APPROXIMATE BAYESIAN INFERENCE FOR ANALYSIS OF SPATIOTEMPORAL FLOOD FREQUENCY DATA

BY ÁRNI V. JÓHANNESSON[1,a], STEFAN SIEGERT[2,c], RAPHAËL HUSER[3,d], HAAKON BAKKA[4,e] AND BIRGIR HRAFNKELSSON[1,b]

[1]*Department of Mathematics, Faculty of Physical Sciences, School of Engineering and Natural Sciences, University of Iceland,* [a]*avj2@hi.is,* [b]*birgirhr@hi.is*

[2]*Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter,* [c]*s.siegert@exeter.ac.uk*

[3]*Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST),* [d]*raphael.huser@kaust.edu.sa*

[4]*Department of Mathematics, University of Oslo,* [e]*bakka@r-inla.org*

Extreme floods cause casualties and widespread damage to property and vital civil infrastructure. Predictions of extreme floods, within gauged and ungauged catchments, is crucial to mitigate these disasters. In this paper a Bayesian framework is proposed for predicting extreme floods, using the generalized extreme-value (GEV) distribution. A major methodological challenge is to find a suitable parametrization for the GEV distribution when multiple covariates and/or latent spatial effects are involved and a time trend is present. Other challenges involve balancing model complexity and parsimony, using an appropriate model selection procedure and making inference based on a reliable and computationally efficient approach. We here propose a latent Gaussian modeling framework with a novel multivariate link function designed to separate the interpretation of the parameters at the latent level and to avoid unreasonable estimates of the shape and time trend parameters. Structured additive regression models, which include catchment descriptors as covariates and spatially correlated model components, are proposed for the four parameters at the latent level. To achieve computational efficiency with large datasets and richly parametrized models, we exploit a highly accurate and fast approximate Bayesian inference approach which can also be used to efficiently select models separately for each of the four regression models at the latent level. We applied our proposed methodology to annual peak river flow data from 554 catchments across the United Kingdom. The framework performed well in terms of flood predictions for both ungauged catchments and future observations at gauged catchments. The results show that the spatial model components for the transformed location and scale parameters as well as the time trend are all important, and none of these should be ignored. Posterior estimates of the time trend parameters correspond to an average increase of about 1.5% per decade with range 0.1% to 2.8% and reveal a spatial structure across the United Kingdom. When the interest lies in estimating return levels for spatial aggregates, we further develop a novel copula-based postprocessing approach of posterior predictive samples in order to mitigate the effect of the conditional independence assumption at the data level, and we demonstrate that our approach indeed provides accurate results.

1. Introduction. Flood predictions are of great importance for the design, management, operation, and protection of vital infrastructure. To aid flood predictions, gauging stations have been built to observe and monitor streams, wells, lakes, canals, and reservoirs. Flood predictions are typically based on analyses of past observed extremes, and they usually rely

on fitting statistical models for extreme floods within gauged catchments. Predictions for ungauged catchments are made possible by using appropriate covariates whose coefficients are learned from gauged catchments and, sometimes, also by exploiting spatial correlation across catchments, using suitable geostatistical models.

One of the first approaches in flood frequency analysis was the index flood method, introduced in Dalrymple (1960). The index flood method is a statistical method that was designed to handle cases, where little or no site-specific information is available, by borrowing strength from similar catchments. The index flood method consists of two steps: (1) Regionalization, which identifies geographically and climatologically similar regions and (2) specification of a regional standardized flood frequency curve for the T-year return level. Various developments have been made to improve the index flood method (GREHYS (1996), Hosking, Wallis and Wood (1985)), and various Bayesian methods have been introduced (Cunnane and Nash (1974), Kuczera (1999), Martins and Stedinger (2000), Rosbjerg and Madsen (1995)). In the United Kingdom (U.K.), the analysis of floods is commonly tackled with the pooling group method (Kjeldsen and Jones (2009a), Robson and Reed (1999)), which is a combination of: (i) the index flood method based on the use of L-moments (Hosking and Wallis (2005)) and (ii) the region of influence (ROI) approach (Burn (1990)). Further related statistical approaches for the modeling, prediction, and uncertainty assessment of annual peak flow data within gauged and ungauged catchments were proposed by, for example, Kjeldsen and Jones (2006, 2009b), Kjeldsen (2010), and Kjeldsen, Ahn and Prosdocimi (2017). More recently, Thorarinsdottir et al. (2018) analyzed annual maximum floods from 203 streamflow stations in Norway, based on a Bayesian latent Gaussian model (LGM) with a generalized extreme-value (GEV) distribution at the data level, which is motivated by extreme value theory (see, e.g., Davison, Padoan and Ribatet (2012) and Davison and Huser (2015)). They used a multivariate link function, specified by identity links for the location and shape parameters, and a logarithmic link for the scale parameter. Their model is fairly simple in the sense that it does not contain any spatially correlated random effects (a priori) and ignores the presence of a time trend, and inference is performed by drawing samples from the posterior distribution, using a Markov chain Monte Carlo (MCMC) algorithm. Bayesian models for extremes with spatially correlated random effects date back to Casson and Coles (1999). More recently, models of this sort were also considered by Huerta and Sansó (2007), Davison, Padoan and Ribatet (2012), Geirsson, Hrafnkelsson and Simpson (2015) and Dyrrdal et al. (2015) in studies of extreme ozone concentrations and/or precipitation, but the datasets in these papers were limited to 19, 36, 40, and 59 stations only, respectively; see, also, Hrafnkelsson, Morris and Baladandayuthapani (2012), who used a similar approach for modeling annual minimum and maximum temperatures observed at 72 stations in Iceland. Sang and Gelfand (2009) modeled extreme precipitation observed on a large regular grid with 1078 grid cells by taking advantage of conditional autoregressive (CAR) priors for the location and scale parameters, along with an additional temporal trend term for the location parameter, and assuming a constant shape parameter. Sang and Gelfand (2010) then relaxed the conditional independence assumption by using a Gaussian dependence structure at the data level. In the same vein, Cooley and Sain (2010) and Jalbert et al. (2017) modeled extreme precipitation data arising from climate model outputs over large grids by exploiting their gridded structure and using intrinsic Gaussian Markov random field (IGMRF) priors at the latent level of the LGM.

In this paper we propose a novel statistical model for annual peak flow data from 554 catchments across the U.K. that efficiently exploits spatial information about the catchments. Similarly to Huerta and Sansó (2007), Sang and Gelfand (2009, 2010), Davison, Padoan and Ribatet (2012), Hrafnkelsson, Morris and Baladandayuthapani (2012), Dyrrdal et al. (2015), Geirsson, Hrafnkelsson and Simpson (2015), Jalbert et al. (2017), and Thorarinsdottir et al. (2018), our proposed model is an LGM that assumes a GEV distribution at the data level with

site-specific location (here, with intercept and time trend), scale, and shape parameters, but its detailed hierarchical structure and our inference procedure have crucial differences that make it more attractive and amenable to complex and large datasets. Moreover, contrary to Sang and Gelfand (2009, 2010), Cooley and Sain (2010), and Jalbert et al. (2017), the data in our application are not regularly located which makes it more challenging from a computational perspective. At the latent level we follow Rigby and Stasinopoulos (2005) and use a generalized additive specification for the transformed location intercept, scale, shape, and trend parameters of the data density. Our proposed model at the latent level contains catchment descriptors as covariates, for example, catchment area and average annual rainfall, along with spatial model components for the transformed location intercept and scale parameters, which are specified as approximate solutions to stochastic partial differential equations (SPDEs, see Lindgren, Rue and Lindström (2011)). These latent SPDE-based model components can capture complex spatially structured residuals more flexibly than the CAR priors of Sang and Gelfand (2009, 2010) and also yield sparse precision matrices which accelerates computations with large spatial datasets. Because the four parameters of the data level (i.e., the GEV location intercept and time trend, scale, and shape parameters) are transformed to the latent level, the proposed model is an LGM with a multivariate link function, referred to here as an extended LGM (Geirsson et al. (2020), Hrafnkelsson et al. (2021)).

The multivariate link function consists of four transformations. The logarithmic transformation is applied to the GEV location parameter, μ. The other three transformations have not been presented before in this context and can thus be considered novel. The shape parameter, ξ, and the time trend parameter, Δ, are assumed to take values only within specified intervals, and their transformations reflect that. The specified ranges for ξ and Δ are selected to eliminate unnecessarily small or large parameter values where the aim is twofold: (i) reducing parameter uncertainty and (ii) stabilizing sitewise likelihood fits. Our proposed link function keeps an intuitive interpretation and desirable properties for the shape parameter while producing reliable estimates when several covariates are involved. Moreover, to prevent overfitting when multiple spatial random effects are involved, we specify relatively informative shrinkage priors for the hyperparameters in order to penalize complex models that depart too much from a simpler counterpart (Simpson et al. (2017)).

In the frequentist framework the `gevlss` function of the R package `mgcv` may be used to fit generalized extreme value location, scale, and shape models. More recently, a fast frequentist method, based on restricted maximum likelihood, was developed for models similar to the model proposed in this paper (Youngman (2019)). Alternatively, in the Bayesian framework, MCMC methods may be exploited. However, instead of inferring our proposed Bayesian extended LGM with an "exact" MCMC sampler, such as the LGM split sampler in Geirsson et al. (2020), we here exploit an accurate Gaussian approximation to the posterior density adapted from the general methodology recently proposed by Hrafnkelsson et al. (2021), referred to as "Max-and-Smooth," in order to better handle the high dimensionality of the data and to improve mixing and convergence of the MCMC chains. The LGM split sampler has generally quite good mixing properties, but it needs to find the mode and Hessian of the conditional posterior of the sitewise parameters in each iteration, while Max-and-Smooth needs to find the mode and Hessian of the likelihood function for the same parameters only once, that is, at the beginning of the simulation. Thus, Max-and-Smooth is computationally more efficient than the LGM split sampler when the number of sites is large while retaining high accuracy. In this paper it is the first time that Max-and-Smooth is applied and carefully validated in the context of large-scale, heavy-tailed, nonstationary, spatially referenced data. For reproducibility and to facilitate future applications of Max-and-Smooth, we make our code available in the Supplementary Material (Jóhannesson et al. (2022b)) along with a simple working example; see also the GitHub link https://github.com/ridivinra/GEV_LGM.

This fast Gaussian-based approximation of the posterior density is also used to break the selection of catchment descriptors down such that the descriptors for one of the four transformed parameters at the latent level are selected separately of the other three. This separation makes it possible to apply the integrated nested Laplace approximation (INLA, Rue, Martino and Chopin (2009)) to each of the transformed parameters and a 10-fold cross-validation experiment for the selection of catchment descriptors becomes feasible. We consider this to be a novel aspect of our analysis. Furthermore, the computation of quantiles (i.e., marginal return levels) is straightforward, and their prediction at ungauged sites can be made efficiently and accurately through the posterior predictive distributions, using site-specific covariates.

The spatial and temporal predictive performances of the proposed latent Gaussian model are evaluated in a detailed cross-validation study by comparing it to other natural alternative approaches, based on the GEV distribution, namely, three frequentist models (constant model; sitewise model; response surface model) and three simplified version of the proposed LGM (without trend, without spatial components, without covariates). These new results provide strong support in favor of our approach and clearly indicate that our proposed model performs well and provides significant improvements over other state-of-the-art approaches.

Whilst our main objective in this work is to estimate marginal return levels of river flow at each individual site, it may sometimes be required for planning or regional flood risk assessment to estimate return levels for spatial aggregates. In this case, both marginal distributions and the spatial dependence structure play a key role. Therefore, while the conditional independence assumption that underpins our proposed model at the data level has a minor effect on marginal return level estimates, it does, however, strongly affect the estimation of spatial return levels. To circumvent this limitation, we here further develop a novel copula-based method, which consists in postprocessing posterior predictive samples simulated at observed sites from our model by simply matching their ranks with those of the observed data, in such a way to accurately "correct" the bias and variance in estimated spatial return levels.

The paper is organized as follows. In Section 2 we present the data and covariates used in our statistical analysis. We then describe our specific flood frequency model in Section 3. In Section 4 we present the exact posterior density and its approximation, and we also introduce the method for selecting covariates for our flood frequency model. The results of the model selection, the posterior inference, a detailed cross-validation study to compare our model to natural alternatives as well as our proposed postprocessing approach of posterior predictive samples for estimating spatial return levels are all presented and discussed in Section 5. The paper concludes with a discussion and perspectives for future research in Section 6.

2. Data and exploratory analysis. The data used in this paper come from the U.K. National River Flow Archive which is hosted by the Centre for Ecology & Hydrology on behalf of the Natural Environment Research Council; see National River Flow Archive (2018). The data consist of annual maxima from daily river flow observations, measured at 554 gauging stations that are located across the U.K. The observations date back to 1851, but most of the observations were obtained from 1970 to 2013. A large fraction of the observational sites have between 50 and 80 annual maxima available. The dataset actually contains 958 gauging stations, but only stations that are deemed suitable for pooling and index flood calculations, due to their homogeneous quality, are used in our analysis. These stations are explicitly flagged as such in the original dataset. Figure 1 shows the location of the 554 stations along with the number of annual maxima available at each station. The dataset used in this study is available upon request from the authors.

The Anderson and Darling (1954) goodness-of-fit test, as presented by Stephens (1974), was applied across the observational sites. It revealed that the GEV distribution with a time trend is an appropriate model for the annual maxima data; see Section 2.2 of the Supplementary Material (Jóhannesson et al. (2022a)) for more details. This is no surprise, because
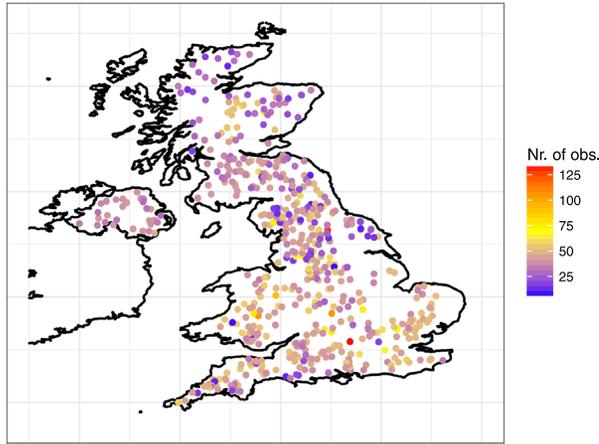
FIG. 1. *Map of the U.K. with the 554 stations used in our flood frequency analysis. The color bar indicates the number of annual maximum flow observations available at each station.*

the GEV distribution is the only possible limiting distribution for renormalized block maxima. The GEV distribution has been applied successfully to a wide range of problems; see, for example, the book of Coles (2001), the review paper by Davison and Huser (2015), and references therein. The GEV cumulative distribution function may be expressed as

$$(1) \qquad \mathrm{GEV}(y \mid \mu, \sigma, \xi) = \begin{cases} \exp[-\{1 + \xi(y - \mu)/\sigma\}_+^{-1/\xi}], & \xi \neq 0; \\ \exp[-\exp\{-(y - \mu)/\sigma\}], & \xi = 0, \end{cases}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ are location, scale, and shape parameters, respectively, and $a_+ = \max(a, 0)$. The support of the GEV distribution is $[-\infty, \mu - \sigma/\xi]$, $[-\infty, +\infty]$, and $[\mu - \sigma/\xi, +\infty]$ when $\xi < 0$, $\xi = 0$, and $\xi > 0$, respectively, and ξ determines the tail heaviness, with a short, light, and heavy upper tail when $\xi < 0$, $\xi = 0$, and $\xi > 0$, respectively.

In order to propose transformations for the GEV parameters μ, σ, and ξ, we first explore their relationships, as suggested by the data. To investigate this, we fit the GEV distribution with a linear time trend in μ to all stations separately by (generalized) maximum likelihood (ML). For completeness, we report these preliminary sitewise ML estimates in Section 2.1 of the Supplementary Material (Jóhannesson et al. (2022a)). The estimated location parameters are all positive and highly right-skewed, and a time trend is often present. Thus, we model the location parameter at location i and time t as $\mu_{it} = \mu_i\{1 + \Delta_i(t - t_0)\}$, where μ_i is a site-specific location intercept at time $t = t_0$, the reference year is set to $t_0 = 1975$, and Δ_i is a site-specific time trend parameter. We use the logarithmic transformation $\psi_i = \log(\mu_i)$ and a scaled logit transformation to transform Δ_i to γ_i such that Δ_i lies in the interval $(-0.008, 0.008)$ for all stations i. This corresponds to 8% increase or decrease in μ_i per decade. Most of the U.K. flood datasets are observed during the years from 1950 to 2013. A trend line with $\Delta = 0.008$ means that μ will change by a factor of 1.63 over the period 1950–2013. The same comparison with $\Delta = -0.008$ gives the factor 0.58. These two extreme scenarios correspond to a substantial increase and decrease over time for a natural process that is driven by meteorological variables. Thus, the interval $[-0.008, 0.008]$ contains plausible values for the trend parameter Δ. Our Bayesian model is fully specified below in Section 3. Moreover, the stationwise ML estimates suggest there is a clear positive linear relationship between the logarithms of the location and the scale parameters; see Section 2.1 of the Supplementary Material (Jóhannesson et al. (2022a)) for more details. Thus, we propose using the transformation $\tau_i = \log(\sigma_i/\mu_i)$. When modeling τ_i with covariates and a spatial effect, we can thus capture the variability in the scale parameter that is not explained by the

location parameter. In Section 3, criteria for the shape parameter suggest that it should lie in the interval $(-0.5, 0.5)$, and we thus propose a suitable function that maps site-specific shape parameter values, ξ_i, from the interval $(-0.5, 0.5)$ to transformed shape parameter values, ϕ_i, which are defined on the whole real line.

The annual peak flow time series contain spatial information, and it turns out that using this information is beneficial for accurate flood predictions. The spatial features in the location, scale, shape, and time trend of the data can be diagnosed by considering the variograms (Cressie (1993)) of the residuals of four basic linear models; see Section 2.1 of the Supplementary Material (Jóhannesson et al. (2022a)) for further details. These variograms clearly indicate that the residuals from the linear models for the ML estimates of ψ_i and τ_i are spatially correlated, while the variograms of the residuals, corresponding to the ML estimates of ϕ_i and γ_i, provide a weak indication of ϕ_i and γ_i being spatially correlated. To determine which of the four parameters need spatial model components, a rigorous model selection procedure is done; see Sections 4.2 and 5.1.

Catchment descriptors have previously been used to model extreme floods in the U.K. (e.g., Kjeldsen (2010), Kjeldsen, Ahn and Prosdocimi (2017)). In Table 1 we present all the catchment descriptors considered in our analysis. We selected a suitable transformation for each catchment descriptor on the basis of to their empirical distribution and their relationship with the ML estimates; see the caption of Table 1. Most of the catchment descriptors were log-transformed. For more details on the catchment descriptors, see Robson and Reed (1999).

The exploratory analysis, briefly outlined above, motivates the specification of our full Bayesian model which is presented in the next section.

3. Model specification. Here, we propose a novel Bayesian extended latent Gaussian model for the annual maximum flow data, where the latent level is composed of four regression models for the transformed location intercept, scale, shape, and time trend parameters of the GEV density. The models for the transformed location intercept and scale parameters also contain spatial model components. We adopt a Bayesian framework to infer the model parameters. The Bayesian approach naturally handles simultaneous inference of multiple parameters and allows us to incorporate latent random effects and prior information.

Let y_{it} be the annual maximum flow at station i in year t. The data from site i are assumed to be independent across years and to follow the GEV distribution with time-dependent location parameter μ_{it}, time-constant scale parameter σ_i, and time-constant shape parameter ξ_i, that is,

$$y_{it} \sim \text{GEV}(\mu_{it}, \sigma_i, \xi_i), \quad \mu_{it}, \sigma_i > 0, \xi_i \in (-0.5, 0.5),$$

where, for all y_{it}, $1 + \xi_i(y_{it} - \mu_{it})/\sigma_i > 0$, $i \in \{1, \ldots, J\}$, J is the number of sites, and $t \in \mathcal{A}_i$ where the set $\mathcal{A}_i$ indexes the years for which observations are available at site i. The location parameter μ_{it} is defined as

$$\mu_{it} = \mu_i\{1 + \Delta_i(t - t_0)\},$$

where μ_i and Δ_i are the location intercept and time trend parameters for site i and $t_0 = 1975$ is a reference year. Note that the domains of each parameter are restricted, according to the findings of our exploratory analysis conducted in Section 2. Moreover, the data are assumed to be conditionally independent given the latent GEV parameters. While this conditional independence assumption may be questioned, it is standard in the literature when the primary goal is to estimate *marginal* return levels, like in this paper, and relaxing it would make the model much more complex and intricate to fit. We here believe that our choice of informative covariates already accounts for most of the spatial variability in the data so that the residual spatial dependence structure is relatively weak in comparison. This is indeed confirmed by our model comparison in Section 5.3. Nevertheless, to rigorously assess the effect of this con-

TABLE 1
Description of the catchment descriptors (covariates). The logarithmic function is used to transform these catchment descriptors, except for BFIHOST, URBEXT, and ASPBAR, which are transformed as BFIHOST^2, $\log(\text{URBEXT}+1)$, *and* $\text{ASPBAR}/100$, *respectively*

			Stats.	
Catchment descriptor	Description	Min	Mean	Max
AREA	Catchment drainage area is the total area, in km², where the river collects water [km²].	1.63	440.65	9930.80
SAAR	Average annual rainfall [mm].	559.00	1114.12	2913.00
FARL	The Flood Attenuation by Reservoirs and Lakes index quantifies the degree of flood attenuation attributable to reservoirs and lakes in the catchment.	0.64	0.96	1.00
BFIHOST	A measure of catchment responsiveness.	0.17	0.49	0.97
FPEXT	Floodplain extent is the fraction of the total area of the catchment that is covered by the 100-year event.	0.00	0.06	0.30
URBEXT	Urban extent is an index that depicts the extent of urban/suburban land cover in the catchment area.	0.00	0.03	0.59
DPLBAR	Is the mean of the distances between each grid node and the catchment outlet [km].	1.30	23.05	139.87
DPSBAR	Mean drainage path slope provides an index of steepness in the catchment, [m/km].	13.30	100.12	441.80
LDP	Longest drainage path [km].	2.21	42.94	286.84
SPRHOST	Standard percentage runoff associated with each HOST soil class.	4.85	37.20	59.85
ASPBAR	The mean aspect or orientation of a catchment in degrees.	1.00	151.95	358.00
ALTBAR	The mean elevation of a catchment area [m].	25.00	220.78	682.00
ASPVAR	The aspect variance in the catchment in degrees.	0.02	0.17	0.59
PROPWET	An index for the proportion of time the soil in the catchment area is wet.	0.23	0.48	0.85

ditional independence assumption, we report theoretical calculations in a simplified setting in Section 4 of the Supplementary Material (Jóhannesson et al. (2022a)), showing that this model simplification, mostly has a moderate effect on the parameters' estimated uncertainty in our context. Moreover, we further show in Section 5.5 how posterior predictive samples from our model can be postprocessed, using a copula-based approach, in order to "correct" their spatial dependence structure which is key when the estimation of return levels for spatial aggregates is required.

Similar to Martins and Stedinger (2000), we use a generalized likelihood function to infer the parameters of the GEV distribution. Martins and Stedinger (2000) introduced a generalized maximum likelihood estimator as an alternative estimator in the frequentist setting for the GEV parameters. The generalized maximum likelihood estimator is based on maxi-

mizing the generalized likelihood function, defined as the product of the likelihood function and a prior density. For flood data, Martins and Stedinger (2000) argued that a beta density shifted to the interval $(-0.5, 0.5)$ with mean 0.10 and standard deviation 0.122 is a reasonable choice; see, also, Cooley and Sain (2010). Here, we adopt a similar approach in our Bayesian inference scheme by replacing the likelihood function with a generalized likelihood function that includes a prior density for each ϕ_i. These individual prior densities at each site i are used in addition to the joint prior density, defined at the latent level for the ϕ parameters. We argue below that ξ should lie in $(-0.5, 0.5)$ which should be reasonable in most environmental applications. In our view the prior distribution of Martins and Stedinger (2000) is slightly too informative for our particular dataset. We opt instead for a prior distribution that is tailored to capture a wider range of tail behaviors than implied by the prior distribution of Martins and Stedinger (2000). Here, in the case of our U.K. flood dataset we select a prior density that is centered around zero with most of its mass on $(-0.3, 0.3)$, as we believe a priori that values of ξ close to -0.5 (very small upper bound) and 0.5 (very heavy tails) are not likely. We, therefore, use a symmetric beta prior density shifted to the interval $(-0.5, 0.5)$ for each ξ_i with parameters $\alpha = 4$ and $\beta = 4$, that is, with mean zero and standard deviation 0.167. This beta prior density for each ξ_i is then transformed to a prior density for each ϕ_i; see Section 1.2 of the Supplementary Material (Jóhannesson et al. (2022a)). This transformation is specified below.

We propose a novel multivariate link function, f, for the parameters of the GEV density, defined as

$$(\psi_i, \tau_i, \phi_i, \gamma_i)^\mathsf{T} := f(\mu_i, \sigma_i, \xi_i, \Delta_i) = \big(\log(\mu_i), \log(\sigma_i/\mu_i), h(\xi_i), d(\Delta_i)\big)^\mathsf{T} \in \mathbb{R}^4,$$

where

$$\gamma_i = d(\Delta_i) = \frac{1}{2}\delta_0\big\{\log(\delta_0 + \Delta_i) - \log(\delta_0 - \Delta_i)\big\}$$

with $\delta_0 = 0.008$. The function d and the selected value of δ_0 ensure that Δ_i lies in the interval $(-0.008, 0.008)$. Most of the floods are observed in the period 1950 to 2013. The values of $\{1 + \Delta_i(t - t_0)\}$ at the beginning and at the end of this period when $\Delta_i = 0.008$ are 0.8 and 1.3 while $\Delta_i = -0.008$ gives the values 1.2 and 0.7. Thus, the interval for Δ_i allows a substantial change over this time period, and we believe that this interval is wide enough. A Gaussian prior density with mean zero and standard deviation $0.5\delta_0$ is selected for γ_i. Hence, there is a priori a 95% chance for γ_i to be in the interval $(-0.008, 0.008)$ which translates into a 95% prior chance for Δ_i to be in the interval $(-0.00609, 0.00609)$. This corresponds to about $\pm 6\%$ change per decade in the location parameter of the GEV distribution. More details about the prior density for Δ_i are given in Section 1.4 in the Supplementary Material (Jóhannesson et al. (2022a)).

The transformed shape parameter is denoted by $\phi = h(\xi)$. Our proposed choice for the function h relies on four natural criteria: (1) the variance of the data distribution describing the floods (i.e., the GEV distribution) should be finite. This leads to the constraint $\xi < 0.5$; (2) the upper bound of the GEV distribution should not be smaller than $\mu + 2\sigma$. This leads to the constraint $\xi > -0.5$; (3) the asymptotic variance of the ML estimator of the transformed shape parameter, ϕ, should vary as little as possible for values of ϕ that correspond to the interval $\xi \in [-0.3, 0.3]$. This ensures that the estimation of parameters for the additive effects in the latent model for ϕ will be affected to a similar degree by the data from each site, regardless of the underlying value of ϕ, which is crucial when multiple covariates are involved in a linear way; (4) the transformation h should be monotonic and such that ϕ is approximately equal to ξ for values of ξ around zero, which facilitates its interpretation. Mathematically, this means that $\phi = h(\xi) \approx \xi$ when ξ is close to zero. Hence,

$$\frac{\mathrm{d}h(\xi)}{\mathrm{d}\xi}\bigg|_{\xi=0} = h'(0) = 1, \quad h(0) = 0.$$
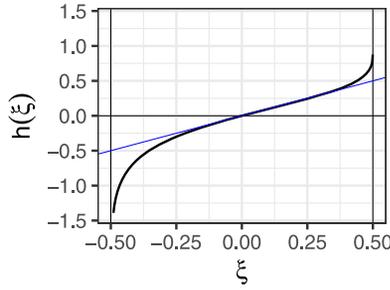
FIG. 2. *The transformation for the shape parameter, $\phi = h(\xi)$ (black curve). The blue line is a reference line with intercept zero and slope one. The black vertical lines show the lower and upper bounds of ξ, -0.5, and 0.5, respectively.*

The constraints stemming from these four criteria motivate the transformation

$$\phi = h(\xi) = a_\phi + b_\phi \log[-\log\{1 - (\xi + 1/2)^{c_\phi}\}],$$

where $c_\phi = 0.8$,

$$b_\phi = -c_\phi^{-1} \log\{1 - (1/2)^{c_\phi}\}\{1 - (1/2)^{c_\phi}\}2^{c_\phi-1} = 0.39563,$$

$$a_\phi = -b_\phi \log[-\log\{1 - (1/2)^{c_\phi}\}] = 0.062376.$$

Furthermore, the inverse of the transformation is

$$(2) \qquad \xi = h^{-1}(\phi) = g(\phi) = \left[1 - \exp\left\{-\exp\left(\frac{\phi - a_\phi}{b_\phi}\right)\right\}\right]^{1/c_\phi} - \frac{1}{2}.$$

The transformation from ξ to ϕ via $h(\cdot)$ is shown in Figure 2. More details are given in Sections 1.2 and 1.3 of the Supplementary Material (Jóhannesson et al. (2022a)).

The generalized likelihood function for ψ_i, τ_i, ϕ_i, and γ_i is the product of the likelihood function for site i and the prior densities for ϕ_i and γ_i. Estimates based on the mode of our proposed generalized likelihood function will simply be referred to as ML estimates hereafter.

At the latent level our proposed regression models for the four transformed parameters ψ, τ, ϕ, and γ, may be expressed in general vector notation as

$$(3) \qquad \begin{aligned} \boldsymbol{\psi} &= X_\psi \boldsymbol{\beta}_\psi + A\boldsymbol{u}_\psi + \boldsymbol{\epsilon}_\psi, \\ \boldsymbol{\tau} &= X_\tau \boldsymbol{\beta}_\tau + A\boldsymbol{u}_\tau + \boldsymbol{\epsilon}_\tau, \\ \boldsymbol{\phi} &= X_\phi \boldsymbol{\beta}_\phi + A\boldsymbol{u}_\phi + \boldsymbol{\epsilon}_\phi, \\ \boldsymbol{\gamma} &= X_\gamma \boldsymbol{\beta}_\gamma + A\boldsymbol{u}_\gamma + \boldsymbol{\epsilon}_\gamma, \end{aligned}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_J)^\mathsf{T}$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)^\mathsf{T}$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)^\mathsf{T}$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\mathsf{T}$ are vectors of the model parameters. The design matrices X_ψ, X_τ, X_ϕ, and X_γ contain covariates, and $\boldsymbol{\beta}_\psi$, $\boldsymbol{\beta}_\tau$, $\boldsymbol{\beta}_\phi$, and $\boldsymbol{\beta}_\gamma$ are the corresponding coefficients. The vectors $\boldsymbol{u}_\psi$, $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\phi$, and $\boldsymbol{u}_\gamma$ are independent spatial random effects with Matérn correlation structure, modeled using finite-dimensional Gaussian Markov random fields (GMRFs) via the stochastic partial differential equation (SPDE) approach of Lindgren, Rue and Lindström (2011). A review of certain types of SPDE models can be found in Bakka et al. (2018) and detailed case studies in Krainski et al. (2019). The SPDE approach requires a triangulated mesh defined over the region of interest used to approximate the continuous-space random field. The mesh constructed for our application is illustrated in Figure 3. The matrix A in (3) describes the projection from the nodes of the triangulation onto the observational sites. The products, $A\boldsymbol{u}_\psi$,
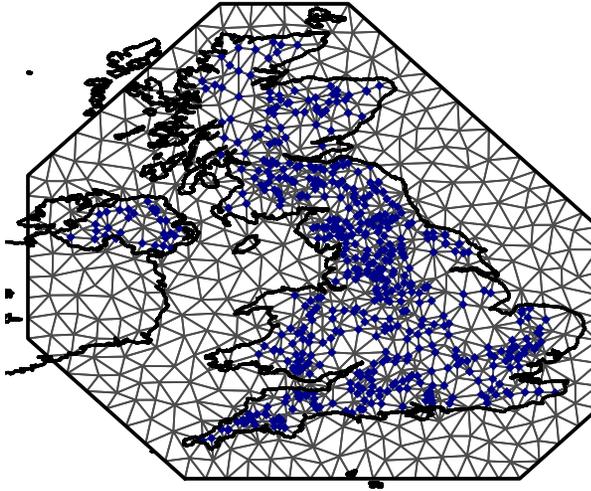
FIG. 3. *Triangulated mesh over the U.K. used to define the SPDE spatial model components at the latent level. The blue dots denote the observational sites.*

$A\boldsymbol{u}_\tau$, $A\boldsymbol{u}_\phi$, and $A\boldsymbol{u}_\gamma$, are thus the effects of the spatially structured model components at the observational sites. The spatial model components $\boldsymbol{u}_\psi$, $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\phi$, and $\boldsymbol{u}_\gamma$, are governed by two hyperparameters each, namely, the range parameters, ρ_l, and the marginal standard deviations, s_l, for $l = \psi, \tau, \phi, \gamma$. The smoothness parameter (usually denoted by ν) of the Matérn correlation structure is here fixed to one in order to produce reasonably smooth realizations. The vectors $\boldsymbol{\epsilon}_\psi$, $\boldsymbol{\epsilon}_\tau$, $\boldsymbol{\epsilon}_\phi$ and $\boldsymbol{\epsilon}_\gamma$ in (3) contain unstructured zero-mean Gaussian model errors with variances $\sigma^2_{\epsilon\psi}$, $\sigma^2_{\epsilon\tau}$, $\sigma^2_{\epsilon\phi}$, and $\sigma^2_{\epsilon\gamma}$, respectively, sometimes called "nugget effects" in classical geostatistics which capture micro-scale variations. That is,

$$\boldsymbol{\epsilon}_\psi \sim \mathrm{N}\big(0, \sigma^2_{\epsilon\psi} I\big), \qquad \boldsymbol{\epsilon}_\tau \sim \mathrm{N}\big(0, \sigma^2_{\epsilon\tau} I\big), \qquad \boldsymbol{\epsilon}_\phi \sim \mathrm{N}\big(0, \sigma^2_{\epsilon\phi} I\big), \qquad \boldsymbol{\epsilon}_\gamma \sim \mathrm{N}\big(0, \sigma^2_{\epsilon\gamma} I\big),$$

where I is the identity matrix. All these error terms are assumed to be mutually independent, both spatially and across parameters.

Combining the four equations in (3) together, the model at the latent level may be rewritten more compactly as

$$(4) \qquad \boldsymbol{\eta} = Z\boldsymbol{v} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\eta} = \big(\boldsymbol{\psi}^\mathsf{T}, \boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\phi}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T}\big)^\mathsf{T},$$
$$\boldsymbol{v} = \big(\boldsymbol{\beta}_\psi^\mathsf{T}, \boldsymbol{u}_\psi^\mathsf{T}, \boldsymbol{\beta}_\tau^\mathsf{T}, \boldsymbol{u}_\tau^\mathsf{T}, \boldsymbol{\beta}_\phi^\mathsf{T}, \boldsymbol{u}_\phi^\mathsf{T}, \boldsymbol{\beta}_\gamma^\mathsf{T}, \boldsymbol{u}_\gamma^\mathsf{T}\big)^\mathsf{T},$$
$$\boldsymbol{\epsilon} = \big(\boldsymbol{\epsilon}_\psi^\mathsf{T}, \boldsymbol{\epsilon}_\tau^\mathsf{T}, \boldsymbol{\epsilon}_\phi^\mathsf{T}, \boldsymbol{\epsilon}_\gamma^\mathsf{T}\big)^\mathsf{T},$$
$$Z = \begin{pmatrix} X_\psi & A_\psi & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & X_\tau & A_\tau & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & X_\phi & A_\phi & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & X_\gamma & A_\gamma \end{pmatrix},$$

where $\mathbf{0}$ denotes zero matrices of appropriate dimensions.

Prior specification for the unknown parameters are such that the covariate coefficients, $\boldsymbol{\beta}_\psi$, $\boldsymbol{\beta}_\tau$, $\boldsymbol{\beta}_\phi$, and $\boldsymbol{\beta}_\gamma$, and the spatial model components, $\boldsymbol{u}_\psi$, $\boldsymbol{u}_\tau$, $\boldsymbol{u}_\phi$, and $\boldsymbol{u}_\gamma$, have Gaussian priors, while the hyperparameters, $\sigma_{\epsilon\psi}$, $\sigma_{\epsilon\tau}$, $\sigma_{\epsilon\phi}$, $\sigma_{\epsilon\gamma}$, s_ψ, s_τ, s_ϕ, s_γ, ρ_ψ, ρ_τ, ρ_ϕ, ρ_γ, are assigned penalized complexity (PC) priors to regularize the random effects of the model and shrink

our complex Bayesian model toward a simpler reference model, thus preventing overfitting (Simpson et al. (2017)). We use the same PC priors for $(s_\psi, \rho_\psi)^\top$, $(s_\tau, \rho_\tau)^\top$, $(s_\phi, \rho_\phi)^\top$, and $(s_\gamma, \rho_\gamma)^\top$, as in Fuglstad et al. (2019). Mutually independent prior densities are assumed for $\sigma_{\epsilon\psi}$, $(s_\psi, \rho_\psi)^\top$, $\sigma_{\epsilon\tau}$, $(s_\tau, \rho_\tau)^\top$, $\sigma_{\epsilon\phi}$, and $\sigma_{\epsilon\gamma}$. The PC priors for $\sigma_{\epsilon\psi}$, $\sigma_{\epsilon\tau}$, $\sigma_{\epsilon\phi}$, and $\sigma_{\epsilon\gamma}$ are exponential distributions, while PC priors for $(s_\psi, \rho_\psi)^\top$, $(s_\tau, \rho_\tau)^\top$, $(s_\phi, \rho_\phi)^\top$, and $(s_\gamma, \rho_\gamma)^\top$, have joint densities

$$\pi(s_l, \rho_l) = \lambda_{\rho,l}\lambda_{s,l}\rho_l^{-2}\exp(-\lambda_{\rho,l}\rho_l^{-1} - \lambda_{s,l}s_l), \quad s_l > 0, \rho_l > 0, l \in \{\psi, \tau, \phi, \gamma\},$$

where $\lambda_{s,l}$ and $\lambda_{\rho,l}$ are calculated after selecting a threshold for s_l and another threshold for ρ_l that are believed to be a priori exceeded with probability 0.05 and 0.95, respectively; see Fuglstad et al. (2019).

Prediction of flow at a set of ungauged sites involves predicting ψ, τ, ϕ, and γ, and requires covariates and coordinates of these sites. Modified versions of the equations in (3) are needed to predict ψ, τ, ϕ, and γ. For example, let $\boldsymbol{\psi}_{\mathrm{un}}$ denote ψ at the ungauged sites. Its formula is

$$\boldsymbol{\psi}_{\mathrm{un}} = X_{\psi,\mathrm{un}}\boldsymbol{\beta}_\psi + A_{\mathrm{un}}\boldsymbol{u}_\psi + \boldsymbol{\epsilon}_{\psi,\mathrm{un}},$$

where $X_{\psi,\mathrm{un}}$ contains the covariates, A_{un} is such that $A_{\mathrm{un}}\boldsymbol{u}_\psi$ is a prediction of the spatial component at the ungauged sites, and $\boldsymbol{\epsilon}_{\psi,\mathrm{un}}$ contains the unstructured random effects of the ungauged sites which all have mean zero and variance $\sigma_{\epsilon\psi}^2$.

Further details about the statistical model can be found in Section 1.5 in the Supplementary Material (Jóhannesson et al. (2022a)). In the next section, we describe—using the same notation—how to fit our proposed extended LGM to potentially very large datasets by exploiting and extending an approximate Bayesian inference scheme called Max-and-Smooth.

4. Approximate posterior inference.

4.1. *Gaussian-based approximation to the posterior density.* To deal with the high dimensionality of the data and the latent parameters, and to enable model selection at the latent level, a fast computational method for posterior inference is required. Hrafnkelsson et al. (2021) recently developed a general approximate posterior inference scheme for extended LGMs, called Max-and-Smooth. While Hrafnkelsson et al. (2021) tested the performance of Max-and-Smooth in a simple Gaussian regression example with spatially-varying mean and log-variance defined on a lattice, we here extend this methodology and validate it in our extreme-value framework with a heavy-tailed non-Gaussian likelihood. It is the first time that this methodology is rigorously tested in such a large-scale, strongly nonstationary, and highly non-Gaussian data setting. This approach is based on a Gaussian approximation to the likelihood function which leads to a Gaussian–Gaussian pseudo model that is inferred instead of the initial extended LGM. This approximation is very accurate when numerous time replicates are available and/or when the likelihood is close to Gaussian, and it significantly facilitates inference, while also substantially reducing the computational burden.

To be more precise, let $\boldsymbol{\theta}$ be the vector containing the hyperparameters of the extended latent Gaussian model in Section 3, namely,

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{12})^\top = (\sigma_{\epsilon\psi}, s_\psi, \rho_\psi, \sigma_{\epsilon\tau}, s_\tau, \rho_\tau, \sigma_{\epsilon\phi}, s_\phi, \rho_\phi, \sigma_{\epsilon\gamma}, s_\gamma, \rho_\gamma)^\top.$$

Then, the posterior density can be written as

$$\begin{aligned}\pi(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{v} \mid \boldsymbol{\theta})\pi(\boldsymbol{\eta} \mid \boldsymbol{v}, \boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{\eta}) \\ &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{v} \mid \boldsymbol{\theta})\pi(\boldsymbol{\eta} \mid \boldsymbol{v}, \boldsymbol{\theta})L(\boldsymbol{\eta} \mid \boldsymbol{y}),\end{aligned}$$

(5)

where $L(\cdot \mid \cdot)$ denotes the generalized likelihood function. We then approximate the generalized likelihood function by a Gaussian density with mean $\hat{\boldsymbol{\eta}}$ and covariance matrix $\Sigma_{\eta y} = Q_{\eta y}^{-1}$, where $\hat{\boldsymbol{\eta}}$ is the mode of the generalized log-likelihood function, $\log L$, and $Q_{\eta y}^{-1}$ is the inverse of the negative Hessian matrix of $\log L$, evaluated at $\hat{\boldsymbol{\eta}}$. Let $\hat{L}(\boldsymbol{\eta} \mid \boldsymbol{y})$ be the Gaussian approximation to the generalized likelihood function. Denoting the approximation to the posterior density by $\hat{\pi}(\cdot \mid \cdot)$, we have

$$
\begin{aligned}
\hat{\pi}(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{v} \mid \boldsymbol{\theta})\pi(\boldsymbol{\eta} \mid \boldsymbol{v}, \boldsymbol{\theta})\hat{L}(\boldsymbol{\eta} \mid \boldsymbol{y}) \\
&\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{v} \mid \boldsymbol{\theta})\pi(\boldsymbol{\eta} \mid \boldsymbol{v}, \boldsymbol{\theta})\mathrm{N}(\boldsymbol{\eta} \mid \hat{\boldsymbol{\eta}}, Q_{\eta y}^{-1}),
\end{aligned}
$$

(6)

where $\mathrm{N}(\cdot \mid \cdot)$ denotes a Gaussian density. Now, consider a pseudo model, that is, such that $\hat{\boldsymbol{\eta}}$ (obtained by maximizing $L(\cdot \mid \boldsymbol{y})$ at each site separately in a first step) is treated as the data (i.e., as noisy measurements of $\boldsymbol{\eta}$), and assume that $\boldsymbol{\eta}$ is modeled with the linear model in (4) used at the latent level of our original model. The proposed data density is now $\pi(\hat{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) = \mathrm{N}(\hat{\boldsymbol{\eta}} \mid \boldsymbol{\eta}, Q_{\eta y}^{-1})$ where $Q_{\eta y}$ is as above and assumed to be known. The posterior density for this model can thus be written as

$$
\begin{aligned}
\pi(\boldsymbol{\eta}, \boldsymbol{v}, \boldsymbol{\theta} \mid \hat{\boldsymbol{\eta}}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{v} \mid \boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}} \mid \boldsymbol{\eta}) \\
&\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{v} \mid \boldsymbol{\theta})\mathrm{N}(\hat{\boldsymbol{\eta}} \mid \boldsymbol{\eta}, Q_{\eta y}^{-1}) \\
&\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{v} \mid \boldsymbol{\theta})\hat{L}(\boldsymbol{\eta} \mid \hat{\boldsymbol{\eta}}),
\end{aligned}
$$

(7)

where $\hat{L}(\boldsymbol{\eta} \mid \hat{\boldsymbol{\eta}}) = \hat{L}(\boldsymbol{\eta} \mid \boldsymbol{y})$. Hence, the above posterior density in (7) is the same as the Gaussian approximation (6) to the *exact* posterior density (5). This observation motivates a two-step approximate Bayesian inference scheme, called Max-and-Smooth in Hrafnkelsson et al. (2021), whereby the estimates $\hat{\boldsymbol{\eta}}$ are obtained first at each site separately (Max) and are then smoothed by inferring the Gaussian–Gaussian pseudo model (Smooth). The main reason for using the pseudo model is that it is fully conjugate, and thus posterior inference takes less computation time than posterior inference for the exact model. Notice that, once the estimates $\hat{\boldsymbol{\eta}}$ are obtained, the computational time required for fitting the pseudo model no longer depends on the number of time replicates. Therefore, Max-and-Smooth benefits from large datasets with a lot of time replicates in two ways: the Gaussian approximation to the likelihood function not only becomes more accurate with more time replicates, but the relative computational cost with respect to an "exact" MCMC approach also decreases dramatically. We also emphasize here that, although our inference approach is done in two consecutive steps (i.e., first, estimating $\hat{\boldsymbol{\eta}}$, and, second, inferring $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ with the Gaussian–Gaussian pseudo model for $\hat{\boldsymbol{\eta}}$), the uncertainty involved in the first step is properly propagated into the second step in a way that provides a valid approximation to the full posterior distribution. Further details about the approximate posterior inference can be found in Section 1.6 in the Supplementary Material (Jóhannesson et al. (2022a)).

We assess the accuracy of the Gaussian approximation to ith likelihood contribution $L(\boldsymbol{\eta}_i \mid \boldsymbol{y}_i)$, where $L(\boldsymbol{\eta} \mid \boldsymbol{y}) = \prod_{i=1}^{J} L(\boldsymbol{\eta}_i \mid \boldsymbol{y}_i)$, in Section 1.7 of the Supplementary Material (Jóhannesson et al. (2022a)) for various parameter settings when using the GEV distribution without the time trend. Expectedly, our results show that the approximation becomes more accurate as the number of time replicates T increases. It is exceedingly accurate for the transformed location and shape parameters ψ_i and ϕ_i when $T = 50$ or $T = 80$ but has a slight negative bias for the transformed scale parameter τ_i in these temporal dimensions or for smaller values of T.

4.2. *Model selection procedure based on INLA.* In order to quickly select covariates and spatial model components to include in the linear regression models for the transformed parameters at the latent level, an additional approximation may be performed. This approximation, detailed below, is only used to guide model selection at a preliminary stage in a data-driven way that respects the core model structure, but it should not be used in the main inference procedure to fit the final model that has been selected. More precisely, for model selection purposes only, the Gaussian–Gaussian pseudo model may be further simplified by setting the off-diagonal elements of $Q_{\eta y}^{-1}$ equal to zero, that is, by neglecting all covariances between the parameters $(\hat{\psi}_i, \hat{\tau}_i, \hat{\phi}_i, \hat{\gamma}_i)^\top$ for each catchment i and instead considering the marginal variances. The simplified version of the data density is thus

$$\pi(\hat{\eta} \mid \eta, \nu, \theta) = \mathrm{N}(\hat{\eta} \mid \eta, \mathrm{diag}(Q_{\eta y}^{-1})),$$

while the pseudo model for $\hat{\psi}$ at the data level then takes the form

$$\pi(\hat{\psi} \mid \psi, \nu, \theta) = \mathrm{N}(\hat{\psi} \mid \psi, Q_{\psi y}^{-1}),$$

where $Q_{\psi y}^{-1}$ is the sample variance of $\hat{\psi}$, obtained from $\mathrm{diag}(Q_{\eta y}^{-1})$. Similarly, the pseudo models for $\hat{\tau}, \hat{\phi},$ and $\hat{\gamma}$ at the data level take the forms

$$\pi(\hat{\tau} \mid \tau, \nu, \theta) = \mathrm{N}(\hat{\tau} \mid \tau, Q_{\tau y}^{-1}),$$

$$\pi(\hat{\phi} \mid \phi, \nu, \theta) = \mathrm{N}(\hat{\phi} \mid \phi, Q_{\phi y}^{-1}),$$

$$\pi(\hat{\gamma} \mid \gamma, \nu, \theta) = \mathrm{N}(\hat{\gamma} \mid \gamma, Q_{\gamma y}^{-1}).$$

This translates into assuming independence between the transformed location intercept, scale, shape, and time trend parameters at the "data level" of the pseudo model such that the linear model for each of these four parameters can be inferred separately. This separation allows applying the very fast integrated nested Laplace approximation (INLA) methodology (Rue, Martino and Chopin (2009)) and exploiting the associated R package to estimate models independently for each of the four parameters, thus making a 10-fold cross-validation feasible. The INLA methodology bypasses MCMC sampling and the associated convergence assessment issues for LGMs with univariate link functions by relying on a fast numerical approximation to the posterior density, which is similar in spirit (albeit with some important differences) to our proposed procedure, detailed in Section 4.1, in the sense that INLA also exploits Gaussian-based approximations; for more details on INLA and its application to spatial models, see Bakka et al. (2018), and for an example of INLA applied to extreme precipitation data, see Opitz et al. (2018).

In Section 6 in the Supplementary Material (Jóhannesson et al. (2022a)), we show that inferring ψ, τ, and ϕ separately for the purpose of model selection, based on three separate models, can indeed be justified, since the posterior means and the marginal posterior intervals of the βs and the us, for say ψ, are numerically similar when separate INLA runs are used (assuming incorrectly that the parameters are independent in the posterior) or when a correct full MCMC approach is used. In other words, information about the βs and us for ψ is found in $\hat{\psi}$, while the other parameters (e.g., $\hat{\tau}$) add negligible knowledge about ψ when $\hat{\psi}$ is already known. However, these independent INLA fits neglect the dependence between parameters, and thus cannot be used to evaluate posterior uncertainty in return levels, which involve the GEV quantile function that combines all parameters, ψ, τ, ϕ, and γ. Neglecting the posterior covariance between $\hat{\psi}$, $\hat{\tau}$, $\hat{\phi}$, and $\hat{\gamma}$ can thus dramatically impact the posterior variability of return levels. Moreover, it turns out that γ is quite sensitive to ignoring the dependence in the likelihood function. In particular, the posterior means of γ have a wider

spread with the INLA approach than with the full correct model, and the corresponding pos-
terior standard deviations are larger when computed with INLA. Thus, the model for γ that
is selected by the INLA approach needs to be interpreted with caution. For further details,
see Section 6 in the Supplementary Material (Jóhannesson et al. (2022a)).

The separate inferences for ψ, τ, ϕ, and γ can be made very efficiently using INLA, tak-
ing two, two, one, and one, minutes, respectively, on a standard laptop for our large dataset.
It is, therefore, possible to select the "best" covariates by cross-validation using a forward
selection procedure in a reasonable amount of time. The 10-fold cross-validation procedure
that we used involves splitting the data into 10 disjoint randomly chosen subsets with the
same number of sites, the data from each site being used only once. The model is then trained
on nine of those subsets and then used to predict the left-out subset and to compute the cor-
responding root mean squared prediction error, defined as the sum of squared differences
between ML estimates of say ψ at the left-out sites and predictions of these based on the
$X_\psi \boldsymbol{\beta}_\psi + A_\psi \boldsymbol{u}_\psi$ part of the model using the posterior mean of $\boldsymbol{\beta}_\psi$ and $\boldsymbol{u}_\psi$. This step is re-
peated until each of the 10 subsets have been used for testing, and the root mean squared
prediction errors of each test set are then averaged to calculate the overall root mean squared
prediction error (also simply called the "test error") for a given model. In our application the
whole model selection procedure took only about seven hours in total for all four parame-
ters. As a result, linear models with good prediction properties were found for each of the
four parameters. Moreover, in order to further motivate the inclusion of the spatial model
components, the forward model selection was performed with and without the spatial model
components for each of ψ, τ, ϕ, and γ.

5. Results.

5.1. *Model selection results.* We applied the forward model selection approach, based
on INLA and described in Section 4.2, to select covariates for the linear models of ψ, τ, ϕ,
and γ that are given by equation (3). Figure 4 shows the relative test error (i.e., the overall
root mean squared prediction error rescaled by that of the best fitted model) for the pseudo
models fitted separately to $\hat{\psi}$, $\hat{\tau}$, $\hat{\phi}$, and $\hat{\gamma}$ as a function of the number of covariates selected
by the forward selection procedure.

When selecting a final model for each of the four parameters, based on the overall root
mean square prediction error, we opt for a parsimonious model with good prediction prop-
erties. Due to the high complexity of the spatial component, we only add it to a final model
if the gain is considerable. In particular, the overall root mean square prediction error is
significantly lower when spatial components are included in ψ and τ, which suggests that
they should be included in the final model, while this is not the case for ϕ and γ. Based on
Figure 4, we selected models for ψ, τ, ϕ, and γ with four, five, one, and one covariate, re-
spectively. There is a negligible improvement in the overall root mean square prediction error
for models with additional covariates. All models also include an intercept. Table 2 shows
the model components (covariates and spatial effects) selected in the final models for ψ, τ,
ϕ, and γ.

5.2. *Parameter estimates and interpretation.* Summary statistics for the estimated hy-
perparameters are presented in Section 3.2 of the Supplementary Material (Jóhannesson et
al. (2022a)). Convergence assessment of the posterior simulation for the full model is given
in Section 3.1 of the Supplementary Material (Jóhannesson et al. (2022a)). A comparison
of the marginal standard deviations of the spatial model components and the unstructured
model components of ψ and τ shows that the marginal standard deviations of the spatial
model components are bigger than those of the random errors $\boldsymbol{\epsilon}_\psi$ and $\boldsymbol{\epsilon}_\tau$ which suggests that
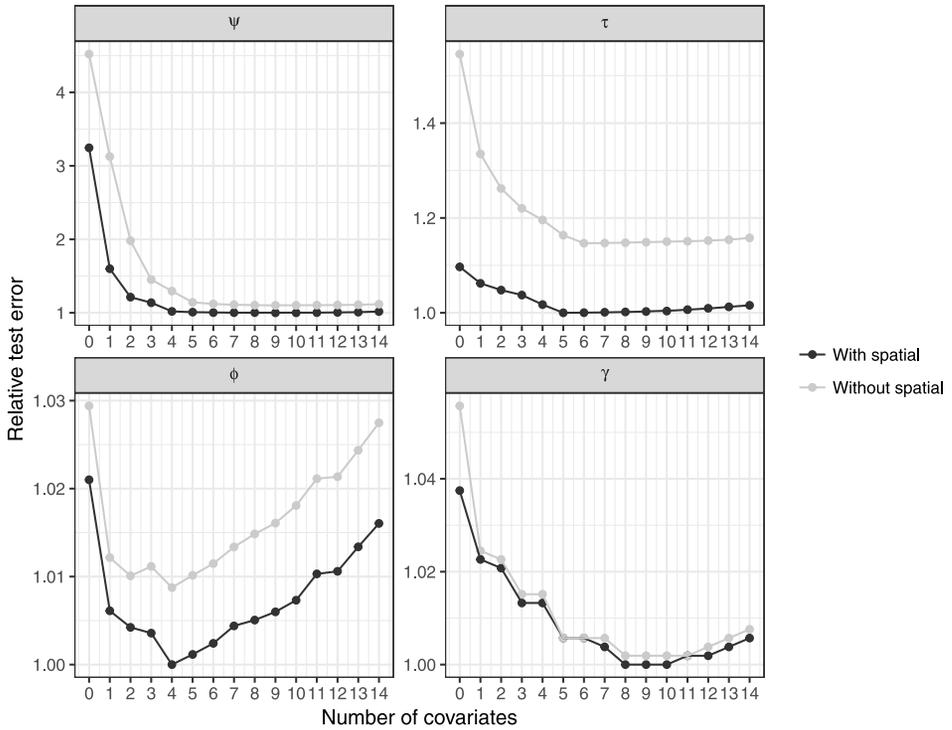
FIG. 4. *Relative test error (i.e., the overall root mean squared prediction error rescaled by that of the best fitted model) plotted as a function of the number of covariates selected by the forward selection procedure in the models for ψ, τ, ϕ, and γ, upper left to lower right, respectively. Black (respectively grey) curves correspond to models, including (respectively, not including) spatially structured model components at the latent level.*

the inclusion of latent spatial random effects helps to borrow strength across locations. Fitting simplified models to ψ and τ that do not contain spatial model components but use the same covariates as in the full model, resulted in the estimates $\sigma_{\psi\epsilon}^* = 0.397$ and $\sigma_{\tau\epsilon}^* = 0.211$. Here, $\sigma_{\psi\epsilon}^*$ and $\sigma_{\tau\epsilon}^*$ denote the posterior means of the standard deviations of the unstructured model components in the simplified models for ψ and τ without the spatial component. This result is not surprising since

$$0.397 = \sigma_{\psi\epsilon}^* \approx \sqrt{s_{\psi,\text{post}}^2 + \sigma_{\psi\epsilon,\text{post}}^2} = \sqrt{0.311^2 + 0.247^2} = 0.397$$

TABLE 2
The latent level of our final model. Model components (covariates or spatial effects) selected in our final model are indicated by check marks (✓) for ψ, τ, ϕ, and γ

Model component	ψ model	τ model	ϕ model	γ model
log(AREA)	✓	✓		
log(SAAR)	✓	✓		
log(FARL)	✓	✓		
log(FPEXT)		✓	✓	
log(URBEXT + 1)		✓		
BFIHOST^2	✓			
log(PROPWET)				✓
Spatial component	✓	✓		
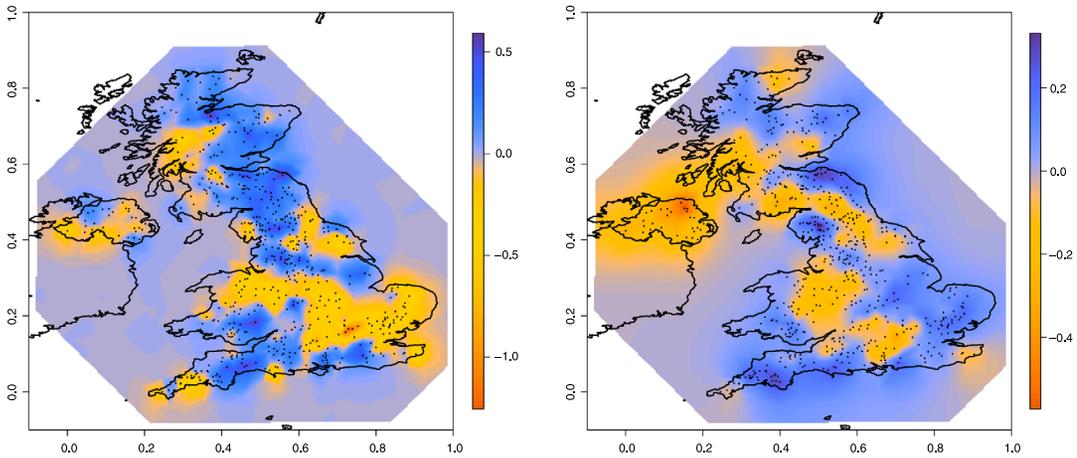
FIG. 5.   *The posterior mean of the spatial model components of ψ (left) and τ (right). Note that the color scales are different in both panels, although zero has the same color.*

and

$$0.211 = \sigma_{\tau\epsilon}^* \approx \sqrt{s_{\tau,\text{post}}^2 + \sigma_{\tau\epsilon,\text{post}}^2} = \sqrt{0.182^2 + 0.133^2} = 0.225,$$

where $s_{\psi,\text{post}}^2$, $\sigma_{\epsilon\psi,\text{post}}^2$, $s_{\tau,\text{post}}^2$, and $\sigma_{\epsilon\tau,\text{post}}^2$ denote the posterior means of s_{ψ}^2, $\sigma_{\epsilon\psi}^2$, s_{τ}^2, and $\sigma_{\epsilon\tau}^2$, respectively, in the final model with the spatial model components. Thus, the calculations above show that the sum of the marginal variances of the spatial model components and the variances of the unstructured model components in the spatial model was approximately equal to the variances of the unstructured error terms in the model that did not have spatial model components. This suggests that the spatial model components explain variability that would otherwise not have been explained by the covariates. Figure 5 displays the posterior mean of the spatial model components of ψ and τ. The spatial model component of ψ takes values mainly between -0.5 and 0.5 which corresponds to a multiplicative factor for μ that varies spatially between 0.61 and 1.65. Similarly, the spatial model component of τ takes values mainly between -0.3 and 0.3, which translates into a multiplicative factor for σ, that varies spatially between 0.74 and 1.35 after taking μ into account. The spatial model component of τ has a range parameter that is greater than the one of ψ by a factor two. This indicates that there is a stronger spatial smoothing for the transformed scale parameter τ, and the effect of this can indeed be seen in Figure 5. The spatial components represent the effect of location not explained by the covariates. When u_{ψ} is positive, then the joint effect of the covariates is not large enough in the sense that extreme flow at these sites is, on average, higher than the covariates would have predicted without the spatial component. The average annual rainfall (SAAR) has the strongest spatial structure of the covariates; thus, to some extent, the spatial components of ψ and τ are compensating for SAAR as a predictor for the GEV location and scale parameters of the extreme flow.

The original GEV location intercept and scale parameters, μ and σ, can be written in terms of ψ and τ for station i as

$$\mu_i = \exp(\psi_i) = \exp(\boldsymbol{x}_{\psi,i}\boldsymbol{\beta}_{\psi} + \boldsymbol{a}_i \boldsymbol{u}_{\psi} + \epsilon_{\psi,i})$$

and

$$\sigma_i = \exp(\psi_i + \tau_i) = \exp(\boldsymbol{x}_{\psi,i}\boldsymbol{\beta}_{\psi} + \boldsymbol{x}_{\tau,i}\boldsymbol{\beta}_{\tau} + \boldsymbol{a}_i(\boldsymbol{u}_{\tau} + \boldsymbol{u}_{\psi}) + \epsilon_{\psi,i} + \epsilon_{\tau,i}).$$
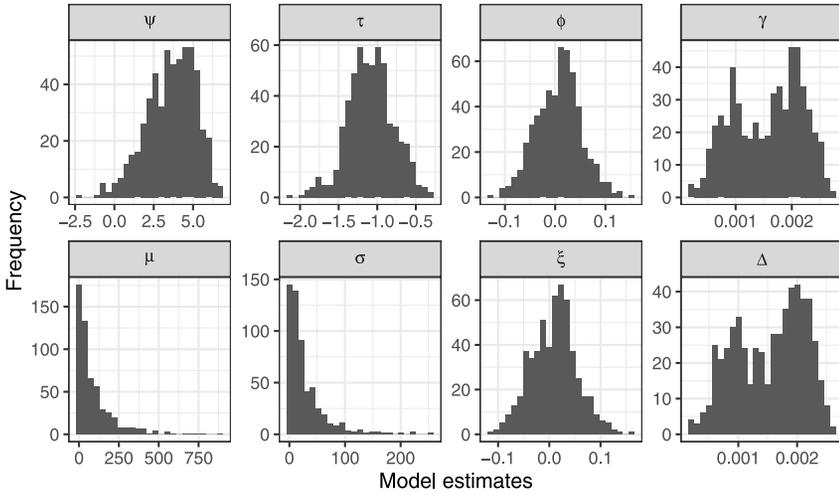
FIG. 6. *Histograms of the posterior means of the GEV parameters, μ, σ, and ξ, and the time trend Δ (bottom) from each of the observational sites along with the posterior means of the transformed parameters, ψ, τ, ϕ, and γ (top).*

Using the posterior means as point estimates for the covariates coefficients, the estimated location intercept, scale, shape, and time trend parameters may be expressed as

$$\mu_i = e^{-11.622}(\text{AREA}_i)^{0.892}(\text{SAAR}_i)^{1.661}(\text{FARL}_i)^{3.616}$$
$$\times e^{-3.272(\text{BFIHOST}_i)^2}e^{a_i u_\psi + \epsilon_{\psi,i}},$$

$$\sigma_i = e^{-8.115}(\text{AREA}_i)^{0.837}(\text{SAAR}_i)^{0.981}(\text{FARL}_i)^{2.578}$$
$$\times (\text{URBEXT}_i + 1)^{-0.750}(\text{FPEXT}_i)^{-0.135}e^{a_i(u_\psi + u_\tau) + \epsilon_{\psi,i} + \epsilon_{\tau,i}},$$

$$\xi_i = \left[1 - \exp\left\{-\exp\left(\frac{-0.102 - 0.037\log(\text{FPEXT}_i) + \epsilon_{\phi,i} - a_\phi}{b_\phi}\right)\right\}\right]^{1/c_\phi} - \frac{1}{2},$$

$$\Delta_i = 2\delta_0[1 + \exp\{-2\delta_0^{-1}(\beta_{\gamma,1} + \beta_{\gamma,2}\log(\text{PROPWET}_i) + \epsilon_{\gamma,i})\}]^{-1} - \delta_0,$$

where $a_\phi = 0.062376$, $b_\phi = 0.39563$, $c_\phi = 0.8$, $\delta_0 = 0.008$, $\beta_{\gamma,1} = 3.09 \cdot 10^{-3}$, and $\beta_{\gamma,2} = 1.97 \cdot 10^{-3}$. The effect of a covariate that is used in both the ψ model and the τ model on σ needs to be handled in a particular way; for example, the effect of AREA on σ_i is governed by the exponent $\beta_{\psi,2} + \beta_{\tau,2}$, which is estimated to be 0.837 since the posterior estimates of $\beta_{\psi,2}$ and $\beta_{\tau,2}$ are 0.892 and -0.055, respectively.

Figure 6 shows histograms of the posterior means of the GEV parameters, μ_i, σ_i, ξ_i, and Δ_i from all observational sites i along with the posterior means of the corresponding transformed parameters, ψ_i, τ_i, ϕ_i, and γ_i. The histograms of the preliminary sitewise ML estimates of the same parameters from these sites are shown in Section 2.1 of the Supplementary Material (Jóhannesson et al. (2022a)). Based on the posterior means of the shape parameters, the data appear to be close to the Gumbel distribution with light tails, as $\xi_i \approx 0$ at all sites i. As for the time trend Δ_i, most sites indicate a slight positive increase of about 1.5% per decade, revealing a temporal effect on river flow extremes in the United Kingdom that may be due to climate change. Figure 7 shows the preliminary sitewise ML estimates of the GEV parameters and the transformed GEV parameters plotted against their final posterior means. The values of μ_i and σ_i vary a lot across catchments. Despite these large differences, our flexible spatial model can accurately capture the data distribution over space thanks to
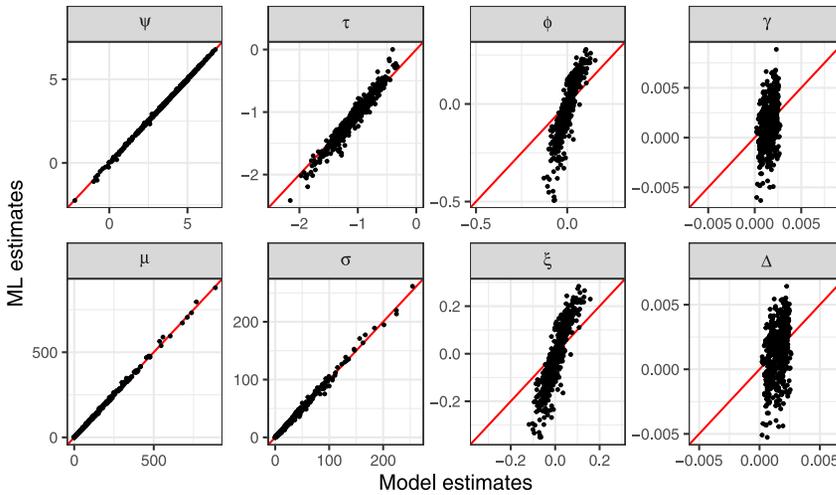
FIG. 7. *The ML estimates of the GEV parameters, μ, σ, and ξ, and the time trend parameter Δ (bottom), and their transformations, ψ, τ, ϕ, and γ (top), as a function of their posterior estimates.*

our choice of informative covariates and the inclusion of latent spatial model components and unstructured model components. There is an excellent match between the preliminary sitewise ML estimates of the location parameter μ_i and their final model-based counterparts, and similarly (albeit to a less degree) for the scale parameter σ_i, and yet the uncertainty of the posterior estimates is substantially lower than the uncertainty associated with the ML estimates. This is due to spatial smoothing, and it implies that our Bayesian spatial model adequately borrows strength across locations for improving the estimation of μ_i and σ_i in the GEV distribution. There are, however, significant differences between the ML estimates and final model-based posterior estimates of the shape parameter ξ_i and the time trend parameter Δ_i. This is mainly due to shrinkage that is introduced by the unstructured model component in the latent models for ϕ and γ, and the corresponding PC prior specified for their standard deviations. While the preliminary sitewise ML estimates of the shape and time trend parameters have very large uncertainties, our fully Bayesian spatial model succeeds in controlling these uncertainties, thus providing significant improvements, that is, with reduced posterior uncertainty and shrinkage toward more reasonable values. Thus, there is overall an excellent goodness-of-fit with a largely reduced estimation uncertainty compared to sitewise estimation uncertainty.

To investigate the effect of model components on sitewise extreme events, for example, the marginal 100-year event, the corresponding quantiles need to be estimated. The quantile function of the GEV distribution has an explicit expression which makes it straightforward to compute. Inverting (1), the expression for the p-quantile is

$$Q(p \mid \mu, \sigma, \xi) = \begin{cases} \mu + \sigma[\{-\log(p)\}^{-\xi} - 1]/\xi, & \xi \neq 0, \\ \mu - \sigma \log\{-\log(p)\}, & \xi = 0, \end{cases} \quad p \in (0, 1),$$

and the 100-year event is defined as $Q(0.99 \mid \mu, \sigma, \xi)$. In our case, return levels change with time because of the time trend. Therefore, in our context the 100-year event is the temporally-varying 0.99-quantile. The multiplication effect of each model component (covariates and spatial effects) on the 100-year event, with all other components being held fixed, is given in Table 3. Note that the numbers reported in Table 3 do not change with time. In this table the parameters are set equal to their posterior means. The unstructured model components, ϵ_ψ, ϵ_τ, and ϵ_ϕ, and the spatial model components, u_ψ and u_τ, are set equal to zero when the effects of the covariates are quantified. Table 3 shows how the 100-year event changes when

TABLE 3
The effects of the covariates and the spatial model components in the final model on the 100-year event (0.99 quantile) of the GEV density. The effect of a given covariate (or a given spatial model component) on the 100-year event is measured by comparing the 100-year event computed with the median value of the covariate (or the spatial model component) to the 100-year event computed with the first quartile and third quartile of the covariate (or the spatial model component) while keeping other covariates and random effects fixed. The table reports the multiplicative effect on the 100-year event

| | 100-year event | |
Covariate	1st	3rd
AREA	0.502	2.035
SAAR	0.747	1.390
FARL	0.910	1.046
BFIHOST	1.071	0.905
URBEXT	1.002	0.990
FPEXT	1.055	0.953
Spatial ψ	0.835	1.176
Spatial τ	0.935	1.058

covariates or spatial model components vary from their median values to their lower or upper quartiles. As above, AREA appears to be the most important covariate, followed by SAAR.

We then illustrate the estimates of marginal return levels and annual maximum flow distributions, under a stationary model, at three stations (namely, stations 54020, 37020, and 25011) selected such that the posterior mean of the shape parameter, ξ, of the GEV distribution is approximately -0.1, 0.0, and 0.1, respectively. The observations at each site were corrected with respect to the posterior estimate of the time trend at the site, that is, the time trend component was removed. This was done to facilitate comparison between a stationary model and the data. The top panel of Figure 8 shows the posterior means of return levels as a function of the return period, with associated 95% credible intervals, and we also display the order statistics (after time trend correction) for each station with their associated uncertainty (i.e., 95% prediction intervals). Notice that the x-axis is expressed on a log-scale. The bottom panel of Figure 8 shows the empirical distribution of the observed annual maxima of the three stations and the posterior mean of the distribution along with 95% credible bands. The relatively good match between the posterior distributions and the empirical distributions indicates that our model fits well overall. Similar results (not shown) were obtained for the other stations. Moreover, by comparing the uncertainty bands in the top panel of Figure 8, we can see that the uncertainty of our model-based estimates is quite moderate, compared to the uncertainty of the (empirical) order statistics.

5.3. *Predictive performance assessment and model comparison.* To assess the performance of the proposed model and other natural alternative approaches, we design an extensive cross-validation study in which we evaluate the out-of-sample skill of the posterior predictive distributions. To allow evaluation in space and in time, we cross-validate by removing from the training set the data after 2000 at all stations as well as the six stations depicted in Figure 9 altogether. To homogenize the data, we only considered stations that have training data before 1980 and complete test data up to 2013, which leaves 351 stations (compared to 554 in the full dataset). Predictions of the left-out data are evaluated using the logarithmic score (also "log-score" or "ignorance," see Lindley (1985), Roulston and Smith (2002)). The log-score is here defined as the negative dual logarithm of the predictive density $p(\cdot)$ assigned to the observed outcome y,

$$(8) \qquad \mathrm{LOGS}(p, y) = -\log_2 p(y)$$
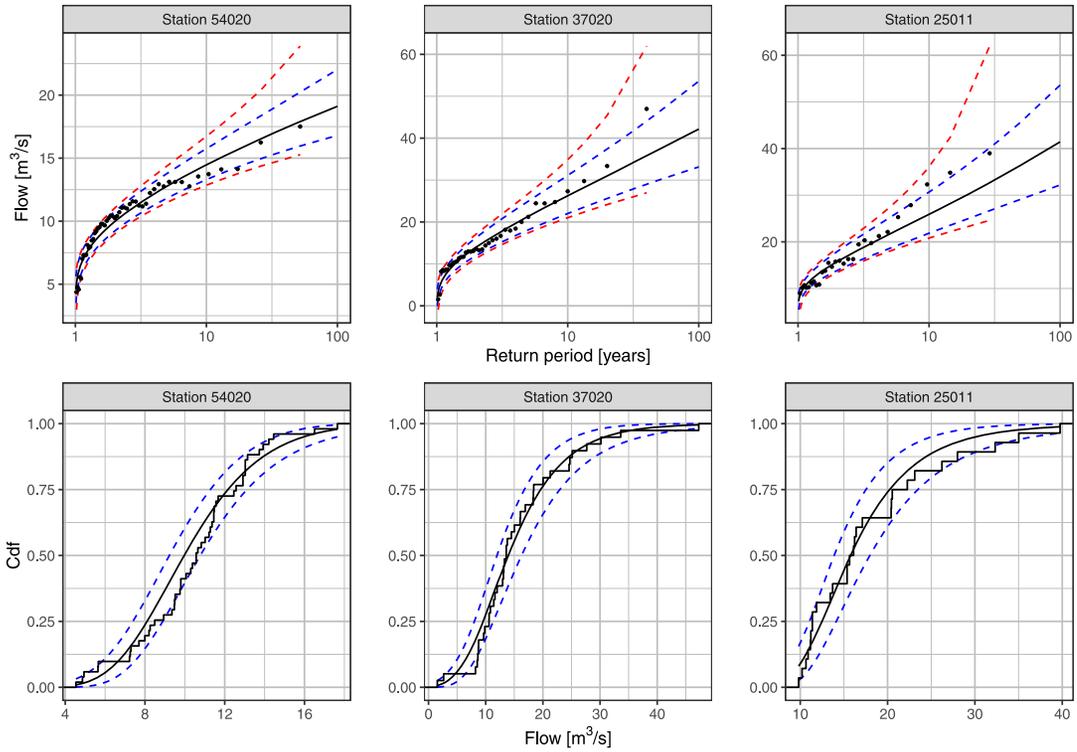
FIG. 8. *Top row*: *Estimates* (*i.e.*, *posterior means* (*solid black line*)) *of return levels as a function of the return period for three randomly selected stations. The observations at each station were corrected with respect to the posterior estimate of the time trend at the station; that is, the trend was removed to match a stationary model. The blue dashed lines show* 95% *credible intervals for the return levels. The ordered observations* (*after trend correction*) *as a function of return period* (*black points*) *are shown along with* 95% *prediction intervals for the ordered observations* (*red dashed lines*) *which represent the variability of the order statistics. Bottom row*: *Empirical* (*black stepwise curves*) *and fitted* (*solid black lines*) *cumulative distribution functions of the annual maximum flow data along with* 95% *credible intervals* (*blue dashed lines*) *for the fitted cumulative distribution functions.*

The log-score is negatively oriented, indicating better predictions by smaller values. We summarize the predictive performance by averaging log-scores over the test data and compare predictive performance between models by average log-score differences. To give an indication of the robustness of the average log-score differences, we also report their estimated standard errors. To account for correlations in the calculation of standard errors, we replaced the full sample size by an effective sample size $n_{\text{eff}} = n_{\text{eff,time}} \times n_{\text{eff,space}}$. We verified that temporal correlation of log-score differences is low and hence chose $n_{\text{eff,time}} = 13$ (the number of years in the test data). The effective spatial sample size $n_{\text{eff,space}}$ was estimated as 50, based on spatial variogram ranges of log-score differences.

We include the following models in the comparison, all fitted to the same training data:

- CONST: A stationary, spatially invariant, and time-constant, GEV distribution fitted to all available training data by maximum-likelihood estimation.
- MLE: Individual GEV distributions fitted to training data at each location separately, using maximum likelihood estimation. This method is not applicable for out-of-site predictions.
- RSM: A spatial response surface GEV model, using the same covariates as in our full model (Table 2), and additional tensor products of cubic splines in μ and σ to account for spatial variability not accounted for by the covariates. More precisely, this tensor product of splines is in 2D-space with respect to longitude and latitude and is specified by default with basis dimension 5^2.
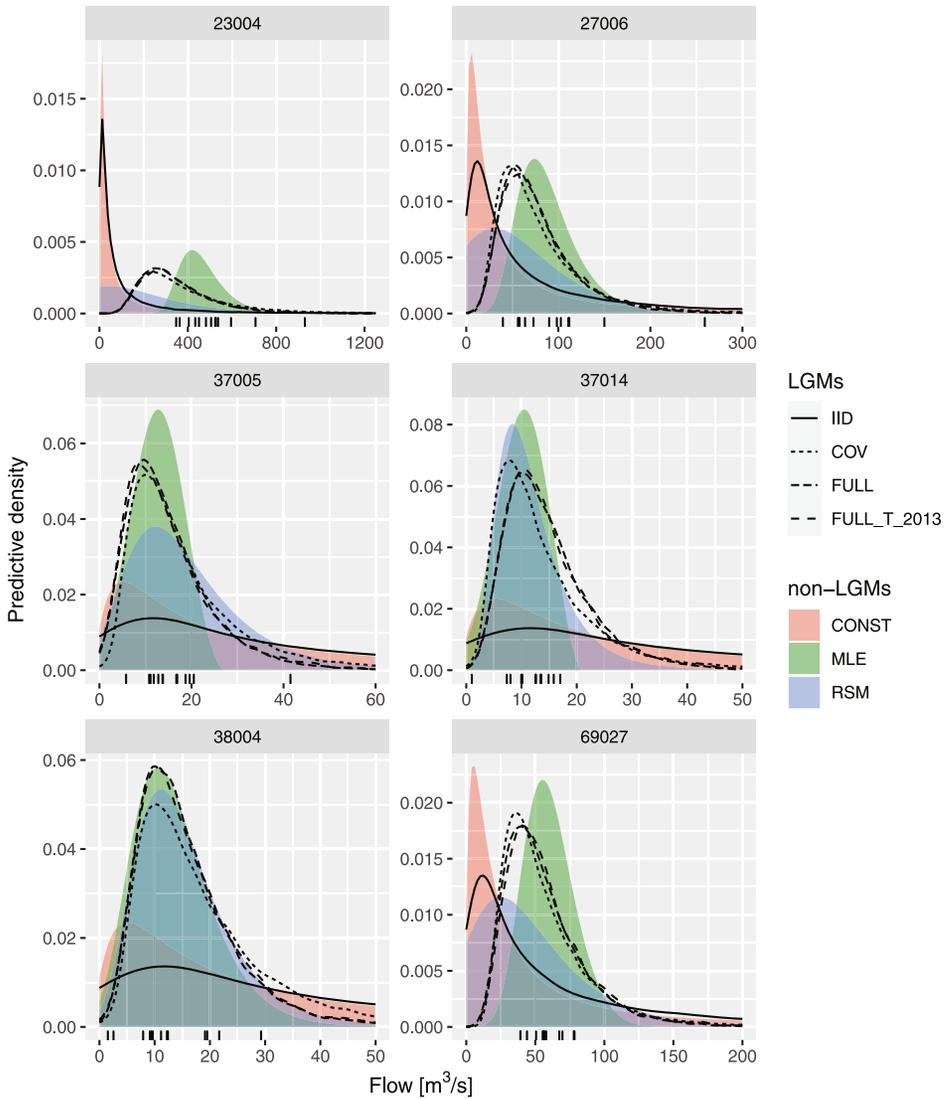
FIG. 9. *Posterior predictive densities based on four LGMs, the constant model and the response surface model at six stations that were removed from the training data. The long-dashed density ("FULL_T_2013") corresponds to the full LGM with trend, that is, FULL+T for the year 2013. The green density corresponds to the ML estimates of the GEV parameters based on sitewise observations and provides a reference (i.e., "ideal" prediction) for the out-of-site-out-of-time predictions.*

- IID: An LGM with unstructured spatial random effects ϵ only included in the GEV model parameters at the latent level.
- COV: An LGM with unstructured spatial random effects ϵ and fixed effects of covariates shown in Table 2 (i.e., IID model plus covariates).
- FULL: The fully spatial LGM with all covariates and structured/unstructured spatial random effects but without any time trend (i.e., COV model plus spatial effects).
- FULL+T: Same as FULL with an additional linear time trend in the location parameter.

CONST is the simplest possible null-model, issuing the same prediction at each site in each year and thus serves a lower bound of goodness-of-fit and predictive skill. The response surface model RSM, which we consider as a competitive benchmark, is included to compare our LGM methodology with a more traditional state-of-the-art frequentist methodology for spatial GEV models which describes GEV parameters in terms of fixed covariate

TABLE 4
Within-site-out-of-time cross validation. Pairwise average log-score differences between models and estimated standard errors in brackets. Positive entries indicate that column model (e.g., LGM_FULL) is "better" than row model (e.g., CONST)

	MLE	RSM	LGM_IID	LGM_COV	LGM_FULL	LGM_FULL_T
CONST	2.20 (0.07)	1.05 (0.04)	2.23 (0.06)	2.23 (0.06)	2.23 (0.06)	2.24 (0.07)
MLE		−1.15 (0.07)	0.03 (0.03)	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)
RSM			1.18 (0.06)	1.19 (0.06)	1.18 (0.06)	1.19 (0.06)
LGM_IID				0.00 (0.01)	0.00 (0.01)	0.01 (0.02)
LGM_COV					−0.00 (0.00)	0.01 (0.02)
LGM_FULL						0.01 (0.02)

effects and smooth spline terms capturing residual spatial variability (Youngman (2019)). The response surface model was fitted by maximum likelihood via the R package `evgam` (Youngman (2020)). The four LGM versions tested in this analysis are nested and increase in complexity from IID to FULL+T. Predictions by all models, except FULL+T, are constant in time. Some of the predictions produced probabilities smaller than 2^{-50}, which we treated as numerically equal to zero, and removed the resulting infinitely large log scores from the analysis.

For out-of-time cross validation, data at station i up to the year 2000 is included in the training data, to make post-2000 predictions at station i. For out-of-time out-of-site cross validation, no past data from station i is used to make predictions at station i. For the time-constant LGMs (IID, COV, FULL) predictive densities were constructed from 32,000 posterior predictive samples by kernel density approximation with a Gaussian kernel and automatic bandwidth selection (using R function `bw.nrd0`). For the time-varying LGM (FULL+T) we used only 3200 posterior predictive samples at each station and each year to reconstruct the predictive density in the same way.

Model CONST was included as a simple null model. All models outperform it significantly. In terms of the within-site-out-of-time predictions, the sitewise MLE estimates predict the observations better than the response surface model but give slightly worse predictions than the LGMs; see Table 4. The four LGMs (IID, COV, FULL, FULL+T) are practically the same and all are adequate for within-site predictions, though FULL+T is marginally better. Thus, a Bayesian framework slightly improves over a likelihood-based framework, but covariates and spatial effects do not improve the within-site predictions. We conclude that, in terms of average log-score comparisons, using covariates and spatial effects is unnecessary if site-specific observations are available while the time trend gives a marginal improvement.

Out-of-time-out-of-site, the latent Gaussian model IID has similar predictive skill as model CONST and is outperformed by the other models; see Table 5. The sitewise MLEs cannot be produced out-of-site but were included in the table as an upper reference point of predictive skill that could be achieved if local data were available. The response surface model RSM outperforms the IID model because of the added information from local covariates and the smooth spatial terms. Similar to the within-site cross validation, all of COV, FULL, and FULL+T outperform RSM. The relatively small improvement of including spatial random effects (FULL vs. COV) might be due to the relatively large number of covariates, whose spatial structures are enough to account for the spatial variation in the model parameters. The FULL+T model is on a par with the FULL model, so there is here no noticeable additional benefit of including a linear trend.

Figures 9 and 10 provide a visual confirmation that the full LGM (with or without the time trend component) does slightly better than the other models in terms of out-of-site predictions, as its densities and return level curves are always closest to the densities and return

TABLE 5
Out-of-site-out-of-time cross validation. Note that MLE is within-site and added here only for reference; see the caption of Table 4 for further details

	MLE	RSM	LGM_IID	LGM_COV	LGM_FULL	LGM_FULL_T
CONST	1.97 (0.16)	1.13 (0.09)	−0.08 (0.05)	1.48 (0.12)	1.54 (0.12)	1.54 (0.13)
MLE		−0.84 (0.11)	−2.05 (0.14)	−0.48 (0.09)	−0.43 (0.08)	−0.43 (0.08)
RSM			−1.21 (0.09)	0.36 (0.07)	0.41 (0.06)	0.41 (0.06)
LGM_IID				1.56 (0.10)	1.62 (0.11)	1.62 (0.11)
LGM_COV					0.06 (0.03)	0.05 (0.03)
LGM_FULL						0.00 (0.01)

level curves stemming from the data. These results are in line with the out-of-site log-score results presented in Table 5. Figure 10 shows that the three LGMs without time trend perform equally well in terms of within-site predictions, while the constant model and the response surface model perform relatively poorly, matching the results in Table 4.

We now further discuss the out-of-site predictions for Station 23004 presented in Figure 9. In this case the predictors, that is, the covariates and the spatial components, of the three most complex LGMs are such that the ϵ term for the log-location parameter ψ is positive and around 0.2 in size, that is, not particularly large values since the standard deviation of ϵ_ψ is estimated as 0.247. However, since the ϵ term is unknown, this implies that the out-of-sample posterior predictive density for this site will be centered at a level about 20% smaller
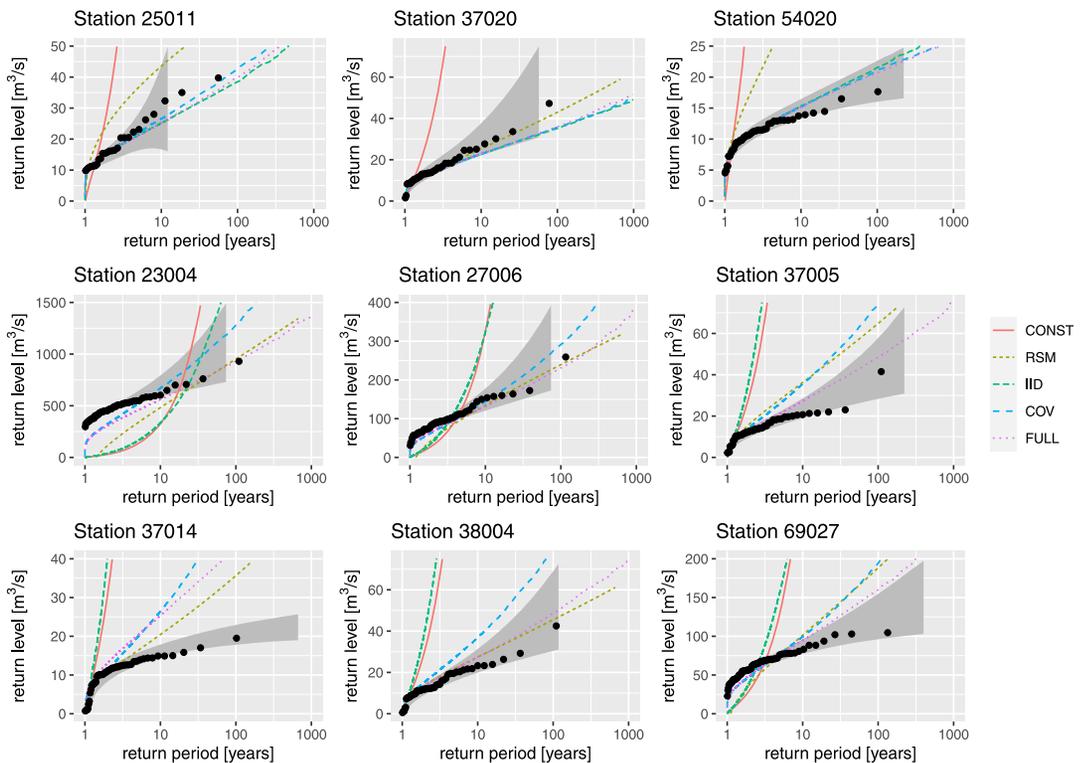
FIG. 10. *Return level plots based on three of the LGMs (LGM with time trend excluded), the constant model and the response surface model for three within-sample stations and six stations that were removed from the training data (same as in Figure 9). The black dots show the ordered data, and the grey bands show 95% confidence intervals for the return level, based on the within-site data.*
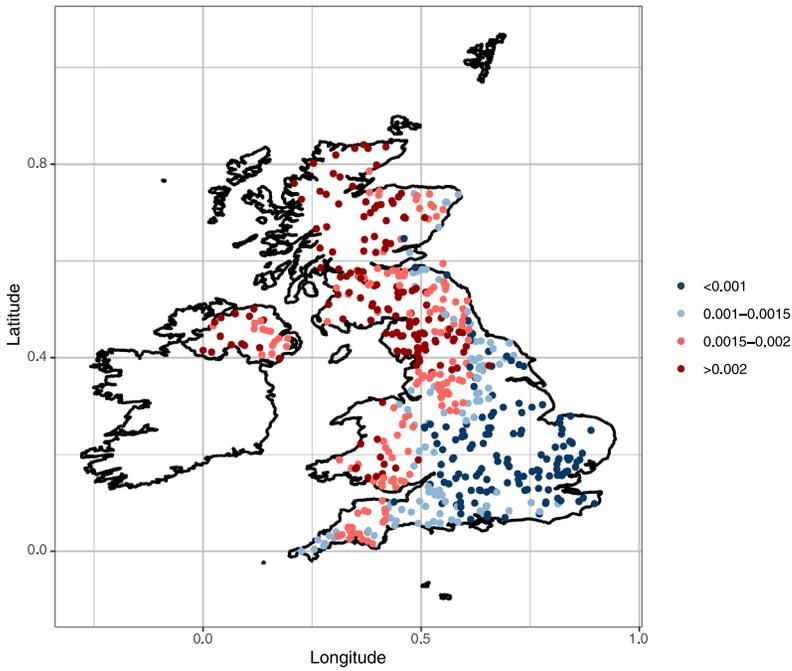
FIG. 11.   *Posterior estimates of Δ_i across the U.K., presented with a four-color scheme that splits trend values at $\Delta_i = 0.0010, 0.0015, 0.0020$, corresponding to a 1.0%, 1.5%, and 2.0% increase per decade.*

than the center of the data density. However, the variability in the data density at the site is smaller than the variability in the posterior predictive density due to the four unknown ϵ terms, namely, $\epsilon_\psi, \epsilon_\tau, \epsilon_\phi$, and ϵ_γ, and thus, most of the observations fall under the posterior predictive density, mainly its upper tail since ϵ_ψ was positive. The LGM with unstructured terms only does poorly here. This is due to the standard deviation of the ϵ being much larger than in the LGMs with covariates and/or spatial components. The response surface model does not perform as well as the three most complex LGMs but does better than the LGM with unstructured terms only, underlining the importance of covariates.

5.4. *Time trend interpretation and assessment.*   Figure 11 shows the spatial structure of the time trend parameter Δ, based on its posterior estimates. The trend estimates correspond to a 0.1% to 2.8% increase per decade with a median value of around 1.5% per decade. In Figure 11 it can be seen that the trend is greater in the northern and southwestern part of the U.K., while smaller trends are observed in the southeastern part. This is in line with the results in Blöschl et al. (2019), who also found greater trends in the northern part of the U.K. and smaller trends in the southeastern part. However, their results for the southwest part and the part below the middle of the U.K. are somewhat different from ours. The analysis of Blöschl et al. (2019) indeed resulted in negative trend values in the middle part of the U.K., while here our trend estimates are always positive. Due to the latent Gaussian structure of our proposed model, in particular, the part involving Δ, the values of the corresponding posterior estimates spread less than those found in Blöschl et al. (2019), the estimates above 0.0015 being usually smaller than those in Blöschl et al. (2019) and none of the estimates below 0.0015 being negative. We believe the main reason for this disagreement is that Blöschl et al. (2019) performed a site-by-site analysis (with independent GEV fits at each site), while our more complex latent Gaussian model incorporates informative covariates and spatial effects at the latent level which allows a drastic reduction in parameter uncertainty through shrinkage by borrowing strength across neighboring sites. This is especially helpful at sites with small sample sizes and can prove critical to avoid estimating unrealistic trend values.

In Dadson et al. (2017), various aspects of flood management in the U.K. were considered. They point out that: (i) winter precipitation has been increasing during the period 1950–2015, and the same is true for the summer precipitation for the period 1975–2015 (Alexander and Jones (2000)); (ii) the annual mean flood index has been increasing during the period 1990–2015 (Wilby and Quinn (2013)). Blöschl et al. (2019) show that an upward trend over the period 1960–2010 in a flood index and in maximum seven-day precipitation in the U.K. Generally speaking, our results are, therefore, in line with the published literature which provides further evidence that our proposed LGM is indeed able to detect the time trend appropriately

5.5. *Spatial dependency at the data level and estimation of return levels for spatial aggregates*. The assumption of independence at the data level is problematic when spatially extended quantities, such as aggregates or coexceedances, are of interest. To reliably predict, say, the annual maximum flow summed over a few neighboring catchments, we have to take into account not only their individual marginal distributions but also their spatial dependency. Working under a modeling framework that ignores spatial dependence at the data level will inevitably provide biased predictions for spatially aggregated quantities.

Accounting for spatial dependence at the data level in a fully Bayesian way leads to a more complicated inference framework which is outside the scope of this paper. However, we can predict annual maximum flow jointly at multiple stations in a way that respects the spatial dependence structure by reordering independently drawn posterior predictive samples, based on the empirical copula obtained from historical observations; see Clark et al. (2004) and Schefzik, Thorarinsdottir and Gneiting (2013) for related approaches. Our proposed method starts from a set of M historical observations available at each of L stations of interest. The rank order of the observations at each site is calculated. Then, M independent samples are drawn from each of the L univariate posterior predictive distributions of the GEV model fitted to each site. These M independent samples are subsequently reordered to have the same rank order as the M historical observations. The reordered samples have the same pairwise Spearman rank correlations as the historical observations and thus mimic the observed spatial dependency. By repeatedly drawing M independent samples at each site and rank-reordering them in the same way, the posterior predictive sample size can be increased indefinitely.

For illustration we have selected 15 nearby stations in the eastern part of the U.K. and considered the sum of their annual flow maxima as a prediction target. The aggregate of annual maxima over several catchments might be of interest to insurers or infrastructure planners as an indicator of compound regional flood risk. After fitting the full LGM (that assumes conditional independence at the data level), we generated independent posterior predictive samples at each location. We estimated the predictive distribution of the spatially aggregated annual maximum flow, based on conditionally independent samples as well as on samples that were rank-reordered to restore spatial correlation as outlined above.

The histogram of observed aggregate maximum flow, and density estimates of predictive samples before and after rank-reordering are shown in Figure 12 (top panel). The posterior predictive distribution of aggregated flow estimated from conditionally independent samples clearly does not reflect the observed variability. Both small values and large values are underrepresented, compared to the empirical histogram. After rank-reordering, the predictive distribution has the appropriate width and correctly reflects the observed variability of aggregate flow maxima. As a result, return levels of aggregate maximum flow calculated from the rank-reordered samples are also much more accurate than the return levels calculated from spatially independent samples (see the bottom panel of Figure 12).

These results suggest that postprocessing independent samples by rank-reordering is a suitable method to restore the previously ignored dependency at the data level. However, it
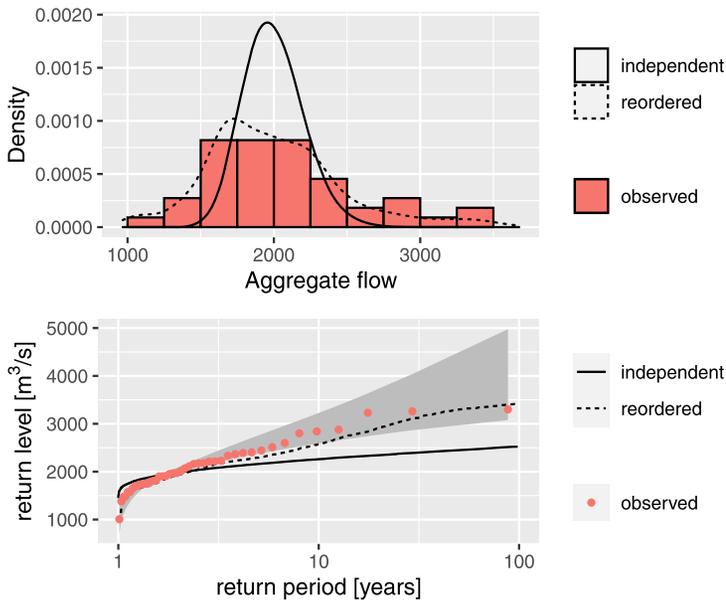
FIG. 12. *Top*: *Predictions of aggregate annual maximum flow taken over* 15 *nearby stations*, *based on conditionally independent samples and rank-reordered samples. After restoring the spatial dependence in posterior predictive samples, the predictive distribution matches the frequency histogram. Bottom*: *Return level plots corresponding to the densities in the top panel show that observed return levels are truthfully represented after rank-reordering.*

is worth noting that, by ignoring that dependency during parameter inference, the posterior distributions of the latent parameters are likely too narrow and thus overconfident. Moreover, rank-reordering can only be applied at locations where historical data are available. If past data is not available, the rank order has to be estimated either by interpolating between nearby observed stations or from a suitable surrogate, such as numerical model simulations.

6. Discussion. In this paper we developed an extended latent Gaussian model, based on the generalized extreme-value distribution, designed for analysis of spatiotemporal flood frequency data, and we illustrated our approach by application to a large dataset of maximum annual peak flow time series from the U.K. Because of the large dimensionality of the data and the high model complexity, GMRF priors were assumed for latent spatial effects. We used Max-and-Smooth (Hrafnkelsson et al. (2021)) for fast posterior inference and thoroughly verified its accuracy when using a GEV distribution at the data level. This inference approach relies on a Gaussian approximation of the likelihood function which simplifies inference and provides significant speed-up. The covariates for the final model were selected with a simplified model setup that made it possible to use INLA which led to further substantial reductions in computation time. The selected covariates were all highly significant in the final model; that is, their 95% posterior intervals did not include zero.

How to structure and parameterize the GEV distribution within a Bayesian hierarchical model is an important and challenging methodological question. In this paper a novel multivariate link function for the three parameters of the GEV distribution was proposed along with a temporal trend parameter included in the GEV location parameter. A standard logarithmic transformation was used for the location μ, and the confounding between the location μ and scale σ was dealt with by transforming them jointly. We also proposed a novel transformation for ξ that was constructed based on four criteria that we believe are reasonable when inferring the tail behavior in a wide range of environmental applications. Moreover, the new

transformation stabilized the inference of ξ, and through an additional beta prior density for each ξ_i, its inference can be tailored to each specific application. The time trend was bounded between $-\delta_0$ and δ_0, with $\delta_0 = 0.008$, which is physically justifiable and seemed reasonable in our context. The posterior estimates of the time trend parameters correspond to an average of 1.5% increase per decade in extreme river flow in the United Kingdom, likely due to climate change. Our results show a spatial structure in the time trend that is similar to that of Blöschl et al. (2019); however, our estimates are usually smaller and all positive. While prior specification and reparametrizations are sometimes regarded as a minor aspect in the entire modeling pipeline, we rather consider it as a crucial element in Bayesian modeling that requires in-depth reasoning and justifiable arguments. The novel transformations and priors proposed in our paper rely on valid physical and statistical considerations, and our paper is a call for a more careful modeling of key parameters (such as the shape ξ and the trend Δ).

The proposed model was inferred with a Bayesian approach, whereby the data were not transformed before inferring the parameters, and thus their uncertainty was properly quantified. Normalization is required for some commonly used flood frequency models, such as the index flood method, and, as a consequence, uncertainty is not properly propagated. Under the extended LGM framework used here, it was straightforward to include covariates and additional random effects at the latent level and to keep track of the uncertainty of the model parameters. PC priors were defined for the hyperparameters of the model. The hyperparameters in the extended LGM setup govern the latent parameters which are a part of an additive regression model, and thus the penalization of increased complexity becomes essential to regularize the random effects at the latent level.

The inference results from the MCMC output showed that all parameters of the model converged quickly, and most parameters were well defined. In particular, all hyperparameters had small posterior standard deviations with respect to their posterior means. The results showed that the spatial model components for ψ and τ explained more than half of the otherwise unexplained variability which indicates that they are essential in our model. The return period of extreme floods (under a stationarity assumption) was investigated by computing quantiles of the fitted GEV distribution and taking their uncertainty into account. Posterior predictive distributions were computed for six gauging stations, and our results showed that our proposed model is useful for predicting extreme flow data within ungauged catchments.

Our analysis demonstrates that the LGM framework is very flexible and powerful for predicting extreme flow data within gauged and ungauged catchments by efficiently borrowing strength across locations. However, the underlying conditional independence assumption of the data with respect to latent parameters makes it unsuitable when dependence across catchments is present and predictions on the joint behavior of the catchments are needed. Sang and Gelfand (2010) suggested relaxing the conditional independence assumption by using a Gaussian dependence structure at the data level, but this approach does not properly capture extremal dependence. For this purpose, more complex and specialized extreme-value models are required. In the case of strong extremal dependence, max-stable processes (Asadi, Davison and Engelke (2015), Davison, Huser and Thibaud (2019), Davison, Padoan and Ribatet (2012), Huser and Davison (2014), Padoan, Ribatet and Sisson (2010), Vettori, Huser and Genton (2019)) have proven to be useful for modeling spatial block maxima, while in the case of weakening extremal dependence, broader classes of max-infinitely divisible processes (Bopp, Shaby and Huser (2021), Huser, Opitz and Thibaud (2021), Huser and Wadsworth (2020)) have recently been proposed. These models, however, are usually much more intensive to fit than our proposed extended LGM, and further research is needed to make them applicable in large dimensions as in this paper. Nevertheless, we computed the effect of neglecting dependence at the data level in a simplified setting (see Section 4 of the Supplementary Material (Jóhannesson et al. (2022a))) and found that this effect was mostly

moderate in our application. These results might also be exploited to correct the uncertainty estimates provided by the conditional independence model. Furthermore, we here also developed a novel copula-based postprocessing approach to correct the dependency in posterior predictive samples and, thereby, accurately estimate return levels of spatial aggregates. Our proposed approach, however, relies on computing the empirical copula of historical data at the target locations, so future research is needed to extend it for the estimation of spatial return levels involving unobserved locations.

We believe that our proposed latent Gaussian model and its inference scheme have features that are important for flood frequency analysis in general. In particular, predictions of extreme events in the case of both gauged and ungauged catchments are improved due to: (i) reducing the uncertainty in the shape and time trend parameters by transforming them and regularizing the transformed parameters with unstructured model components, (ii) modeling the location and scale parameters spatially and utilizing the correlation between these two parameters, and (iii) applying the fast approximate inference scheme, called Max-and-Smooth, making it possible to evaluate large number of potential regression models and to perform extensive cross-validation to achieve good prediction properties. In our view, future research efforts in flood frequency analysis should focus on the development of fast inference schemes for models that take extremal dependence into account.

SUPPLEMENTARY MATERIAL

Supplementary material for the paper "Approximate Bayesian inference for analysis of spatiotemporal flood frequency data" (DOI: 10.1214/21-AOAS1525SUPPA; .pdf). We first provide further details on the statistical model fitted in our application and on the approximate Bayesian inference scheme that we used. In particular, we demonstrate the accuracy of the Gaussian likelihood approximation when the GEV distribution is used at the data level. We then provide further details on our preliminary data analysis, and we describe additional results regarding the convergence diagnostics for our MCMC chains, and the practical interpretation of parameter estimates. We also study the effect of ignoring the spatial dependence at the data level on the estimated parameter uncertainty, and we show that this effect is generally quite moderate. A sensitivity analysis to assess the effect of the predetermined intervals for Δ and ξ on their posterior inference is also described. Finally, the effect of using R-INLA for inferring the four model parameters independently is evaluated.

R **code for the paper "Approximate Bayesian inference for analysis of spatiotemporal flood frequency data"** (DOI: 10.1214/21-AOAS1525SUPPB; .zip). We provide the R code for fitting our model with Max-and-Smooth, as well as for drawing inferences from the MCMC chains. We also provide a detailed simulation data example (similar to our real dataset) to illustrate our inference method step by step.

REFERENCES

ALEXANDER, L. V. and JONES, P. D. (2000). Updated precipitation series for the U.K. and discussion of recent extremes. *Atmospheric Science Letters* **1** 142–150.

ANDERSON, T. W. and DARLING, D. A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.* **49** 765–769. MR0069459

ASADI, P., DAVISON, A. C. and ENGELKE, S. (2015). Extremes on river networks. *Ann. Appl. Stat.* **9** 2023–2050. MR3456363 https://doi.org/10.1214/15-AOAS863

BAKKA, H., RUE, H., FUGLSTAD, G.-A., RIEBLER, A., BOLIN, D., ILLIAN, J., KRAINSKI, E., SIMPSON, D. and LINDGREN, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdiscip. Rev.: Comput. Stat.* **10** e1443. MR3873676 https://doi.org/10.1002/wics.1443

BLÖSCHL, G., HALL, J., VIGLIONE, A., PERDIGÃO, R. A. P., PARAJKA, J., MERZ, B., LUN, D., ARHEIMER, B., ARONICA, G. T. et al. (2019). Changing climate both increases and decreases European river floods. *Nature* **573** 108–111.

BOPP, G. P., SHABY, B. A. and HUSER, R. (2021). A hierarchical max-infinitely divisible spatial model for extreme precipitation. *J. Amer. Statist. Assoc.* **116** 93–106. MR4227677 https://doi.org/10.1080/01621459.2020.1750414

BURN, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour. Res.* **26** 2257–2265.

CASSON, E. and COLES, S. (1999). Spatial regression models for extremes. *Extremes* **1** 449–468.

CLARK, M., GANGOPADHYAY, S., HAY, L., RAJAGOPALAN, B. and WILBY, R. (2004). The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5** 243–262.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics*. Springer London, Ltd., London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

COOLEY, D. and SAIN, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *J. Agric. Biol. Environ. Stat.* **15** 381–402. MR2787265 https://doi.org/10.1007/s13253-010-0023-9

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151

CUNNANE, C. and NASH, J. (1974). Bayesian estimation of frequency of hydrological events. *Mathematical Models in Hydrology* **1**.

DADSON, S., HALL, J., MURGATROYD, A., ACREMAN, M., BATES, P., BEVEN, K., HEATHWAITE, A., HOLDEN, J., HOLMAN, I. et al. (2017). A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **473** 20160706.

DALRYMPLE, T. (1960). Flood-frequency analyses, Manual of Hydrology: Part 3. Technical Report USGPO.

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235.

DAVISON, A., HUSER, R. and THIBAUD, E. (2019). Spatial extremes. In *Handbook of Environmental and Ecological Statistics* M. Fuentes, J. A. Hoeting and R. L. Smith, eds.) *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 711–744. CRC Press, Boca Raton, FL. MR3889918

DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980 https://doi.org/10.1214/11-STS376

DYRRDAL, A. V., LENKOSKI, A., THORARINSDOTTIR, T. L. and STORDAL, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics* **26** 89–106. MR3324904 https://doi.org/10.1002/env.2301

FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *J. Amer. Statist. Assoc.* **114** 445–452. MR3941267 https://doi.org/10.1080/01621459.2017.1415907

GEIRSSON, Ó. P., HRAFNKELSSON, B. and SIMPSON, D. (2015). Computationally efficient spatial modeling of annual maximum 24-h precipitation on a fine grid. *Environmetrics* **26** 339–353. MR3366968 https://doi.org/10.1002/env.2343

GEIRSSON, Ó. P., HRAFNKELSSON, B., SIMPSON, D. and SIGURDARSON, H. (2020). LGM split sampler: An efficient MCMC sampling scheme for latent Gaussian models. *Statist. Sci.* **35** 218–233. MR4106602 https://doi.org/10.1214/19-STS727

GREHYS (1996). Presentation and review of some methods for regional flood frequency analysis. *J. Hydrol.* **186** 63–84.

HOSKING, J. R. M. and WALLIS, J. R. (2005). *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge Univ. Press, Cambridge.

HOSKING, J. R. M., WALLIS, J. R. and WOOD, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27** 251–261. MR0797563 https://doi.org/10.2307/1269706

HRAFNKELSSON, B., MORRIS, J. S. and BALADANDAYUTHAPANI, V. (2012). Spatial modeling of annual minimum and maximum temperatures in Iceland. *Meteorol. Atmos. Phys.* **116** 43–61.

HRAFNKELSSON, B., SIEGERT, S., HUSER, R., BAKKA, H. and JÓHANNESSON, Á. V. (2021). Max-and-Smooth: A two-step approach for approximate Bayesian inference in latent Gaussian models. *Bayesian Anal.* **16** 611–638. MR4255342 https://doi.org/10.1214/20-ba1219

HUERTA, G. and SANSÓ, B. (2007). Time-varying models for extreme values. *Environ. Ecol. Stat.* **14** 285–299. MR2405331 https://doi.org/10.1007/s10651-007-0014-3

HUSER, R. and DAVISON, A. C. (2014). Space–time modelling of extreme events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 439–461. MR3164873 https://doi.org/10.1111/rssb.12035

HUSER, R., OPITZ, T. and THIBAUD, E. (2021). Max-infinitely divisible models and inference for spatial extremes. *Scand. J. Stat.* **48** 321–348. MR4233175 https://doi.org/10.1111/sjos.12491

HUSER, R. and WADSWORTH, J. L. (2020). Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev.: Comput. Stat.* e1537.

JALBERT, J., FAVRE, A.-C., BÉLISLE, C. and ANGERS, J.-F. (2017). A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 941–962. MR3715590 https://doi.org/10.1111/rssc.12212

JÓHANNESSON, A. V., SIEGERT, S., HUSER, R., BAKKA, H. and HRAFNKELSSON, B. (2022a). Supplement to "Approximate Bayesian inference for analysis of spatiotemporal flood frequency data" https://doi.org/10.1214/21-AOAS1525SUPPA

JÓHANNESSON, A. V., SIEGERT, S., HUSER, R., BAKKA, H. and HRAFNKELSSON, B. (2022b). R code for the paper "Approximate Bayesian inference for analysis of spatio-temporal flood frequency data" https://doi.org/10.1214/21-AOAS1525SUPPB

KJELDSEN, T. R. (2010). Modelling the impact of urbanization on flood frequency relationships in the UK. *Hydrology Research* **41** 391–405.

KJELDSEN, T. R., AHN, H. and PROSDOCIMI, I. (2017). On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrol. Sci. J.* **62** 1354–1363.

KJELDSEN, T. R. and JONES, D. A. (2006). Prediction uncertainty in a median-based index flood method using L moments. *Water Resour. Res.* **42** W07414.

KJELDSEN, T. R. and JONES, D. A. (2009a). A formal statistical model for pooled analysis of extreme floods. *Hydrology Research* **40** 465–480.

KJELDSEN, T. R. and JONES, D. A. (2009b). An exploratory analysis of error components in hydrological regression modeling. *Water Resour. Res.* **45**.

KRAINSKI, E. T., GÓMEZ-RUBIO, V., BAKKA, H., LENZI, A., CASTRO-CAMILIO, D., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press/CRC, New York.

KUCZERA, G. (1999). Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resour. Res.* **35** 1551–1557.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LINDLEY, D. V. (1985). *Making Decisions*, 2nd ed. Wiley, London. MR0892099

MARTINS, E. S. and STEDINGER, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.* **36** 737–744.

NATIONAL RIVER FLOW ARCHIVE (2018). NERC CEH, Wallingford.

OPITZ, T., HUSER, R., BAKKA, H. and RUE, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21** 441–462. MR3855716 https://doi.org/10.1007/s10687-018-0324-x

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. MR2757202 https://doi.org/10.1198/jasa.2009.tm08577

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. MR2137253 https://doi.org/10.1111/j.1467-9876.2005.00510.x

ROBSON, A. and REED, D. (1999). Flood estimation handbook. Institute of Hydrology, Wallingford.

ROSBJERG, D. and MADSEN, H. (1995). Uncertainty measures of regional flood frequency estimators. *J. Hydrol.* **167** 209–224.

ROULSTON, M. S. and SMITH, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130** 1653–1660.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SANG, H. and GELFAND, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environ. Ecol. Stat.* **16** 407–426. MR2749848 https://doi.org/10.1007/s10651-007-0078-0

SANG, H. and GELFAND, A. E. (2010). Continuous spatial process models for spatial extreme values. *J. Agric. Biol. Environ. Stat.* **15** 49–65. MR2755384 https://doi.org/10.1007/s13253-009-0010-1

SCHEFZIK, R., THORARINSDOTTIR, T. L. and GNEITING, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28** 616–640. MR3161590 https://doi.org/10.1214/13-STS443

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

STEPHENS, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.* **69** 730–737.

THORARINSDOTTIR, T. L., HELLTON, K. H., STEINBAKK, G. H., SCHLICHTING, L. and ENGELAND, K. (2018). Bayesian regional flood frequency analysis for large catchments. *Water Resour. Res.* **54** 6929–6947.

VETTORI, S., HUSER, R. and GENTON, M. G. (2019). Bayesian modeling of air pollution extremes using nested multivariate max-stable processes. *Biometrics* **75** 831–841. MR4012089 https://doi.org/10.1111/biom.13051

WILBY, R. L. and QUINN, N. W. (2013). Reconstructing multi-decadal variations in fluvial flood risk using atmospheric circulation patterns. *J. Hydrol.* **487** 109–121.

YOUNGMAN, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *J. Amer. Statist. Assoc.* **114** 1865–1879. MR4047306 https://doi.org/10.1080/01621459.2018.1529596

YOUNGMAN, B. D. (2020). evgam: Generalised Additive Extreme Value Models. R package version 0.1.4.