Use of a national flood mark database to estimate flood hazard in the distant past

Benjamin Renard

Use of a National Flood Mark Database to Estimate Flood Hazard in the Distant Past

Benjamin Renard [a, b] *

[a]*INRAE, RiverLy, Lyon, France ;* [b]*INRAE, Aix Marseille University, RECOVER, Aix-En-Provence, France*

benjamin.renard@inrae.fr

A national flood mark database spanning several centuries contains valuable information to characterize ancient flood events. However, the network of flood mark sites is distinct from the network of hydrometric stations, making this information difficult to use. This work describes a probabilistic model jointly describing flood marks at sites and flood peaks at stations. The model is based on the estimation of Hidden Climate Indices driving both flood marks and peaks: this allows transferring information between the two variables despite them being measured on distinct networks. The model is applied to about 300 flood mark sites (1705-2015) and 200 stations (1904-2015) in France. Results demonstrate that flood marks allow estimating the time-varying probability of exceeding some high discharge threshold at stations during the whole period 1705-2015, which largely predates the existence of stations. The resulting probability maps provide quantitative information on the extent and spatial structure of ancient floods.

Keywords: Historical floods; Flood marks; Bayesian modeling; Space-time variability; Hidden Climate Indices; Mixed data

1 Introduction

Understanding the past variability of floods and quantitatively assessing flood hazard is challenged by the short length of hydrometric series. It is therefore common to complement local series with other temporal, spatial and causal sources of information (Merz and Blöschl 2008). In this context, the recent development of a national flood marks database in France offers an opportunity to spatially and temporally expand the information provided by the standard hydrometric network. The aim of this work is to demonstrate how data from both

sources can be used together, and to assess the resulting added value for flood frequency analysis (FFA) and for the characterization of flood variability in the distant past.

The interest of using historical and paleoflood information dating back several centuries for improving flood hazard estimation has long been recognized (e.g. Benito et al. 2004; Brázdil et al. 2006). Historical and paleoflood data allow studying the past variability of floods at a temporal scale that is out of reach for systematic data from hydrometric networks. They have often been used to characterize low-frequency variability (i.e. hydro-climatic oscillations, flood-rich/flood-poor periods) and to reconstruct spatio-temporal patterns of flood occurrences (e.g. Barriendos and Martin-Vide 1998 in Spain; Pichard et al. 2017 in France; Blöschl et al. 2020; Brönnimann et al. 2022 in Europe; Wilhelm et al. 2022 in the European Alps).

Historical data are also frequently used in combination with systematic data from hydrometric stations to improve FFA (Stedinger and Cohn 1986). In this case, historical information is used to reconstruct peak discharges at the target station, or at least to assess whether historical floods exceeded some high threshold. Such reconstructions are affected by large uncertainties that need to be accounted for (Reis and Stedinger 2005; Neppel et al. 2010). When historical and systematic data are available at several stations, this approach can also be applied within a regional FFA procedure (e.g. Jin and Stedinger 1989; Gaume et al. 2010; Nguyen et al. 2014; Sabourin and Renard 2015). Combined historical+systematic records have also been used in time-varying models incorporating climate information (e.g. Machado et al. 2015; Xiong et al. 2020).

Databases collecting historical flood marks provide valuable long-term and spatially-distributed information that may be used in this context. Such a database has been recently developed at the national scale in France. The French National Flood Marks Platform (Piotte

et al. 2016, https://www.reperesdecrues.developpement-durable.gouv.fr) collects qualitative and quantitative information on high-water marks that may be used for a variety of purposes such as hazard assessment, flood mapping, hydraulic modeling or public education. Typical information includes the localization and date of the mark, the description of the site, but also sometimes the mark elevation, photos, and other qualitative comments. Many data correspond to temporary marks such as mud or debris lines that have been collected fairly recently during post-flood field campaigns (Koenig et al. 2016; Galiana et al. 2017). However, the database also includes thousands of geolocalized, perennial flood marks such as plaques or markers carved on walls, sometimes dating back several centuries: these are the focus of this study.

Using these historical flood marks in combination with systematic data from hydrometric stations is not a straightforward task. The first difficulty is that the network of flood mark sites is distinct from the network of hydrometric stations. While some sites could probably be paired with existing stations, this cannot be done systematically since flood marks exist outside of the hydrometric network. Moreover, for sites that could be paired with stations, transforming flood marks into peak discharges is a significant undertaking that requires setting up hydraulic models reflecting the evolution of river topography and flow conditions (Benito et al. 2004): such information is highly site-specific, so that the reconstruction of peak discharges from flood marks cannot be automated on hundreds of sites. The approach taken in this paper to face these issues is to avoid reconstructing peak discharges altogether: instead, the flood marks database is only used to provide the date and localization of flood occurrences. The methodological challenge can then be expressed as follows: how to transfer information between two distinct networks (sites vs. stations) measuring distinct variables (mark occurrences vs. flood peaks)?

In this study, this challenge is addressed by means of a recently-proposed modeling framework for multi-variable space-time data (Renard et al. 2021). Its principle is to uncover

a set of hidden time series, called Hidden Climate Indices (HCI), that drive the temporal variability of all studied variables. The flexibility of the framework allows considering any type of variables (discrete or continuous), measured on distinct networks, and with varying data availability. Moreover, considering common HCIs as drivers of distinct variables provides a mechanism to transfer information between variables.

The work described in this paper therefore aims at deriving a probabilistic HCI model for both flood peaks at hydrometric stations and flood marks at sites, and to use it for flood hazard estimation. More specifically, we aim at assessing the added value brought by flood marks with respect to the following objectives:

1. FFA, i.e. the task of estimating the marginal distribution of flood peaks and the associated T-year events at hydrometric stations.
2. characterizing historical floods in the distant past by estimating the probability that they exceeded some high threshold at hydrometric stations.

The remainder of this paper is organized as follows. Section 2 describes the two main datasets used in this work, namely flood peaks at hydrometric stations and flood marks at sites. Section 3 describes the probabilistic models, their inference and their use to estimate flood probabilities. Section 4 describes the main results of this work and Section 5 discusses them in terms of interpretations, limitations and avenues for future work. The key conclusions are finally summarized in Section 6.

2 Data

2.1 Flood Peaks at Hydrometric Stations, 1904-2015

Data from 207 hydrometric stations forming the French reference hydrological network (Giuntoli et al. 2012) are used (Figure 1a). These stations monitor near-natural catchments

where direct human influence on floods is deemed negligible. Measurement quality and homogeneity is considered suitable for high flow analysis, and time series are at least 40-year long. For more than 90% of stations, catchment size is between 20 and 2,000 km^2.

Annual maxima are extracted at each station from the daily streamflow series (expressed in mm), with the hydrological year starting on the first of September (i.e. hydrological year 1998 is the period from 1st September 1998 to 31st August 1999). The average series length is 50 years, with the majority of series starting during the 1960's, and very few available data before 1950 (Figure 1a). The oldest series starts in 1904.

2.2 Flood Marks at Sites, 1705-2015

Flood marks are taken from the National Flood Marks Platform. For the purpose of this work, we focus on perennial historical flood marks such as plaques or markers carved on walls. We also restrict to marks with a known localization and corresponding to river flooding (as opposed to coastal flooding or urban runoff, for instance). At each site, a series of flood occurrences is derived by making the following hypotheses:

1. The site is considered active from its first to its last recorded flood mark (included).

2. During the activity period, years with at least one recorded flood mark are associated with flood occurrences (1), while all other years are associated with non-occurrences (0).

The first hypothesis above implies that at least two flood marks are necessary for a site to be included in the analysis. However, we increased this requirement to 4 flood marks to facilitate sensitivity analyses that will require removing the first and the last marks (see section 4.6). Moreover, we restrict to the period 1705-2015, as preliminary analyses suggested that including earlier years led to inference difficulties due to the number of active sites becoming too small. The resulting dataset comprises 327 sites recording 1604 flood marks as shown in

Figure 1b. The average duration of the activity period is 96 years and the average number of flood marks is about 5. Note that there is no need to define a perception threshold above which flood marks would be recorded, as done in FFA with historical data (see introduction section): this threshold will rather be treated through an unknown parameter in the probabilistic model for flood marks (see section 3.1.4).

The two hypotheses described above are fairly strong and probably inaccurate to some degree. Regarding the first hypothesis, the consideration of an activity period is necessary because it is clear that not all sites recorded flood events as early as 1705, and some sites stopped recording events at some point in time. While the definition used here is conservative in the sense that it assumes no information before the first mark and after the last, it is still disputable. The second hypothesis corresponds to assuming that there are neither false detection nor missed events during the activity period. While assuming that flood marks correspond to genuine flood events seems reasonable, assuming that the absence of flood mark corresponds to no flood occurrence is more disputable: events may have been missed during the activity period. Sensitivity analyses will be performed in section 4.6 to assess the influence of these hypotheses.

3 Methods

The methodology used here is built upon existing statistical frameworks (Datta et al. 2016; Banerjee 2017; Renard and Thyer 2019; Renard et al. 2021). All models used as part of this study are described in the following sections, but technical or implementation aspects are not exhaustively detailed as they can be found in the aforementioned references or in the Appendix sections A1 and A2.

3.1 Models

Four models of increasing complexity are used and compared in this work. The first three models focus on flood peaks at hydrometric stations over the period 1904-2015, while the fourth model jointly describes flood peaks at stations and flood marks at sites.

3.1.1 M1: Local GEV Model for Peaks

Let $Q(s,t)$ denote annual maxima at station s and time t and $GEV(\theta_1, \theta_2, \theta_3)$ denote the Generalized Extreme Value distribution with parameters θ_1 (location), θ_2 (scale) and θ_3 (shape). The first model considered in this paper is a purely local GEV model with no prior information:

$$Q(s,t) \sim GEV\left(e^{\mu(s)}, e^{\mu(s)} \times e^{\gamma(s)}, \xi(s)\right) \quad (1)$$

An exponential function is used to ensure that the location parameter $\theta_1 = e^{\mu(s)}$ is positive. The scale parameter is not estimated directly: instead, it is computed as the product of the location parameter and a coefficient of variation (CV) $e^{\gamma(s)}$. The motivation behind this reparameterization is that the CV parameter is expected to vary more smoothly in space than the scale parameter (Prosdocimi and Kjeldsen 2021), which may be beneficial with the spatial models introduced next.

3.1.2 M2: Spatial GEV Model for Peaks

The second model is a spatialized version of the first one: the location, CV and shape parameters are assumed to be realizations from a spatial process. This allows reflecting the spatial consistency in these parameters that the data may suggest, i.e. the fact that nearby sites may tend to have similar parameter values. Spatial Gaussian processes are typically used for this purpose (e.g. Diggle et al. 1998; Cooley et al. 2007). In this work we used a particular

class of processes called Nearest-Neighbors Gaussian Processes (NNGP, Datta et al. 2016; Banerjee 2017). A NNGP can essentially be interpreted as a standard spatial Gaussian process modified for computational efficiency. Formally, model M2 can be written as follows:

$$\begin{cases} Q(s,t) \sim GEV\left(e^{\mu(s)}, e^{\mu(s)} \times e^{\gamma(s)}, \xi(s)\right) \\ \boldsymbol{\mu} \sim NNGP(\mathbf{m}_\mu, \mathbf{V}_\mu) \\ \boldsymbol{\gamma} \sim NNGP(\mathbf{m}_\gamma, \mathbf{V}_\gamma) \\ \boldsymbol{\xi} \sim NNGP(\mathbf{m}_\xi, \mathbf{V}_\xi) \end{cases} \quad (2)$$

where $\boldsymbol{\mu} = \left(\mu(s)\right)_{s=1\ldots S}$ with S the number of stations, and similarly for other parameters. For each parameter, the NNGP is parameterized by a mean vector $\mathbf{m} = (m_1, \ldots, m_S)$ and a covariance matrix $\mathbf{V} = \left(V_{i,j}\right)_{i,j=1\ldots S}$ defined as follows:

$$\begin{cases} m_i = \alpha, \ \forall i = 1 \ldots S \\ V_{i,j} = \eta_0^2 exp\left(-d_{i,j}/\eta_1\right) \ \forall i,j = 1 \ldots S \end{cases} \quad (3)$$

where $d_{i,j}$ is the distance between stations i and j. Hyperparameters α (constant-mean), η_0 (sill) and η_1 (range) are unknown and need to be inferred. The range parameter η_1 controls spatial smoothness: large range values correspond to smooth spatial fields, while spatial independence can be obtained as a limiting case by letting the range tend to zero. The prior distributions used for the hyperparameters are given in Table 1.

The use of NNGPs is motivated by computational aspects: NNGPs avoid computations using the full covariance matrix $\mathbf{V}$ and replace them by computations using many smaller sub-matrices corresponding to a small number of neighboring stations. NNGP's thus remain practical when a large number of stations are used. We refer to the aforementioned references and to section A1 for technical details.

3.1.3 M3: HCI Model for Peaks

The third model builds on the previous spatialized GEV model, but now assumes that the location parameter is also varying in time:

$$\begin{cases} Q(s,t) \sim GEV\left(e^{\mu_0(s)} \times \left(1 + \sum_{k=1}^{K} \mu_k(s)\tau_k(t)\right), e^{\mu_0(s)} \times e^{\gamma(s)}, \xi(s)\right) \\ \boldsymbol{\mu}_0 \sim NNGP\left(\mathbf{m}_{\mu_0}, \mathbf{V}_{\mu_0}\right) \\ \boldsymbol{\gamma} \sim NNGP\left(\mathbf{m}_{\gamma}, \mathbf{V}_{\gamma}\right) \\ \boldsymbol{\xi} \sim NNGP\left(\mathbf{m}_{\xi}, \mathbf{V}_{\xi}\right) \\ \boldsymbol{\mu}_k \sim NNGP\left(\mathbf{m}_{\mu_k}, \mathbf{V}_{\mu_k}\right) \ \forall k = 1 \dots K \\ \boldsymbol{\tau}_k \overset{iid}{\sim} \mathcal{N}(0,1) \ \forall k = 1 \dots K \end{cases} \quad (4)$$

In eq. (4), time variability is induced by a set of K time series $\tau_k(t)$. This is similar to the covariate modeling approach widely used in the literature (e.g. Maraun et al. 2011; Prosdocimi et al. 2015), with the major difference that these covariates are here assumed unknown and are hence treated as latent variables that need to be inferred. The time series $\tau_k(t)$ are termed Hidden Climate Indices (HCI) because, apart from their hidden nature, they are similar to climate indices such as the Southern Oscillation Index (e.g. Sun et al. 2014) or the North Atlantic Oscillation index (e.g. Whan and Zwiers 2017) frequently used in this context. We refer to the paper by Renard et al. (2021) for a thorough description of the HCI modeling framework.

As in eq. (2), NNGPs are used to model the spatial variability of CV and shape parameters γ and ξ and of the intercept term μ_0. Parameters $\mu_1, \dots, \mu_K$ represent the effect of each HCI and are also assumed to vary in space following a NNGP. Note that no log transformation is used here because HCI effects can be positive or negative. Finally, each HCI time series is assumed to be composed of iid realizations from a standard normal distribution.

3.1.4 M4: HCI Model for Peaks and Marks

Let $O(r,t)$ denote the occurrence of a flood mark at time t and site r (note the distinction with the index s used for stations). Moreover, let $\mathcal{B}(\theta)$ denote the Bernoulli distribution with probability of occurrence θ. The fourth model uses the HCI model of eq. (4) to describe flood peaks at stations, and complements it with the following HCI model for flood marks at sites:

$$\begin{cases} O(r,t) \sim \mathcal{B}\left(g\left(\lambda_0(r) + \sum_{k=1}^{K} \lambda_k(r)\tau_k(t)\right)\right) \\ \boldsymbol{\lambda}_0 \overset{iid}{\sim} \mathcal{N}\left(m_{\lambda_0}, v_{\lambda_0}\right) \\ \boldsymbol{\lambda}_k \sim NNGP\left(\mathbf{m}_{\lambda_k}, \mathbf{V}_{\lambda_k}\right) \ \forall k = 1 \dots K \end{cases} \quad (5)$$

where $g(x) = 1/(1+e^{-x})$ is the standard logistic function, used to ensure the parameter of the Bernoulli distribution remains between 0 and 1.

It is stressed that this fourth model uses both equations (4) and (5) and is hence a joint model for flood peaks and flood marks. Importantly, the HCIs $\tau_k(t)$ used in equations (4) and (5) are the same, corresponding to assuming that the temporal variability of food peaks and marks is driven by a set of common HCIs. This assumption of common HCIs is of prime importance in the context of this paper since it enables the transfer of information between flood marks and flood peaks, as will be described in section 3.3 (see also Renard et al. 2021 for a general discussion).

As in eq. (4), NNGPs are used to model the spatial variability of HCI effects $\lambda_k(r)$. The intercept term $\lambda_0(r)$ controls the marginal probability of having a flood mark at site r. This probability strongly depends on the local hydraulic configuration of the site, and there is little reason to expect this parameter to vary smoothly in space. We therefore specified an independent Gaussian process for $\lambda_0(r)$.

3.2 Inference

For all models described previously, inference is performed by deriving the posterior distribution and exploring it with a Markov Chain Monte Carlo (MCMC) sampler. Posterior distributions are given in the appendix section A2. They are derived under the assumption that data are independent in both space and time, conditionally on the inferred parameter. Note that for HCI models M3 and M4, this is not equivalent to assuming (unconditional) space-time independence: indeed, the use of HCIs with spatially structured effects can introduce dependence, and can even be considered as an indirect way to model it (see Renard et al. 2021 for practical illustrations).

The MCMC sampler used here is an adaptive block Metropolis algorithm, carefully implemented to avoid unnecessary computations and minimize computation time (Renard and Thyer 2019). Models M1-M2 correspond to standard hierarchical models in a spatial context and the MCMC exploration of their posterior distributions poses no particular difficulty. HCI models M3-M4 are more challenging because the estimation of both spatial and temporal latent variables leads to non-identifiability issues that need to be addressed to make inference feasible. A two-part solution is used here (see Renard and Thyer 2019 for technical details). First, each HCI time series $\left(\tau_k(t)\right)_{t=1\ldots T}$ is forced to have mean zero and variance one: this implies that only $T-2$ values need to be inferred, the remaining two being derived from the two constraints. Second, a stepwise inference is performed, with the model being estimated one HCI component at a time.

Note that the models used in this paper require estimating a large number of parameters, but they also use a large volume of data. Models M1 and M2 require estimating 621 and 630 parameters, respectively, but using 10,452 data points, leading to a ratio of 16.8 and 16.6 data point per inferred parameter. Model M3 uses the same data but introduces additional

complexity to describe the HCI time series and its spatial effect: for one component (i.e. at one step of the stepwise inference), this leads to 950 inferred parameters, i.e. 11.0 data point per parameter. Finally model M4 adds even more parameters for the Bernoulli distribution of equation (5), but it also uses additional flood mark occurrence data: for one component, this leads to 1,808 parameters for 41,310 data points, i.e. 22.8 data point per parameter. While counting inferred parameters vs. data points has limitations and is by no means sufficient to discard potential inference difficulties, it at least shows that the models used in this paper are not inherently over-parameterized. We also refer to the synthetic case studies of (Renard and Thyer 2019) for more detailed analyses on this topic.

MCMC sampling is performed on a high-performance computing cluster allowing to run many chains in parallel. For each of the four models, 10 chains are run in parallel during 500,000 iterations, with a computing time of approximately 1 day for models M3-M4 (much less for simpler models M1-M2). This high number of iterations was found to be necessary because of a relatively slow mixing for some of the inferred parameters, typically HCIs $\tau_k(t)$ associated with early years having poor data coverage. To avoid storage issues, only one iteration every 500 is saved, so that a total of 10,000 iterations is available across the 10 chains. The first 20% of each chain is further discarded as a burn-in period. MCMC convergence is assessed by monitoring the Gelman-Rubin criterion (Gelman and Rubin 1992) and visualizing MCMC traces.

3.3 Estimating Flood Probabilities

The HCI models M3-M4 allow computing time-varying flood probabilities at hydrometric stations. For a given station s and time step t, let $\boldsymbol{\Theta}(s) = \left(\mu_0(s), \mu_1(s), \dots, \mu_K(s), \gamma(s), \xi(s)\right)$ denote the values of all spatially-varying parameters at this station, and $\boldsymbol{\tau}(t) = \left(\tau_1(t), \dots, \tau_K(t)\right)$ denote the values of HCIs at this time step. The probability of not exceeding

a particular streamflow value q can be computed from the following GEV cumulative distribution function (cdf):

$$Pr(Q(s,t) \leq q | \boldsymbol{\Theta}(s), \boldsymbol{\tau}(t)) = F_{GEV}\left(q; e^{\mu_0(s)} \times \left(1 + \sum_{k=1}^{K} \mu_k(s)\tau_k(t)\right), e^{\mu_0(s)} \times e^{\gamma(s)}, \xi(s)\right) \quad (6)$$

This equation can be applied to any station and time step belonging to the estimation dataset. For model M3 using flood peaks data only, this means that flood probabilities can be estimated at all stations over the period 1904-2015. More interestingly, since model M4 also uses flood marks available over a much longer period, it allows computing flood probabilities at all stations over the period 1705-2015. As explained in section 3.1.4, this is made possible by the assumption that a common set of HCIs is driving the time variability of both flood peaks and flood marks: in a nutshell, ancient flood marks allow identifying the values taken by HCIs during the earlier years of the period, and this information can be transferred to stations by means of eq. (6), even before the availability of any flood peak data.

A limitation of eq. (6) is that it is conditional on the parameters $\boldsymbol{\Theta}(s)$ and $\boldsymbol{\tau}(t)$. In practice, point-estimates for these parameters can be derived from the MCMC samples (typically, the posterior mode or median) and used in eq. (6). However, this approach ignores estimation uncertainty which may be large, as will be shown in the case study. In a Bayesian context, this is typically addressed by using the predictive distribution which integrates out parameter uncertainty. Denoting by **D** the estimation dataset (flood peaks for M3 and peaks+marks for M4), this can be formalized as follows:

$$Pr(Q(s,t) \leq q|\mathbf{D}) = \int \underbrace{Pr(Q(s,t) \leq q|\boldsymbol{\Theta}(s),\boldsymbol{\tau}(t))}_{\text{eq. (6)}} \underbrace{p(\boldsymbol{\Theta}(s),\boldsymbol{\tau}(t)|\mathbf{D})}_{\text{posterior pdf}} d\boldsymbol{\Theta}(s)d\boldsymbol{\tau}(t) \quad (7)$$

In practice, the integration in eq. (7) is analytically intractable but can readily be estimated from the MCMC samples. Indeed, eq. (7) can be interpreted as the posterior expectation of eq. (6), and can hence be approximated by computing eq. (6) for each MCMC sample and averaging the results.

While models M3-M4 are time-varying in nature, they are also able to provide a time-invariant marginal distribution by integrating out the HCIs with respect to their hyperdistributions as shown below. This marginal distribution may be requested by typical FFA applications such as engineering design (as further discussed in section 5.1), but in the context of this paper it also useful to compare all four models M1-M4.

$$Pr(Q(s) \leq q|\boldsymbol{\Theta}(s)) = \int \underbrace{Pr(Q(s) \leq q|\boldsymbol{\Theta}(s),\boldsymbol{\tau})}_{\text{eq. (6)}} \underbrace{\prod_{k=1}^{K} f_{\mathcal{N}}(\tau_k;0,1)}_{\text{hyperdistributions}} d\boldsymbol{\tau} \quad (8)$$

As previously, this integral is intractable but can estimated as described above, with HCI values $\boldsymbol{\tau}$ being sampled from their $\mathcal{N}(0,1)$ hyperdistributions. Finally, a predictive version of this marginal distribution can be obtained by further integrating out parameters $\boldsymbol{\Theta}(s)$:

$$Pr(Q(s) \leq q|\mathbf{D}) = \int \underbrace{Pr(Q(s) \leq q|\boldsymbol{\Theta}(s),\boldsymbol{\tau})}_{\text{eq. (6)}} \underbrace{p(\boldsymbol{\Theta}(s)|\mathbf{D})}_{\text{posterior pdf}} \underbrace{\prod_{k=1}^{K} f_{\mathcal{N}}(\tau_k;0,1)}_{\text{hyperdistributions}} d\boldsymbol{\Theta}(s)d\boldsymbol{\tau} \quad (9)$$

4 Results

The four models of section 3.1 are applied to the datasets of section 2, and the results description is organized as follows. Estimated parameters are first compared between the four models in order to understand how the models' hypotheses affect estimation. For models M3-M4, estimated HCIs and their effects are also described and interpreted. The model comparison is then extended to the marginal distributions and associated flood quantiles. Flood probabilities at all stations are then estimated over the period 1705-2015 thanks to the joint modeling of flood peaks and flood marks by M4. The reliability of these ancient estimates is evaluated by means of a cross-validation exercise. Finally, a sensitivity analysis is carried out to assess how deviations from some key assumptions made in this work affect the results.

4.1 Estimated Parameters

The top row of Figure 2 compares GEV parameters at all stations. For the location parameter e^{μ} (M1-M2) or its intercept e^{μ_0} (M3-M4), differences between models are barely noticeable, while stronger differences are found for the CV and shape parameters. Focusing on the comparison between M1 and M2 first, M2 estimates of CV and shape appear to be smoothed versions of M1 estimates, which indicates that these parameters show some spatial consistency that can be taken into account in M2 through the spatial models of equation (2). On the other hand, the strong similarity between M1 and M2 location estimates indicates that spatial consistency is weak for this parameter. This result illustrates that the use of a spatial model in equation (2) does not necessarily lead to spatially smooth estimates - it does only if the data suggest it should. Note that the smoothing effect is particularly strong for the shape parameter, which is highly variable when estimated locally (M1), reflecting its well-known sensitivity to sampling uncertainty. Models M3-M4 also show systematic differences in CV

and shape estimates compared with M1-M2. In particular, CVs are on average 50% higher with M1-M2 than with M3-M4. This can be attributed to the fact that M3-M4 are HCI models: part of the temporal variability that CV and shape parameters represent are accounted for by the temporal variability of HCIs. This is akin to covariates explaining part of the data variability in regression.

While estimated values of CV and shape parameters highlight systematic differences between M1-M2 and M3-M4, their uncertainties separate M1 from the three other models (bottom row of Figure 2): M1 leads to the highest uncertainties, while M2-M3-M4 are very similar in this respect. This result suggests that much of the uncertainty reduction is achieved thanks to the explicit modeling of spatial variability in models M2-M3-M4. By contrast, the use of HCIs (M3-M4) or the inclusion of flood marks (M4) does not make any noticeable difference in terms of GEV parameter uncertainty.

4.2 Estimated HCIs and their effects

In addition to the GEV parameters discussed above, models M3-M4 also estimate HCI time series $\tau_k(t)$ and their effects at stations ($\mu_k(s)$, eq. (4)) and, for M4, sites ($\lambda_k(r)$, eq. (5)). Figure 3 shows the M4 estimates for the first three HCIs. A similar figure is provided for HCIs 4 to 6 (see Supplementary material, Figure S1). The total number of 6 HCIs was selected because the magnitude of HCI effects drops at the seventh HCI (not shown - see Renard et al. 2021 for additional discussion on how to select the number of components).

For the first component shown in Figure 3, effects at both stations and sites are positive everywhere, with a northwest-southeast decreasing gradient. This means that high values of the first HCI $\tau_1(t)$ are associated with higher-than-usual flood peaks and more-frequent-than-usual flood marks, especially in the northwestern half of the country. The HCI time series itself is characterized by a highly variable uncertainty (pink band). During the recent period

1951-2015, for which many stations and sites provide peaks and marks data, uncertainty is very small. Interestingly, the HCI estimated from model M3 (blue band), which uses peaks data only, is very similar to the M4 one. This indicates that the estimation of the HCI is mostly driven by the peaks data. For the period 1904-1950, the number of stations providing peaks data decreases, and the HCI uncertainty hence increases. However, this uncertainty increase is much weaker for M4 than for M3, which illustrates the additional information brought by flood marks compared to flood peaks alone. For the pre-1904 period, only flood marks are available and the HCI is hence only available for M4. Since the number of active sites decreases as one moves back in time, the HCI uncertainty increases until reaching a high level before the 19th century. Such a high level of uncertainty cannot reasonably be ignored in further computations, which justifies the use of predictive distributions as described in section 3.3.

Similar comments can be made for the second and third HCIs in Figure 3 in terms of uncertainty. The effects of HCI2 are much more localized than for HCI1, with high positive values being mostly restricted to the northwest and northeast tips of the country (Brittany and Lorraine regions). HCI3 has both positive effects in the southern Mediterranean region and negative effects in the northeast.

4.3 Marginal Distributions

Figure 4 compares the marginal distributions (represented as quantile curves) estimated at three stations. For models M1-M2, the GEV distributions of eq. (1)-(2) are used. For models M3-M4, marginal distributions are derived by integrating out HCIs with respect to their hyperdistributions, as described in section 3.3 (eq. (8)). For these three particular stations, there is no discernible difference between M3 and M4. This suggests that the inclusion of flood marks data does not change the estimated marginal distribution: this may appear

surprising and will be further discussed in section 5.1. Moderate differences appear with model M2, but overall M2-M3-M4 quantile curves remain compatible with each other and have comparable uncertainties. This similarity in terms of uncertainty between the three models was already noted in section 4.1 and can be explained by the fact that they share the same spatial models for GEV parameters. Finally, model M1 leads to strong differences with other models, notably a much higher uncertainty for the first two stations.

The comparison above is extended to all 207 stations in Figure 5 which compares 100-year flood estimates and their uncertainties. Overall, the variability of Q100 across stations is very similar with the four models. However, larger differences appear when expressed in terms of deviation from a reference model (taken as M2). This is particularly the case with local model M1, with Q100 being sometimes more than twice larger than the reference. By contrast, Q100 from models M3-M4 remain similar to M2, with deviations being mostly within $\pm 20\%$ (median:-6%). This is a remarkable result given that marginal distributions were obtained in very different ways: M2 directly models the GEV marginal distribution, while M3-M4 integrate out HCIs from conditional GEV distributions, resulting in a marginal distribution that does not even belong to the GEV family. Strong differences between methods also appear in terms of Q100 uncertainties: they are mostly in the range 10-15% with M2-M3-M4, but they are on average 2.5 times higher with M1, mostly in the range 15-60%. This is a direct consequence of the larger estimation uncertainty affecting M1 parameters, and in particular the shape parameter, as discussed in previous section 4.1.

4.4 Estimated Flood Probabilities, 1705-2015

As explained in section 3.3, model M4 allows estimating the time-varying distribution of flood peaks at all stations over the period 1705-2015, by transferring the information provided by flood marks at sites. Figure 6a illustrates this for a station located in Brittany (northwest).

The predictive distribution (eq. (7)) strongly varies in time, so that for some years the probability of exceeding the 10-year flood is close to zero or one. This corresponds to sharp estimations and denotes an efficient transfer of information from the marks dataset to peaks at this station. By contrast, estimations made at the station shown in Figure 6b have low sharpness: the predictive distribution does not vary much in time, and the probability of exceeding the 10-year flood always remains close to 0.1. These two examples suggest that sharpness can be quantified by computing the range of estimated probabilities over the 1705-2015 period. The map in Figure 6c indicates that sharpness is particularly high in Brittany (northwest) while it is very low around the Mediterranean (southeast). This sharpness gradient reflects the fact that HCI effects at stations tend to be higher in the northwest than in the southeast (Figure 3). Possible reasons behind this gradient are discussed in section 5.1.

In addition to the time series of flood probabilities at one particular station, it is of interest to derive the probability map computed for one particular year: this provides information on the spatial structure of floods. Figure 7 shows examples of such probability maps for a few selected years. The complete movie for all years 1705-2015 is provided in the supplementary material (Video S1) and the underlying dataset is released in an online repository (see “data availability” section).

For hydrological year 1710 (September 1710 to August 1711), only 14 sites are active but 8 of them record a flood mark, all associated with the same event in February 1711. 5 marks are located in the downstream part of the Loire catchment (northwest), while 3 are located in the quite distant Rhône catchment (center-east): this may be the sign of a large-scale event that affected an important part of the country. The associated probability map indeed shows high probabilities of exceeding a 10-year event in the northern half of the country, and especially in Brittany (northwest). These high probabilities result from the high values taken by the first two HCIs for this year, and the fact that they have large effects in these regions (Figure 3).

Hydrological years 1855 and 1866 are both characterized by a large number of marks, mostly located along two large rivers: the Loire River (approximately flowing from East to West in the northern half of France) and the Rhône River (flowing North-South in the Southeast). By contrast most stations are located on smaller upstream tributaries, which explains why probability patterns and flood marks patterns are quite different. Comparing maps for 1855 and 1866 yields interesting interpretations. For 1866, all marks relate to the same event in September-October 1866, and they are mostly located in the Loire river catchment, with a few additional marks in the Southwest. For 1855, almost all marks relate to the same event in May-June 1856. As for 1866, many are located along the Loire River, but unlike 1866, many marks are also found along the Rhône River (Southeast). As a result, the area with high flood probability is much more widespread in 1855 than in 1866. As for 1710, the presence of flood marks in both the Loire and the Rhône catchments hints toward a generalized event that affected a large part of the country.

Marks for hydrological year 1874 are related to the event of June 1875 and are located in southwest France, mostly in the Garonne catchment. This example is interesting because few stations are located in the immediate vicinity of flood mark sites. Despite this, an area with high flood probability is identified in the upstream Pyrenees Mountain range, despite the lack of flood marks in this area. This indicates that the model was able to identify the covariability between flood peaks at upstream stations and flood marks at downstream sites, and to use it to transfer information spatially.

Finally, hydrological years 1887 and 1890 illustrate cases where no high-probability area is detected in the country. While this seems to be a sensible estimation for 1887 given the absence of any flood mark, this is more questionable for 1890. Indeed, several marks are recorded for September 1890, which constitutes a reference flood in the western Mediterranean area, even one of the largest known flood in some catchments (Naulet et al.

2005). However, the model fails to associate these flood marks with a high flood probability at stations. More generally, flood probabilities in the Mediterranean area show little variation in time and remain close to 0.1 for all years: this is a consequence of the low sharpness found in this area (Figure 6c), whose possible causes are further discussed in section 5.1.

4.5 Are ancient flood probabilities reliable?

Flood probabilities during the 18th and 19th centuries rely on a restricted number of flood mark sites: for instance, only 14 sites were active in 1710. In order to assess the reliability of these estimates, a cross-validation exercise is set up by applying Model M4 on reduced datasets mimicking the information that was available during ancient flood events. Dataset V1 is derived by removing all flood peaks data during the decade 1981-1990, leaving 138 active sites available to estimate flood probabilities during this decade. This is comparable with the information that was available for the flood of hydrological year 1855 (141 sites, 0 stations). Dataset V2 is made more challenging by only retaining, among the 138 sites above, the 10 that started during the 18th century. This is similar to the situation in 1710 (14 sites, 0 stations).

Figure 8 compares the estimates obtained with reduced datasets V1 and V2 against the ones obtained with the full dataset V0. Figure 8a shows that during the validation period 1981-1990, the removal of all station data (V1 and V2) leads to large uncertainties for estimated HCIs. V2-estimated HCIs are more uncertain than V1 ones due to the further removal of many flood mark sites. V0 estimates, which use the station data of 1981-1990, are much less uncertain. While different, the HCIs estimated with the three datasets remain compatible in the sense that the V0 HCIs are mostly included in the V1/V2 uncertainty bands. The three estimates are very similar before and after the validation period 1981-1990.

Figure 8b illustrates how the differences discussed above translate in terms of probability

maps. For hydrological year 1982, high flood probabilities are estimated in northern France with all three datasets, driven by the large estimated value of HCI τ_1. Station data confirm that many floods occurred (black dots in Figure 8b). Importantly, these data were included in V0 but not in V1 and V2 datasets, hence providing a validation of the V1/V2 probability maps estimated using flood marks only. Hydrological year 1986 corresponds to a case where low flood probabilities are estimated with the three datasets, and indeed very few events occurred. For hydrological year 1987, the three probability maps differ more markedly: the V0 map indicates an area of high flood probability in the northwest that V1 and V2 maps fail to detect. This is because the floods that indeed occurred were probably not large enough to materialize as flood marks, which is the only source of information available in V1/V2 during the validation period.

The reliability of estimated flood probabilities can be evaluated more formally and for all validation years by means of the reliability diagram (Laio and Tamea 2007). This diagram is based on probability-transformed values $F_{s,t}(q_{s,t})$, where $q_{s,t}$ is the observed peak flow at site s and year t, and $F_{s,t}$ is the cdf of the corresponding predictive distribution (equation (7)). These probability-transformed values should be uniformly distributed between 0 and 1 if the predictive distribution is reliable, which is assessed in Figure 8c for each year of the decade 1981-1990. For most years the V1/V2 reliability curves remain close to the V0 curve and close to the diagonal, indicating high reliability. A slight loss of reliability is observed for a few years including year 1987, which is shown in Figure 8b and discussed in the previous paragraph. Such a slight loss of reliability is expected as V1/V2 estimates are evaluated in a validation context, while evaluation data $q_{s,t}$ are included in V0. Overall, the cross-validation exercise carried out in this section indicates that estimating probability maps based on a restricted number of flood mark sites can deliver an acceptably reliable information on ancient flood events.

4.6 Sensitivity Analyses

As explained in section 2.2, the interpretation of the flood marks dataset in term of flood occurrence/non-occurrence relies on two strong hypotheses: (i) at each site, the activity period spans from the first to the last mark; (ii) the absence of flood mark is interpreted as 'no flood'. In order to assess the sensitivity of the results to departures from these hypotheses, three synthetic datasets are created by altering the original dataset as follows:

D1: At each site, the first and last occurrences are removed. This implies that the activity period starts and ends with a generally long series of non-occurrences, instead of starting and ending with an occurrence.

D2: Years with no flood marks are randomly associated with a flood occurrence with probability 0.1. This mimics a 10% miss rate affecting the original dataset.

D3: same alteration as D2 but only applied to the period 1800-1825. This introduces a non-homogeneity in the flood detection process that affects all sites.

Sensitivity is assessed by monitoring the first HCI $\tau_1(t)$ and its effects at sites ($\lambda_0(r)$, $\lambda_1(r)$) and stations ($\mu_0(s)$, $\mu_1(s)$). The first row of Figure 9 indicates that HCI estimation is not sensitive to the definition of the activity period (D1 is barely distinguishable from the reference). Likewise, it is weakly sensitive to a constant miss rate: the D2 interval remains very similar to the reference, despite being somewhat noisier. However, sensitivity is higher with a non-constant miss rate: the D3 interval clearly departs from the reference during 1800-1825, while it remains very similar outside of this period. Overall, these results suggest that the existence of missed events is not problematic as long as the associated recording process remains homogeneous. Conversely, non-homogenities or trends in the miss rate may be picked up by the estimated HCIs and be wrongly interpreted as a hydroclimatic signal.

The second row of Figure 9 indicates that spatial parameters related to sites are highly sensitive to departures from the hypotheses. Parameter λ_0 controls the marginal flood mark probability and is hence directly related to the number of flood marks. Consequently, it is slightly lower than the reference with D1 due to the systematic removal of the first and last marks. After logistic transformation (see eq. (5)), λ_0 values correspond to flood mark probabilities around 1% (D1) vs. 3% (reference). At the opposite, λ_0 values are much higher than the reference with D2 (corresponding to probabilities around 13%), due to the addition of many flood marks. λ_0 values are only slightly higher with D3 because additional flood marks are only added during the 1800-1825 period. Parameters λ_1 control the effect of the first HCI on the flood mark probability and are very similar to the reference with datasets D1 and D3. However, effects are much smaller than the reference with dataset D2, although the pattern remains similar. This can be explained by the fact that the additional flood marks of D2 have been added randomly and are hence not controlled by the HCI, which weakens its apparent effect.

Finally, the last row of Figure 9 indicates that spatial parameters related to stations (μ_0 and μ_1) are virtually insensitive to departures from the hypotheses. This is an important result in the context of this paper because estimations made at stations such as the marginal distributions of Figure 4 or the flood probabilities of Figure 6 and Figure 7 only depend on these parameters and the HCIs (see section 3.3). Conversely, the site-specific parameters λ_0 and λ_1 play no role in these estimations, so that their high sensitivity is not problematic. Consequently, the sensitivity analysis carried out in this section indicates that the estimations made at stations are not sensitive to the definition of the activity period at sites or to the existence of a non-zero but constant miss rate. A non-constant miss rate is more problematic as it will be picked up by the estimated HCIs and may hence affect time-varying flood probabilities (but not marginal distributions).

5 Discussion

5.1 Interest of jointly modeling flood peaks and flood marks

Compared with usual approaches that require transforming historical information into flood discharge at hydrometric stations, the model used in this paper preserves the original localization of flood marks and their original nature as time series of occurrence/non-occurrences. The results presented in the previous sections indicate that this simplified approach is sufficient to transfer information on ancient flood events from sites to stations and to reconstruct the associated probability maps. These maps provide information on the intensity that historical floods may have reached at hydrometric stations, well before these stations even existed. They also provide valuable information on the spatial structure of historical floods, which may quantitatively complement the spatial structure suggested by more detailed but event-specific historical analyses of various sources. However, the interest of these maps appears to be limited in Mediterranean regions (southeast) due to a low sharpness. This is likely due to the combination of two factors. First, Mediterranean flood events are highly localized and may be intrinsically less amenable to prediction from flood marks than large-scale oceanic events. Second, the density of flood marks is rather low in the southeast, which further affects predictability but may be remedied by improving data collection.

While flood marks play a pivotal role in the characterization of ancient flood events, results suggest that they do not induce any noticeable change in the estimation of marginal distributions at stations. At first sight this is a disappointing result, since a key motivation for including historical information in FFA is to improve the estimation of this distribution. It is also a surprising one given that previous research demonstrated the usefulness of historical data (e.g., Payrastre et al. 2011 amongst many others). We stress, however, that this result is

not in contradiction with existing literature since it only holds within the context of the specific model used in this work. More precisely, the lack of influence of flood marks is a consequence of keeping the sites network and the hydrometric stations network separated, with no attempt at pairing sites and stations. Indeed, under this setup, flood marks can only influence the location parameter of the GEV distribution in an indirect way, through the common HCIs in eq. (4)-(5). However, flood marks have no connection with the scale and shape GEV parameters, which are the most difficult to estimate. By contrast, if sites and stations were paired, it would be possible to express the probability of recording a flood mark in eq. (5) as a function of the GEV parameters at the paired station. Flood marks would therefore have a more direct influence on the estimation of all three GEV parameters. However, pairing hundreds of sites and stations is not obvious as explained in the introduction, and since flood marks may occur outside of the hydrometric network, many sites would still be left 'unpaired'.

Finally, the case study delivered an additional interesting result in terms of FFA: the distribution obtained by 'integrating out' conditional, time-varying GEV distributions (models M3-M4) is similar to the one obtained by directly modeling a marginal, time-invariant GEV distribution (model M2). This finding was not obvious beforehand considering the quite different assumptions underlying the two approaches and their distinct parameterizations and degrees of freedom. This similarity indicates that modeling conditional distributions may be of interest not only for the purpose of making time-varying estimations, but also for deriving the marginal distribution used in FFA, or alternatively to design structures based on their reliability over a given lifetime as suggested by Read and Vogel (2015). The pros and cons of each approach and their link with related methods such as the derived distribution approach (e.g., Michele and Salvadori 2002) need to be further studied in future work.

5.2 Limitations and future developments

The analyses performed in this work rely on a few strong hypotheses regarding how flood marks can be interpreted as flood occurrences, and more critically, their absence as flood non-occurrences. The sensitivity analysis of section 4.6 is reassuring since it indicates that estimations made at hydrometric stations are quite robust to departures from these hypotheses, despite the fact that site-specific estimates are not. This constitutes an advantage of keeping the two networks separated: if sites and stations were paired, estimations made at a given station would likely be more sensitive to misinterpretations of the mark recording process at the associated site. The sensitivity analysis suggests that the most problematic situation would be a strong inhomogeneity in the flood detection process affecting a large part of the country - for instance, decades during which flood events failed to materialize into flood marks for some historical reasons. Modifying the model to account for such inhomogeneity is feasible but would require making specific historical hypotheses on homogeneity periods and the associated flood detection probabilities.

A possible limitation of the analysis described in this paper is the use of a full hydrological year as time step. In some cases, marks from the same hydrological year may correspond to distinct and unrelated flood events, but this information would be lost since the data used is the occurrence of a flood mark at any time during the year. This situation occurs quite rarely in the studied dataset: as described in section 4.4, marks within a year are generally related to a single event. For alternative datasets where this could be problematic, an easy and potentially sufficient strategy would be to apply the model at a seasonal, rather than annual, time scale. A more challenging approach would be to develop an event-based model.

Estimated flood probabilities for ancient events could be further improved by growing the flood marks dataset and making a better use of its content. First, this database is part of a ‘citizen science’ collaborative platform: anybody can register and contribute flood marks,

with new contributions being validated by professional staff before publication. There is therefore scope to improve the spatial and temporal coverage of flood marks, which would be of particular interest for estimations in poorly-covered regions (southeast, southwest and northwest coastal areas). In addition, the mark elevation is also available for about 70% of them, but this information was completely disregarded in this paper. The added value of using mark elevation compared with occurrence only remains to be evaluated. It may provide valuable information on the ranking of flood events at each site, but it may also be overly sensitive to changes in the local hydraulic configuration.

Finally, an area of promising future development is to complement the analysis performed in this work with large-scale climate information. Indeed, the availability of long reconstructions such as 20CR (Compo et al. 2011) provides an opportunity to jointly use climate and historical data over nearly two centuries. Climate proxies such as dendroclimatic data may allow extending this period even further (Steinschneider et al. 2018). This joint use of climate, historical and systematic flood data may shed light on the climate mechanisms driving the occurrence of large flood events and improve the understanding of natural flood variability. It may also offer the opportunity to build and calibrate a downscaling tool between large-scale climate and floods that may be used for various applications such as seasonal forecasting or future projections.

6 Conclusion

The availability of a national flood mark database collecting geolocalized information spanning several centuries constitutes an opportunity to better understand flood natural variability in France and to improve the characterization of the associated hazard. This work aimed at building a model to jointly analyze flood mark occurrences at sites and flood peaks measured at stations from the standard hydrometric network. This model keeps the site and

station networks separated, and uses the idea that common Hidden Climate Indices drive both flood marks and peaks to transfer information between them. Following this approach, flood marks at sites were used to estimate flood probabilities at stations during a more-than-300-year period starting in the early 18th century. These estimates were gathered into probability maps providing useful information on the extent and the spatial structure of ancient flood events. A sensitivity analysis suggested that these estimates are reasonably robust to misinterpretations of the mark recording process, thanks to the separation of the site and station networks.

The same model feature - separation of site and station networks- was found to make the model less useful for FFA purposes. Indeed, the inclusion of old flood marks did not noticeably change the marginal FFA distribution compared with the same model using flood peaks at stations only. It is stressed that this conclusion highlights a limitation of the particular model used in this paper rather than an intrinsic inability of flood marks to improve FFA. In particular, pairing flood mark sites and hydrometric stations is likely to make flood marks more influential for FFA purposes, although it may also come with drawbacks such as a greater sensitivity to misinterpretations of the mark recording process. By contrast to the use of flood marks, the use of spatial models to smooth some of the GEV parameters was found to be an effective way to reduce parameter and quantile uncertainties.

Systematic data from hydrometric networks are the cornerstone of flood hazard assessment, but their temporal coverage is limited. In order to understand the long-term natural variability of floods and to characterize ancient flood events, it is therefore valuable to complement systematic data with other sources of information, including but not limited to historical information. The development of flexible statistical models that can handle a wide variety of data types and recognize their specificities is an important step in this endeavor.

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 HIICS cluster at INRAE.

Flood marks were taken from the 'Plateforme nationale des repères de crues' (National Flood Marks Platform, https://www.reperesdecrues.developpement-durable.gouv.fr), developed by the flood forecasting network 'Vigicrues' from the Ministry of the Ecological Transition. Anne-Marie Fromental is gratefully acknowledged for her insights on this dataset.

I also thank the Editor and an anonymous Reviewer for their insightful comments.

Data availability

The original datasets used in this article can be accessed online as cited in the text. The following repositories have been created to complement the article:

- All data used in this article (flood marks at sites and flood peaks at stations) and estimated flood probabilities over the period 1705-2015 are available in a Zenodo repository (Renard 2022, https://doi.org/10.5281/zenodo.6793501) and can be visualized at https://vimeo.com/815008124.
- R scripts used for setting up models, analyzing results and preparing figures are available in a Zenodo repository (Renard 2023, https://doi.org/10.5281/zenodo.7853033).
- 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)

Appendix

A1. Nearest-Neighbors Gaussian Processes

A standard Gaussian Process (GP) is characterized by the fact that its joint pdf at a set of S locations is the multivariate normal distribution:

$$f_{GP}(\mathbf{x}; \mathbf{m}, \mathbf{V}) = \frac{1}{\sqrt{(2\pi)^S det(\mathbf{V})}} exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})\right) \quad \text{(A1)}$$

where $\mathbf{x} = (x_1, \dots, x_S)$ is the vector at which the pdf is evaluated, $\mathbf{m}$ is the mean vector and $\mathbf{V}$ is the covariance matrix. The computational bottleneck in equation (A1) is the inversion of the $S \times S$ matrix $\mathbf{V}$.

This joint pdf can be decomposed as follows using conditional probability algebra:

$$\begin{aligned} p(x_1, \dots, x_S) &= p(x_1)p(x_2, \dots, x_S|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \dots, x_S|x_1, x_2) = \cdots \\ &= p(x_1)\prod_{i=2}^{S} p\left(x_i|x_1, \dots, x_{i-1}\right) = \prod_{i=1}^{S} p\left(x_i|\mathbf{x}_{\boldsymbol{\pi}_i}\right) \end{aligned} \quad \text{(A2)}$$

where the notation $\boldsymbol{\pi}_i$ introduced in the last line denotes the indices $1, \dots, i-1$ (and, by convention, $\boldsymbol{\pi}_1 = \emptyset$). In the Gaussian case, each conditional pdf $p\left(x_i|\mathbf{x}_{\boldsymbol{\pi}_i}\right)$ is a normal distribution with mean $\tilde{m}_i$ and variance $\tilde{v}_i$ defined as:

$$\begin{cases} \tilde{m}_i = m_i + \mathbf{V}_{i,\boldsymbol{\pi}_i}\mathbf{V}_{\boldsymbol{\pi}_i,\boldsymbol{\pi}_i}^{-1}\left(\mathbf{x}_{\boldsymbol{\pi}_i} - \mathbf{m}_{\boldsymbol{\pi}_i}\right) \\ \tilde{v}_i = V_{i,i} - \mathbf{V}_{i,\boldsymbol{\pi}_i}\mathbf{V}_{\boldsymbol{\pi}_i,\boldsymbol{\pi}_i}^{-1}\mathbf{V}_{\boldsymbol{\pi}_i,i} \end{cases} \quad \text{(A3)}$$

Equation (A2) allows decomposing the multivariate joint pdf (A1) into a product of univariate conditional pdf's. However deriving each conditional pdf using equation (A3) still requires inverting the matrix $\mathbf{V}_{\boldsymbol{\pi}_i,\boldsymbol{\pi}_i}$, whose size may be as high as $(S-1) \times (S-1)$: the computational bottleneck hence still holds. The idea behind Nearest-Neighbors Gaussian

Processes is to replace the growing list of conditioning stations $\boldsymbol{\pi}_i = (1, \dots, i-1)$ by a smaller list of neighboring stations $\boldsymbol{\pi}_i^* = (i_1, \dots, i_k) \subset \boldsymbol{\pi}_i$ with maximum size k. The corresponding joint pdf is therefore defined as:

$$f_{NNGP}(\mathbf{x}; \mathbf{m}, \mathbf{V}) = \prod_{i=1}^{S} p\left(x_i \middle| \mathbf{x}_{\boldsymbol{\pi}_i^*}\right) \tag{A4}$$

where each conditional distribution in the product is Gaussian with mean and variance given in equation (A3). Note that since all $\boldsymbol{\pi}_i^*$ have size at most k, evaluating the NNGP joint pdf only requires inverting $k \times k$ matrices ($k = 5$ was used in this paper). We refer to Datta et al. (2016) for a more in-depth description of the properties of NNGP's and for thorough analyses demonstrating their efficiency even with small values of k.

A2. Posterior Distributions

Let $\mathbf{q} = (q(s,t))_{s=1\dots S, t=1\dots T}$ denote the flood peaks dataset (Figure 1a) and $\mathbf{o}$ similarly denote the $R \times T$ flood marks dataset (Figure 1b). Both datasets include missing values.

For model M1, the posterior distribution is simply proportional to a GEV likelihood. If a data $q(s,t)$ is missing, the corresponding term is simply dropped from the double product.

$$p(\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\xi} | \mathbf{q}) \propto \prod_{s=1}^{S} \prod_{t=1}^{T} f_{GEV}\left(q(s,t); e^{\mu(s)}, e^{\mu(s)} \times e^{\gamma(s)}, \xi(s)\right) \tag{A5}$$

For model M2, additional terms appear in the equation. They correspond to the NNGP's used to model the spatial variability of GEV parameters and to the priors for NNGP's hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$.

$$
\begin{aligned}
&p(\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1 | \mathbf{q}) \propto \\
&\prod_{s=1}^{S} \prod_{t=1}^{T} f_{GEV}\left(q(s,t); e^{\mu(s)}, e^{\mu(s)} \times e^{\gamma(s)}, \xi(s)\right) \times \\
&f_{NNGP}(\boldsymbol{\mu}; \mathbf{m}_{\mu}, \mathbf{V}_{\mu}) f_{NNGP}(\boldsymbol{\gamma}; \mathbf{m}_{\gamma}, \mathbf{V}_{\gamma}) f_{NNGP}(\boldsymbol{\xi}; \mathbf{m}_{\xi}, \mathbf{V}_{\xi}) \times \\
&p(\boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)
\end{aligned}
\tag{A6}
$$

In model M3, the GEV pdf is now varying with time, and additional terms are introduced for HCIs $\boldsymbol{\tau}_k$ and their effects $\boldsymbol{\mu}_k$:

$$
\begin{aligned}
&p(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1 | \mathbf{q}) \propto \\
&\prod_{s=1}^{S} \prod_{t=1}^{T} f_{GEV}\left(q(s,t); e^{\mu_0(s)} \times \left(1 + \sum_{k=1}^{K} \mu_k(s) \tau_k(t)\right), e^{\mu_0(s)} \times e^{\gamma(s)}, \xi(s)\right) \times \\
&f_{NNGP}(\boldsymbol{\mu}_0; \mathbf{m}_{\mu_0}, \mathbf{V}_{\mu_0}) \prod_{k=1}^{K} f_{NNGP}(\boldsymbol{\mu}_k; \mathbf{m}_{\mu_k}, \mathbf{V}_{\mu_k}) \prod_{k=1}^{K} \prod_{t=1}^{T} f_{\mathcal{N}}(\tau_k(t); 0,1) \times \\
&f_{NNGP}(\boldsymbol{\gamma}; \mathbf{m}_{\gamma}, \mathbf{V}_{\gamma}) f_{NNGP}(\boldsymbol{\xi}; \mathbf{m}_{\xi}, \mathbf{V}_{\xi}) p(\boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)
\end{aligned}
\tag{A7}
$$

Finally, model M4 also considers flood marks data $\mathbf{o}$ and its posterior hence comprises additional terms for the data likelihood ($f_{\mathcal{B}}$ stands for the Bernoulli probability mass function) and for the HCI effects $\boldsymbol{\lambda}_k$.

$$
\begin{aligned}
&p(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_K, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K, \boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1 | \mathbf{q}, \mathbf{o}) \propto \\
&\prod_{s=1}^{S}\prod_{t=1}^{T} f_{GEV}\left(q(s,t); e^{\mu_0(s)} \times \left(1 + \sum_{k=1}^{K} \mu_k(s)\tau_k(t)\right), e^{\mu_0(s)} \times e^{\gamma(s)}, \xi(s)\right) \times \\
&f_{NNGP}(\boldsymbol{\mu}_0; \mathbf{m}_{\mu_0}, \mathbf{V}_{\mu_0}) \prod_{k=1}^{K} f_{NNGP}(\boldsymbol{\mu}_k; \mathbf{m}_{\mu_k}, \mathbf{V}_{\mu_k}) \prod_{k=1}^{K}\prod_{t=1}^{T} f_{\mathcal{N}}(\tau_k(t); 0,1) \times \\
&f_{NNGP}(\boldsymbol{\gamma}; \mathbf{m}_{\gamma}, \mathbf{V}_{\gamma}) f_{NNGP}(\boldsymbol{\xi}; \mathbf{m}_{\xi}, \mathbf{V}_{\xi}) \times \\
&\prod_{r=1}^{R}\prod_{t=1}^{T} f_{\mathcal{B}}\left(o(r,t); g\left(\lambda_0(r) + \sum_{k=1}^{K} \lambda_k(r)\tau_k(t)\right)\right) \times \\
&\prod_{r=1}^{R} f_{\mathcal{N}}(\lambda_0(r); m_{\lambda_0}, v_{\lambda_0}) \prod_{k=1}^{K} f_{NNGP}(\boldsymbol{\lambda}_k; \mathbf{m}_{\lambda_k}, \mathbf{V}_{\lambda_k}) \times p(\boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)
\end{aligned}
\tag{A8}
$$

Note that we made a slight abuse of notation in these equations: for models M1-M3, since only flood peaks data are used, the study period is 1904-2015 and therefore $T = 112$. For model M4, the study period is 1705-2015 and $T = 311$. Similarly, we used the unique notation $(\boldsymbol{\alpha}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$ to denote all hyperparameters of the model, but the precise content of these vectors differs between models M2-M4.

References

Banerjee, Sudipto. 2017. "High-Dimensional Bayesian Geostatistics." *Bayesian Anal.* 12 (2). https://doi.org/10.1214/17-BA1056R.

Barriendos, M, and Javier Martin-Vide. 1998. "Secular Climatic Oscillations as Indicated by Catastrophic Floods in the Spanish Mediterranean Coastal Area (14th–19th Centuries)." *Climatic Change*, 473–91. https://doi.org/10.1023/A:1005343828552.

Benito, Gerardo, Michel Lang, Mariano Barriendos, M. Carmen Llasat, Felix Francés, Taha Ouarda, Varyl Thorndycraft, et al. 2004. "Use of Systematic, Palaeoflood and Historical Data for the Improvement of Flood Risk Estimation. Review of Scientific Methods." *Natural Hazards*. https://doi.org/10.1023/B:NHAZ.0000024895.48463.eb.

Blöschl, Günter, Andrea Kiss, Alberto Viglione, Mariano Barriendos, Oliver Böhm, Rudolf Brázdil, Denis Coeur, et al. 2020. “Current European Flood-Rich Period Exceptional Compared with Past 500 Years.” *Nature* 583 (7817). https://doi.org/10.1038/s41586-020-2478-3.

Brázdil, Rudolf, Zbigniew W. Kundzewicz, and Gerardo Benito. 2006. “Historical Hydrology for Studying Flood Risk in Europe.” *Hydrological Sciences Journal* 51: 739–64. https://doi.org/10.1623/hysj.51.5.739.

Brönnimann, S., P. Stucki, J. Franke, V. Valler, Y. Brugnara, R. Hand, L. C. Slivinski, et al. 2022. “Influence of Warming and Atmospheric Circulation Changes on Multidecadal European Flood Variability.” *Climate of the Past*. https://doi.org/10.5194/cp-18-919-2022.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, et al. 2011. “The Twentieth Century Reanalysis Project.” *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.776.

Cooley, D., D. Nychka, and P. Naveau. 2007. “Bayesian Spatial Modeling of Extreme Precipitation Return Levels.” *Journal of the American Statistical Association* 102 (479).

Datta, Abhirup, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand. 2016. “Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets.” *Journal of the American Statistical Association* 111 (514). https://doi.org/10.1080/01621459.2015.1044091.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. “Model-Based Geostatistics.” *Journal of the Royal Statistical Society Series C-Applied Statistics* 47.

Galiana, Claire, Celine Perherin, Morgane Rocher, and Jean-Luc Souladadie. 2017. “Collecte d’informations Sur Le Terrain Suite à Une Inondation.” Cerema, Bron. https://doc.cerema.fr/doc/SYRACUSE/15907.

Gaume, E., L. Gaál, A. Viglione, J. Szolgay, S. Kohnová, and G. Blöschl. 2010. “Bayesian MCMC Approach to Regional Flood Frequency Analyses Involving Extraordinary Flood Events at Ungauged Sites.” *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2010.01.008.

Gelman, Andrew, and Donald B. Rubin. 1992. “Inference from Iterative Simulation Using Multiple Sequences.” *Statistical Science* 7 (4). https://doi.org/10.2307/2246093.

Giuntoli, I., B. Renard, and M. Lang. 2012. “Floods in France.” In *Changes in Flood Risk in Europe.*, edited by Z. W. Kundzewicz. IAHS Press.

Jin, Minghui, and Jery R. Stedinger. 1989. “Flood Frequency Analysis with Regional and Historical Information.” *Water Resources Research*. https://doi.org/10.1029/WR025i005p00925.

Koenig, Todd A., Jennifer L. Bruce, Jim O’Connor, Benton D. McGee, Robert R. Holmes Jr., Ryan Hollins, Brandon T. Forbes, et al. 2016. “Identifying and Preserving High-Water Mark Data.” Report 3-A24. Reston, VA: USGS. https://doi.org/10.3133/tm3A24.

Laio, F., and S. Tamea. 2007. “Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables.” *Hydrology and Earth System Sciences* 11 (4).

Machado, M. J., B. A. Botero, J. López, F. Francés, A. Díez-Herrero, and G. Benito. 2015. “Flood Frequency Analysis of Historical Flood Data Under Stationary and Non-Stationary Modelling.” *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-19-2561-2015.

Maraun, D., T. J. Osborn, and H. W. Rust. 2011. “The Influence of Synoptic Airflow on UK Daily Precipitation Extremes. Part I: Observed Spatio-Temporal Relationships.” *Climate Dynamics*. https://doi.org/10.1007/s00382-009-0710-9.

Merz, R., and G. Bloschl. 2008. “Flood Frequency Hydrology: 1. Temporal, Spatial, and Causal Expansion of Information.” *Water Resources Research* 44. https://doi.org/10.1029/2007WR006744.

Michele, C.De, and G. Salvadori. 2002. “On the Derived Flood Frequency Distribution: Analytical Formulation and the Influence of Antecedent Soil Moisture Condition.” *Journal of Hydrology*. https://doi.org/10.1016/S0022-1694(02)00025-2.

Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay. 2005. “Flood Frequency Analysis on the Ardeche River Using French Documentary Sources from the Last Two Centuries.” *Journal of Hydrology* 313. https://doi.org/10.1016/j.jhydrol.2005.02.011.

Neppel, L., B. Renard, M. Lang, P. A. Ayral, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K. Pobanz, and F. Vinet. 2010. “Flood Frequency Analysis Using Historical Data: Accounting for Random and Systematic Errors.” *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*. https://doi.org/10.1080/02626660903546092.

Nguyen, C.C., E. Gaume, and O. Payrastre. 2014. “Regional Flood Frequency Analyses Involving Extraordinary Flood Events at Ungauged Sites: Further Developments and Validations.” *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2013.09.058.

Payrastre, O., E. Gaume, and H. Andrieu. 2011. “Usefulness of Historical Information for Flood Frequency Analyses: Developments Based on a Case Study.” *Water Resources Research*. https://doi.org/10.1029/2010WR009812.

Pichard, G., G. Arnaud-Fassetta, V. Moron, and E. Roucaute. 2017. “Hydro-Climatology of the Lower Rhône Valley: Historical Flood Reconstruction (AD 1300–2000) Based on Documentary and Instrumental Sources.” *Hydrological Sciences Journal*. https://doi.org/10.1080/02626667.2017.1349314.

Piotte, Olivier, Céline Boura, Anaïs Cazaubon, Carine Chaléon, Dominique Chambon, Gwenaël Guillevic, Fabien Pasquet, Céline Perherin, and Emmanuel Raimbault. 2016. "Collection, Storage and Management of High-Water Marks Data: Praxis and Recommendations." Edited by M. Lang, F. Klijn, and P. Samuels. *E3S Web of Conferences*. https://doi.org/10.1051/e3sconf/20160716003.

Prosdocimi, I., T. R. Kjeldsen, and J. D. Miller. 2015. "Detection and Attribution of Urbanization Effect on Flood Extremes Using Nonstationary Flood-Frequency Models." *Water Resources Research*. https://doi.org/10.1002/2015WR017065.

Prosdocimi, Ilaria, and Thomas Kjeldsen. 2021. "Parametrisation of Change-Permitting Extreme Value Models and Its Impact on the Description of Change." *Stochastic Environmental Research and Risk Assessment*, no. 2. https://doi.org/10.1007/s00477-020-01940-8.

Read, Laura K., and Richard M. Vogel. 2015. "Reliability, Return Periods, and Risk Under Nonstationarity." *Water Resources Research*. https://doi.org/10.1002/2015WR017089.

Reis, D. S., and J. R. Stedinger. 2005. "Bayesian MCMC Flood Frequency Analysis with Historical Information." *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2005.02.028.

Renard, Benjamin. 2021a. "STooDs-Tools/RSTooDs: RSTooDs Package V0.1.1." https://doi.org/10.5281/zenodo.5075760.

Renard, Benjamin. 2021b. "STooDs-Tools/STooDs: STooDs Engine V0.1.0." https://doi.org/10.5281/ZENODO.5075586.

Renard, Benjamin. 2022. "Reconstruction of Flood Probabilities in France, 1705-2015." Zenodo. https://doi.org/10.5281/ZENODO.6793501.

Renard, Benjamin. 2023. "Codes and Data Related to the Article: Renard (2023). Use of a National Flood Mark Database to Estimate Flood Hazard in the Distant Past. Hydrological Sciences Journal." Zenodo. https://doi.org/10.5281/ZENODO.7853033.

Renard, B., and M. Thyer. 2019. “Revealing Hidden Climate Indices from the Occurrence of Hydrologic Extremes.” *Water Resources Research*. https://doi.org/10.1029/2019WR024951.

Renard, B., M. Thyer, D. McInerney, D. Kavetski, M. Leonard, and S. Westra. 2021. “A Hidden Climate Indices Modeling Framework for Multi-Variable Space-Time Data.” *Water Resources Research*. https://doi.org/10.1029/2021WR030007.

Sabourin, Anne, and Benjamin Renard. 2015. “Combining Regional Estimation and Historical Floods: A Multivariate Semiparametric Peaks-over-Threshold Model with Censored Data.” *Water Resources Research*. https://doi.org/10.1002/2015WR017320.

Stedinger, J. R., and T. A. Cohn. 1986. “Flood Frequency-Analysis with Historical and Paleoflood Information.” *Water Resources Research* 22 (5): 785–93.

Steinschneider, Scott, Michelle Ho, A. Park Williams, Edward R. Cook, and Upmanu Lall. 2018. “A 500-Year Tree Ring-Based Reconstruction of Extreme Cold-Season Precipitation and Number of Atmospheric River Landfalls Across the Southwestern United States.” *Geophysical Research Letters*. https://doi.org/10.1029/2018GL078089.

Sun, X., M. Thyer, B. Renard, and M. Lang. 2014. “A General Regional Frequency Analysis Framework for Quantifying Local-Scale Climate Effects: A Case Study of ENSO Effects on Southeast Queensland Rainfall.” *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2014.02.025.

Whan, Kirien, and Francis Zwiers. 2017. “The Impact of ENSO and the NAO on Extreme Winter Precipitation in North America in Observations and Regional Climate Models.” *Climate Dynamics*. https://doi.org/10.1007/s00382-016-3148-x.

Wilhelm, B., W. Rapuc, B. Amann, F. S. Anselmetti, F. Arnaud, J. Blanchet, A. Brauer, et al. 2022. “Impact of Warmer Climate Periods on Flood Hazard in the European Alps.” *Nature Geoscience*. https://doi.org/10.1038/s41561-021-00878-y.

Xiong, Bin, Lihua Xiong, Shenglian Guo, Chong-Yu Xu, Jun Xia, Yixuan Zhong, and Han Yang. 2020. “Nonstationary Frequency Analysis of Censored Data: A Case Study of the Floods in the Yangtze River from 1470 to 2017.” *Water Resources Research*. https://doi.org/10.1029/2020WR027112.

List of Tables

List of Figures

Sharpness at all stations, as quantified by the range of estimated probabilities. The two stations in (a) and (b) are denoted by crossed circles.

Figure 7. Probability maps estimated with model M4: for each selected year, the map on the left shows occurrences of flood marks and site status, the map on the right shows the probability of exceeding a 10-year flood at stations.

Figure 8. Cross-validation by means of datasets recreating for the decade 1981-1990 the information that was available in 1855 (V1) and 1710 (V2). (a) First three HCIs (90% uncertainty intervals) for a few years bracketing the validation period 1981-1990 (dashed lines). (b) Probability maps for 3 selected years in 1981-1990. Small black dots denote stations where a 10-year flood did occur. (c) Reliability diagrams showing the empirical cdf of $F_{s,t}(q_{s,t})$ values. A curve close to the diagonal denotes uniform values between 0 and 1 and therefore reliable estimates.

Figure 9. Sensitivity analysis for model M4: 90% posterior intervals for the first HCI and its effects are compared between the original dataset (reference) and the three modified datasets (see text for details). First row: HCI time series; second row: intercept and HCI effect on flood mark occurrences at sites; third row: log-location intercept and HCI effect on flood peaks at sites.

Supplementary material

Supplementary Figure 1. HCIs numbered 4 to 6 and their effects at stations and sites for model M4. In the left panels (HCI time series), the line is the posterior median, the dark area is the 90% uncertainty band. For comparison, the light area is the 90% uncertainty band for model M3, which only applies to the period 1904-2015

Supplementary Video 1. Probability maps estimated with model M4 for all years 1705-2015.

The map on the left shows occurrences of flood marks and site status, the map on the right shows the probability of exceeding a 10-year flood at stations.

Table 1. Prior distributions for the hyper-parameters of spatial processes.

Spatial process	Applies to equation(s)	Constant-mean α	Sill η_0	Range η_1 [km]
Log-location μ or μ_0	(2) and (4)	$\mathcal{N}(log(20), 2^2)$	$log\mathcal{N}(log(2), 1^2)$	$log\mathcal{N}(log(500), 1^2)$
Log-CV γ	(2) and (4)	$\mathcal{N}(log(0.5), 0.5^2)$	$log\mathcal{N}(log(0.5), 1^2)$	$log\mathcal{N}(log(500), 1^2)$
Shape ξ	(2) and (4)	$\mathcal{N}(0, 0.15^2)$	$log\mathcal{N}(log(0.15), 1^2)$	$log\mathcal{N}(log(500), 1^2)$
HCI effects at stations μ_k	(2)(4)	$\mathcal{N}(0, 0.25^2)$	$log\mathcal{N}(log(0.5), 1^2)$	$log\mathcal{N}(log(500), 1^2)$
Intercept λ_0	(5)	$\mathcal{U}(-\infty, \infty)$	$\mathcal{U}(0, \infty)$	Not applicable
HCI effects at sites λ_k	(5)	$\mathcal{N}(0, 1^2)$	$log\mathcal{N}(log(1), 1^2)$	$log\mathcal{N}(log(500), 1^2)$

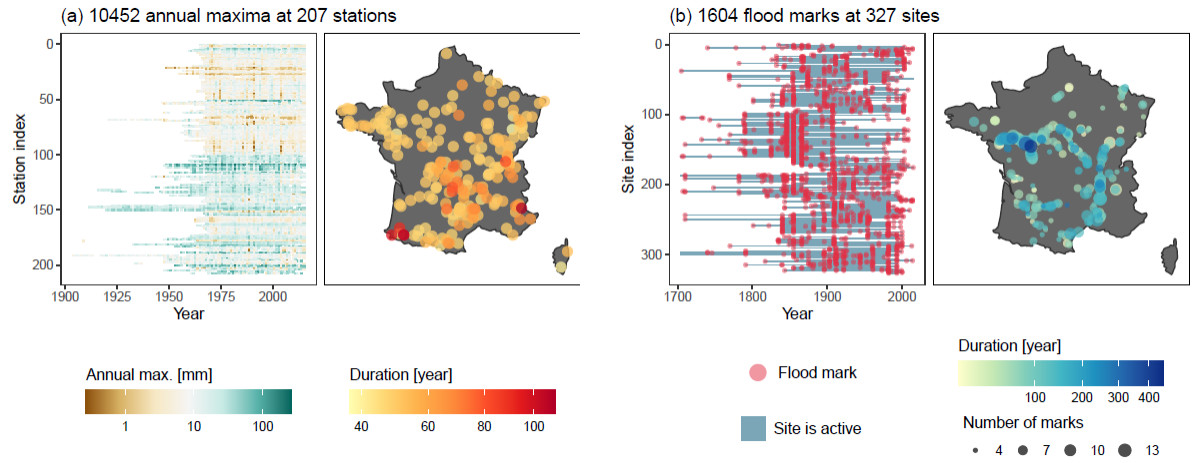


Figure 1. Data used in this article: (a) annual maxima at hydrometric stations, 1904-2015; (b) flood marks at flood sites, 1705-2015.

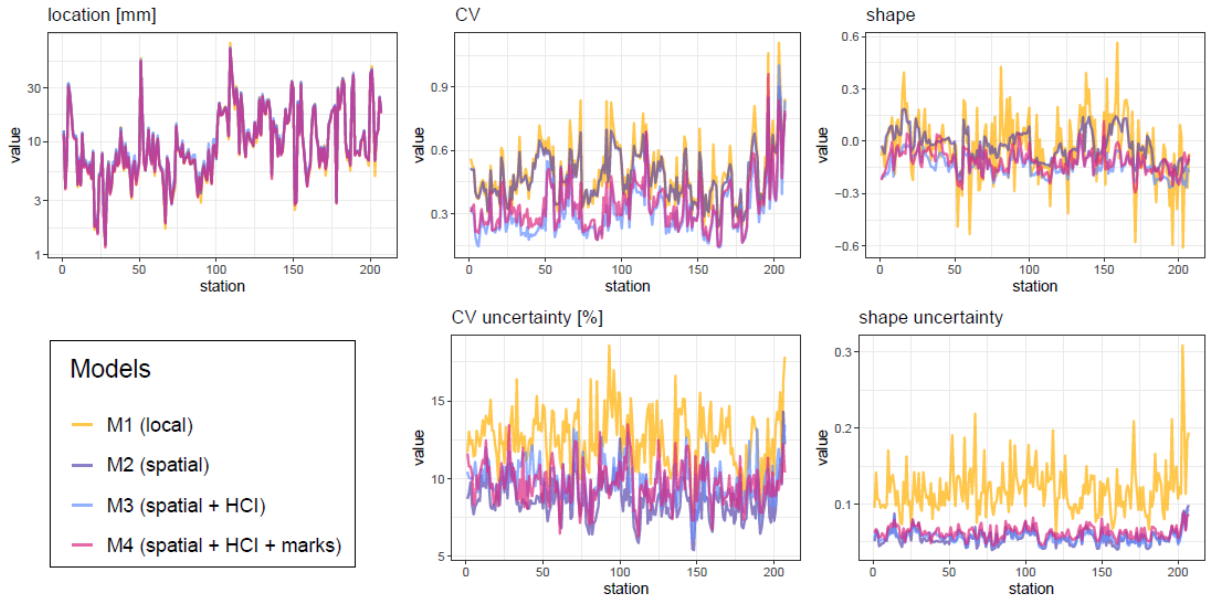


Figure 2. Comparison of GEV parameters at 207 stations: estimated values (posterior mean) for location e^{μ}, CV e^{γ} and shape ξ, and uncertainty (posterior standard deviation) for CV and shape.

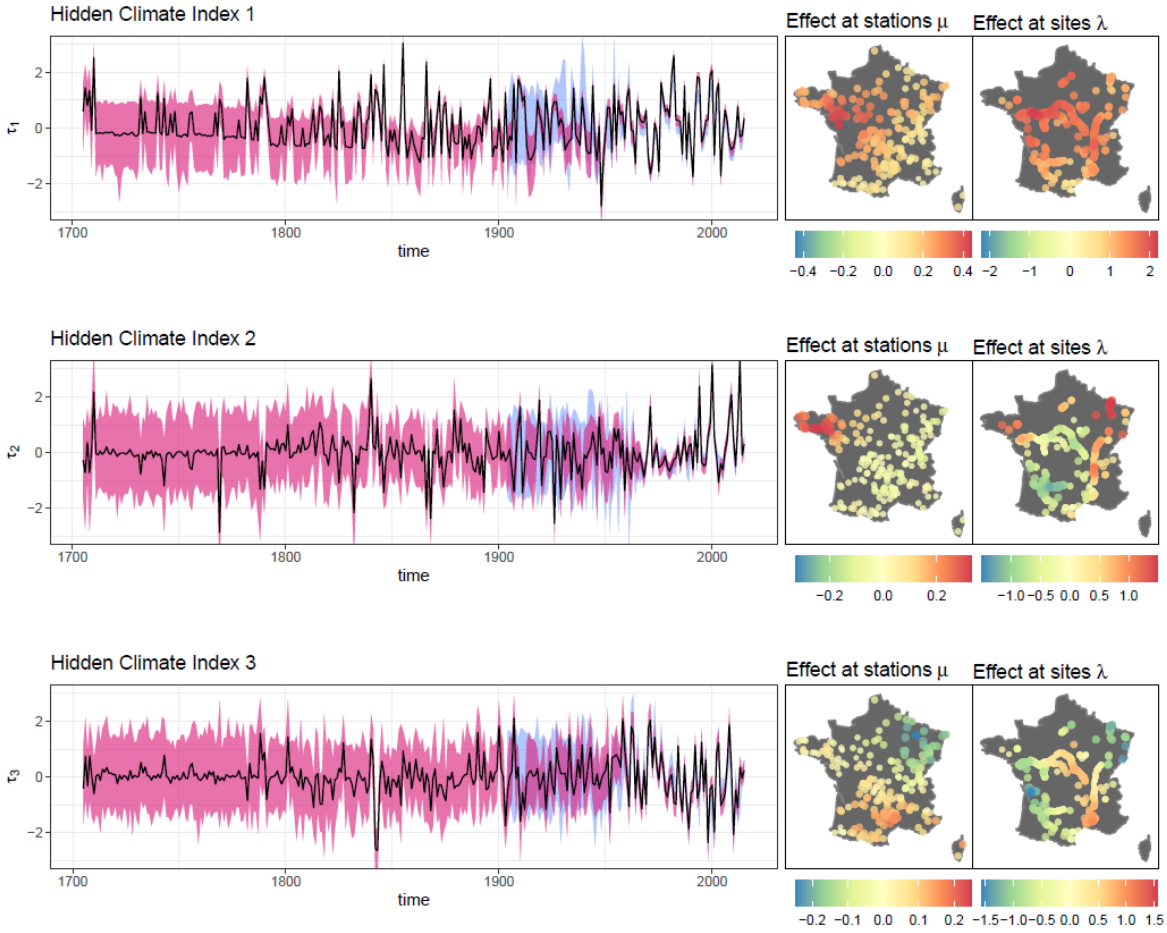


Figure 3. First three HCIs and their effects at stations and sites for model M4. In the left panels (HCI time series), the line is the posterior median and the dark area is the 90% uncertainty band. For comparison, the light area is the 90% uncertainty band for model M3, which only applies to the period 1904-2015.

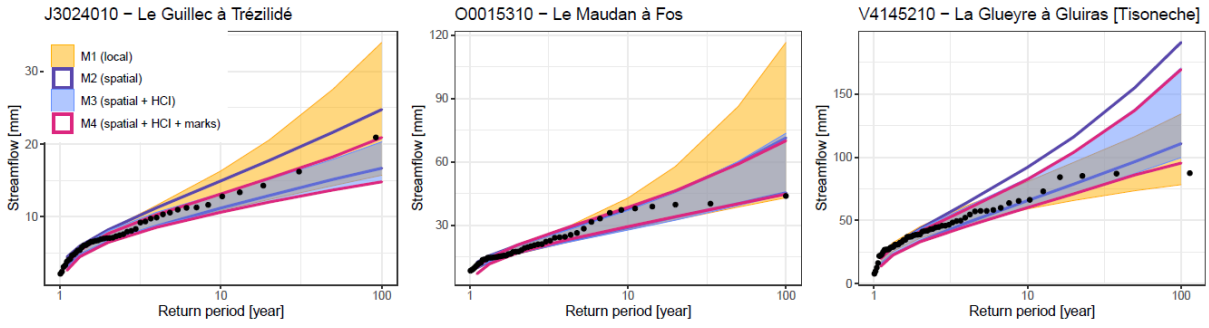


Figure 4. Comparison of 90% uncertainty intervals around quantile curves for the four models at three selected stations.

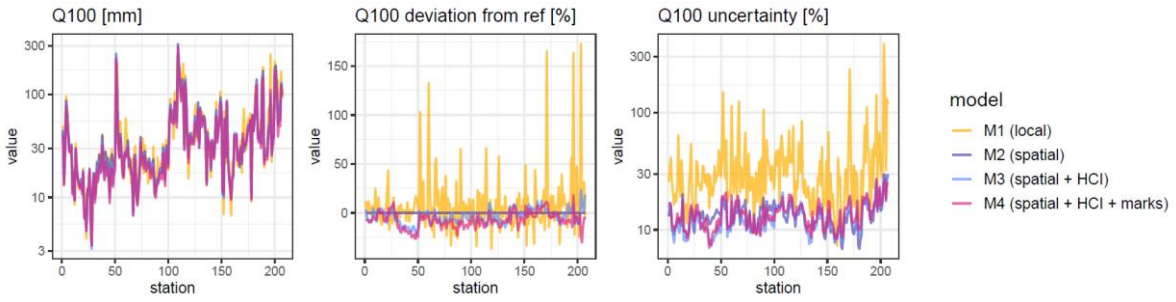


Figure 5. Comparison of 100-year floods at 207 stations: estimated values (posterior mean), estimated values expressed as a deviation from model M2 used as a reference, and uncertainty (posterior standard deviation).

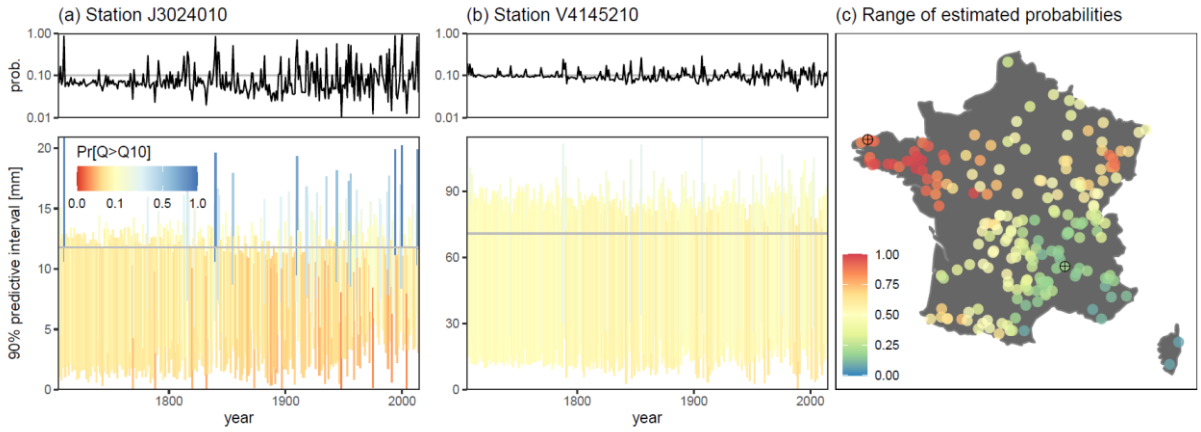


Figure 6. Time-varying distribution of flood peaks, 1705-2015. (a) Station J3024010 located in northwest France. In the bottom panel, vertical lines represent 90% probability intervals from the predictive distribution, the horizontal line is the 10-year flood Q10 (estimated from the marginal distribution of Figure 4). The top panel shows the corresponding probability of exceeding Q10. (b) Same as (a) for station V4145210 located in southeast France. (c) Sharpness at all stations, as quantified by the range of estimated probabilities. The two stations in (a) and (b) are denoted by crossed circles.

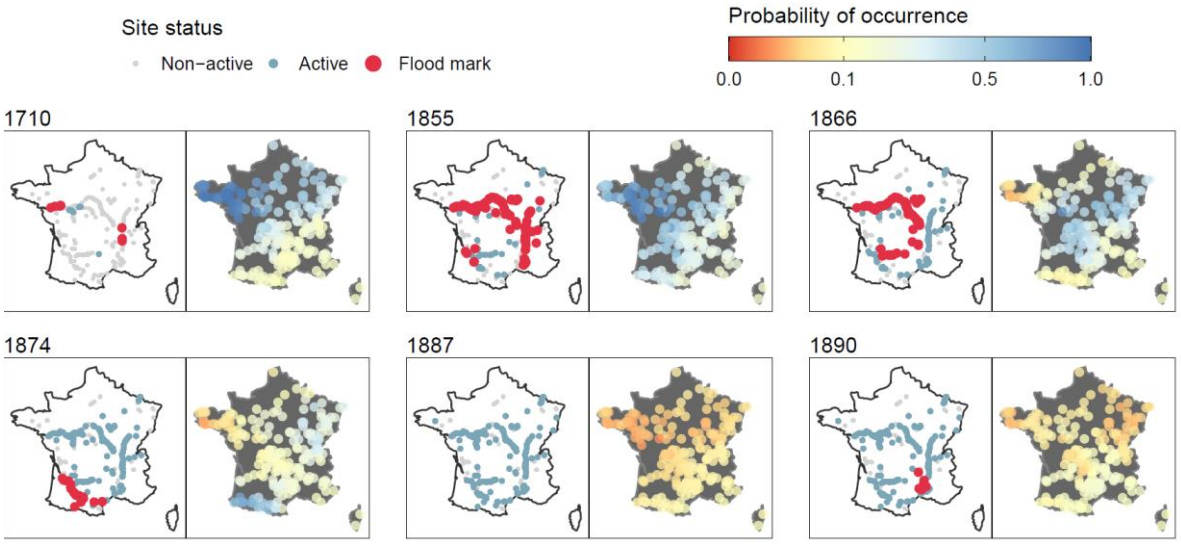


Figure 7. Probability maps estimated with model M4: for each selected year, the map on the left shows occurrences of flood marks and site status, the map on the right shows the probability of exceeding a 10-year flood at stations.

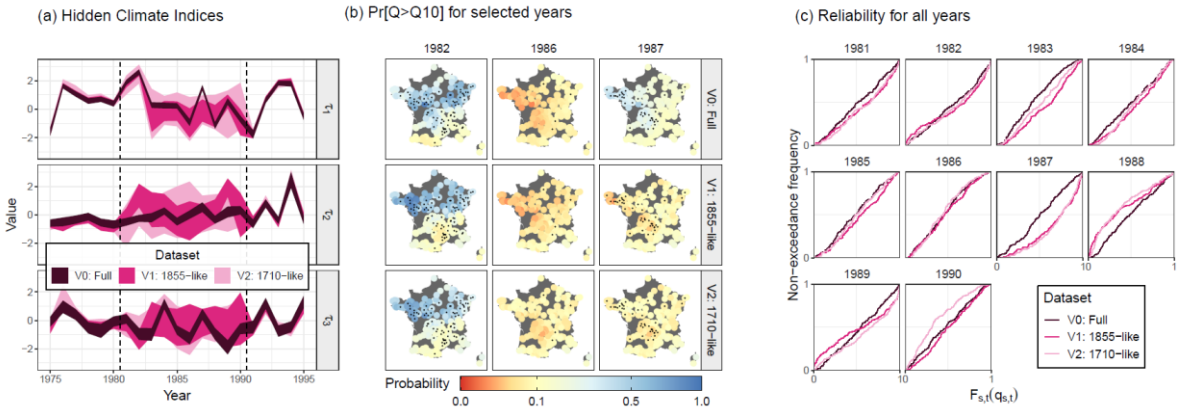


Figure 8. Cross-validation by means of datasets recreating for the decade 1981-1990 the information that was available in 1855 (V1) and 1710 (V2). (a) First three HCIs (90% uncertainty intervals) for a few years bracketing the validation period 1981-1990 (dashed lines). (b) Probability maps for 3 selected years in 1981-1990. Small black dots denote stations where a 10-year flood did occur. (c) Reliability diagrams showing the empirical cdf of $F_{s,t}(q_{s,t})$ values. A curve close to the diagonal denotes uniform values between 0 and 1 and therefore reliable estimates.

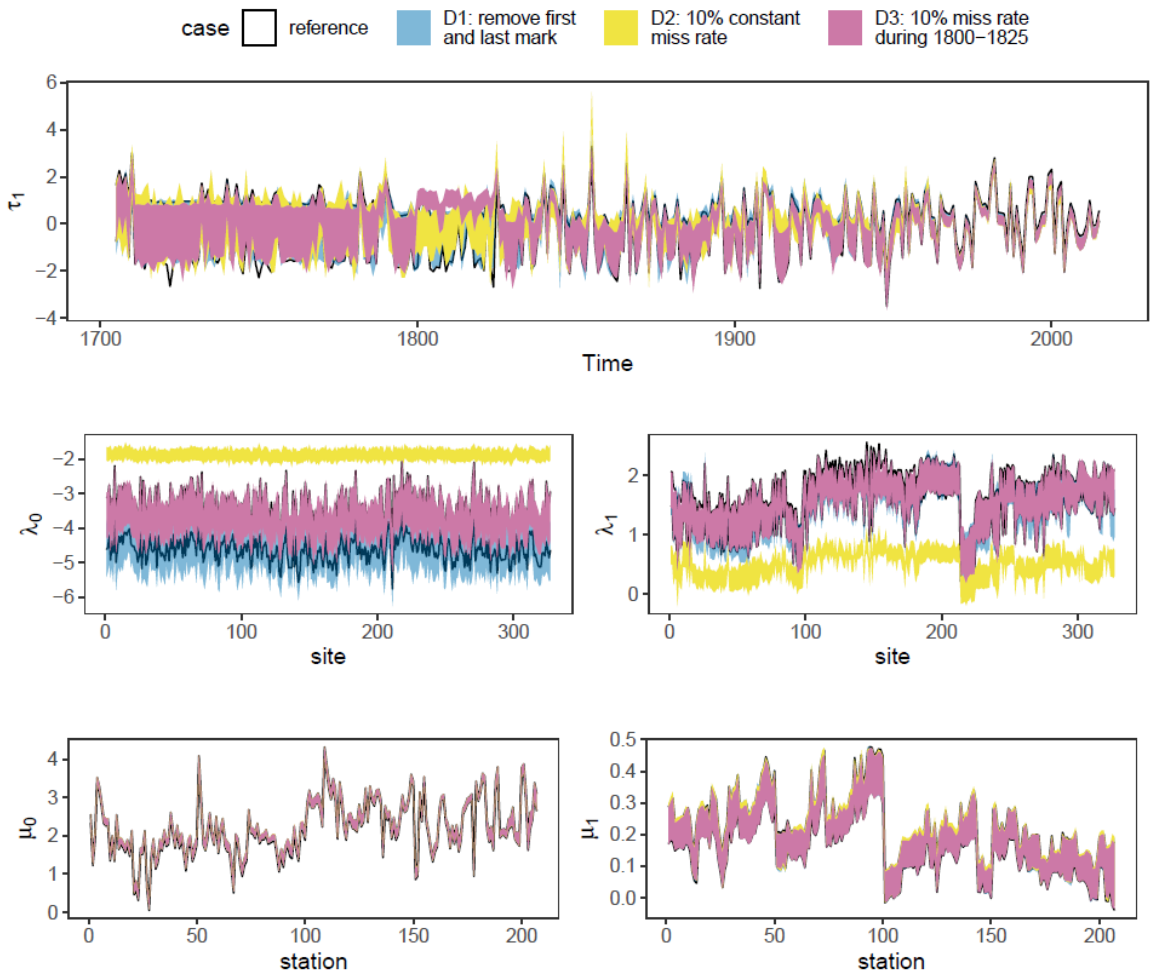
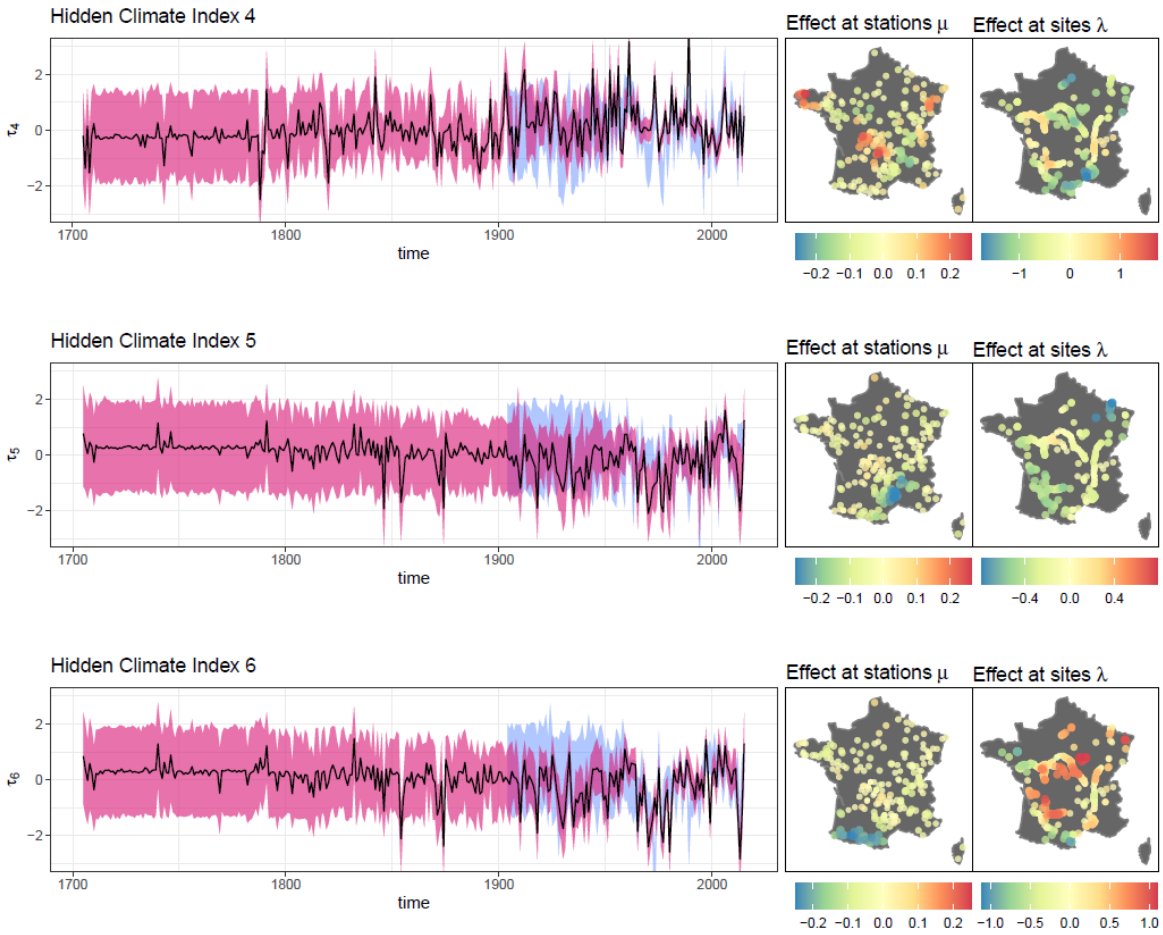


Figure 9. Sensitivity analysis for model M4: 90% posterior intervals for the first HCI and its effects are compared between the original dataset (reference) and the three modified datasets (see text for details). First row: HCI time series; second row: intercept and HCI effect on flood mark occurrences at sites; third row: log-location intercept and HCI effect on flood peaks at sites.



Supplementary Figure 1. HCIs numbered 4 to 6 and their effects at stations and sites for model M4. In the left panels (HCI time series), the line is the posterior median, the dark area is the 90% uncertainty band. For comparison, the light area is the 90% uncertainty band for model M3, which only applies to the period 1904-2015.