; Q. Sun et al., 2021). This overall increase is consistent with the larger water-holding capacity of a warmer atmosphere, but regional differences indicate that dynamic changes (e.g. change in storms trajectory) may play a role as well. In contrast, flood changes do not show such a consistent signal. Continental-scale studies generally find a mixture of increasing and decreasing trends, with many regions showing no discernible signal at all (e.g. Berghuijs et al., 2017; Hodgkins et al., 2017; Do et al., 2017; Blöschl, Hall, et al., 2019; Gudmundsson et al., 2019; L. Slater et al., 2021). While the discrepancy between the consistent signal found for precipitation and the lack thereof for floods may appear surprising at first sight, it can be explained by the diversity and the complexity of flood-generating mechanisms (Sharma et al., 2018). For instance, Tramblay et al. (2019) showed that antecedent moisture conditions could resolve an apparent contradiction between increasing heavy precipitation and decreasing floods in Mediterranean France. Brunner et al. (2021) also demonstrated the existence of a catchment-specific threshold below which flood changes do not reflect precipitation changes due to the confounding effect of land surface processes. Alternatively, one of the few robust flood signals is the change in flood timing for snowmelt regimes (e.g. Blöschl et al., 2017; Burn & Whitfield, 2017; Dudley et al., 2017), which is temperature-driven rather than precipitation-driven.

Although trends have been the focus of a majority of papers studying the historical variability of floods and heavy precipitation, other forms of temporal variability have also been studied. For instance, the tendency of events to cluster into flood-rich and flood-poor periods has attracted attention (Hall et al., 2014; Blöschl, Bierkens, et al., 2019) and has been highlighted in some regions of Australia (Franks & Kuczera, 2002; Liu & Zhang, 2017) or Europe (Merz et al., 2016; Lun et al., 2020). Such a low frequency variability, also referred to as persistence, may result from the influence of oceanic modes of climate variability such as the Pacific Decadal Oscillation (Wei et al., 2021).

Detecting trends, persistence or any other type of temporal variability using station data faces several methodological challenges, as reviewed by L. J. Slater et al. (2020). The most typical approach used in the literature is to analyze each site separately, and then to look for coherent patterns using, for example, mapping or kriging of at-site results. This is the simplest approach but the limited length of many station series may induce a large sampling uncertainty and hence limits the power to detect trends or the ability to model more complex temporal structures (Bertola et al., 2020). The analysis is also generally restricted to a common period for all sites in order to make at-site results comparable, hence discarding valuable older data.

An alternative approach is to aggregate local series at the level of predefined regions, typically using spatial averaging (e.g. Papalexiou & Montanari, 2019) or by counting events (e.g. Hodgkins et al., 2017; Najibi & Devineni, 2018). The rationale behind this aggregation is to reduce the variability of local series in order to increase statistical power. However this approach still requires working with a short common period to avoid inhomogeneities due to a varying number of aggregated sites. Limitations for detecting a low-frequency signal using short series hence remain. The regions also need to be defined carefully since opposite trends within a region may cancel out.

A third approach is based on spatial modeling. The principle is to use the original at-site series within a probabilistic model that makes explicit assumptions on how trends or other variability components vary across stations (e.g. Renard et al., 2006, 2008; Aryal et al., 2009; X. Sun & Lall, 2015; Bertola et al., 2020, 2021). This reduces estimation uncertainties at the cost of making assumptions that need scrutiny. It is also more complex to implement than the previous approaches because it requires accounting for spatial dependence and missing data, and it typically leads to a high-dimensional inference problem.

Beyond these methodological challenges, analyzing the historical variability of floods and heavy precipitation also faces the difficulty of handling station datasets with highly irregular data availability. It is striking to observe that most contributions to the latest IPCC report use analysis periods starting around 1960 and rarely before 1950 (IPCC, 2021, see also a few examples in Table 1). There exist a few exceptions using $\sim$100-year long periods (e.g. Mediero et al., 2015; Burn & Whitfield, 2018; Q. Sun et al., 2021) but with a drastically reduced number of stations. In other words, most studies restrict themselves to short periods common to many sites or long periods common to a few sites, whereas station datasets often increase in data availability as the measurement network grows (see Figure 1 for an illustration). As discussed in previous paragraphs, this restriction often results from methodological constraints and is hence not unavoidable. For instance, the Hidden Climate Indices (HCI) approach proposed by Renard et al. (2021) accommodates such growing datasets, and leads to estimates related to sparsely represented regions or periods being affected by larger uncertainties.

Another approach to alleviate the limitations of short and irregular datasets is to build reconstructed series, generally by downscaling long reanalyses such as 20CR (Compo et al., 2011). In France for instance, daily precipitation and temperature series have been reconstructed since 1871 (Radanovics et al., 2013; Caillouet et al., 2016; Devers et al., 2020, 2021), and have been transformed into catchment-scale streamflow series by hydrologic modeling (Caillouet et al., 2017; Bonnet et al., 2017; Caillouet et al., 2021). At a larger continental or global scale, a related approach uses the outputs of global hydrologic models (Stahl et al., 2012). However, the existence of large inconsistencies between observed and modeled flood trends (Do et al., 2020) casts doubt on the adequacy of global hydrologic models to represent extremes in small to moderately-sized catchments. The latter generally constitute the majority of catchments monitored in station datasets and may also represent major interests such as operational monitoring, flood warning, reservoir management, agricultural or environmental application.

In an attempt to overcome the limitations identified in the previous paragraphs, this study undertakes a global-scale analysis of the joint historical variability of floods and heavy precipitation, with the following main objectives:

1. Analyze a long 100-year period, and evaluate whether the detected trend and persistence components differ from those identified in the literature.
2. Provide a 180-year reconstruction of probabilities of occurrence at precipitation/streamflow stations, with a global extent.

To achieve these objectives, this study uses a probabilistic model belonging to the recently-developed Hidden Climate Indices framework (Renard et al., 2021). HCIs are used in a similar way to standard climate indices such as Nino 3.4 or NAO (among many others) to explain the temporal variability of data. An important difference, however, is that HCIs are not predefined time series but instead are inferred from the data. They are conceptually similar to the principal components extracted from a space-time dataset using Principal Component Analysis (also known as Empirical Orthogonal Functions analysis, e.g. Hannachi et al., 2007).

A key strength of this HCI-based model is that it allows analyzing floods and heavy precipitation jointly, and distinguishing between: (i) trend and persistence components that affect both floods and heavy precipitation, and (ii) components that are specific to only one of them. The model also handles varying data availability and does not rely on predefined geographical regions. The joint analysis of floods and heavy precipitation over a long period (objective 1) constitute the first innovation, as illustrated by Table 1. The 180-year reconstruction (objective 2) is also innovative, since no similar global-extent reconstructions of extreme probabilities computed at the scale of stations exist as far as our knowledge goes.

The remainder of this paper is organized as follows. Section 2 describes the precipitation, streamflow and atmospheric datasets. Section 3 describes the models used for analyzing floods and heavy precipitation and for reconstructing their probabilities of occurrence from atmospheric variables (pressure, wind and temperature). Results for the 100-year analysis and the 180-year reconstruction are described in Section 4. Section 5 compares the main findings of this analysis with literature results, and discusses limitations and avenues for future work. Finally, the concluding Section 6 summarizes the key insights from this work.

2 Data

2.1 Precipitation

Precipitation data are taken from HadEX2 (Donat et al., 2013) and its successor HadEX3 (Dunn et al., 2020) datasets, which are reference global-scale datasets for detecting changes in temperature and precipitation extremes (see IPCC, 2021, Chapter 11). HadEX datasets exist in two versions. The 'station' dataset contains time series of extreme indices derived from daily station measurements, for instance the time series of monthly maxima of daily precipitation (Rx1day). The 'gridded' dataset is a spatial interpolation of these extreme indices on a regular grid. The 'station' dataset is used in this work to avoid any smoothing effect induced by spatial interpolation and any temporal inhomogeneity induced by the varying number of available stations. Statistical analyses are based on seasonal maxima of daily precipitation, with the four seasons being defined as DJF, MAM, JJA and SON. The time series associated with each season is analyzed separately.

A subset of 1721 stations from HadEX datasets is used (Figure 1). The selection procedure is described in detail in the Supporting Information Text S1, and can be broadly summarized as follows:

1. Remove stations with less than 20 years of data: a higher threshold would result in many stations from Africa and South-East Asia being excluded from the study.
2. Remove stations containing suspicious outliers (see Supporting Information Text S1 for details).

Reference	Var.[a]	Extent	# stations × period	Key findings
Papalexiou & Montanari (2019)	P	Global[b]	8730 × 1964-2013	Overall increase in frequency
Dunn et al. (2020)	P	Global	gridded[c] × 1950-2018	Overall increase, with regional differences
Q. Sun et al. (2021)	P	Global	7293 × 1950-2018 or 1974 × 1900–2018	Significant increases dominate, with regional differences
Hodgkins et al. (2017)	Q	Europe and North America	1204 × 1961-2010 or 322 × 1931–2010	No compelling evidence for increase in major floods
Blöschl, Hall, et al. (2019)	Q	Europe	3738 × 1960-2010	Region-dependent, with both increases and decreases
Gudmundsson et al. (2019)	Q	Global	($\approx$ 7000) × (40-year periods in 1951-2010)	Region-dependent, with both increases and decreases
This article	P & Q	Global	3141 × 1916-2015	See Section 6

Table 1. Properties of large-scale analyses of floods and heavy precipitation for a few selected recent references.

[a] Variable: P for heavy precipitation, Q for floods.

[b] data are available on all continents (Antarctica excluded) but density may strongly vary.

[c] 1.875° × 1.25° longitude-latitude grid.

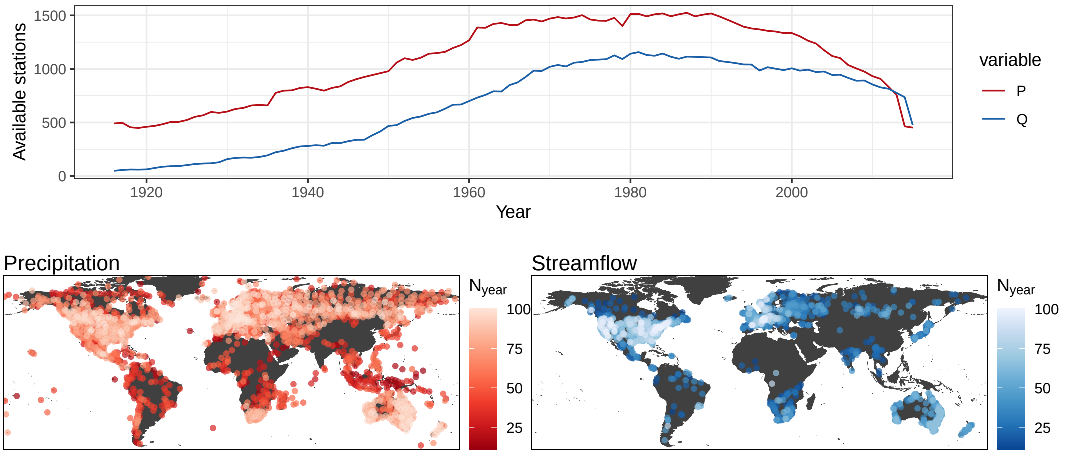

Figure 1. Data availability: evolution of the number of precipitation (P) and streamflow (Q) stations (top) and maps of their location (bottom). The figure shows all selected stations as described in Sections 2.1 (P, 1721 stations) and 2.2 (Q, 1420 stations). Note however that the number of stations effectively used in each of the four seasonal analyses will be smaller due to the season-specific constraint described in Section 3.2.1. Zoomable versions of these maps are available online at https://hydroapps.recover.inrae.fr/HEGS-paper.

3. Remove sets of stations sharing more than 10% of identical non-zero values: these are likely affected by an infilling procedure used in some countries where a single series is used to infill many others.
4. Merge HadEX2 and HadEX3 by favoring the HadEX3 version whenever a station appears in both datasets: this allows preserving large parts of South America, Africa and Southeast Asia that had data in HadEX2 but not in HadEX3.
5. Apply spatial subsampling by selecting the single longest station in a 2×2 degrees box: this reduces large inhomogeneities in the spatial density of stations and makes their number more computationally manageable for the same global coverage.

2.2 Streamflow

Streamflow data are taken from the GSIM dataset (Do et al., 2018; Gudmundsson et al., 2018b), which contains time series of streamflow indices (e.g. monthly mean, min and max) at more than 30,000 stations worldwide. GSIM includes the GRDC dataset, which has been frequently used in large-scale hydrologic analyses (Global Runoff Data Centre, 2015), as well as 11 regional or national datasets. As for precipitation, statistical analyses are based on time series of seasonal (DJF, MAM, JJA, SON) maxima of daily streamflow, with the four seasons being treated separately.

GSIM is probably the most complete streamflow dataset in terms of spatial coverage, but it includes highly regulated catchments that are not suited to the analysis of climate-driven variability. The usual approach to avoid this challenge is to use 'Reference Hydrologic Networks' (RHN, Whitfield et al., 2012; Burn et al., 2012), but RHNs are restricted to a few countries and do not have, to date, a global extent. In order to favor RHN or RHN-like stations while preserving the global extent of the GSIM dataset, the following strategy for selecting stations is implemented:

1. In countries where a known RHN exists, only GSIM stations belonging to the RHN are used. This applies to the European and North-American countries studied in the flood trend analysis of Hodgkins et al. (2017), plus Australia (Bureau of Meteorology, 2020) and Brazil.
2. In countries that do not have a known RHN, stations are selected using GSIM metadata (series length and homogeneity, missing value rate, reliability of catchment delineation, population density, total dam volume and land cover type).
3. For France and Australia, GSIM data are replaced with a more recent version of the RHN datasets: this allows improving space and time coverage and, in the case of Australia, to resolve an issue linked to the handling of quality flags (Gudmundsson et al., 2018a).
4. As for precipitation, spatial subsampling is implemented but with a 0.5 degrees grid box.

This selection procedure results in the subset of 1420 stations shown in Figure 1. The Supporting Information Text S2 provides more details on this procedure, and in particular on the metadata-based criteria used in point 2 to judge the 'RHN-ness' of stations in countries with no formal RHN.

2.3 Atmospheric Variables

In this work, atmospheric variables are used as predictors to reconstruct flood and heavy precipitation probabilities in the distant past. Two long reanalysis products can be used for this purpose: the ERA-20C (Poli et al., 2016) and the 20th Century (20CR, Compo et al., 2011) reanalyses. We opted for the latter in its third version (20CRv3, Slivinski et al., 2019) because it is an ensemble reanalysis, with multiple members representing uncertainty, and it also starts earlier (1836 vs. 1900 for ERA-20C).

Four variables are used in this study: temperature, zonal and meridional wind components at 850 hPa (T850, U850, V850) and mean sea level pressure (PRMSL). For each variable, data are averaged over the season of interest and subsampled on a 2.8125 degree grid (1/4 of the original resolution) to avoid unnecessary storage and computing time issues. The 80 individual members provided by 20CRv3 to represent uncertainty are used rather than the ensemble mean (`https://portal.nersc.gov/archive/home/projects/incite11/www/20C_Reanalysis_version_3/`, accessed January 2022). These variables were chosen because they are frequently used to study large-scale climate variability and derive climate indices. Likewise, seasonal averaging is frequently applied when using climate-informed models for floods or heavy precipitation (e.g. X. Sun et al., 2015; Lee et al., 2018). However, we note that alternative choices could be made on both aspects: this will be further discussed in Section 5.4.

3 Methods

The study methodology uses two probabilistic models to implement three main tasks as summarized in Figure 2. We start by providing a short and intuitive introduction to the HCI modeling framework upon which the two probabilistic models are built, referring to Renard & Thyer (2019) and Renard et al. (2021) for an in-depth description of technical aspects. We then describe the three tasks implemented in this work. The first task analyses the precipitation+streamflow dataset in order to identify a set of HCIs that drive their temporal variability (Model 1). In the second task, the effects of the same HCIs on atmospheric variables are estimated (Model 2). Finally, the third task uses these two models to reconstruct flood and heavy precipitation probabilities from atmospheric data.

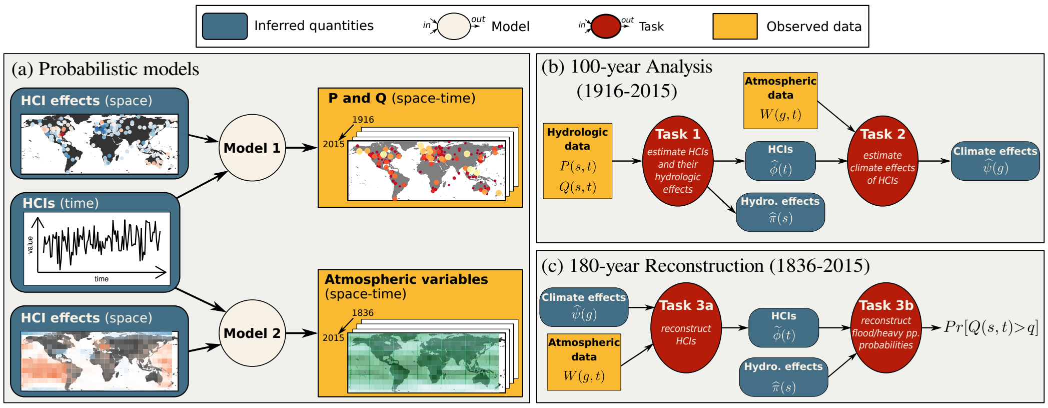

Figure 2. Methodological overview. (a) Two probabilistic models used in this study for describing hydrologic extremes (floods and heavy precipitation) and atmospheric variables (pressure, wind, temperature). Note that the two models share the same Hidden Climate Indices (HCIs) as input. (b) Tasks applied to implement the 100-year analysis; (c) Tasks applied to perform the 180-year reconstruction.

3.1 A Short Introduction to HCI modeling

Consider a space-time dataset such as the one shown in Figure 3a, representing standardized streamflow anomalies at $S = 42$ stations during $T = 45$ years (1970-2014, see Renard & Thyer, 2019). Let $Y(s,t)$ denote the random variable generating the observation at site s and time t. A common way to describe the temporal variability of such data is to use a linear regression to model the influence of a time-varying covariate $\tau(t)$ at each site:

$$Y(s,t) = \lambda(s)\tau(t) + \varepsilon(s,t), \text{ with } \varepsilon(s,t) \sim \mathcal{N}(0, \sigma(s)) \tag{1a}$$

$$\text{or equivalently: } Y(s,t) \sim \mathcal{N}(\lambda(s)\tau(t), \sigma(s)) \tag{1b}$$

A climate index such as Nino 3.4, for instance, is often used as the covariate $\tau(t)$. However, it is also possible to consider that the climate index is *hidden* by treating it as an unknown time series that needs to be inferred from the data. This cannot be achieved at a single site because the number of datapoints (T) is smaller than the number of unknown quantities ($T+2$). However, inference becomes feasible when all sites are considered together, since the number of datapoints ($T \times S$) becomes large compared with the number of unknowns ($T + 2S$).

The Gaussian HCI model of Equation (1) is closely related to Principal Component Analysis (PCA), as shown by Tipping & Bishop (1999). As an illustration, Figure 3b shows the estimated HCI $\hat{\tau}(t)$ (as described in Renard et al., 2021), and compares it with the first component of a standard PCA applied to the same data: the two time series are nearly identical. PCA therefore provides a convenient analogy to interpret the outcomes of an HCI model: the estimated HCI time series $\hat{\tau}(t)$ can be thought of as the principal component driving the temporal variability of the dataset. The associated spatial parameters $\hat{\lambda}(s)$ (Figure 3c, called 'effects' in statistical terminology) are similar to PCA loadings and control the strength of the HCI influence at each site: data from sites where $\hat{\lambda}(s)$ is large closely follow the HCI $\hat{\tau}(t)$ (or its opposite if $\hat{\lambda}(s)$ is negative), while data from sites where $\hat{\lambda}(s) \approx 0$ follow an unrelated pattern.

While the similarity with PCA is convenient for interpretation, we stress that HCI modeling has important advantages over PCA that will be exploited in this work:

1. It is based on an explicit probabilistic model, which provides a natural framework to make probabilistic predictions.
2. Probabilistic assumptions such as the regression formula or the normality assumption in Equation (1) can be modified as needed.
3. The treatment of missing values is straightforward with likelihood and Bayesian estimation methods (Renard et al., 2021) and does not require infilling; this is particularly useful for the datasets shown in Figure 1.
4. Additional probabilistic assumptions can be made to model the time series $\tau(t)$ (e.g. trend, autocorrelation) and the spatial process $\lambda(s)$ (e.g. spatial correlation).

3.2 Step 1: Identifying HCIs from Precipitation and Streamflow Data

3.2.1 Expressing Data as Nonexceedance Probabilities

Raw series of seasonal maxima are expressed in mm (P) or $m^3.s^{-1}$ (Q), and in the case of streamflow they may vary by several orders of magnitude between sites. The usual approach of expressing streamflow in mm cannot be applied because catchment areas are unreliable for a non-negligible fraction of the dataset (see Do et al., 2018, for details

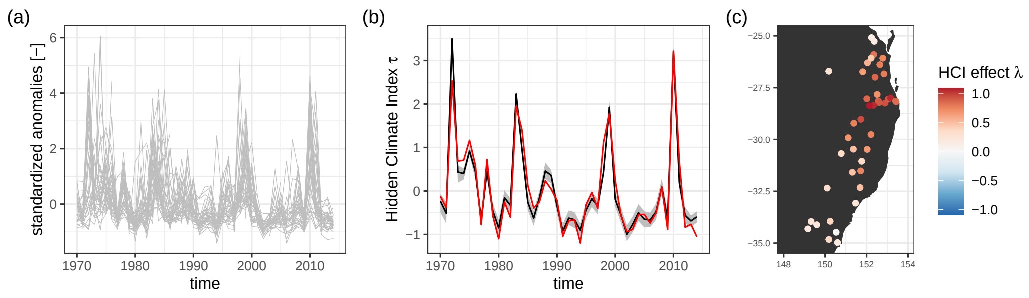

Figure 3. Illustration of a simple Hidden Climate Index (HCI) model and its relationship with Principal Component Analysis (PCA). (a) Standardized streamflow anomalies during the austral spring (SON) at 42 stations in Eastern Australia (one line per station). (b) Estimated HCI $\hat{\tau}(t)$ (black line) and 90% uncertainty interval (gray area). The red line is the standardized first component of a PCA applied to the same data. (c) Effect of the HCI at each site $\hat{\lambda}(s)$.

on this issue). Some form of standardization is therefore desirable to facilitate the derivation of a spatial model. Given the focus on extremes, we decided to consider the return period associated with each seasonal maxima, or equivalently but more conveniently, to transform seasonal maxima into nonexceedance probabilities (Figure 4). This is achieved at each site as follows:

1. Extract the time series of annual maxima.
2. Estimate a Generalized Extreme Value (GEV) distribution using the L-Moment method.
3. Apply the cumulative distribution function (cdf) of this estimated GEV to seasonal maxima.

Note that the GEV is estimated using annual maxima, but is applied to seasonal maxima. Consequently, nonexceedance probabilities will all be close to zero at a station where extremes never occur during the considered season, as illustrated in Figure 4(b). An additional constraint is used to avoid such situations which are not representative of floods or heavy precipitation: at a given station, at least one probability value should exceed 0.8 (i.e. at least one 5-year event should have occurred during the season). If this does not hold, the station is removed from the analysis for this season. This brings the number of stations effectively used in the analyses to 1406 (P) and 818 (Q) for SON, 947

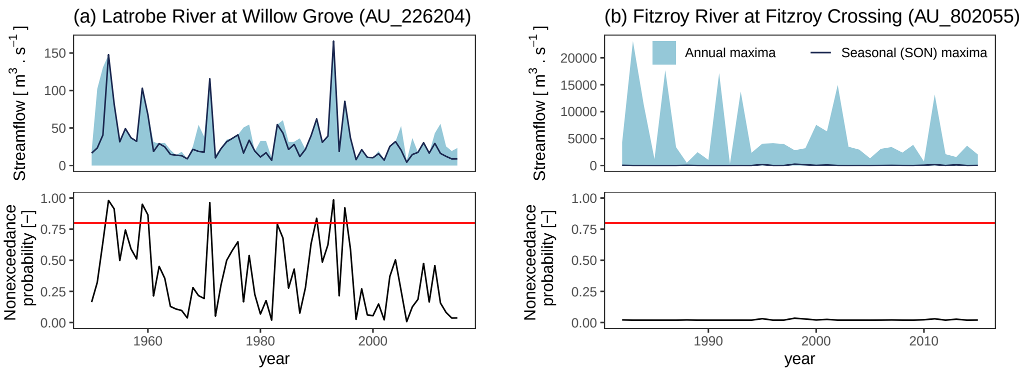

Figure 4. Illustration of the transformation from raw data to nonexceedance probabilities using two Australian streamflow stations. In case (a), the maximum daily streamflow during the SON season (line) often coincides with the annual maximum (shaded area). This indicates that floods often occur during the SON season at this station, leading nonexceedance probabilities to exceed the 0.8 threshold (red line). By contrast, no floods occur during the SON season in case (b), and as a result, all probabilities are well below the 0.8 threshold: this station will therefore be excluded from the analysis for the SON season.

(P) and 834 (Q) for DJF, 1219 (P) and 1179 (Q) for MAM, 1406 (P) and 881 (Q) for JJA.

The use of probability-transformed values does not constitute a limitation in the context of this work. Indeed, the physical values (in mm or $m^3.s^{-1}$) taken by extreme events at stations strongly depend on local factors (e.g. windward / leeward location for P, catchment size for Q), but probability-transformed values are sufficient to study the regional covariability of extremes and its modulation by the large-scale climate. Besides, nonexceedance probabilities can always be transformed back into mm (P) or $m^3.s^{-1}$ (Q) by applying the quantile function of the estimated GEV distribution.

3.2.2 HCI Model

The model described in this section applies to data for one given season, and will be used four times to separately analyze SON, DJF, MAM and JJA. Let $P(s,t)$ and $Q(s,t)$ denote precipitation and streamflow data at site s and time t, expressed as nonexceedance probabilities as described in the previous section. A natural distribution for such data belonging to the interval (0;1) is the Beta distribution $Beta\,(a, b)$, where a and b are two shape parameters. In this work, a reparameterized version $Beta\,(\mu, \nu)$ is favored, where $\mu \in (0; 1)$ is the mean and $\nu > 0$ is a concentration parameter (the larger ν, the smaller the variance). This reparameterized version makes the model more convenient to build and use since mean/concentration parameters are easier to interpret than shape parameters. The formulas to move between parameterizations are the following:

$$\begin{cases} \mu = a/(a+b) \\ \nu = a+b \end{cases} \iff \begin{cases} a = \mu\nu \\ b = (1-\mu)\nu \end{cases} \tag{2}$$

Precipitation and streamflow data are then assumed to be realizations from Beta distributions whose parameters vary in space and time as follows:

$$\text{Distributions of } P \text{ and } Q\text{:} \quad \begin{cases} P\,(s,t) \sim Beta\,(\mu_P\,(s,t)\,, \nu_P\,(s,t)) \\ Q\,(s,t) \sim Beta\,(\mu_Q\,(s,t)\,, \nu_Q\,(s,t)) \end{cases} \tag{3a}$$

$$\text{Space-time model for } \mu\text{:} \quad \begin{cases} logit\,(\mu_P\,(s,t)) = \zeta_{\mu_P}(s) + \sum_{k=1}^{K} \lambda_{k,P}(s)\tau_k(t) + \sum_{k=1}^{K} \theta_{k,P}(s)\delta_k(t) \\ logit\,(\mu_Q\,(s,t)) = \zeta_{\mu_Q}(s) + \sum_{k=1}^{K} \lambda_{k,Q}(s)\tau_k(t) + \sum_{k=1}^{K} \theta_{k,Q}(s)\omega_k(t) \end{cases} \tag{3b}$$

$$\text{Space model for } \nu\text{:} \quad \begin{cases} log\,(\nu_P\,(s,t)) = \zeta_{\nu_P}(s) \\ log\,(\nu_Q\,(s,t)) = \zeta_{\nu_Q}(s) \end{cases} \tag{3c}$$

Equation (3) can be interpreted as a generalization of the simple HCI model of Equation (1b), using a different distribution (Beta rather than Gaussian) and more complex regression formulas. Equation (3b) describes how the mean of precipitation and streamflow data varies in space and time and is at the core of the model. The *logit* transformation is used to ensure that the mean remains in the interval (0;1). For each variable, the first term ($\zeta_{\mu_P}(s)$ or $\zeta_{\mu_Q}(s)$) is a site-specific constant (a.k.a. intercept). The second term models time variability by means of a set of K HCI time series $\tau_k(t)$. The effect of these HCIs at each site is controlled by a set of K spatial processes ($\lambda_{k,P}(s)$ or $\lambda_{k,Q}(s)$). Importantly, the same time series $\tau_k(t)$ are used for both P and Q variables: the second term of equation (3b) therefore represents the temporal variability common to P and Q. By contrast, the third term models time variability in a similar way but uses distinct time series $\delta_k(t)$ and $\omega_k(t)$ for P and Q, respectively. This third term therefore represents the temporal variability specific to P or Q. Finally, equation (3c) states that the concentration parameters vary in space but not in time, with the *log* transformation ensuring they remain positive.

In addition, it is assumed that any spatial or temporal dependence in precipitation and streamflow, or any cross-dependence between them, is induced by the HCIs and their effects. In statistical terms, this corresponds to making an assumption of conditional independence. We refer to Renard et al. (2021) for a thorough analysis of this assumption and its consequences, but one important point in the context of this work is that conditional independence makes the treatment of missing values straightforward: datasets presenting highly irregular availability, such as those in Figure 1, can hence easily be accommodated.

Model specification is completed with additional assumptions on the time and space variability of HCIs and their effects. Starting with the latter, all spatial processes in equation (3) are assumed to follow Nearest-Neighbor Gaussian Processes (NNGP, Datta et al., 2016a). Using the generic notation $\boldsymbol{\pi} = (\pi(s))_{s=1:S}$ to denote any of the spatial processes in equation (3) (ζ's, λ's or θ's):

$$\begin{cases} \boldsymbol{\pi} \sim NNGP\left(\boldsymbol{m}, \boldsymbol{V}\right) \\ m_i = \alpha, \; \forall i = 1 \ldots S \\ V_{i,j} = \eta_0^2 exp\left(-d_{i,j}/\eta_1\right) \; \forall i,j = 1 \ldots S \end{cases} \tag{4}$$

Equation (4) corresponds to a constant-mean process with intersite covariance decreasing exponentially as a function of intersite distance. The NNGP is essentially a standard Gaussian Process that has been modified to make it computationally tractable with a large number of sites. It does so by avoiding the need to use the whole covariance matrix $\boldsymbol{V}$ (whose inversion/multiplication involves $\mathcal{O}(n^3)$ operations), replacing it by the use of many smaller $m \times m$ matrices representing covariances betwen the m nearest neighbors of each site ($m = 5$ is used in this study). We refer to the papers by Datta et al. (2016a,b) and Banerjee (2017) for technical details.

Similar to the spatial effects, all HCI time series are assumed to follow NNGPs. Using as previously a generic notation $\boldsymbol{\phi} = (\phi(t))_{t=1:T}$ to denote any of the HCI time series in equation (3) (τ's, δ's or ω's):

$$\begin{cases} \boldsymbol{\phi} \sim NNGP\left(\boldsymbol{m}, \boldsymbol{V}\right) \\ m_i = \beta\left(i - \frac{T}{2}\right), \ \forall i = 1 \ldots T \\ V_{i,j} = \gamma_0^2 exp\left(-|i-j|/\gamma_1\right) \ \forall i, j = 1 \ldots T \end{cases} \tag{5}$$

Two parameters are of particular interest in equation (5) and will be specifically monitored in the results: β represents a trend affecting the HCI, while γ_1 controls its autocorrelation (the lag-1 autocorrelation is equal to e^{-1/γ_1}). The latter can be used to detect the existence of low-frequency variability (extreme-rich, extreme-poor periods). It is noted that many alternative models could be used to describe low-frequency variability (Henley et al., 2011), but the simple model of equation (5) is considered fit for the purpose of first detecting its existence.

3.2.3 Inference

The model described in Section 3.2.2 requires estimating the intercepts $\boldsymbol{\zeta}$, the HCIs $\boldsymbol{\tau}$, $\boldsymbol{\delta}$, $\boldsymbol{\omega}$ and their effects $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$ along with the parameters of their hyper-distributions $\boldsymbol{\alpha}$, $\boldsymbol{\eta}_0$, $\boldsymbol{\eta}_1$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$. This is achieved by deriving the posterior distribution of these unknown parameters and exploring it with a Monte Carlo Markov Chain (MCMC) sampler. We refer to Renard & Thyer (2019) and Renard et al. (2021) for a complete technical description. In a nutshell, the key ingredients are:

1. Identifiability constraints that make the estimation of HCIs feasible: each HCI has mean zero and variance one;
2. A stepwise approach: the model is first estimated with a single component ($K = 1$ in equation (3)), yielding estimates for $\tau_1(t)$, $\delta_1(t)$ and $\omega_1(t)$; the second component (K=2) is then estimated conditionally on the first-component estimates, etc.;
3. A customized MCMC algorithm that avoids unnecessary computations.

Prior distributions need to be specified for hyper-parameters. For $\boldsymbol{\eta}_1$ and $\boldsymbol{\gamma}_1$ that control decorrelation distance and time, exponential priors with parameters 1000 km and

10 years, respectively, are used to set their order of magnitude. Flat priors are used for all other hyper-parameters.

MCMC sampling is performed by running 40 chains in parallel, corresponding to 10 chains for each of the 4 seasons. Each chain is run for 30,000 iterations and the first third is discarded as burn-in. Computing time is case-dependent, but as a rough order of magnitude, 36 hours are needed to generate 30,000 MCMC samples (i.e. one chain) on a high-performance computing cluster. This is for one step of the stepwise approach described previously, and it therefore needs to be multiplied by the number of components considered, which is set to $K = 5$ in this study.

3.3 Step 2: Estimating HCI effects on Atmospheric Variables

Estimated HCI time series $\widehat{\boldsymbol{\tau}}_k$, $\widehat{\boldsymbol{\delta}}_k$ and $\widehat{\boldsymbol{\omega}}_k$ are available for all $k = 1 \dots K$ after the completion of Step 1 (Section 3.2). As illustrated in Figure 2b, the next step is to estimate their effects on the atmospheric variables described in Section 2.3 (pressure, U and V wind and temperature). As previously, a generic notation $\widehat{\boldsymbol{\phi}}_k = \left(\widehat{\phi}_k(t)\right)_{t=1:T}$ is used to denote any of these HCI time series. Let $W_v(g,t)$ denote the value taken by the vth atmospheric variable at gridpoint g and time t (belonging to the calibration period used to estimated the HCIs). Each variable is centered and scaled to unit standard deviation, i.e. standardized anomalies are considered. It is assumed that the space-time variability of variables $\boldsymbol{W}$ is influenced by the same HCIs as the one controlling precipitation and streamflow data according to the following model:

$$\begin{aligned} &W_v(g,t) \sim \mathcal{N}\left(\mu_v(g,t), \sigma_v(g)\right) \\ &\text{with } \mu_v(g,t) = \psi_{0,v}(g) + \sum_{k=1}^{K} \psi_{k,v}(g)\widehat{\phi}_k(t) \end{aligned} \tag{6}$$

For a given variable v and a given gridpoint g, this equation is a standard linear regression model, which allows estimating the effects $\boldsymbol{\psi}$ using standard regression formulas. More precisely, let $\boldsymbol{w}$ denote observations of the atmospheric variables for the T calibration time steps, arranged in a matrix with T rows and $G \times V$ columns (this assumes that all V variables are observed on the same spatial grid of size G, but this can easily be generalized). Moreover let the estimated HCIs be arranged in a $T \times (K+1)$ matrix $\boldsymbol{\Upsilon}$ as follows:

$$\boldsymbol{\Upsilon} = \begin{pmatrix} 1 & \widehat{\phi}_1(t_1) & \dots & \widehat{\phi}_K(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \widehat{\phi}_1(t_T) & \dots & \widehat{\phi}_K(t_T) \end{pmatrix} \tag{7}$$

Estimation of the effects $\boldsymbol{\psi}$ can then be performed using the ordinary least square estimator:

$$\Bigg(\underbrace{\widehat{\boldsymbol{\psi}}_0, \overbrace{\widehat{\boldsymbol{\psi}}_1, \dots, \widehat{\boldsymbol{\psi}}_K}^{\widehat{\boldsymbol{\Psi}}}}_{GV \times (K+1)} \Bigg)^{\mathsf{T}} = \underbrace{(\boldsymbol{\Upsilon}^{\mathsf{T}}\boldsymbol{\Upsilon})^{-1}}_{(K+1)\times(K+1)} \times \underbrace{\boldsymbol{\Upsilon}^{\mathsf{T}}}_{(K+1)\times T} \times \underbrace{\boldsymbol{w}}_{T \times GV} \tag{8}$$

Note that the formula in equation (8) applies to observed atmospheric variables $\boldsymbol{w}$. However, as explained in Section 2.3, the 20CRv3 reanalysis provides 80 realizations of atmospheric variables $\boldsymbol{w}$ to represent the uncertainty affecting the reanalysis. This uncertainty can be propagated forward to the effects $\boldsymbol{\psi}$ by simply reapplying equation (8) to each of the 80 realizations.

3.4 Step 3: Reconstructing Flood and Heavy Precipitation Probabilities from Atmospheric Variables

As illustrated in Figure 2c, the objective of this third step is to use the atmospheric variables described in Section 2.3 to reconstruct the HCI time series which, in turn, can be used to estimate flood and heavy precipitation probabilities using the HCI model of Step 1. This is of particular interest to extend the analysis period from 1916-2015 to 1836-2015.

For a particular time step t^*, the task is therefore to estimate the HCIs $\boldsymbol{\phi}^*$ using atmospheric data $\boldsymbol{w}^*$. Equation (6) can once again be used for this purpose, but in an 'inverted' setup: effects $\widehat{\boldsymbol{\psi}}$ are known from equation (8) and HCIs $\boldsymbol{\phi}^*$ are sought, which is the opposite of step 2. Since standard deviations $\widehat{\sigma}_v(g)$ have also been estimated in the previous step, a weighted least square estimator can be used to compute the reconstructed $\widetilde{\boldsymbol{\phi}}^*$:

$$\widetilde{\boldsymbol{\phi}}^* = \left(\widetilde{\phi}_1(t^*), \dots, \widetilde{\phi}_K(t^*)\right)^{\mathsf{T}} = \underbrace{\left(\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Psi}}\right)^{-1}}_{K \times K} \times \underbrace{\widehat{\boldsymbol{\Psi}}^{\mathsf{T}}}_{K \times GV} \times \underbrace{\widehat{\boldsymbol{\Omega}}}_{GV \times GV} \times \underbrace{\left(\boldsymbol{w}^* - \widehat{\boldsymbol{\psi}}_0\right)}_{GV \times 1} \tag{9}$$

where $\widehat{\mathbf{\Omega}}$ is a $GV \times GV$ matrix containing $1/\widehat{\sigma}_v^2(g)$ on its diagonal and $\widehat{\mathbf{\Psi}}$ is defined in equation (8). The reconstructed $\widetilde{\boldsymbol{\phi}}^*$ can then be used in the HCI model of equation (3) to reconstruct the distribution of P and Q and any related quantities (e.g. probability of exceeding some threshold, task 3b in Figure 2c). As previously, this process can be repeatedly applied to the 80 20CRv3 members to propagate the associated uncertainty.

4 Results

This section follows the steps outlined in Figure 2. HCI time series and their spatial effects are first identified from P and Q data and their properties are described. The effects of these HCIs on atmospheric variables are then estimated, and the associated regression model is finally used to reconstruct flood and heavy precipitation distributions since 1836. The latter analysis also includes an assessment of the reliability and sharpness of the probabilistic reconstructions using a cross-validation exercise. Detailed results are shown only for the SON season in the paper. Results for other seasons are available through an online app `https://hydroapps.recover.inrae.fr/HEGS-paper` (see also Section 7) and are only summarized herein.

4.1 Hidden Climate Indices

4.1.1 MCMC convergence

MCMC convergence is assessed with the Gelman-Rubin (GR) criterion (Gelman & Rubin, 1992) and by visualizing MCMC traces (not shown). For most inferred quantities, the GR criterion is well below 1.2 and the MCMC traces show that the chains are mixing well, indicating good convergence. Overall, convergence is much faster for the P-specific HCIs $\delta_k(t)$ than for Q-specific and common HCIs $\omega_k(t)$ and $\tau_k(t)$. Further analysis of the GR values reveals that convergence difficulties mostly pertain to HCI values $\omega_k(t)$ and $\tau_k(t)$ at the beginning of the period, which can be explained by the scarcity of streamflow data prior to 1950 (Figure 1).

4.1.2 HCIs and their effects in SON

Figure 5 shows the estimated HCIs and their effects for the first component (additional components are illustrated in the online app). The P-specific HCI δ_1 shows a slight decreasing trend (the 90% interval for β does not contain zero) but no strong autocorrelation. Its effects are concentrated in central North America and are mostly negative: high values of δ_1 are hence associated with lower-than-usual heavy precipitation in this area. Note that the decreasing trend should be interpreted in relation to the sign of HCI effects: here the combination of a decreasing HCI trend and negative effects translates into increasing heavy precipitation.

The Q-specific HCI ω_1 shows a slight increasing trend and no noticeable autocorrelation. Its effects reveal a dipole structure across the North-Atlantic: high values of ω_1 are associated with higher-than-usual floods in the Eastern US, but lower-than-usual ones in Western Europe. Note that these effects are approximately twice as large (in absolute value) as those estimated for the P-specific HCI (compare color scales in Figure 5). Given the model in equation (3b) and the fact that HCIs are standardized to unit standard deviation, this indicates that the distribution of Q may show larger temporal variations than that of P.

The common $P+Q$ HCI τ_1 shows no strong trend or autocorrelation. It mostly affects Australia, indicating that heavy precipitation and floods are affected by a common temporal signal in this region. This shared variability suggests a close association between heavy precipitation and floods, indicating that typical confounding factors such as antecedent moisture or snowpack play a limited role during the SON season.

Finally, Figure 5 shows that uncertainty intervals around the HCIs are fairly tight, indicating that HCIs can be precisely identified from the data. For Q-specific HCI ω_1 and common $P+Q$ HCI τ_1, intervals are about twice larger at the beginning of the period than at the end, reflecting the strongly decreasing availability of streamflow data.

4.1.3 HCI properties for all seasons

Figure 6 evaluates the existence of trend or autocorrelation in the HCIs for all seasons. Note that it makes sense to compare trend or autocorrelation values across seasons and HCIs because all HCIs have the same standard deviation equal to one (see identifiability constraints in Section 3.2.3). Marked trends are found for the P-specific HCIs. For each of the four seasons, a large trend makes one component stand out. Figure 7 shows for instance the second P-specific HCI in SON and its effects: the upward trend is indeed clearly visible, and moreover the HCI effects are widespread, suggesting that the trend affects many areas of the world. A similar observation can be made for other sea-

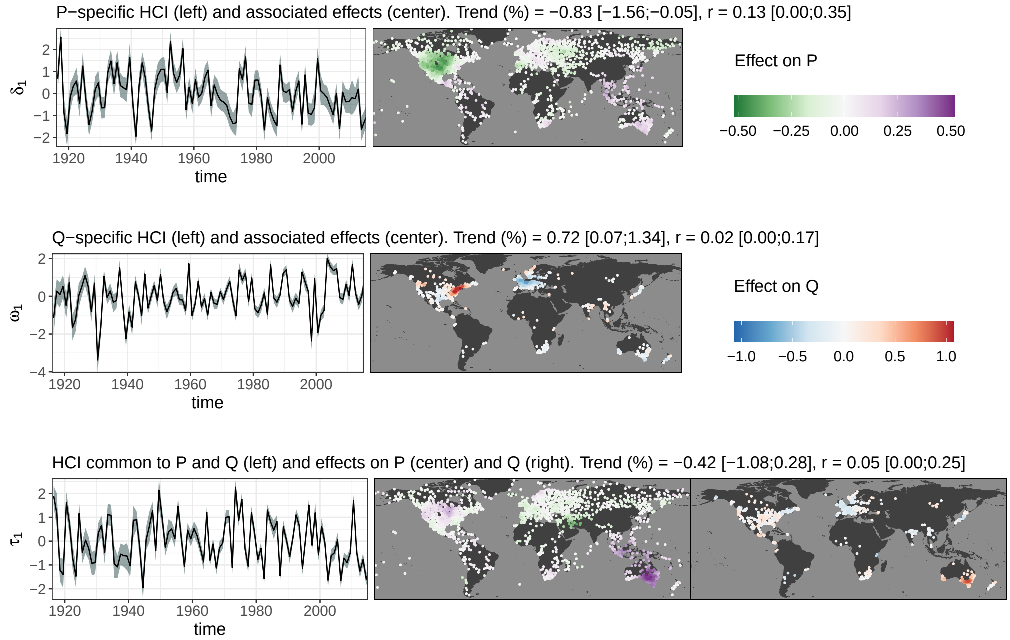

Figure 5. Hidden Climate Indices (HCIs) and their effects for the first component (SON season). Rows correspond to the HCI type (P-specific, Q-specific or common to P and Q). Panels on the left show the HCI time series, with 90% posterior intervals shown in gray. Center and right panels show the associated spatial effects on P and/or Q. For each row, the title gives estimated values and 90% posterior intervals for the trend parameter β and the lag-1 autocorrelation $r = e^{-1/\gamma_1}$.

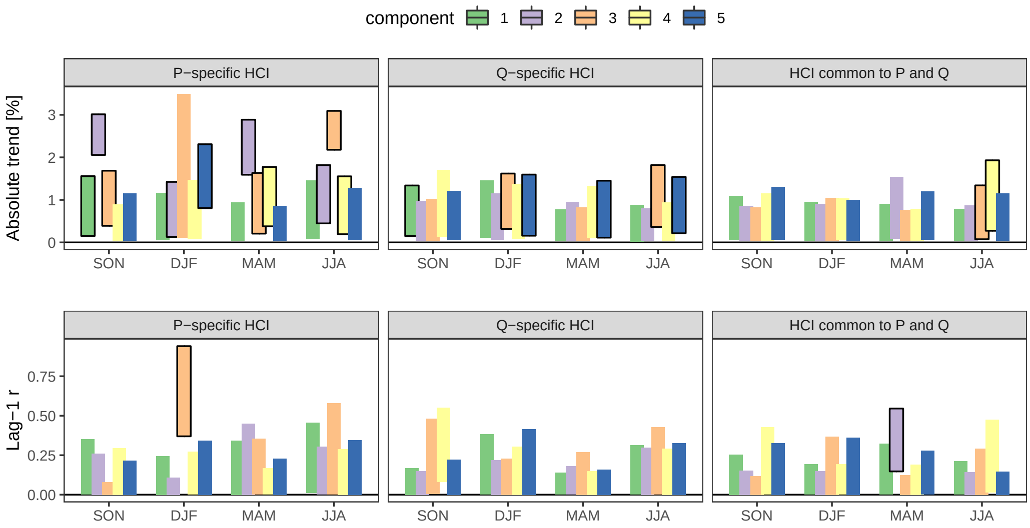

Figure 6. Summary of HCI properties for all 5 components and 4 seasons. Boxes denote 90% posterior intervals for the absolute trend $|\beta|$ (top) and the lag-1 autocorrelation $r = e^{-1/\gamma_1}$ (bottom). Outlined boxes highlight 'large' trends and autocorrelations, and correspond to β-intervals not containing 0 or r-intervals above 0.1.

sons (see online app). A few trends are found for the Q-specific HCIs, but they are much smaller than those affecting heavy precipitation, and the associated effects are also much less widespread (see online app). Finally, trends are barely noticeable for the common $P+Q$ HCIs. Overall, these results are consistent with the literature finding that heavy precipitation shows some sign of global increase over land areas, whereas floods do not show such a consistent signal.

The bottom row of Figure 6 indicates that most HCIs do not show noticeable autocorrelation, suggesting that they represent modes of interannual, rather than low-frequency, variability. The strongest autocorrelation is detected for the third P-specific HCI during DJF, but closer inspection reveals a step-change behavior rather than a low-frequency oscillation (see online app). The second $P+Q$ HCI in MAM also shows some moderate autocorrelation, and it mostly affects the East Coast of Australia (see online app).

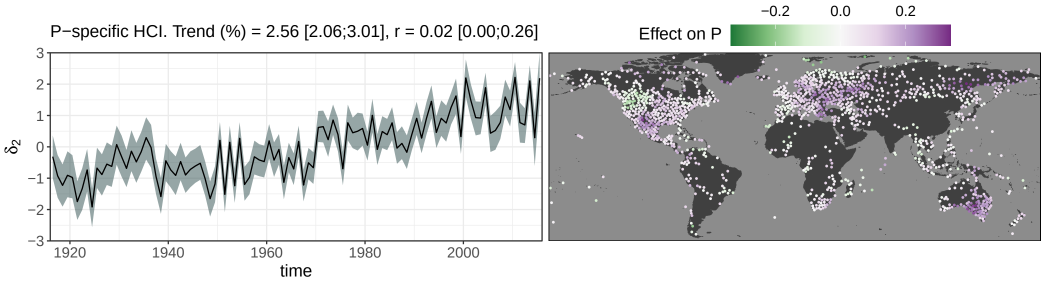

Figure 7. Same as Figure 5 for the P-specific HCI of the second component. This HCI is characterized by a large and wide-ranging increasing trend.

4.2 HCI Effects on Atmospheric Variables

The interest in quantifying the effect of HCIs on atmospheric variables is twofold: first, it can shed light on the origin of the HCIs, and hence on the variability of floods and heavy precipitation, in terms of large-scale circulation; second, it sets up the regression model that will be used in Step 3 for reconstruction. Figure 8 maps the effects of the HCIs described in Figure 5 on the four atmospheric variables (corresponding to $\widehat{\boldsymbol{\psi}}_k$ in equation (8)). These effects are referred to as 'HCI atmospheric effects' in this section, as opposed to the 'HCI hydrologic effects' that were described in Figure 5. HCI atmospheric effects can be compared both in space and between variables since atmospheric variables have been centered and scaled.

Hydrologic effects of P+Q HCI τ_1 are essentially restricted to Australia (see Section 4.1.2 and Figure 5), and the associated atmospheric effects shown in Figure 8 (bottom row) reflect well-known drivers of floods and heavy precipitation in this region. More precisely, strong westerly winds in the equatorial Indian Ocean, negative pressure anomaly in the Eastern Indian Ocean and cold anomaly in the Western Indian Ocean are all typical fingerprints of the negative phase of the Indian Ocean Dipole (IOD). Likewise, the cold anomaly pattern in the equatorial Pacific is typical of La Niña events. This single HCI can therefore be seen as the combination of the two most influential standard climate indices in this area, namely IOD and ENSO.

Atmospheric effects of Q-specific HCI ω_1 (middle row) highlight well-structured patterns of pressure and winds. For atmospheric pressure, the key features are widespread positive effects over the Pacific Ocean, and a dipole over the Eastern US and Western Europe, similar to the one observed for hydrologic effects (Figure 5). The latter can be easily interpreted: high values of ω_1 are associated with positive (resp. negative) pressure anomalies over Western Europe (resp. Eastern US), and hence with less (resp. more) floods. For zonal wind, banded patterns crossing the North Atlantic may be associated with the trajectory of storms reaching Western Europe, with a similar interpretation as above (less westerly winds over Western Europe mean less floods). For meridional wind, fairly localized poles are found in the Tropical Atlantic. Since SON is the hurricane season, these may correspond to wind patterns that favor the landfall of tropical storms and hurricanes in the Eastern US.

Hydrologic effects of P-specific HCI δ_1 are concentrated in the central US (Figure 5). The associated atmospheric effects (top row of Figure 8) are less clearly structured than for other HCIs and are hence more difficult to interpret. Pressure and temperature dipoles are found over Alaska and the western US. The negative anomaly in meridional wind located in the southern US may reflect the influence of moisture transport from the south (less southerly winds means less heavy precipitation in the central US). The atmospheric effects of other HCIs and other seasons are illustrated in the online app.

4.3 Reconstruction

4.3.1 Reconstructing Time-Varying Distributions

Figure 9 shows the HCIs reconstructed from atmospheric variables, as described in Section 3.4. Overall they are in good agreement with the HCIs that were directly estimated from P and Q data over the period 1916-2015 (average correlations: 0.71, 0.68 and 0.77 for δ_1, ω_1 and τ_1, respectively). The added value of the reconstruction is that it extends back to 1836, at the cost however of an increased uncertainty: the dispersion of the 80 members of 20CRv3 is 3 to 4 times larger at the beginning of the period than at the end.

The reconstructed HCIs can then be used in the model of equation (3) to derive the time-varying distributions of P and Q over the period 1836-2015 and at all sites. Figure 10 illustrates these distributions for two sites in Australia, while the corresponding reconstructions for all sites and all seasons are released as an open dataset (see Section 7). In any given year, the variance of the distribution represents the uncertainty in the re-

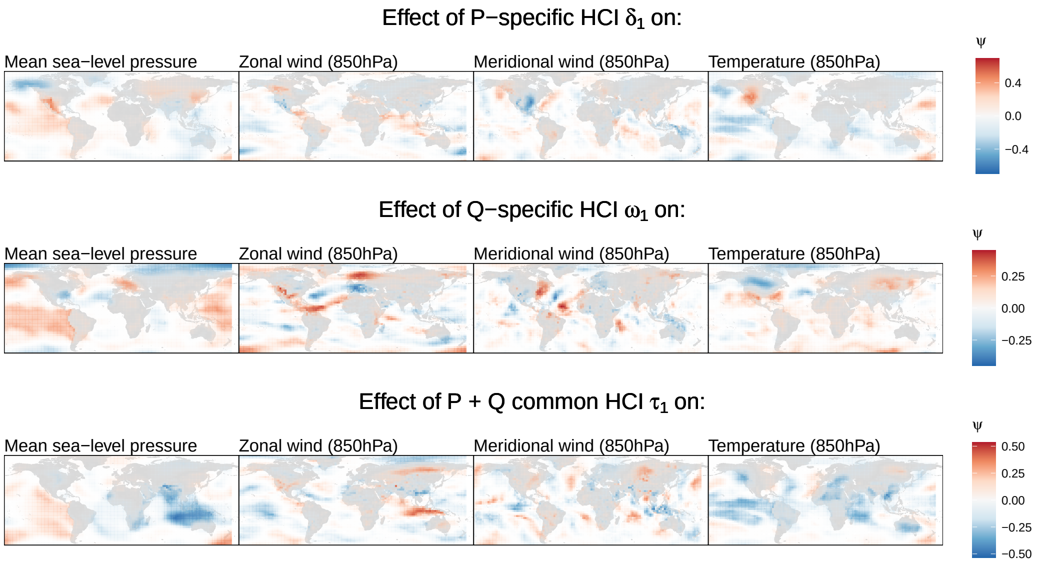

Figure 8. Effects of Hidden Climate Indices (HCIs) on atmospheric variables for the first component (SON season). As in Figure 5, rows correspond to the HCI type (P-specific, Q-specific or common to P and Q), columns to the atmospheric variable the effect applies to. Effects in this figure are computed with equation (8) using atmospheric data provided by the first member of the 20CRv3 reanalysis.

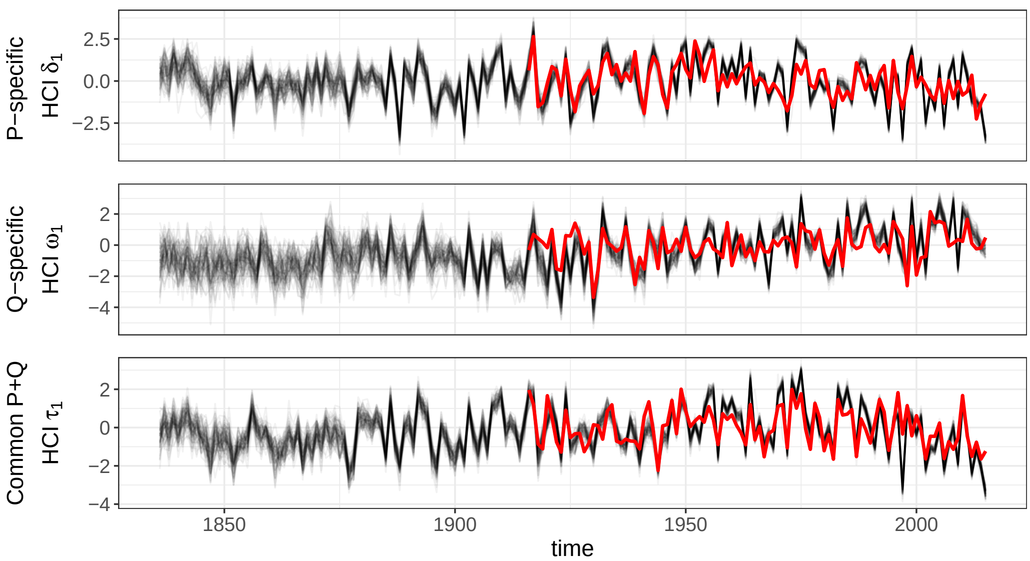

Figure 9. Hidden Climate Indices (HCIs) reconstructed from atmospheric data (pressure, wind, temperature) for the first component (SON season). In each panel, the red line corresponds to the HCI estimated from floods and heavy precipitation data (as shown in Figure 5). Each thin black line is a reconstruction based on one member of the 20CRv3 reanalysis, using equation (9).

construction, which is affected by both the uncertainty in reconstructed HCIs as discussed in the previous paragraph, but also by the uncertainty in the estimation of all spatial terms in equation (3). For the precipitation site, the time-varying distribution has a large variance, resulting in a 90% probability interval that covers an important part of the (0;1) y-axis interval (average width: 0.74). In contrast, the streamflow time-varying distribution is less uncertain (average width: 0.57), which allows highlighting years with well above-average flood probabilities: 1975 or 1992, during which major floods indeed occurred, but also 1910 or 1916, before the availability of any streamflow data at this site, or even anywhere in Australia.

The time-varying distributions can be further appraised by evaluating reliability and sharpness. Reliability is based on the Probability Integral Transform (PIT) diagram (Laio & Tamea, 2007) which evaluates the consistency between the time-varying distributions (with cdf $F_t(x)$) and the observations o_t through the uniformity of PIT values $F_t(o_t)$. A reliability index can be computed using the area between the PIT curve shown in Figure 10 and the diagonal (Renard et al., 2010). The sharpness index is proportional to the interannual variance of the distribution's mean (Renard et al., 2021). Both indices are scaled between 0 (poor) and 1 (good).

For the precipitation site, the time-varying distribution is very reliable but not very sharp (Figure 10): it does not strongly vary between years. In contrast, the streamflow time-varying distribution is slightly less reliable but much sharper. This is a consequence of HCI effects tending to be larger for Q-specific HCIs than for P-specific ones, as discussed in Section 4.1.2. Reliability and sharpness indices at all sites are reported in Figure 11. Overall reliability indices are similar for both variables. The lack of marked spatial structures suggests that the reliability of reconstructions is similar across regions. By contrast, sharpness varies much more both spatially and between variables. Overall sharpness is markedly lower for P than for Q. Southeast Australia is the area where P-reconstructions are the sharpest, probably due to the strong influence of large-scale modes of climate variability that can be predicted from atmospheric variables. Sharpness strongly varies in space for Q-reconstructions: for instance it is much higher in Australia than in Japan, and this cannot be blamed on data availability since the station density is similar in both cases. Also note that the properties of reconstructions may also vary across seasons (not shown): for instance, during MAM and JJA, reconstructed distributions of streamflow have high reliability and sharpness in large parts of the West-

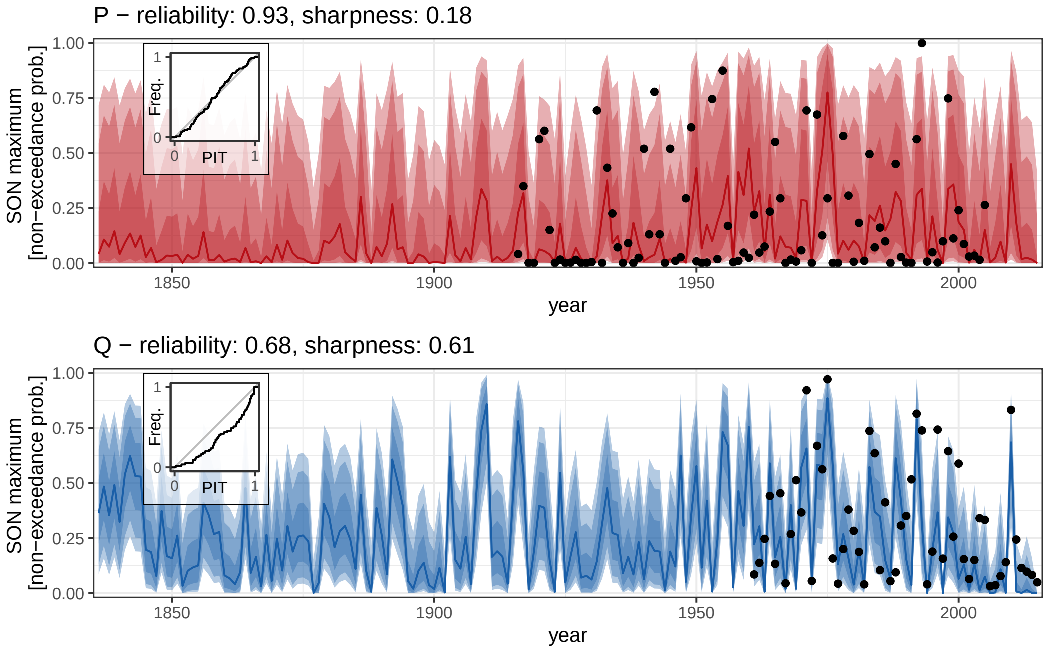

Figure 10. Time-varying distributions derived from reconstructed Hidden Climate Indices for one precipitation (top) and one streamflow (bottom) site, both located in Northern Victoria, Australia (SON season). The solid line denotes the median, stacked colored bands represent 50, 80 and 90% probability intervals, dots represent observed values. The title gives reliability and sharpness indices, ranging between 0 (poor) and 1 (good). The subplot panel shows the PIT diagram used to evaluate reliability (see Section 4.3.1 for details).

ern US, probably linked to snowmelt-induced flows. The sharpness of precipitation reconstructions also appears to be higher in DJF.

4.3.2 Cross-Validation

A cross-validation experiment is used to complement the previous assessment of reliability and sharpness in a predictive context. The estimation sample comprises even-numbered years and is used to estimate HCI atmospheric effects (regression model used in Section 4.2). The validation sample comprises odd-numbered years and is used to compare observed values with reconstructed time-varying distributions. Figure 12 summarizes the results for both heavy precipitation and floods, with reliability and sharpness indices computed on the validation sample only, or on the full dataset as in Section 4.3.1.

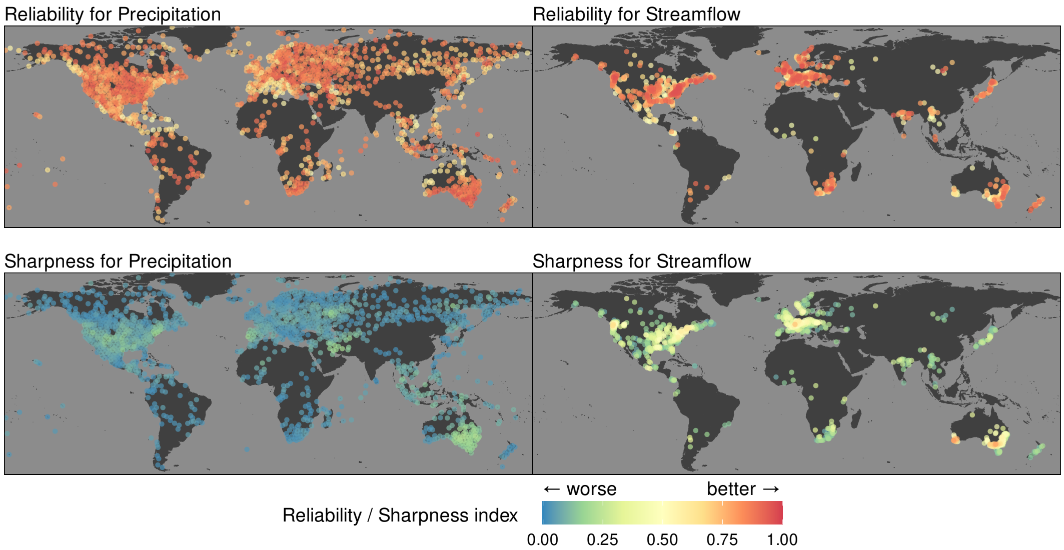

Figure 11. Reliability and sharpness indices associated with the reconstructed time-varying distributions (as shown in Figure 10 for two sites), for all precipitation and streamflow sites (SON season).

PIT diagrams in Figure 12a indicate a good overall reliability for both P and Q and confirm that reliability remains good in validation. Figure 12b breaks down this assessment at the station scale by showing the distribution of reliability indices. Reliability is again acceptable for both P and Q (although slightly better for the former) and there is no marked reliability loss with the validation sample. It is also of interest to compute the reliability index for each year rather than for each station in order to assess whether the reconstruction quality remains stable in time. Figure 12d suggests that this is indeed the case: reliability is stable and high (mostly above 0.9) after 1960 in all cases. It is more variable before 1960 for variable Q, but this may be attributed to sampling variability: streamflow data are indeed scarce before 1960 (see Figure 1), so that reliability indices are computed on a small number of stations for earlier years. Finally, Figure 12c shows the distribution of sharpness indices across stations. It confirms that Q reconstructions are much sharper than P ones, and it also suggests a noticeable loss of sharpness for the validation sample.

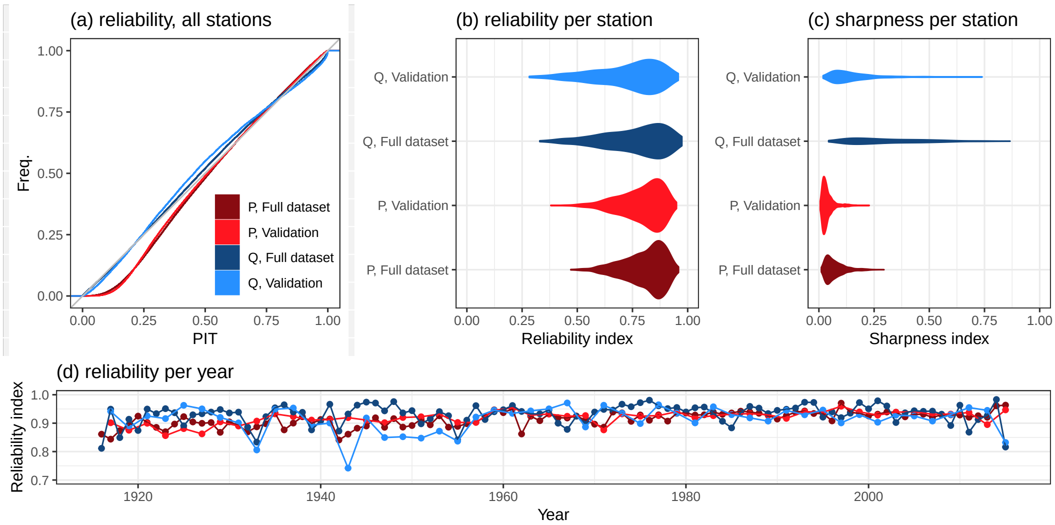

Figure 12. Assessment of the reliability and sharpness of reconstructed time-varying distributions in a cross-validation exercise (SON season). (a) PIT diagrams for all stations; (b) distribution of reliability indices computed by station; (c) distribution of sharpness indices computed by station; (d) time series of reliability indices computed by year.

4.3.3 Reconstructing Probability Maps

A possible way to use the time-varying distributions of Section 4.3.1 is to compute the probability of exceeding the T-year quantile at each site and in any given year. Using Figure 10 as an illustration, this corresponds to the probability of exceeding the value $1-1/T$ according to the time-varying distributions. These probabilities are released as an open dataset (see Section 7) for the four seasons and for return periods $T=$2, 10 and 100 years. The corresponding maps can be browsed through in the online app. Figure 13a shows an example of such a map for the 10-year quantile (i.e. $T = 10$) in SON 1903. At each site, the probability can be compared to $1/T = 0.1$, which is an upper bound for the marginal (i.e. long-term average) probability. It is only an upper bound because the map refers to seasonal rather than annual maxima (the marginal probability would be equal to 0.1 if annual maxima systematically fell in SON). For this particular year, the P map does not highlight strong exceedances of the value 0.1, which is a consequence of the low sharpness of P-reconstructions. At the opposite, the Q map suggests a 'flood hotspot' in the Northeastern US, where the probability of a 10-year flood exceeds 0.4, and to a lesser extent, in Northwestern US, Western Europe and Southern Australia.

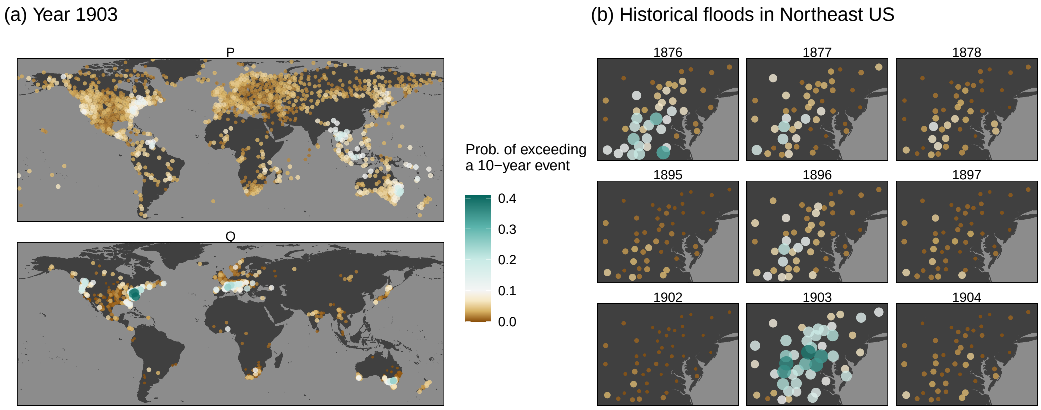

Figure 13. Reconstructed probabilities of exceeding a 10-year event during the SON season. (a) Example of global maps for both heavy precipitation (top) and floods (bottom) during SON 1903. (b) Regional zoom for floods during 9 selected years. Each row shows three consecutive years, with the one in the middle column corresponding to the occurrence of a major historical flood in SON (1877, 1896 and 1903).

Figure 14 provides a synthetic view of these probability maps in SON for the whole period 1836-2015 by sorting the stations according to the AR6 region they belong to (Iturbide et al., 2020). For heavy precipitation, the most prominent feature is the clustering of higher-than-usual occurrence probabilities after ∼1950 in most regions. This indicates that atmospheric conditions have been more favorable to the occurrence of heavy precipitation events in the recent decades, in line with the widespread increase detected in station data (Section 4.1.3). Similar high-probability clusters can also be found during the 19th century in a few regions such as Eastern and Western North America (ENA and WNA). The figure for floods is quite distinct from the precipitation one: it does not highlight any widespread trend but rather region-specific patterns. In particular, high-probability clusters are visible during the mid-19th century in Western and Central Europe (WCE), in the Mediterranean (MED) and in some regions of Asia (EAS and SAS). Conversely atmospheric conditions have been less favorable to the occurrence of floods during the most recent decades in these regions. The opposite pattern is observed in Northern Europe (NEU) and in North America (WNA and ENA), with high-probability clusters appearing in recent decades.

It is also of interest to inspect in more detail specific areas, in particular those showing good reliability and sharpness (Figure 11). Here we focus on a region of the North-

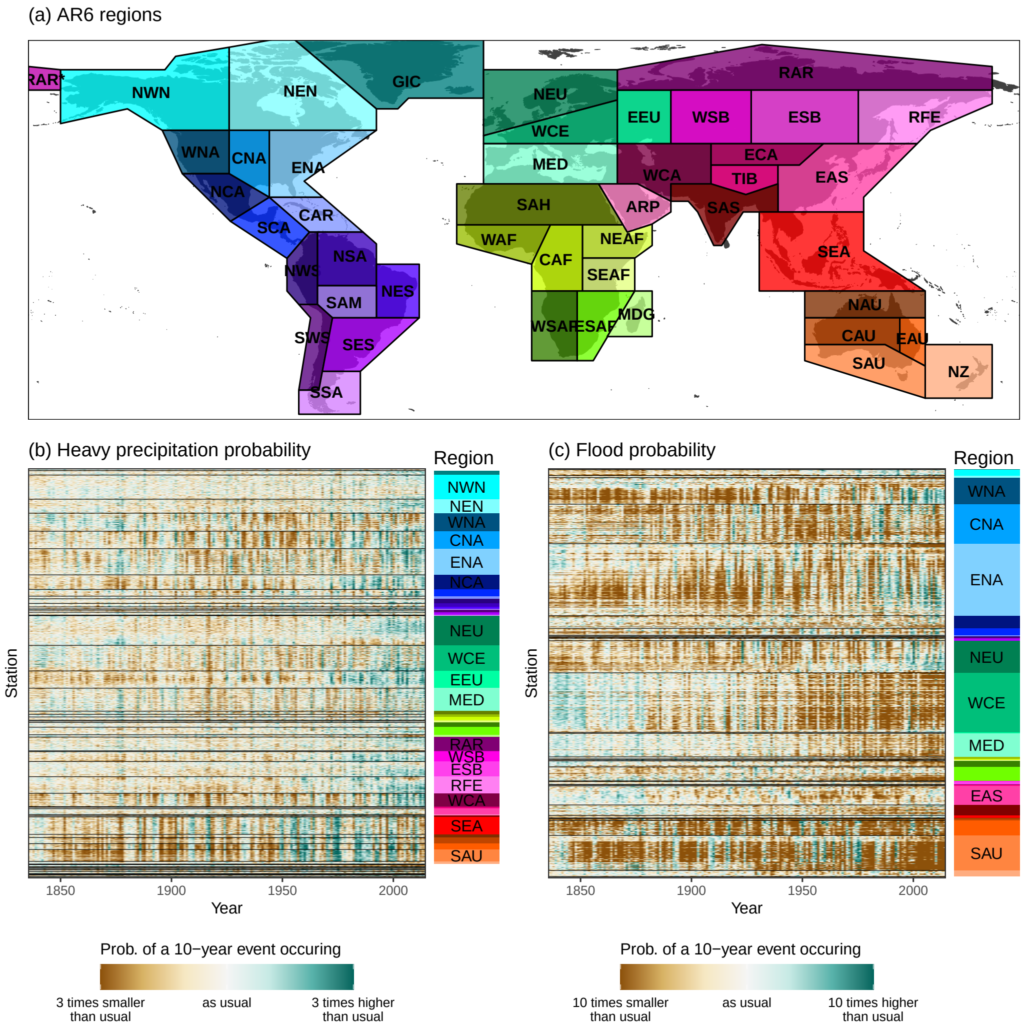

Figure 14. Synthetic illustration of the 180-year reconstruction for the SON season. (a) AR6 regions as defined by Iturbide et al. (2020). (b) Reconstructed probabilities of exceeding a 10-year precipitation during the SON season, for all years (columns) and stations (rows, sorted by AR6 region then by latitude within each region). Colors and acronyms in the right stripe correspond to the AR6 regions shown in panel (a). (c) Same as (b) for probabilities of exceeding a 10-year flood.

eastern US delimited by the Appalachian Mountains to the west, North Carolina to the south and the State of New York to the north (Figure 13b). This region was selected due to the availability of an inventory of historical floods provided by the NOAA-NWS Middle Atlantic River Forecast Center (`https://www.weather.gov/marfc/Flood_Frequency`). According to this inventory, major flooding occurred during the SON season in 1877, 1896 and 1903. Figure 13b shows the associated flood probabilities reconstructed from atmospheric variables (and hence not directly using streamflow information since the P and Q datasets started in 1916). These three particular years are indeed characterized by probabilities above 0.1 (middle column), and up to 4 times above it in 1903. By contrast, the non-flood years before and after 1896 and 1903 show probabilities close to 0. The case of 1877 is different since the previous year also shows high probabilities, but the inventory does not mention any flood in 1876.

5 Discussion

5.1 How do Results from the 100-year Analysis Compare with Literature?

The joint modeling of floods and heavy precipitation and the use of a 100-period make this study stand out from other large-scale analyses in the literature, as illustrated in Table 1. It is therefore of interest to assess whether these specific features yield insights that differ from those of the literature.

Overall, the results in terms of trends (or lack thereof) are remarkably consistent with the literature. The wide-ranging trends found in P-specific HCIs are in agreement with IPCC's statement that heavy precipitation has increased since the mid-20th century (IPCC, 2021, chapter 11): the statement hence also holds since the early 20th century, and it still holds for each of the four seasons (Figure 6). In contrast to heavy precipitation, trends affecting Q-specific HCIs are smaller and have much more localized effects. This is also in line with the lack of globally-consistent flood trend reported in the literature, suggesting that this negative result is not due to the relative short period used in most flood analyses (Table 1). Finally, trends affecting common $P{+}Q$ HCIs are barely noticeable, confirming that floods and heavy precipitation should not be expected to change in the same way (Sharma et al., 2018), unlike annual streamflow and precipitation (McCabe & Wolock, 2011).

It is also of interest to make this comparison at a smaller regional scale, for instance using the AR6 regions shown in Figure 14(a) and used in the recent analyses of Q. Sun et al. (2021, heavy precipitation) and Gudmundsson et al. (2019, floods). To achieve this, the time-varying mean of the Beta distribution ($\mu(s,t)$ in Equation (3)) is computed for each individual station over the whole period 1916-2015. The resulting time series are grouped by AR6 region and the common regional trend is computed for each region. The corresponding figures are shown in the Supporting Information (Figures S1 to S8). For heavy precipitation (Figures S1 to S4), the trends are remarkably consistent with the results described by Q. Sun et al. (2021, in particular their Table 1). These authors reported mostly increasing trends in annual maxima of daily precipitation in several regions of North America (CNA, ENA, NCA), Europe (NEU, EEU) and Asia (WSB, RFE). For all these regions, increasing trends are also discernible over the period 1916-2015 and for most seasons (Figures S1 to S4). Conversely, regions where trends were reported as less consistent (SAU, RAR, NWN) also show no clear increasing trend in our results. The only notable discrepancy is the MED region, for which Q. Sun et al. (2021) reported rather inconsistent trends while our results show a discernible increasing trend, especially in SON which is the most extreme-prone season (Figure S1). For floods, the comparison with the results of Gudmundsson et al. (2019, in particular their Figure 3) is not as conclusive. One of the strongest result reported by these authors was a decrease in streamflow of the MED region, including for annual maxima, but our results highlight no clear trend in the main flood seasons (DJF and SON, Figures S6 and S5). On the other hand, the clear decreasing trend reported by Gudmundsson et al. (2019) for SAU since the 1970's is also visible for 3 seasons in our results (Figures S5 to S7), but not in JJA which is the most extreme-prone season in this region (Figures S8). Several reasons may explain this mostly inconclusive comparison for floods. First, the 100-year time period used here differs from those used in the literature (see Table 1), and many authors reported that flood trends are highly sensitive to the selected period (see e.g. Hodgkins et al., 2017; Gudmundsson et al., 2019). Moreover, we performed four separate seasonal analyses, while other comparable global-scale trend analyses worked at the annual scale, thus complicating direct comparisons. Finally, flood trends are overall quite weak and spatially inconsistent, making them more sensitive to data or methodological differences between studies.

Results in terms of low-frequency variability are only partly consistent with the literature. Indeed, several studies have highlighted significant clustering of flood events in time in some regions of Australia (e.g. Franks & Kuczera, 2002) or Europe (e.g. Lun et al., 2020), which should result in the presence of autocorrelation in HCIs. Some confirming evidence is found in the case of Australia: the second $P{+}Q$ HCI during the MAM season has a noticeable autocorrelation (Figure 6, bottom right panel), and it mostly affects Eastern Australia. However, no noticeable autocorrelation is detected for other HCIs affecting Europe. This failure to detect flood clustering might be partly due to the lack of power of the HCI model used in this study to detect such variability. We stress, however, that the long 100-year analysis period used in this paper is beneficial in terms of detection power. Moreover, the HCI framework is not inherently unable to detect low-frequency variability, as demonstrated by Renard & Thyer (2019) using a synthetic experiment. Finally, we applied the HCI model used in this study to Sea Surface Temperature data (SST, not shown), and the model identified components with a very clear low-frequency signal. Our interpretation is therefore that low frequency variability may exist but it only accounts for a small part of the temporal variability of floods and heavy precipitation, at least when they are considered at the global scale over the last 100 years.

5.2 Originality of the 180-year Reconstruction

A key contribution of this work is the global reconstruction of flood and heavy precipitation probabilities since 1836. This reconstruction allows highlighting periods during which atmospheric pressure, wind and temperature conditions were favorable to the occurrence of extremes in specific regions. The widespread increase in heavy precipitation probabilities is in line with their expected behavior under a warming climate and with the increasing trends revealed by the 100-year analysis. Regarding floods, the high-probability period affecting Western, Central and Southern Europe during the mid-19th century is worth a particular note since it predates the availability of station data and is hence purely identified from atmospheric information. Interestingly, this period is consistent with one of the flood-rich period identified by Blöschl et al. (2020) using historical information. The release of the reconstruction as an open dataset makes it open to further appraisal by means of local historical data or other sources of information.

In addition to its length, the uniqueness of the reconstruction lies in the fact that it reaches a global extent while operating on station data (i.e. streamflow measured at hydrometric stations and precipitation measured at raingauges). As far as we know, similar long and station-based reconstructions have been limited to a national extent so far (e.g. Caillouet et al., 2017; Devers et al., 2020, 2021, in France). Alternatively, global-extent hydrologic reconstructions are generally shorter and operate on relatively large gridcells, which makes them relevant for large catchments only. As an illustration, the reconstruction of Alfieri et al. (2020) (1980-2018) was calibrated on catchments larger than 5,000 km^2, which only represents around 10% of the catchments we used in this work. The 180-year reconstruction therefore fills a gap in the landscape of hydrologic reconstructions. A drawback of this uniqueness is that a detailed quantitative comparison with existing products is difficult.

From a methodological standpoint, this reconstruction also constitutes a proof of concept for a 'bottom-up' approach that starts from hydrologic data observed on operational station networks and attempts to uncover sources of predictability from the larger-scale climate (Figure 2). This approach is generic and could be applied to other surface variables and other spatial or temporal scales. The 'bottom-up' approach is to be compared with the more standard 'top-down' method that transforms climate inputs into streamflow by means of hydrologic modeling (see Prudhomme et al., 2010, for a similar discussion in the context of future projections).

5.3 Improving Reconstructions using Historical Information

In this study historical information is used to identify the dates of remarkable flood events that could be compared against reconstructed flood probabilities. While this is the most straightforward use of this information, it does not fully take advantage of its richness to better understand flood risk (Brázdil et al., 2006). In particular, historical information goes back in time much further than reanalyses. As a few examples, the flood inventory used in Section 4.3.3 goes back to 1687; historical floods of large European rivers such as the Rhône (Pichard et al., 2017) or the Rhine (Wetter et al., 2011) have been documented since around 1300; the European historical dataset collated by Blöschl et al. (2020) goes back 500 years; paleofloods even allow considering millennial time scales (Wilhelm et al., 2022). In addition, regional historical datasets provide information on the spatial structure and extent of large-scale flood events. Finally, historical data may include information on flood intensity, albeit a possibly qualitative one.

A promising research avenue would therefore be to derive reconstructions of flood and heavy precipitation probabilities based on the joint use of three sources of information: station measurements, long atmospheric reanalyses and paleo-historical data. In addition to the availability of large-scale, well-documented and homogeneous datasets, a necessary ingredient to achieve this is a flexible probabilistic model that can be properly adapted to the specificity of such a mixed dataset. This includes the joint use of different types of data (qualitative, quantitative both discrete and continuous), the handling of missing and censored values, the ability to account for the complex space-and-time-varying availability of historical sources, etc. The HCI framework used in this study has been built with such a flexibility as a core objective, and could hence be adapted to perform this analysis. This has the potential to improve both the quality and the temporal extent of long-term reconstructions of floods and heavy precipitation.

5.4 Further Improving Historical Reconstructions

Several promising directions exist to improve the sharpness of probabilistic reconstructions, globally for heavy precipitation and at least in some regions for floods. A first direction would be to consider alternative predictor variables. For instance, atmospheric variables such as vertical temperature gradient or vertical wind shear may be important for extreme-generating phenomena such as hurricanes and medicanes (Cavicchia et al., 2014). Alternatively, surface variables describing antecedent moisture and snowmelt may also be of interest for floods (Blöschl, Hall, et al., 2019).

A second direction would be to avoid the seasonal averaging of atmospheric predictors. Indeed, this averaging is likely to 'smooth out' features that are important for floods in small catchments and for local precipitation. The use of seasonal quantiles rather than averages may be considered. An alternative solution would be to preserve the daily resolution of atmospheric fields and to look for specific dynamic patterns that are associated with floods and heavy precipitation, using for instance a lag-embedding approach (Giannakis & Majda, 2012).

Finally, a third direction to improve historical reconstructions would be to leverage recent progress in Machine Learning (ML), in particular in neural network approaches tailored to large spatiotemporal datasets (e.g. Nielsen et al., 2022). We note that the methods used in this work already share many similarities with ML approaches. For example, the HCI model can be viewed as an extension of probabilistic principal component analysis (Renard et al., 2021). The prediction method described in Section 3.4 is known in ML as the inverse regression approach (see Devijver & Perthame, 2020, and reference therein for details). The idea of using HCI time series as intermediate variables when both predictor and predictand variables are highly dimensional (thousands of gridpoints/sites) is similar to the encoder-decoder approach used in ML (Murphy, 2012).

All these avenues for improvement notwithstanding, we note that there may also be intrinsic predictability limits related to the nature of floods and heavy precipitation: their high variability in both space and time make them much more difficult to predict from large-scale climate than e.g. seasonally-averaged precipitation/streamflow or smoother variables such as temperatures. As an illustration, applying the exact same framework as in this study to SST predictand yielded much sharper reconstructions than those obtained with floods and heavy precipitation (not shown).

5.5 The Importance of Global Station-Based Datasets

The use of large global-scale datasets does not allow performing a thorough analysis of data quality at every site. However, the quality checks and screening procedures implemented by data owners, dataset providers and ourselves provide confidence that data errors, while certainly not absent, are isolated. A more challenging issue is the adequacy of the selected streamflow stations to monitor climate-driven variability. Indeed, HCIs may compensate for omitted time-varying factors affecting the data, including anthropogenic influences (e.g. a catchment moving from natural to regulated). The main safeguard against this issue is our attempt at selecting 'RHN-like' stations in countries with no known RHN (Section 2.2). This procedure is far from infallible, so that regulated catchments likely made it into the analyzed dataset. However, we are confident that they did not strongly affect the results for two reasons. First, the majority of stations used in this study (66%) do come from a formal RHN. The second reason is methodological: the spatial model used for HCI effects (Equation (4)) favors the identification of HCIs having a smooth and consistent effect at the regional scale. Isolated stations affected by non-climatic changes are hence unlikely to be picked up by the first few HCIs, unless these changes have a wide-ranging spatial effect (e.g. a change in the measurement process affecting a whole country).

The challenges discussed above apply to any study trying to identify climate-driven trends or variability in hydrologic regimes. Consequently, initiatives aimed at collating global station-based datasets and documenting their properties are of prime importance. As an illustration, the recent ROBIN initiative (`https://www.ceh.ac.uk/our-science/projects/robin`) is an important step toward collating existing RHNs at the global scale. More generally, a perennial approach to collating and managing multi-national streamflow datasets - RHN or not - is needed to avoid recurring difficulties such as homogenizing quality flags, documenting infilling procedures, detecting duplicates, performing regular updates etc. We therefore second the call by Gudmundsson et al. (2018b) for '*the hydrological community [...] to collectively improve the organization of initiatives for coordinated systems that facilitate updating, storage and documentation of existing data, and to lobby for existing closed databases to be made open and accessible*'.

6 Conclusion

Understanding how floods and heavy precipitation may evolve in a changing climate requires characterizing their historical space-time variability as well as their co-variability. The overarching aim of this study was to contribute to this characterization by means of two long and global-scale analyses. The first analysis jointly explores floods and heavy precipitation station data over a 100-year period. The second analysis provides a 180-year reconstruction of flood and heavy precipitation probabilities derived from atmospheric information.

The 100-year analysis highlights wide-ranging increasing trends affecting heavy precipitation, whereas flood trends are weaker, may be upward or downward and affect smaller regions. These results mostly confirm literature findings (e.g. Sharma et al., 2018; IPCC, 2021) and put them on firmer ground by extending the analysis period (100-year vs. the typical 50-to-60-year used in the literature) and jointly analyzing floods and heavy precipitation. Despite its length, the analysis does not detect strong persistence components affecting the data, suggesting that low-frequency variability accounts for a small fraction of the temporal variability of floods and heavy precipitation.

The second analysis provides a 180-year, global-scale reconstruction of flood and heavy precipitation probabilities, based on atmospheric pressure, wind and temperature variables taken from the 20CRv3 reanalysis. This reconstruction was found to be reliable for both floods and heavy precipitation, but sharpness is much higher for the former than for the latter. In general, higher-than-usual precipitation probabilities were found to cluster in the latest decades, reflecting atmospheric conditions favorable to the occurrence of heavy precipitation events, as expected under a warming climate (IPCC, 2021). Flood probabilities patterns did not follow such a general behavior and were found to be much more region- and season-specific. The reconstruction allowed identifying regions with abnormally high flood probabilities in the distant past, for years well before the establishment of perennial station networks. The reconstruction is released as an open dataset, which may enable more in-depth analyses at smaller spatial scales, using local historical datasets or other sources of information.

From a methodological standpoint, the HCI approach used in this study has several decisive advantages for analyzing station-based datasets. It naturally accommodates varying data availability: this avoids restricting the analysis to either a short period common to many stations or a long period for a few stations. The approach also allows analyzing the covariability of several variables measured on distinct networks by assuming that they are under the influence of common HCIs. Finally, it simplifies the derivation of relationships between highly dimensional predictor and predictand variables by using the HCI time series as low-dimensional intermediate variables. The HCI approach is very general and could hence be applied to study the historical variability of other phenomena at a large spatial scale. This includes other aspects of the hydrologic regime such as water resources and droughts, but also other variables characterizing the state of ecosystems in the context of a changing climate.

Station datasets originating from long-term monitoring networks constitute a most valuable asset to understand the historical variability of hydro-climatic variables. The statistical models used to analyze these datasets should be flexible enough to adapt to their peculiarities and make the best possible use of available data. This may improve not only the characterization of natural variability, but also the ability to derive predictive methods for past reconstructions or future projections.

7 Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 835496.

This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide and the HIICS cluster at INRAE.

Support for the Twentieth Century Reanalysis Project dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER) program, by the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Physical Sciences Laboratory.

Robert Dunn, Nicholas Herold, Hong Xuan Do and Lukas Gudmundsson are gratefully acknowledged for their insights on HadEX and GSIM datasets. We also thank the Editor and three anonymous Reviewers for their insightful comments.

Open Research

All data used in this article originate from open datasets, as cited in the text. The following repositories have been created to complement the article:

- The 180-year reconstruction and the station data for streamflow and precipitation seasonal maxima are available in a Zenodo repository (Renard, 2023b, `https://doi.org/10.5281/zenodo.7680097`)
- R scripts used for setting up models, analysing results, preparing figures and the interactive app are available in a Zenodo repository (Renard, 2023a, `https://doi.org/10.5281/zenodo.7680594`)
- MCMC simulations have been performed with the following computing codes:
  - STooDs v0.1.0 (Renard, 2021b, https://github.com/STooDs-tools/STooDs)
  - R interface RSTooDs v0.1.1 (Renard, 2021a, https://github.com/STooDs-tools/RSTooDs)

The interactive app to browse through the results for all seasons and variables is also available online at `https://hydroapps.recover.inrae.fr/HEGS-paper`

References

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., & Salamon, P. (2020). A global streamflow reanalysis for 1980–2018. *Journal of Hydrology X*. doi: 10.1016/j.hydroa.2019.100049

Aryal, S. K., Bates, B. C., Campbell, E. P., Li, Y., Palmer, M. J., & Viney, N. R. (2009). Characterizing and modeling temporal and spatial trends in rainfall extremes. *Journal of Hydrometeorology*, *10*(1). doi: 10.1175/2008JHM1007.1

Banerjee, S. (2017). High-dimensional bayesian geostatistics. *Bayesian Anal.*, *12*(2). doi: 10.1214/17-BA1056R

Berghuijs, W. R., Aalbers, E. E., Larsen, J. R., Trancoso, R., & Woods, R. A. (2017). Recent changes in extreme floods across multiple continents. *Environmental Research Letters*, *12*(11). doi: 10.1088/1748-9326/aa8847

Bertola, M., Viglione, A., Lun, D., Hall, J., & Blöschl, G. (2020). Flood trends in europe: are changes in small and big floods different? *Hydrology and Earth System Sciences*, *24*(4). doi: 10.5194/hess-24-1805-2020

Bertola, M., Viglione, A., Vorogushyn, S., Lun, D., Merz, B., & Blöschl, G. (2021). Do small and large floods have the same drivers of change? a regional attribution analysis in europe. *Hydrology and Earth System Sciences*, *25*(3). doi: 10.5194/hess-25-1347-2021

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., … Zhang, Y. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal*. doi: 10.1080/02626667.2019.1620507

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., … Živković, N. (2017). Changing climate shifts timing of european floods. *Science*, *357*(6351). doi: 10.1126/science.aan2506

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., … Živković, N. (2019). Changing climate both increases and decreases european river floods. *Nature*, *573*(7772). doi: 10.1038/s41586-019-1495-6

Blöschl, G., Kiss, A., Viglione, A., Barriendos, M., Böhm, O., Brázdil, R., … Wetter, O. (2020). Current european flood-rich period exceptional compared with past 500 years. *Nature*, *583*(7817). doi: 10.1038/s41586-020-2478-3

Bonnet, R., Boé, J., Dayon, G., & Martin, E. (2017). Twentieth-century hydrometeorological reconstructions to study the multidecadal variations of the water cycle over france. *Water Resources Research*, *53*(10). doi: 10.1002/2017WR020596

Brunner, M. I., Swain, D. L., Wood, R. R., Willkofer, F., Done, J. M., Gilleland, E., & Ludwig, R. (2021). An extremeness threshold determines the regional response of floods to changes in rainfall extremes. *Communications Earth & Environment*. doi: 10.1038/s43247-021-00248-x

Brázdil, R., Kundzewicz, Z. W., & Benito, G. (2006). Historical hydrology for studying flood risk in europe. *Hydrological Sciences Journal*, *51*, 739–764. doi: 10.1623/hysj.51.5.739

Bureau of Meteorology. (2020). *Hydrologic reference stations - update 2020.* Retrieved from `http://www.bom.gov.au/water/hrs/update_2020.shtml`

Burn, D. H., Hannaford, J., Hodgkins, G. A., Whitfield, P. H., Thorne, R., & Marsh, T. (2012). Reference hydrologic networks II. using reference hydrologic networks to assess climate-driven changes in streamflow. *Hydrological Sciences Journal*. doi: 10.1080/02626667.2012.728705

Burn, D. H., & Whitfield, P. H. (2017). Changes in cold region flood regimes inferred from long-record reference gauging stations. *Water Resources Research*, *53*(4). doi: 10.1002/2016WR020108

Burn, D. H., & Whitfield, P. H. (2018). Changes in flood events inferred from centennial length streamflow data records. *Advances in Water Resources*, *121*. doi: 10.1016/j.advwatres.2018.08.017

Caillouet, L., Vidal, J. P., Sauquet, E., Devers, A., & Graff, B. (2017). Ensemble reconstruction of spatio-temporal extreme low-flow events in france since 1871. *Hydrology and Earth System Sciences*, *21*(6). doi: 10.5194/hess-21-2923-2017

Caillouet, L., Vidal, J.-P., Sauquet, E., Devers, A., Lauvernet, C., Graff, B., & Vannier, O. (2021). Intercomparaison des évènements d'étiage extrême en france depuis 1871. *LHB*, *107*(1). doi: 10.1080/00186368.2021.1914463

Caillouet, L., Vidal, J. P., Sauquet, E., & Graff, B. (2016). Probabilistic precipitation and temperature downscaling of the twentieth century reanalysis over france. *Climate of the Past*, *12*(3). doi: 10.5194/cp-12-635-2016

Cavicchia, L., von Storch, H., & Gualdi, S. (2014). A long-term climatology of medicanes. *Climate Dynamics*, *43*(5). doi: 10.1007/s00382-013-1893-7

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., ... Worley, S. J. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654). doi: 10.1002/qj.776

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016a). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514). doi: 10.1080/01621459.2015.1044091

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016b). On nearest-neighbor gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(5), 162–171. doi: 10.1002/wics.1383

Devers, A., Vidal, J., Lauvernet, C., Graff, B., & Vannier, O. (2020). A framework for high-resolution meteorological surface reanalysis through offline data assimilation in an ensemble of downscaled reconstructions. *Quarterly Journal of the Royal Meteorological Society*, *146*(726), 153–173. doi: 10.1002/qj.3663

Devers, A., Vidal, J.-P., Lauvernet, C., & Vannier, O. (2021). FYRE Climate: a high-resolution reanalysis of daily precipitation and temperature in France from 1871 to 2012. *Climate of the Past*. doi: 10.5194/cp-17-1857-2021

Devijver, E., & Perthame, E. (2020). Prediction regions through inverse regression. *Journal of Machine Learning Research*, *21*(113). Retrieved from `http://jmlr.org/papers/v21/19-535.html`

Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The global streamflow indices and metadata archive (GSIM) – part 1: The production of a daily streamflow archive and metadata. *Earth Syst. Sci. Data*, *10*(2). doi: 10.5194/essd-10-765-2018

Do, H. X., Westra, S., & Leonard, M. (2017). A global-scale investigation of trends in annual maximum streamflow. *Journal of Hydrology*, *552*. doi: https://doi.org/10.1016/j.jhydrol.2017.06.015

Do, H. X., Zhao, F., Westra, S., Leonard, M., Gudmundsson, L., Boulange, J. E. S., ... Wada, Y. (2020). Historical and future changes in global flood magnitude – evidence from a model–observation investigation. *Hydrology and Earth System Sciences*, *24*(3). doi: 10.5194/hess-24-1543-2020

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., ... Kitching, S. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research: Atmospheres*, *118*(5). doi: 10.1002/jgrd.50150

Dudley, R. W., Hodgkins, G. A., McHale, M. R., Kolian, M. J., & Renard, B. (2017). Trends in snowmelt-related streamflow timing in the conterminous united states. *Journal of Hydrology*, *547*. doi: http://dx.doi.org/10.1016/j.jhydrol.2017.01.051

Dunn, R. J. H., Alexander, L. V., Donat, M. G., Zhang, X., Bador, M., Herold, N.,

... Bin Hj Yussof, M. N. (2020). Development of an updated global land in situ-based data set of temperature and precipitation extremes: HadEX3. *Journal of Geophysical Research: Atmospheres*, *125*(16). doi: 10.1029/2019JD032263

Franks, S. W., & Kuczera, G. (2002). Flood frequency analysis: Evidence and implications of secular climate variability, new south wales. *Water Resources Research*, *38*(5). doi: 10.1029/2001WR000232

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4). doi: 10.2307/2246093

Giannakis, D., & Majda, A. J. (2012). Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(7). doi: DOI10.1073/pnas.1118984109

Global Runoff Data Centre. (2015). *Twelfth Meeting of the GRDC Steering Committee: 18-19 June 2015, Koblenz, Germany* (Tech. Rep.). Bundesanstalt für Gewässerkunde, Koblenz. Retrieved 2023-01-20, from `http://doi.bafg.de/BfG/2015/GRDC_Report_46.pdf` doi: 10.5675/GRDC_REPORT_46

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018a). *The global streamflow indices and metadata archive (GSIM) - part 2: Time series indices and homogeneity assessment.* PANGAEA - Data Publisher for Earth & Environmental Science. Retrieved from `https://doi.pangaea.de/10.1594/PANGAEA.887470`

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018b). The global streamflow indices and metadata archive (GSIM) – part 2: Quality control, time-series indices and homogeneity assessment. *Earth Syst. Sci. Data*, *10*(2). doi: 10.5194/essd-10-787-2018

Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., & Seneviratne, S. I. (2019). Observed trends in global indicators of mean and extreme streamflow. *Geophysical Research Letters*, *46*(2). doi: 10.1029/2018GL079725

Hall, J., Arheimer, B., Borga, M., Brázdil, R., Claps, P., Kiss, A., ... Blöschl, G. (2014). Understanding flood regime changes in europe: a state-of-the-art assessment. *Hydrology and Earth System Sciences*, *18*(7). doi: 10.5194/hess-18-2735-2014

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Jour-*

nal of Climatology, *27*(9). doi: 10.1002/joc.1499

Henley, B. J., Thyer, M. A., Kuczera, G., & Franks, S. W. (2011). Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data. *Water Resources Research*. doi: 10.1029/2010WR010034

Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., ... Wilson, D. (2017). Climate-driven variability in the occurrence of major floods across north america and europe. *Journal of Hydrology*, *552*. doi: https://doi.org/10.1016/j.jhydrol.2017.07.027

IPCC. (2021). *Climate change 2021: The physical science basis. contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change.*

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., ... Vera, C. S. (2020). An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*. doi: 10.5194/essd-12-2959-2020

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*(4).

Lee, D., Ward, P., & Block, P. (2018). Attribution of Large-Scale Climate Patterns to Seasonal Peak-Flow and Prospects for Prediction Globally. *Water Resources Research*. doi: 10.1002/2017WR021205

Liu, J., & Zhang, Y. (2017). Multi-temporal clustering of continental floods and associated atmospheric circulations. *Journal of Hydrology*, *555*. doi: 10.1016/j.jhydrol.2017.10.072

Lun, D., Fischer, S., Viglione, A., & Blöschl, G. (2020). Detecting flood-rich and flood-poor periods in annual peak discharges across europe. *Water Resources Research*, *56*(7). doi: 10.1029/2019WR026575

McCabe, G. J., & Wolock, D. M. (2011). Independent effects of temperature and precipitation on modeled runoff in the conterminous United States: EFFECTS OF TEMPERATURE AND PRECIPITATION ON RUNOFF. *Water Resources Research*. doi: 10.1029/2011WR010630

Mediero, L., Kjeldsen, T., Macdonald, N., Kohnova, S., Merz, B., Vorogushyn, S., ... Porarinsson, O. (2015). Identification of coherent flood regions across europe by using the longest streamflow records. *Journal of Hydrology*, *528*. doi:

10.1016/j.jhydrol.2015.06.016

Merz, B., Nguyen, V. D., & Vorogushyn, S. (2016). Temporal clustering of floods in germany: Do flood-rich and flood-poor periods exist? *Journal of Hydrology*, *541*. doi: https://doi.org/10.1016/j.jhydrol.2016.07.041

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

Najibi, N., & Devineni, N. (2018). Recent trends in the frequency and duration of global floods. *Earth System Dynamics*. doi: 10.5194/esd-9-757-2018

Nielsen, A. H., Iosifidis, A., & Karstoft, H. (2022). Forecasting large-scale circulation regimes using deformable convolutional neural networks and global spatiotemporal climate data. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-12167-8

Papalexiou, S. M., & Montanari, A. (2019). Global and regional increase of precipitation extremes under global warming. *Water Resources Research*, *55*(6). doi: 10.1029/2018wr024067

Pichard, G., Arnaud-Fassetta, G., Moron, V., & Roucaute, E. (2017). Hydroclimatology of the lower rhône valley: historical flood reconstruction (AD 1300–2000) based on documentary and instrumental sources. *Hydrological Sciences Journal*, *62*(11). doi: 10.1080/02626667.2017.1349314

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., … Fisher, M. (2016). ERA-20C: An Atmospheric Reanalysis of the Twentieth Century. *Journal of Climate*, *29*(11). doi: doi:10.1175/JCLI-D-15-0556.1

Prudhomme, C., Wilby, R., Crooks, S., Kay, A., & Reynard, N. (2010). Scenario-neutral approach to climate change impact studies: Application to flood risk. *Journal of Hydrology*. doi: 10.1016/j.jhydrol.2010.06.043

Radanovics, S., Vidal, J. P., Sauquet, E., Ben Daoud, A., & Bontron, G. (2013). Optimising predictor domains for spatially coherent precipitation downscaling. *Hydrology and Earth System Sciences*, *17*(10). doi: 10.5194/hess-17-4189-2013

Renard, B. (2021a). *STooDs-tools/RSTooDs: RSTooDs package v0.1.1 [software].* Retrieved from `https://zenodo.org/record/5075760` doi: 10.5281/zenodo.5075760

Renard, B. (2021b). *STooDs-tools/STooDs: STooDs engine v0.1.0 [software].* Retrieved from `https://zenodo.org/record/5075586` doi: 10.5281/ZENODO.5075586

Renard, B. (2023a). *Codes and data related to the article: 'Renard et al. Floods* *and Heavy Precipitation at the Global Scale: 100-year Analysis and 180-year Reconstruction. Journal of Geophysical Research - Atmospheres' (version 1.0.0) [software].* Zenodo. Retrieved from `https://zenodo.org/record/7680594` doi: 10.5281/ZENODO.7680594

Renard, B. (2023b). *Global reconstruction of flood and heavy precipitation probabilities, 1836-2015 (version 1.1.0) [dataset].* Zenodo. Retrieved from `https://zenodo.org/record/7680097` doi: 10.5281/ZENODO.7680097

Renard, B., Garreta, V., & Lang, M. (2006). An application of bayesian analysis and markov chain monte carlo methods to the estimation of a regional trend in annual maxima. *Water Resources Research*, *42*(12). doi: 10.1029/2005WR004591

Renard, B., Kavetski, D., Thyer, M., Kuczera, G., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*.

Renard, B., Lang, M., Bois, P., Dupeyrat, A., Mestre, O., Niel, H., ... Gailhard, J. (2008). Regional methods for trend detection: Assessing field significance and regional consistency. *Water Resources Research*, *44*(8). doi: 10.1029/2007WR006268

Renard, B., & Thyer, M. (2019). Revealing hidden climate indices from the occurrence of hydrologic extremes. *Water Resources Research*. doi: 10.1029/2019WR024951

Renard, B., Thyer, M., McInerney, D., Kavetski, D., Leonard, M., & Westra, S. (2021). A hidden climate indices modeling framework for multi-variable space-time data. *Water Resources Research*. doi: 10.1029/2021WR030007

Sharma, A., Wasko, C., & Lettenmaier, D. P. (2018). If precipitation extremes are increasing, why aren't floods? *Water Resources Research*, *54*(11). doi: 10.1029/2018wr023749

Slater, L., Villarini, G., Archfield, S., Faulkner, D., Lamb, R., Khouakhi, A., & Yin, J. (2021). Global Changes in 20-Year, 50-Year, and 100-Year River Floods. *Geophysical Research Letters*. doi: 10.1029/2020GL091824

Slater, L. J., Anderson, B., Buechel, M., Dadson, S., Han, S., Harrigan, S., ... Wilby, R. L. (2020). Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management. *Hydrology and Earth System Sciences*. doi: 10.5194/hess-2020-576

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., ... Wyszyński, P. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the twentieth century reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724). doi: 10.1002/qj.3598

Stahl, K., Tallaksen, L. M., Hannaford, J., & van Lanen, H. A. J. (2012). Filling the white space on maps of european runoff trends: estimates from a multi-model ensemble. *Hydrology and Earth System Sciences*, *16*(7). doi: 10.5194/hess-16-2035-2012

Sun, Q., Zhang, X., Zwiers, F., Westra, S., & Alexander, L. V. (2021). A global, continental, and regional analysis of changes in extreme precipitation. *Journal of Climate*, *34*(1). doi: 10.1175/JCLI-D-19-0892.1

Sun, X., & Lall, U. (2015). Spatially coherent trends of annual maximum daily precipitation in the united states. *Geophysical Research Letters*, *42*(22). doi: doi:10.1002/2015GL066483

Sun, X., Renard, B., Thyer, M., Westra, S., & Lang, M. (2015). A global analysis of the asymmetric effect of ENSO on extreme precipitation. *Journal of Hydrology*. doi: http://dx.doi.org/10.1016/j.jhydrol.2015.09.016

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. doi: 10.1111/1467-9868.00196

Tramblay, Y., Mimeau, L., Neppel, L., Vinet, F., & Sauquet, E. (2019). Detection and attribution of flood trends in mediterranean basins. *Hydrology and Earth System Sciences*, *23*(11). doi: 10.5194/hess-23-4419-2019

Wei, W., Yan, Z., & Li, Z. (2021). Influence of pacific decadal oscillation on global precipitation extremes. *Environmental Research Letters*, *16*(4). doi: 10.1088/1748-9326/abed7c

Westra, S., Alexander, L. V., & Zwiers, F. W. (2012). Global increasing trends in annual maximum daily precipitation. *Journal of Climate*. doi: 10.1175/jcli-d-12-00502.1

Wetter, O., Pfister, C., Weingartner, R., Luterbacher, J., Reist, T., & Trösch, J. (2011). The largest floods in the high rhine basin since 1268 assessed from documentary and instrumental evidence. *Hydrological Sciences Journal*, *56*. doi: 10.1080/02626667.2011.583613

Whitfield, P. H., Burn, D. H., Hannaford, J., Higgins, H., Hodgkins, G. A., Marsh, T., & Looser, U. (2012). Reference hydrologic networks i. the status and potential future directions of national reference hydrologic networks for detecting trends. *Hydrological Sciences Journal*. doi: 10.1080/02626667.2012.728706

Wilhelm, B., Rapuc, W., Amann, B., Anselmetti, F. S., Arnaud, F., Blanchet, J., ... Wirth, S. B. (2022). Impact of warmer climate periods on flood hazard in the european alps. *Nature Geoscience*, *15*(2). doi: 10.1038/s41561-021-00878-y