



Post-processing of wind gusts from COSMO-REA6 with a spatial Bayesian hierarchical extreme value model

Philipp Ertz and Petra Friederichs

Institute of Geosciences, University of Bonn, Auf dem Hügel 20, 53121 Bonn, Germany

Correspondence: Philipp Ertz (pertz@uni-bonn.de)

Received: 16 May 2025 – Revised: 24 October 2025 – Accepted: 27 October 2025 – Published: 27 November 2025

Abstract. The aim of this study is to provide a probabilistic gust analysis for the region of Germany that is calibrated with station observations and with an interpolation to unobserved locations. To this end, we develop a spatial Bayesian hierarchical model (BHM) for the post-processing of surface maximum wind gusts from the COSMO-REA6 reanalysis. Our approach uses a non-stationary extreme value distribution for the gust observations at the top level, with parameters that vary according to a linear model using COSMO-REA6 predictor variables. To capture spatial patterns in surface extreme wind gust behavior, the regression coefficients are modeled as 2-dimensional Gaussian random fields with a constant mean and an isotropic covariance function that depends only on the distance between locations. In addition, we include an elevation offset in the distance metric for the covariance function to account for differences in topography. This allows us to include data from mountaintop stations in the training process and to utilize all available information. The training of the BHM is carried out with an independent data set from which the data at the station to be predicted are excluded. We evaluate the spatial prediction performance at the withheld station using Brier score and quantile score, including their decomposition, and compare the performance of our BHM to climatological forecasts and a non-hierarchical, spatially constant baseline model. This is done for 109 weather stations in Germany. Compared to the spatially constant baseline model, the spatial BHM significantly improves the estimation of local gust parameters. It shows up to 5 % higher skill for prediction quantiles and provides a particularly improved skill for extreme wind gusts. In addition, the BHM improves the prediction of threshold levels at most of the stations. Although a spatially constant approach already provides high skill, our BHM further improves predictions and improves spatial consistency.

1 Introduction

Wind warnings are among the most frequent types of weather warning issued by the German Meteorological Service (DWD). Key users are emergency managers, air and rail traffic, energy companies, and the general public. Improving the quality of extreme wind warnings is therefore an important task for, e.g., meteorological services such as the DWD. A central characteristic of atmospheric wind speed is its high spatial variability with a comparatively sparse observation network for wind speed measurements. In addition, severe wind gusts are recurring but rare, so they are inherently extreme events. This makes forecasting a challenge and complicates the verification of wind gust warnings. A suitable representation of wind gusts that includes uncertainties in re-

analysis models is therefore essential for the development of an effective warning strategy.

State-of-the-art numerical weather prediction (NWP) models provide a deterministic wind gust diagnostic based on the current simulated turbulent and convective state (e.g. Schulz and Heise, 2003; Schulz, 2008). Brasseur (2001) proposed a peak wind speed diagnostic, which assumes that gusts are generated by the deflection of high-flowing air parcels in the upper boundary layer and are brought down by turbulent eddies. The deflection of high-flowing parcels to the surface occurs when the vertical component of the turbulent kinetic energy (TKE) associated with a given air parcel is strong enough to overcome the buoyancy forces. Another mechanism for the generation of surface wind gusts is

described by Nakamura et al. (1996). They suggested that the gusts originate from a convective downdraft that is deflected horizontally when it reaches the surface. Both theories motivate a diagnosis in NWP models for wind gust forecasts based on turbulent kinetic energy and the stability of the vertical stratification of the atmosphere. However, these are mostly deterministic in nature and do not incorporate uncertainty in the gust forecasts.

One approach for addressing uncertainties in the description of wind gusts is a probabilistic post-processing providing a conditional probability distribution for surface wind gusts based on predictor variables from the reanalysis. Given that wind gusts represent extreme values, extreme value statistics is a suitable probabilistic model (Coles, 2001; Beirlant et al., 2004). A basic application of extreme value statistics to maximum wind gusts can be found in Walshaw (1994), who develop a threshold model for wind gusts. The model is expanded with spatial covariates by Walshaw and Anderson (2000). They demonstrate that the model performance at single locations can be enhanced by the use of predictor variables such as the local mean wind speed. These simple models can be readily adapted into a post-processing approach by taking time series of the covariates for the model parameters. As shown in Friederichs et al. (2009) and Friederichs et al. (2018), these simple approaches already possess high skill.

Wessel et al. (2025) present an interesting approach for the post-processing of extreme wind speed, based on ensemble model output statistics (Gneiting et al., 2005). They do not use extreme value theory but train truncated normal and logistic distributions using threshold weighted scoring rules to enhance the predictive performance for extremes. To increase the training data size, data from stations with similar wind characteristics, based on a clustering approach, is pooled together and one model is trained for each cluster of stations. However, it is also possible to include a spatial structure directly into the model by modeling the dependency structure between stations.

Modeling spatial dependency in extreme value theory poses the challenge that there are no parametric forms for multivariate extreme value distributions (Coles, 2001). Davison et al. (2012) present several approaches to incorporate the spatial dependency structure into statistical extreme value models. These approaches include latent spatial process variables (Cooley et al., 2007; Apputhurai and Stephenson, 2013; Stephenson et al., 2016), copula models (Sang and Gelfand, 2010) and max-stable processes (Oesting et al., 2016; Davison and Gholamrezaee, 2012). Latent process variables are introduced in a Bayesian hierarchical model (BHM) framework, while assuming conditional independence between gusts at different locations, whereas copula models and max-stable processes explicitly model dependencies between gusts at different locations. Davison et al. (2012) demonstrate that latent process variables offer the greatest flexibility, while possessing a high degree of accuracy. However, the

assumptions on stationarity and the covariance function may result in the generation of implausible extremal patterns.

The general theory for hierarchical spatial process models is outlined, for example, in Banerjee et al. (2003). The use of a BHM allows for the pooling of information across locations and time, facilitating an optimal spatial and temporal representation. A groundbreaking example of a latent Gaussian process model for extreme weather can be found in Cooley et al. (2007). The authors construct a hierarchical Bayesian extreme value model for maximum precipitation in Colorado. At the initial level, they utilize a generalized Pareto distribution whose parameters vary as a function of covariates and a latent spatial Gaussian process in climate space, which is defined by elevation and mean precipitation. More recent examples of spatial modeling of extreme values in atmospheric sciences include the model developed by Apputhurai and Stephenson (2013), which was used to predict extreme precipitation in Western Australia. This model employed a spatio-temporal hierarchical model with anisotropic Gaussian random fields (GRFs) based on spatial explanatory variables and included trends in the model parameters. Dyrddal et al. (2015) model extreme precipitation return levels in Norway in a similar vein.

In Friederichs et al. (2018), the authors fit a regression model to the predictor variables of the NWP model using a conditional Gumbel distribution for the marginals. Subsequently, the non-explained residuals are modeled as a bivariate Brown-Resnick process to account for the remaining spatial dependency structure. Similarly, Oesting et al. (2016) use a bivariate, maximally stable Brown-Resnick process and Gumbel margins to directly model the dependency between forecasts and observed wind gust fields.

Baran and Lakatos (2024) demonstrate, with reference to the example of 10 m wind speed, that the spatial interpolation of post-processing schemes can achieve a high level of skill at locations where observations are lacking. The interpolated model demonstrates superior skill compared to both a globally trained model and a locally trained model at each individual location. Consequently, it can be concluded that the spatial interpolation of post-processing approaches effectively addresses the challenge of limited training data for post-processing gridded data sets, due to the restricted number and irregular distribution of observation stations.

This paper presents an integration of the post-processing approach originally proposed by Friederichs et al. (2018) with the spatial hierarchical modeling techniques derived from the precipitation model developed by Cooley et al. (2007). We utilize the flexible implementation of latent process models, assuming conditional independence between maxima at two different locations. As wind gusts are typically very localized phenomena, the spatial dependency of maxima at two locations can be disregarded during modeling. Furthermore, our model is designed for spatial interpolation and incorporates spatio-temporal predictors for the generalized extreme value distribution (GEV) parameters, en-

abling us to interpolate the model to unobserved locations via kriging (Stein, 1999). This provides comprehensive post-processing across the entire NWP model domain.

Our methodology is similar to the spatially adaptive Bayesian post-processing approach by Möller et al. (2016) for ensemble temperature forecasts. It employs Gaussian Markov random fields and integrated nested Laplace approximations (INLA) of stochastic partial differential equations (SPDE) (Rue et al., 2009; Lindgren et al., 2011). However, as we investigate wind gusts, we are examining non-Gaussian distributed observations, and in addition, we are incorporating a spatially variable and non-stationary variance. Furthermore, our post-processing approach is currently not designed for ensemble predictions, but it can be readily extended to ensemble post-processing through the selection of suitable predictors, such as the ensemble mean.

The remainder of this paper is structured as follows. Section 2 presents the data used in this study and in Sect. 3, we present our model formulation. The methods used for training and evaluating the model are explained in Sect. 4. Section 5 will show and discuss the results of the verification on observational data. Finally, we will conclude with a summary of our findings in Sect. 6.

2 Data

2.1 Station data

The wind gust observations FX were selected from a total of 109 synoptic observation stations (SYNOP), including mountain top stations. Please refer to Fig. 2 in Sect. 5.1 for a map of the corresponding locations. Additionally, Fig. 2 provides an overview over mean wind characteristics at these locations. Data are retrieved from the DWD Climate Data Center (Climate Data Center, 2023). DWD performs basic quality control of the incoming data. The stations are selected based on the least number of missing observations between 1995 and 2018. The observation time series are available in hourly resolution.

The two fundamental mechanisms of wind gusts production are shear-driven turbulent wind gusts and convective wind gusts driven by thunderstorm activity (Bradbury et al., 1994). Given the increased convective activity during the summer months, it is expected that the behavior of wind gust will differ from that observed in winter, when the characteristics of gusts are largely influenced by the occurrence of winter storms. Accordingly, the data set employed in this study comprises observations from the warm months between May and October, thus avoiding the need to consider the annual cycle of deep convection. Similarly, wind gusts display a pronounced diurnal cycle. During nocturnal hours, a stable boundary layer forms and lower wind gust values are observed, whereas stronger peak gusts occur during the convective maximum in the late afternoon. To include the peak gusts of the day, our investigation window is focused on the

afternoon hours from 13:00–18:00 UTC. Then, we generate daily time series using the maximum value inside the window for each day.

The availability of wind gust measurements from the years preceding 2001 is limited, and thus these years have been excluded from the analysis. Additionally, to ensure spatial homogeneity within the data set, days with fewer than four measurements within the specified investigation window at a given station have been omitted.

2.2 COSMO-REA6

COSMO-REA6 (Bollmeyer et al., 2015) is a regional re-analysis model for the CORDEX-EURO11 domain (Giorgi et al., 2009) based on the COntsortium for Small scale MOdelling (COSMO) climate model COSMO-CLM (Rockel et al., 2008). It was developed through a collaborative effort between the Hans-Ertel-Center for Weather Research (HERZ), Climate Monitoring and Diagnostics, from Bonn and Cologne Universities and DWD. The horizontal grid spacing is 0.055° , which is approximately 6 km, and the model operates on 40 vertical levels. The data included in this study encompass the 10 m values for mean horizontal wind, obtained from the horizontal components U_{10M} and V_{10M} , and the 10 m horizontal maximum wind speed $VMAX_{10M}$. The deterministic gust diagnostic in COSMO-CLM, $VABSMX_{10M}$, is not publicly available in COSMO-REA6. In the following sections, mean horizontal wind will be referred to as V_m and $VMAX_{10M}$ as V_{max} .

The 2D fields from COSMO-REA6 are available in 15 min intervals from 1995 to 2019. However, we retrieved hourly values from COSMO-REA6 between 12:00–18:00 UTC from May to October 2001 to 2018, to be in line with the SYNOP observations. The reanalysis time series were aligned with the weather stations by selecting the closest grid cell to the station in terms of distance on a sphere (distance along great circles). Subsequently, the reanalysis data underwent the same preprocessing as the observational data, with daily maxima employed for V_{max} and daily mean values for V_m .

For the purpose of our simulation study, we assume that the predictor data from COSMO-REA6 and the wind gust observations are independent. This might not necessarily be the case, as mean wind observations from the SYNOP stations are assimilated into COSMO-REA6 and usually, wind gust and mean wind observations are collinear to some degree. However, as the gust observations are not assimilated directly and we do not use any other observational data, the assumption of independence should hold. However, it should be noted that the spatial alignment of weather stations and model grid cells is also imperfect, and that a statistical post-processing is precisely used to address model deficiencies in the reanalysis.

3 Model formulation

We propose a Bayesian hierarchical model based on the extreme precipitation model of Cooley et al. (2007) for the spatial post-processing of wind gusts. The complete version of the spatial Bayesian hierarchical model, henceforth referred to as SpatBHM, is depicted as a directed acyclic graph in Fig. 1. For readers who are unfamiliar with Bayesian hierarchical models, we recommend consulting Gelman et al. (2014, Chap. 5).

Our model comprises three levels: a data level, which describes the actual modeling of wind gust observations using extreme value statistics; an underlying predictor level which includes the predictor variables from COSMO-REA6; and a prior level, which contains our prior knowledge about the model hyperparameters (cf. Fig. 1), as required for Bayesian inference. To accommodate the spatial dependency structure, we introduce an additional level between the predictor level and the prior level. This level will henceforth be referred to as the spatial dependency level or process level.

3.1 Data level

Extreme value theory provides an asymptotic theory for the distribution for sample maxima (Coles, 2001; Beirlant et al., 2004). As gust measurements refer to the highest 3 s average wind speed recorded over the course of a 1 h interval, a generalized extreme value (GEV) distribution serves as a theoretically consistent model for gust observations. The GEV distribution is characterized by three parameters: location μ , scale ς and shape ξ . These parameters may vary both spatially and temporally. However, allowing for non-stationarity in ξ largely increases the sampling uncertainty not only for ξ , but also for the other parameters (Friederichs et al., 2009). Therefore we decided to use a constant shape parameter. To obtain an estimate of a reasonable ξ -value, we conducted single station fits for all weather stations. These fits showed either no clear signal or resulted in slightly negative values, implying a Weibull-type GEV (Appendix D). The Weibull-type GEV imposes an unrealistic upper bound for wind gusts extremes (Perrin et al., 2006). In order to obtain stable parameter estimates and avoid making assumptions regarding an upper bound for wind gusts, we assume $\xi = 0$, resulting in a Gumbel distribution. We model the gust measurements y_{it} at location $\mathbf{r}_i$, $i \in \{1, \dots, q\}$, and time $t \in \{1, \dots, n\}$ as a Gumbel-distribution

$$G(y_{it}) = \exp\left(-\exp\left(-\frac{y_{it} - \mu_{it}}{\varsigma_{it}}\right)\right) \quad (1)$$

with non-stationary location μ_{it} and scale ς_{it} . As indicated by the indices $i = 1, \dots, q$ for location and $t = 1, \dots, N$ for time, the Gumbel parameters are allowed to vary in space and time. The reader is referred to Table A in Appendix A for a thorough list of notations.

3.2 Predictor level

On the next underlying level, the predictor level, we introduce spatio-temporal predictor variables $x_{j,ik}$ as covariates for μ_{it} and ς_{it} . The Gumbel parameters μ_{it} and ς_{it} are constructed as a linear model with regression coefficients μ_i^j and ς_i^j for the j th predictor variable. The vector of the predictor variables is constructed including a constant intercept, so that $\mathbf{x}_{it}^\mu = (1, x_{1,it}, \dots, x_{m_\mu,it})$ and $\mathbf{x}_{it}^\varsigma = (1, x_{1,it}, \dots, x_{m_\varsigma,it})$. With $\boldsymbol{\beta}_i^\mu = (\mu_i^0, \dots, \mu_i^{m_\mu})^T$ and $\boldsymbol{\beta}_i^\varsigma = (\varsigma_i^0, \dots, \varsigma_i^{m_\varsigma})^T$, we obtain

$$\begin{aligned} \mu_{it} &= \mathbf{x}_{it}^\mu \boldsymbol{\beta}_i^\mu \\ \varsigma_{it} &= \exp(\mathbf{x}_{it}^\varsigma \boldsymbol{\beta}_i^\varsigma). \end{aligned} \quad (2)$$

An exponential link for ς_{it} ensures positive definiteness of the scale parameter. The regression coefficient vectors $\boldsymbol{\beta}_i^\mu$ and $\boldsymbol{\beta}_i^\varsigma$ are stationary over time, but vary in space.

3.3 Spatial dependency structure

To address the spatial dependency structure, we assume that the regression coefficients μ_i^j and ς_i^j are realizations of a spatial random process observed at locations $\mathbf{r}_i$, $i = 1, \dots, q$. These spatial random processes will be referenced by $\mu^j(\mathbf{r})$ and $\varsigma^j(\mathbf{r})$, respectively, where j indicates the respective covariate. We therefore assume that the regression coefficients, which in turn determine the GEV parameters, are spatially dependent fields. This approach enables the integration of data within the spatial domain, thereby facilitating the incorporation of information regarding the distribution at neighboring stations. However, the degree of the spatial dependency is estimated in the BHM. In order to model the spatial dependency structure, a Gaussian process level is introduced below the predictor level. This means that each spatial regression coefficient is assigned a spatial process in the form of a GRF, namely $\mu^j(\mathbf{r})$ and $\varsigma^j(\mathbf{r})$, where $\mathbf{r}$ represents the spatial coordinates and $\mu_i^j = \mu^j(\mathbf{r}_i)$ and $\varsigma_i^j = \varsigma^j(\mathbf{r}_i)$.

If $\mu^j(\mathbf{r})$ is a GRF, then $\boldsymbol{\mu}^j = (\mu^j(\mathbf{r}_1), \dots, \mu^j(\mathbf{r}_q))^T$ is multivariate normally distributed random vector. This applies analogously to $\boldsymbol{\varsigma}^j = (\varsigma^j(\mathbf{r}_1), \dots, \varsigma^j(\mathbf{r}_q))^T$. We assume the GRFs as spatially homogeneous with constant expectation values $\alpha_{\mu^j}(\mathbf{r}) = \alpha_{\mu^j}$ and $\alpha_{\varsigma^j}(\mathbf{r}) = \alpha_{\varsigma^j}$, respectively. The process level for the regression coefficients then reads

$$\boldsymbol{\mu}^j \sim \mathcal{MVN}(\alpha_{\mu^j} \mathbf{I}, \boldsymbol{\Sigma}_{\mu^j}) \quad (3)$$

and

$$\boldsymbol{\varsigma}^j \sim \mathcal{MVN}(\alpha_{\varsigma^j} \mathbf{I}, \boldsymbol{\Sigma}_{\varsigma^j}). \quad (4)$$

Here, $\mathbf{I}$ is a unit vector of length q . We further assume spatially homogeneous and isotropic covariance functions such that $\text{Cov}(\mu^j(\mathbf{r}_i), \mu^j(\mathbf{r}_k)) = C_{\mu^j}(h_{ik})$ and $\text{Cov}(\varsigma^j(\mathbf{r}_i), \varsigma^j(\mathbf{r}_k)) = C_{\varsigma^j}(h_{ik})$, where $h_{ik} = |\mathbf{r}_i - \mathbf{r}_k|$ is the distance between two locations.

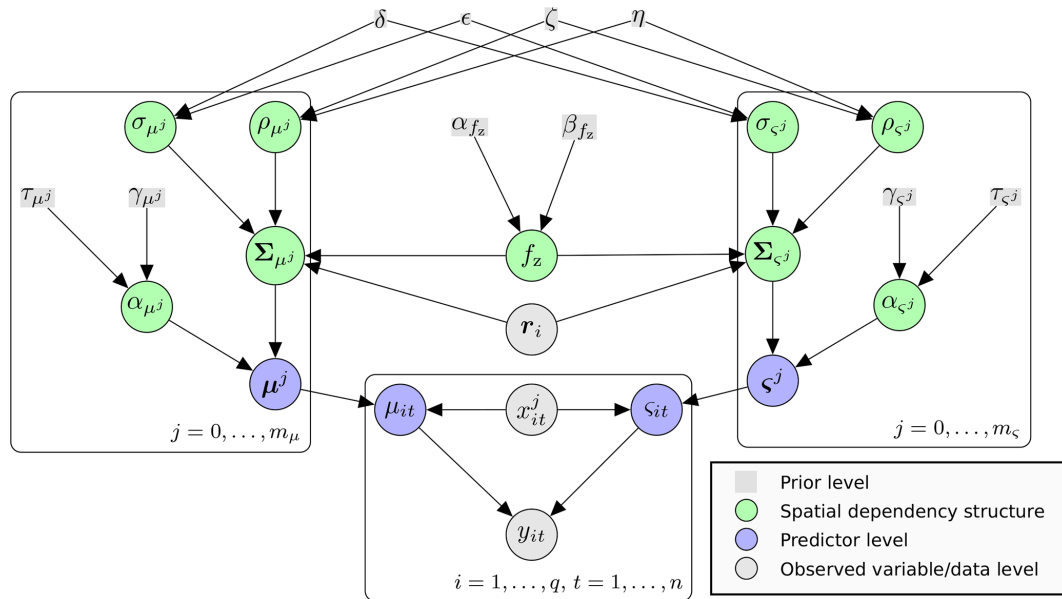


Figure 1. Directed acyclic graph for SpatBHM with predictand y_{ik} , covariates $x_{j,ik}$ and station coordinates $\mathbf{r}_i$, for location i and time t . m_μ/ζ refer to the respective number of covariates for location and scale. Please refer to Table A for a comprehensive notation reference.

In our study we use a Matérn-class isotropic covariance function with parameters $\lambda = (\sigma, \nu, \rho)$. In general, the Matérn covariance function takes a functional form involving the gamma function $\Gamma(x)$ and modified Bessel functions of the second kind $\mathcal{K}_\nu(x)$. In the case of half-integer values for the smoothness parameter ν , this expression can be simplified as a product of a polynomial and an exponential term (Rasmussen and Williams, 2008; Gneiting et al., 2012). Therefore, we decided to fix the smoothness parameter at $\nu = 3/2$ for numerical convenience. With the definitions above for the parameters we obtain the following formulation for the covariance functions

$$\text{Cov}_{\mu^j}(h) = (\sigma_{\mu^j})^2 \left(1 + \frac{\sqrt{3}h}{\rho_{\mu^j}} \right) \exp\left(-\frac{\sqrt{3}h}{\rho_{\mu^j}}\right) \quad (5)$$

and

$$\text{Cov}_{\zeta^j}(h) = (\sigma_{\zeta^j})^2 \left(1 + \frac{\sqrt{3}h}{\rho_{\zeta^j}} \right) \exp\left(-\frac{\sqrt{3}h}{\rho_{\zeta^j}}\right). \quad (6)$$

The Matérn-class covariance models are recommended by Stein (1999) for the statistical interpolation of spatial data, due to their mean squares differentiability and their suitability for application in more than one dimension. Moreover, preliminary tests demonstrated a higher numerical stability for the Matérn-model than for the exponential covariance function.

The introduction of the covariance function results in the incorporation of two additional parameters for each covariate at the process level. In the following, the two parameters, σ_{μ^j/ζ^j} and ρ_{μ^j/ζ^j} , are referred to as the sill and range of

the covariance function for $\mu^j(\mathbf{r})$ and $\zeta^j(\mathbf{r})$, although these terms do not directly correspond to the terms in the exponential covariance function. Therefore, terms “process variance” and “correlation length” also exist (Gneiting et al., 2012). In contrast to the approach proposed by Stein (1999), we directly model the covariance function using the expression $(\sigma_{\mu^j})^2$ and $(\sigma_{\zeta^j})^2$ (i.e. in form of a variance), rather than the more conventional form of a standard deviation. This is done for the purpose of ensuring consistency in the variance of the GRF at a given location and the variance of the marginal normal distribution at that location.

3.4 Distance metric for the spatial dependency structure

We model the spatial fields on the Earth’s surface in geographical coordinates with longitude λ , latitude φ and altitude z , yielding $\mathbf{r} = (\lambda, \varphi, z)$. Therefore, the distance between the training locations is calculated using the distance on a sphere, often called great-circle distance, which we calculate as

$$d_{\text{gc}}(\mathbf{r}_i, \mathbf{r}_k) = 2R_E \arcsin \sqrt{\frac{1 - \cos \Delta\varphi + \cos \varphi_i \cos \varphi_k (1 - \cos \Delta\lambda)}{2}}, \quad (7)$$

for $\Delta\varphi = \varphi_j - \varphi_i$ and $\Delta\lambda = \lambda_j - \lambda_i$ with the Earth’s radius R_E that we take at 6371 km. The inclusion of mountain top stations into the training data poses a challenge for the spatial interpolation due to the existence of valley locations with nearby mountain stations (e.g. Braunlage and Brocken in the Harz Mountains). Despite considerable differences in gust behavior due to the different topographical altitudes, the two

stations are treated as if they were close to each other. This deteriorates the estimation of the dependency structure at all stations. It is therefore necessary to achieve a separation of valley and mountain stations to improve the estimation of the spatial dependency structure. To this end, the station altitude is included in the distance metric used for the covariance structure of the GRFs. This is achieved by introducing an elevation offset to the distance between mountain and valley locations. The altitude difference between two stations is scaled by a factor f_z and added as an increment to the distance along a great circle yielding

$$d_v(\mathbf{r}_i, \mathbf{r}_k) = f_z(|z_k - z_i|), \quad (8)$$

where $|\cdot|$ represents the absolute value, and z_i the station altitude at location $\mathbf{r}_i$. f_z is the scaling factor, which we treat as an additional parameter to be estimated. As the Matérn-covariance function requires a Euclidean metric, the full distance between locations i, k is calculated as

$$d(\mathbf{r}_i, \mathbf{r}_k) = \sqrt{d_{\text{gc}}^2(\mathbf{r}_i, \mathbf{r}_k) + f_z^2(z_k - z_i)^2}. \quad (9)$$

This approach artificially increases the distance between two stations that are in close proximity but at markedly distinct elevations, while maintaining the general spatial dependency structure for locations in flat terrain.

As noted by Gneiting (2013), Matérn covariance functions are not positive definite on the sphere when combined with the great-circle distance and a smoothness parameter $\nu > 0.5$. Nevertheless, we adopted the Matérn-3/2 covariance model because its local smoothness properties yielded numerically more stable covariance matrices for the short distances present in our data. In addition, since we only model a small portion of the sphere, the numerical distortions that could otherwise result in invalid covariance matrices, are unlikely to arise.

To verify that our covariance matrices remain positive definite in practice, we conducted an experiment using the station coordinates and the prior distributions described in Sect. 4.1. As shown in Appendix E, this analysis revealed a parameter region in which the covariance matrices became non-positive definite, specifically when $\rho \sim \mathcal{O}(10^5)$ km. However, as the posterior ranges in our application remain within $\mathcal{O}(10^3)$ km, we conclude that the chosen combination of distance metric and covariance function is valid for the spatial domain under consideration.

4 Model training and verification

4.1 Specifying prior distributions

In order to perform Bayesian inference, it is necessary to specify prior distributions for the model parameters. These prior distributions summarize the researcher's prior knowledge and assumptions regarding the distribution of these parameters. To ensure a fast convergence of the Markov chains,

Table 1. Prior distributions of the model parameters used for SpatBHM. The parameters of the spatial dependency structure, σ and ρ , are simulated using the same prior distribution for all regression coefficients μ^j and ζ^j .

Parameter	Prior SpatBHM
μ^z	$\mathcal{N}(\gamma_{\mu^z}, \tau_{\mu^z}^2) \equiv \mathcal{N}(0.23, 0.1)$
f_z	$\text{Gamma}(15, 0.15)$
α_{μ^j/ζ^j}	$\mathcal{N}(\gamma_{\mu^j/\zeta^j}, \tau_{\mu^j/\zeta^j}^2)$
σ_{μ^j/ζ^j}	$\text{InvGamma}(\delta = 5, \epsilon = 2)$
ρ_{μ^j/ζ^j}	$\text{InvGamma}(\zeta = 1.05, \eta = 100)$

we use informative priors for the means of the GRFs, α_{μ^j/ζ^j} and weakly informative priors for the parameters of the spatial dependency structure $\sigma_{\mu^j/\zeta^j}, \rho_{\mu^j/\zeta^j}$. The impact of the priors in our configuration is small, as the extensive data set reduces the contribution of the prior belief on the inference process. Thus, the primary information is drawn from the training data. A summary of the prior distributions used in this study is presented in Table 1.

For the expectation values α_{μ^j/ζ^j} of the GRFs, we choose normal priors with parameters γ_{μ^j/ζ^j} and τ_{μ^j/ζ^j}^2 . We first estimate the μ^j, ζ^j from the linear model individually at each station. Subsequently, the prior distribution assumes the expectation values for each coefficient from the fit of the linear model on all stations simultaneously. The prior variances were informed by the variability of the results from individual station fits and rounded to facilitate interpretation and reporting. The resulting prior distributions for the model defined in Sect. 5.1 are $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\gamma} = (4.74, 1.49, -0.89, 0.22)^T, \boldsymbol{\tau} = (2, 1, 0.3, 0.1)^T)$ and $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\gamma} = (0.53, 0.4, -0.23)^T, \boldsymbol{\tau} = (0.2, 0.2, 0.1)^T)$.

In the case of positive definite parameters, such as the sill and range of the GRFs, inverse gamma priors are selected. We use the same priors for all spatial dependency structures, independent of μ^j/ζ^j . Consequently, a single prior value is set for the shapes δ, ζ and the scales ϵ, η of the inverse gamma distributions. All sill parameters are assigned $\sigma_{\mu^j/\zeta^j} \sim \text{InvGamma}(5, 2)$, constraining them to a reasonable value range regarding the variability of the parameters. All range parameters are assigned $\rho_{\mu^j/\zeta^j} \sim \text{InvGamma}(1.05, 100)$, favoring values between 0 and 500 km but otherwise not effecting major influence on the parameter estimation.

The scaling factor f_z for the elevation offset in Eq. (8) is assigned a Gamma-distributed prior, to ensure positive definiteness. The expectation value of f_z is derived from a quasi-geostrophic scale analysis and is set to the ratio of the Coriolis force and the buoyancy force on synoptic scales. This ratio is approximated as $\frac{N}{f}$ with the Brunt-Väisälä frequency N and the Coriolis parameter f . In the mid-latitudes, typ-

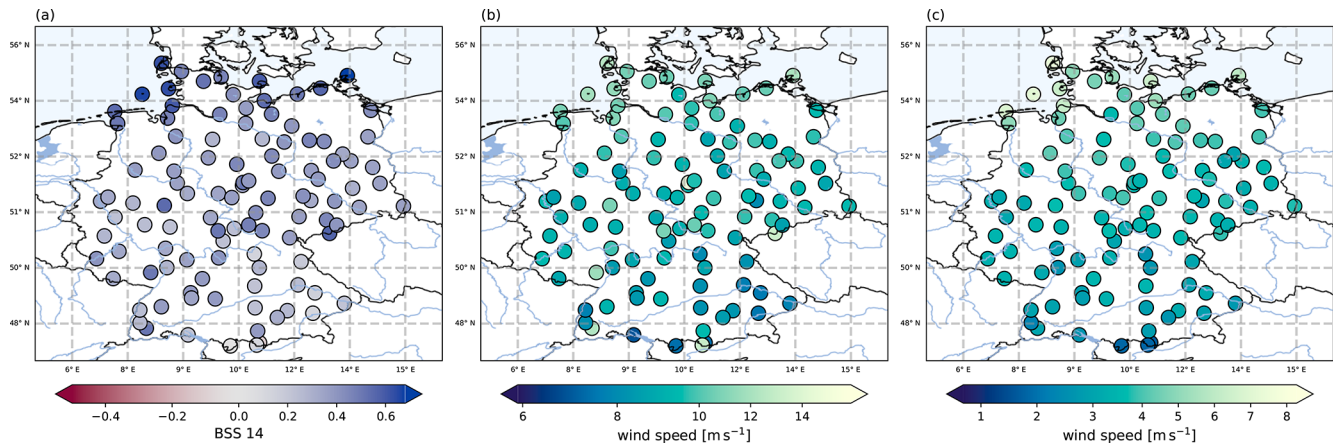


Figure 2. (a) Spatial distribution of the skill of LocMod against climatology for exceedances of the 14 m s^{-1} threshold (BSS 14). Blue colors indicate positive skill, red colors indicate negative skill. Each dot is one SYNOP station used for training. (b) Mean FX value used for training. (c) Mean V_m value used for training.

ical values are found around 100. Hence, we model f_z as $\text{Gamma}(15, 0.15)$.

4.2 Parameter estimation

The Bayesian parameter estimation was performed using Stan (Stan Development Team, 2022). Stan is a statistical software package designed for Bayesian inference based on Markov chain Monte Carlo (MCMC) techniques. In MCMC, the posterior distribution of the parameters is approximated through the sampling of Markov chains, whose transition properties are tuned such that they converge to a target distribution (Gelman et al., 2014; Gilks et al., 1995). We accessed Stan via its python interface, pystan (Riddell et al., 2021). Stan makes use of the No-U-Turn-sampler (NUTS, Hoffman and Gelman, 2014), which presents an adaptive and tuning-easy implementation of Hamiltonian Monte Carlo, leading to reasonably quick convergence of the Markov chains while remaining user-friendly.

For the training of SpatBHM, the sampling parameters for the tuning of the step size and the acceptance rate of the new proposals are set to the default values proposed by Hoffman and Gelman (2014). A chain length of 1500 was simulated, with the first 500 iterations discarded as burn-in. The sampler is initialized with the expectation value from the fit of the linear model with constant coefficients μ^j, ζ^j (Sect. 5.1) and assuming a value of one for all sills and a value of 60 km for all ranges of the GRF. While the initialization is identical for all covariance parameters, the posterior estimates of the GRFs show distinguishable spatial dependency structures, depending on the parameter and model in question. Therefore, it can be concluded that it is fair to use the same priors and initialization for all covariance structures, as differences can successfully be inferred from the data. Furthermore, all parameters converge to stable distributions within the first 500 iterations, indicating that the Markov chains

have reached convergence and that the resulting posterior samples can be assumed as representative of their limit distributions.

All spatio-temporal predictors are standardized before training. They enter the parameter estimation process as deviations from the spatio-temporal mean, expressed in units of the standard deviation of the entire training data set. This normalization is performed with respect to the spatio-temporal mean rather than the local station means to preserve the spatial information contained in the local climatologies of the predictor variables.

4.3 Spatio-temporal prediction

4.3.1 Predictive distribution of gusts

In the Bayesian framework, probabilistic predictions are generated via a posterior predictive distribution. This is the conditional distribution of new observations, given the training observations. The posterior predictive distribution can be derived from the posterior distributions of the parameters through integration over the uncertainties. In practice for a BHM, this integration is achieved by a multi-stage sampling strategy (cf. Gelman et al., 2014, Chap. 5).

In our gust model, the posterior distributions of the parameters are represented by the MCMC samples. The uncertainties associated with the parameter estimates are integrated by iteratively drawing from the posterior samples, using a target sample size of $N = 10000$. For each $k = 1, \dots, N$, one regression coefficient vector β_i is drawn from the MCMC samples, for location i . Subsequently, the Gumbel parameters μ_{it}, ζ_{it} are calculated via the regressions specified in Eq. (23) and the current local observations of the predictor variables $x_{it}^{\mu/\zeta}$. Ultimately, a single wind gust value is sampled from the resulting Gumbel distribution. This procedure is repeated N times to generate a predictive sample of

wind gusts, conditional on location and predictor variables. In consequence, the predictive distribution, as represented by the predictive sample, describes the whole uncertainty inherent to the model predictions. From the predictive samples, predictive quantiles and threshold exceed probabilities are calculated via the empirical distribution function, a method which is consistent with our verification framework outlined in Sect. 4.4 (Krüger et al., 2021). The quantiles are calculated using the median-unbiased method, as defined by Hynman and Fan (1996). The threshold excesses are calculated as the fraction of draws from the posterior distribution that exceed the threshold. As our sample size is large, the effect of the ensemble size on the estimation of the threshold excess probabilities is negligible (Ferro, 2014).

4.3.2 Spatial interpolation of model parameters

In order to make predictions at p unobserved locations $s_l, l = 1, \dots, p$, the spatial fields for the regression coefficients $\mu^j(\mathbf{r})$ and $\zeta^j(\mathbf{r})$ need to be simulated at the respective prediction location, while conditioning on the parameter estimates at the training locations $\mathbf{r}_i, i = 1, \dots, q$. The conditional simulation yields a multivariate Gaussian distribution for the prediction locations.

Following Stein (1999), we can write the joint distribution of $\boldsymbol{\vartheta}_r$ and the predictions $\boldsymbol{\vartheta}_s = \mu^j/\zeta^j(s_1, \dots, s_p)$ as

$$\begin{pmatrix} \boldsymbol{\vartheta}_s \\ \boldsymbol{\vartheta}_r \end{pmatrix} \sim \mathcal{N} \left[\begin{pmatrix} \boldsymbol{\alpha}_{\boldsymbol{\vartheta}} \\ \boldsymbol{\alpha}_{\boldsymbol{\vartheta}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_s & \boldsymbol{\Sigma}_{sr} \\ \boldsymbol{\Sigma}_{rs} & \boldsymbol{\Sigma}_r \end{pmatrix} \right]. \quad (10)$$

The covariance matrices $\boldsymbol{\Sigma}_r$, $\boldsymbol{\Sigma}_s$, and $\boldsymbol{\Sigma}_{rs}$ are calculated using the Bayesian parameter estimates σ_{μ^j/ζ^j} and ρ_{μ^j/ζ^j} corresponding to the regression coefficient in question and the Matérn covariance function from Eq. (5). $\boldsymbol{\Sigma}_r$ is the covariance matrix at the observed locations, $\boldsymbol{\Sigma}_s$ the covariance at the predicted locations and $\boldsymbol{\Sigma}_{rs} = (\boldsymbol{\Sigma}_{sr})^T$ represents the cross-covariance matrix between the estimated locations and the predicted locations. Representations of the predictive parameters are obtained by sampling from $\mathcal{N}(\boldsymbol{\vartheta}_{s|r}, \boldsymbol{\Sigma}_{s|r})$, where

$$\boldsymbol{\vartheta}_{s|r} = \boldsymbol{\alpha}_{\boldsymbol{\vartheta}} + \boldsymbol{\Sigma}_{sr} \boldsymbol{\Sigma}_r^{-1} (\boldsymbol{\vartheta}_r - \boldsymbol{\alpha}_{\boldsymbol{\vartheta},0}) \quad (11)$$

represents the estimate of the conditional mean and

$$\boldsymbol{\Sigma}_{s|r} = \boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_{sr} \boldsymbol{\Sigma}_r^{-1} \boldsymbol{\Sigma}_{rs} \quad (12)$$

the conditional covariance matrix. This process is repeated for all 1000 elements of the MCMC samples to obtain a posterior sample of the spatial parameter fields at the prediction locations. This posterior sample is then used for the multi-stage sampling strategy outlined in Sect. 4.3.1.

4.4 Verification

The resulting predictive distributions are compared to observation values and evaluated by scoring rules to assess

the models' predictive quality. To this end, we apply cross-validation in space and time. In time, the data is separated in two approximately equal-sized data sets, one for training and one for evaluation. This separation ensures that all predictions used for model verification are out-of-sample with respect to the training period. To account for possible autocorrelation and an annual cycle, the data set is partitioned by year, with odd-numbered years designated for training and even-numbered years for verification. To guarantee that the spatial prediction is out-of-sample, that is, that no data from the verifying location was used for model training, we employ leave-one-out cross-validation by always withholding the data for one station. Subsequently, the μ^j/ζ^j are predicted in space for the withheld location and the wind gusts are predicted using the days from the evaluation years. This way, time series of probabilistic predictions and observations are generated for all stations and dates.

The post-processing models are evaluated and compared by out-of-sample predictive performance. Therefore, the verification is conducted via proper scoring rules (Gneiting and Raftery, 2007). We use Brier score (Brier, 1950) for the verification of threshold exceedance probabilities and the quantile score (Koenker, 2005; Koenker and Machado, 1999; Friederichs and Hense, 2007), also referred to as general piecewise linear score (Gneiting et al., 2023), for the verification of predictive quantiles. Brier score (BS) is given by (Wilks, 2019)

$$S_B^u(F, y) = (p_u - \mathbb{I}\{y > u\})^2, \quad (13)$$

and quantile score (QS) is given by Bentzien and Friederichs (2014) as

$$S_Q^\tau(F, y) = \rho_\tau(y - F^{-1}(\tau)), \quad (14)$$

where $\rho_\tau(u) = \tau u$ if $u \geq 0$, and $\rho_\tau(u) = (\tau - 1)u$ otherwise. Both scores are negatively oriented, so that lower values imply better predictive performance.

The decomposition of the scoring rules into uncertainty, reliability and resolution components is also calculated (Murphy, 1973; Bröcker, 2009). We refer to the decomposition components as “uncertainty”, “miscalibration” and “resolution”. With respect to miscalibration (MCB), we follow Gneiting et al. (2023). In Murphy's original decomposition (Murphy, 1973), a lower reliability value implies a higher reliability of the model prediction. Hence, to avoid semantic ambiguity, we use the term “miscalibration” instead. With the term “resolution” (RES), we adhere to the standard nomenclature within the calibration-refinement factorization of the joint distribution of forecasts and observations (Wilks, 2019). The uncertainty (UNC) only depends on the observations and is defined as the score of the climatology forecast and quantifies the data-immanent difficulty to forecast the variable in question. UNC and MCB are negatively oriented, so that higher values correspond to lower skill. RES is positively oriented, so that higher values correspond to larger

skill. The total score S can be decomposed as

$$S = \text{MCB} - \text{RES} + \text{UNC}. \quad (15)$$

Estimating the decomposition from real-valued data presents challenges regarding the binning of the forecast-observation pairs (Bentzen and Friederichs, 2014; Atger, 2004). In this work, we follow the approach by Dimitriadis et al. (2021), who develop a stable estimation of the decomposition terms of BS that produces consistent, optimally binned, reproducible and PAV (pool adjacent violators)-based (CORP) results. Similarly, the CORP approach for general metrics, including QS, was derived by Jordan et al. (2022). This procedure performs an isotonic regression on the forecast and observation values using the pool-adjacent-violators algorithm (PAVA, Ayer et al., 1955; Miles, 1959; Best and Chakravarti, 1990; de Leeuw et al., 2009), thereby conditionally calibrating the forecast values on the observations. For BS, the threshold exceedance forecasts are calibrated to the mean of the corresponding binary-converted observations. For QS, the forecast values are calibrated to the respective quantile functional of the corresponding gust observations. Miscalibration is calculated from the original score S and the recalibrated score S_{rc} as

$$\text{MCB} \equiv \text{REL} =: S - S_{\text{rc}}, \quad (16)$$

and the discrimination/resolution part is obtained from the recalibrated score and the marginal score as

$$\text{RES} \equiv \text{DSC} =: S_{\text{rc}} - S_{\text{mg}} \equiv S_{\text{rc}} - \text{UNC}, \quad (17)$$

where the marginal score S_{mg} refers to the score obtained from the climatology forecast (Gneiting et al., 2023). The marginal score is identical to the uncertainty. In this work, we follow the CORP-based QS decomposition from Gneiting et al. (2023), using their own implementation (Wolfram, 2022), and the BS decomposition from Dimitriadis et al. (2021), which is implemented in R (R Development Core Team, 2023) in the package `reliabilitydiag`.

The skill of a model can be assessed through derived skill scores (Wilks, 2019). Skill scores are obtained from a scoring rule S via

$$\text{SS} = \frac{S - S_{\text{ref}}}{S_{\text{perf}} - S_{\text{ref}}} = 1 - \frac{S}{S_{\text{ref}}}, \quad (18)$$

where S_{perf} refers to the perfect score. The simplification on the right-hand side of Eq. (18) is possible, as both BS and QS have a perfect score of $S_{\text{perf}} = 0$. If the climatology is used as reference, the skill score can be obtained from the decomposition components as

$$\text{SS} = \frac{\text{RES} - \text{MCB}}{\text{UNC}}. \quad (19)$$

Skill scores are positively oriented, indicating higher skill with higher values. The perfect forecast has a value of 1.

Positive values indicate superior skill to the reference model, whereas negative values indicate lower skill. If we refer to skill against climatology in this work, we refer to a locally calculated climatology of the gust observations at each station.

4.4.1 Verification of the posterior distribution

Additionally, we assess the quality of the parameter estimation using the continuous ranked probability score (CRPS, Hersbach, 2000). CRPS evaluates the whole probability distribution and can be regarded as either the integral of BS over all real-numbered thresholds (Hersbach, 2000) or the integral of QS over all possible quantiles (Gneiting and Raftery, 2011). We use CRPS in the form given by Gneiting and Raftery (2007) as

$$S_{\text{CRP}} = \mathbb{E}_{\text{F}}|Y - o| - \frac{1}{2}\mathbb{E}_{\text{F}}|Y - Y'|, \quad (20)$$

for forecast Y and observations o , with Y' being an independent random copy of Y . The expectation values $\mathbb{E}_{\text{F}}$ can be approximated numerically or calculated in closed form under distributional assumptions. As the marginal posterior distributions of our gust model are sufficiently close to Gaussian, we approximate CRPS using the closed form CRPS (Gneiting et al., 2005)

$$S_{\text{CRP}}(\mu, \sigma^2, o) = \sigma \left\{ \frac{o - \mu}{\sigma} \left[2\Phi\left(\frac{o - \mu}{\sigma}\right) - 1 \right] + 2\phi\left(\frac{o - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\} \quad (21)$$

for mean μ and variance σ^2 of a Gaussian distribution. Φ and ϕ denote the cumulative distribution function and the probability density function of the standard normal distribution. For a joint evaluation of the posterior distribution for the regression coefficients, we apply the energy score (ES, Gneiting and Raftery, 2007), which is the multivariate extension of the CRPS. ES is given as

$$S_{\text{E}} = \mathbb{E}_{\text{F}}\|Y - o\| - \frac{1}{2}\mathbb{E}_{\text{F}}\|Y - Y'\|, \quad (22)$$

where $\|\cdot\|$ denotes an euclidean distance. We scale the regression coefficients by the inverse of their spatial variance obtained from maximum likelihood fits on the data from each station to ensure that they contribute homogeneously to the distance. In the case of ES, we approximate the expectation values $\mathbb{E}_{\text{F}}$ numerically by averaging over all representations from the MCMC samples.

5 Results

5.1 Spatially constant and local models

The benefit of SpatBHM for wind gust post-processing is assessed against a spatially constant baseline model (ConstMod), where all regression coefficients μ^j and ς^j are

spatially constant. Although ConstMod is technically not a Bayesian hierarchical model, we treat it as such and use the same Bayesian inference methods as for SpatBHM. To ensure consistency in the model formulation, the priors of the corresponding GRF mean values are assigned to the constant regression coefficients of ConstMod. Provided that the covariates are available at the prediction location, ConstMod also allows the post-processing of wind gusts at unobserved locations, since it is assumed that the regression coefficients do not vary in space in this case.

In search of a strong baseline, we conducted a selection of predictors and predictand for the linear model, based on predictive skill (Appendix B). Models using $FX - V_m$ as predictand showed higher skill as models predicting FX directly. By testing various combinations of predictors, we find the best linear model for the location μ_{it} and the scale σ_{it} parameter as

$$\begin{aligned}\mu_{it} &= \mu^0 + \mu^1 V_{\max, it} + \mu^2 V_{m, it} + \mu^z \Delta z_i \\ \log(\varsigma_{it}) &= \varsigma^0 + \varsigma^1 V_{\max, it} + \varsigma^2 V_{m, it},\end{aligned}\quad (23)$$

for locations $i = 1, \dots, q$ and times $t = 1, \dots, n$, and predicting $FX - V_m$. $V_{\max}$ represents the maximum wind speed in COSMO-REA6, V_m represents the mean horizontal wind from COSMO-REA6 and Δz_i is a spatial covariate for the influence of the station altitude, given by $\Delta z_i = z_{\text{synop}, i} - z_{\text{COSMO}, i}$. For the training of ConstMod, we simulated Markov chains of length $N = 1250$, using only 250 iterations as burn-in because of the lower complexity of the model.

The results are further contextualized by comparing them to the results from a local model (LocMod), where individual regression coefficients are estimated at each station. While in the cases of ConstMod and SpatBHM, the local data are excluded for each station in cross-validation, LocMod is specifically trained on the local data only. Consequently, we expect LocMod to represent the maximum attainable skill from the data with the type of linear model utilized.

As visible from the map in Fig. 2a, even LocMod demonstrates negative skill against the climatology at some locations. The stations with negative skill are mostly situated in mountainous terrain, except for QSS 0.999, where a more general decrease in skill is observed (not shown). This suggests that the COSMO-REA6 predictor data lack explanatory power for the wind gust observations at these stations, pointing to deficiencies in the representation of topography in COSMO-CLM. Additionally, there is a coastward gradient with lower possible skill in southern Germany and higher skill near the coast.

Figure 3 shows that the maximum attainable skill depends on the local mean wind characteristics. LocMod skill against climatology is strongly correlated with the mean wind at each station over all investigated threshold exceedance probabilities and quantiles, except for the most extreme quantile. This mean wind dependency of the attainable skill leads to a more successful post-processing at windy locations, which

explains why the highest level of attainable skill is found near the coastlines and in mountainous areas. Particularly, the station with lowest attainable skill is Garmisch-Partenkirchen, which is also the station recording the lowest average gust speed among the 109 locations.

Figure 4 shows the cross-validated skill scores for ConstMod compared to the local climatology and LocMod for the 14 and 18 m s^{-1} threshold (cf. Deutscher Wetterdienst, 2024) and the 0.75, 0.95, 0.99 and 0.999-quantiles obtained at all investigated locations. ConstMod demonstrates a high level of skill against the local climatology. The skill is high over all investigated distributional features with average skill score values between 10 % and 45 % and approaches the skill of LocMod. The highest median skill is found for the 0.75-quantile at values greater than 40 %. The 0.75-quantile also has the least spread in skill, implying that the prediction skill for this quantile is spatially homogeneous. For higher thresholds and outer quantiles, we observe less skill and also locations with negative skill against climatology. However, the number of instances with negative skill remains similar to the instances with negative skill found from LocMod. Moreover, the 18 m s^{-1} threshold is rarely exceeded within the data set, resulting in a large dispersion of BSS values. Likewise, the 0.999-quantile is extreme and therefore suffers from the generally low predictability of extreme events. The comparison to LocMod shows that ConstMod already closely approaches the maximum attainable skill, and therefore offers an excellent baseline model that is hard to beat.

5.2 Effect of the elevation offset

We evaluate the benefit of including the elevation offset from Eq. (8) in the distance metric for the covariance calculation in a simple version of SpatBHM. In this simple version only one GRF is assigned to μ^0 , while the rest of the regression coefficients are kept spatially constant. The model is trained once using Eq. (7) as distance metric and once using Eq. (9). As a result of the elevation offset, the estimated values for the range parameter $\rho_{\mu, j}$ for SpatBHM become larger and show higher stability during cross-validation. The larger range parameter values lead to a smoother spatial dependency structure, as the elevation offset effectively mitigates the impact of the mountain stations on their direct neighbors. This enhances the discernibility of the large-scale spatial dependency structure.

The added value in terms of miscalibration for SpatBHM is displayed in Fig. 5. For BS 14 (Fig. 5a), the elevation offset has a positive impact on the calibration of SpatBHM, which is improved particularly for stations with large MCB. For QS 0.75 (Fig. 5b), we see a similar effect, although it is not as pronounced as for BS 14. There are cases where the introduction of the elevation offset has a negative impact, but these stations generally show a good calibration, regardless of the elevation offset. For the more extreme 0.99 quantile (Fig. 5c), mean MCB is higher than without the offset,

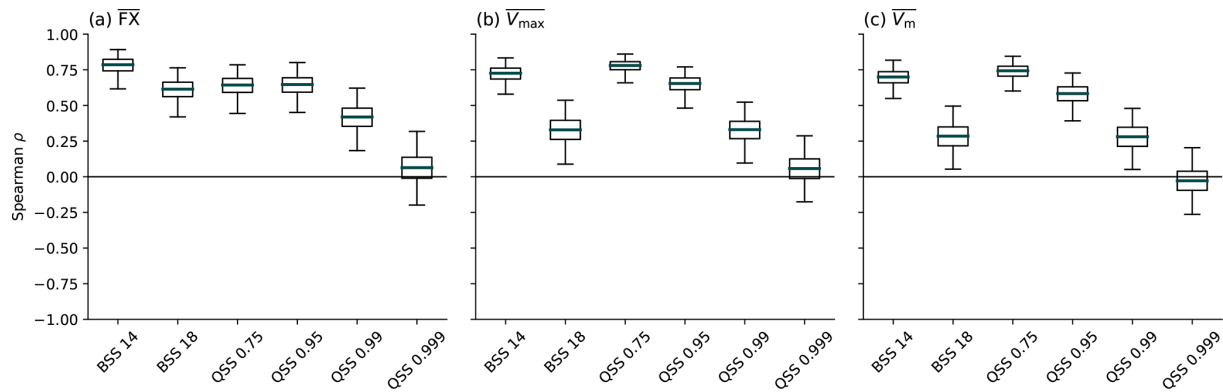


Figure 3. Spearman rank correlation coefficient between various variables and LocMod skill against climatology, shown for exceedance probabilities of the 14 m s^{-1} (BSS 14) and 18 m s^{-1} (BSS 18) threshold, and the 0.75 to 0.999 quantiles (QSS 0.75 to QSS 0.999). Correlations are shown (a) for the gust observations FX, (b) the maximum wind speed $V_{\max}$ and (c) the mean wind V_m . Boxes represent the interquartile range and whiskers extend to the 0.01 and 0.99-quantiles. Uncertainty estimates are based on 5000 bootstrap iterations. In each iteration, both the dates for the calculation of the mean wind and the stations to calculate the correlation are resampled.

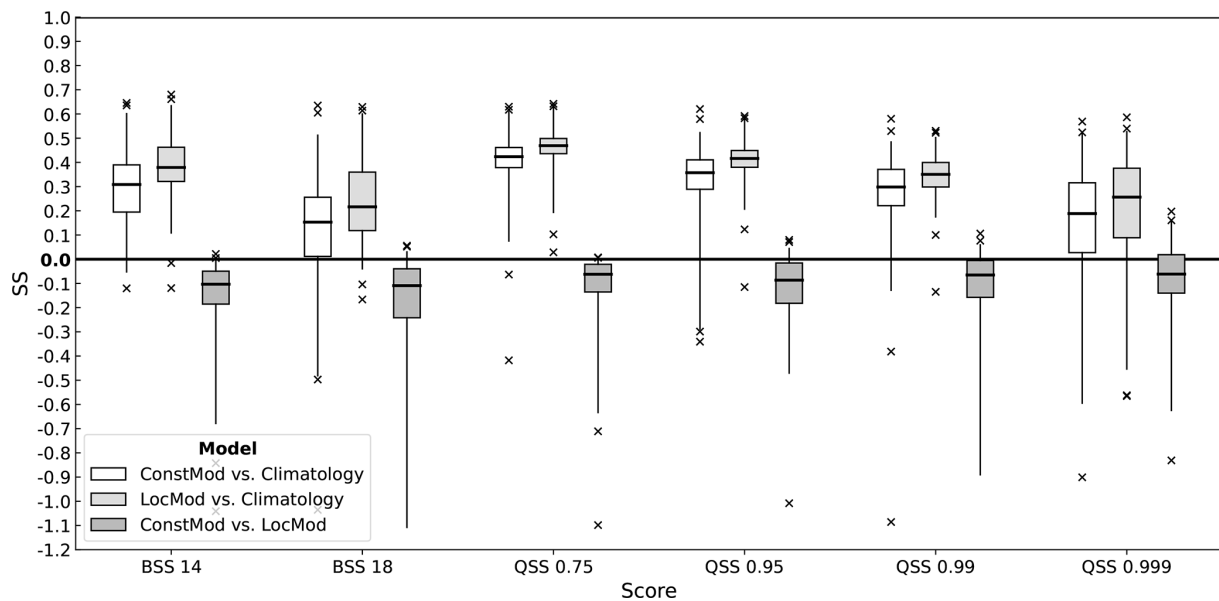


Figure 4. Cross-validated skill scores of ConstMod and LocMod against climatology, and ConstMod against LocMod for exceedance probabilities of the 14 m s^{-1} (BSS 14) and 18 m s^{-1} (BSS 18) thresholds, and the 0.75 to 0.999 quantiles (QSS 0.75 to QSS 0.999). Each box plot contains skill scores at 109 locations. Boxes represent the interquartile range and whiskers extend to the 0.01 and 0.99-quantiles. Bold lines mark the median skill scores. Median score values are given in Table B1.

but this is mainly caused by the deterioration at one single station, i.e. Zugspitze. Since this station repeatedly appears as an outlier in our investigation, one possible explanation is that the Zugspitze station has a significantly different gust climatology compared to other locations and might be poorly represented by its corresponding grid cell in COSMO-REA6.

In summary, the elevation offset enables the mountain stations to be retained in the training data set and mostly improves the calibration of the model without unduly influencing the model skill at already well represented locations.

5.3 Predictive performance of SpatBHM

We conducted a non-exhaustive model selection in order to find a good combination of spatial fields for the parameters (Appendix C). For each GRF, all local representations at the training locations have to be estimated, leading to a significant increase in the number of parameters. Thereby, the computational burden for the parameter estimation increases substantially with more GRFs, without necessarily improving the predictive skill. We decided for a compromise using a model with only a limited number of spatial param-

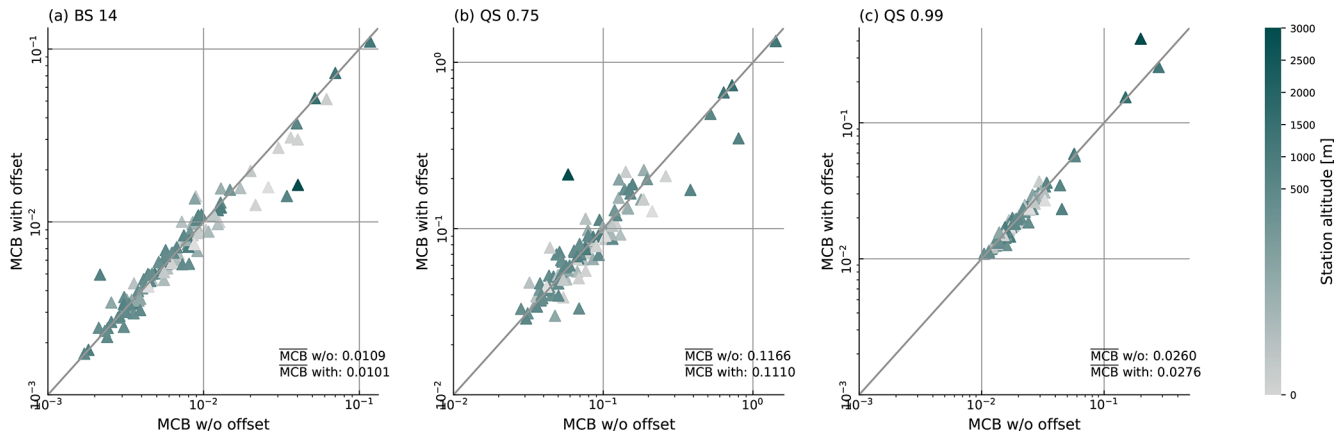


Figure 5. Miscalibration of SpatBHM with and without the elevation offset in the covariance function. Each data point represents the miscalibration at one station. The coloring represents station altitude of the station in question. Miscalibration values are shown for (a) BS 14, (b) QS 0.75, and (c) QS 0.99. The gray line represents the identity function. Note that the axes are scaled logarithmically, so that smaller values are overrepresented.

ters that shows high skill, while keeping the computations for the parameter estimation practicable. The SpatBHM version that was found best has two spatial regression coefficients: the intercept for the location parameter μ^0 and the regression coefficient for V_m for the location parameter μ^2 . Please note that μ^3 which controls a purely spatial covariate, is always assumed as spatially constant. There is limited evidence that a spatial field for the scale parameter improves the estimation of the threshold exceedance probabilities. However, it also leads to a significant loss in QS, which is the reason why it was discarded. The fitted values of the model parameters are shown in Table C2.

Figure 6 shows the skill score values of SpatBHM against ConstMod for the 14 and 18 ms^{-1} threshold and the 0.75, 0.95, 0.99 and 0.999-quantiles obtained at all investigated locations. The BSS values of SpatBHM compared to ConstMod for the threshold exceedance probabilities are close to 0. However, the median and the larger part of the interquartile range are found above 0, pointing to a slight improvement in prediction quality compared to ConstMod. The median BSS is found around 2 %. For the prediction quantiles, there is a clear improvement in skill, especially for the higher quantiles. The median skill score values range from 2 %–5 %, depending on the exact quantile. For the 0.95 and 0.99-quantiles, almost the whole interquartile range is positive and for the 0.999-quantile, SpatBHM skill nearly achieves the same skill as LocMod. At least for the representation of the prediction quantiles, the relative skill of SpatBHM compared to ConstMod can be regarded as a proof of concept of the spatial hierarchical approach.

However, there are some instances, where SpatBHM performs poorly compared to ConstMod. These are mostly stations, where the local wind gust characteristics change rapidly over space. In some cases, the spatial mean of the GRF is more representative for the local gust model than the

coefficients at the surrounding stations. Therefore, the spatial interpolation procedure introduces a bias and the skill of SpatBHM is reduced compared to ConstMod. Nevertheless, the skill against the climatology forecast remains high at the majority of these locations.

Given the spatial model's notable predictive skill for higher quantiles, we proceed to evaluate its performance in stronger wind conditions. Absolute extreme events are evaluated in terms of the exceedance probability of high thresholds. Due to the scarcity of observations for high threshold exceedances, it is not reasonable to calculate BS for each station individually. Instead, we compute the average BS over all stations. Thereby, we increase the number of observations and obtain interpretable values for BSS, which are shown in Fig. 7. As higher thresholds, the next three warning levels by the DWD, i.e. 25, 29 and 33 ms^{-1} (Deutscher Wetterdienst, 2024), are selected.

SpatBHM performs very similar to ConstMod on the higher thresholds. Only for the highest threshold (33 ms^{-1}), it has visibly higher skill. The skill is lowest for the $u = 18$ and $u = 25 \text{ ms}^{-1}$ thresholds. The skill against the climatology in Fig. 7b provides context on the skill for the threshold exceedance probabilities: Owing to the extreme nature of the events, neither SpatBHM nor ConstMod show any significant skill against the climatology for the highest threshold, and the skill for all thresholds above 25 ms^{-1} is generally low. In summary, the spatial approach does not significantly improve the prediction of threshold excess probabilities from the spatially constant approach, but neither does it perform significantly worse.

In order to quantify SpatBHM's performance for quantile predictions in higher wind conditions, we stratify the data and calculate QS for the largest 25 % of the forecast values. The stratification is based on the forecast values, rather than the observations, in order to avoid the forecaster's dilemma

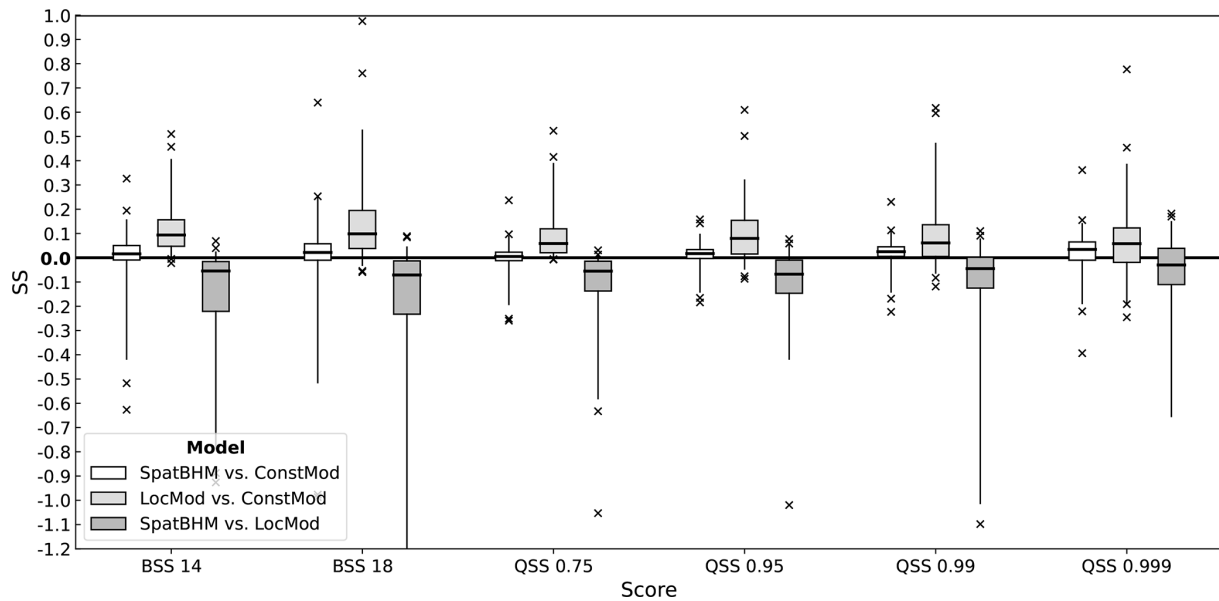


Figure 6. As Fig. 4 but for cross-validated skill scores against of SpatBHM and LocMod against ConstMod and SpatBHM against LocMod.

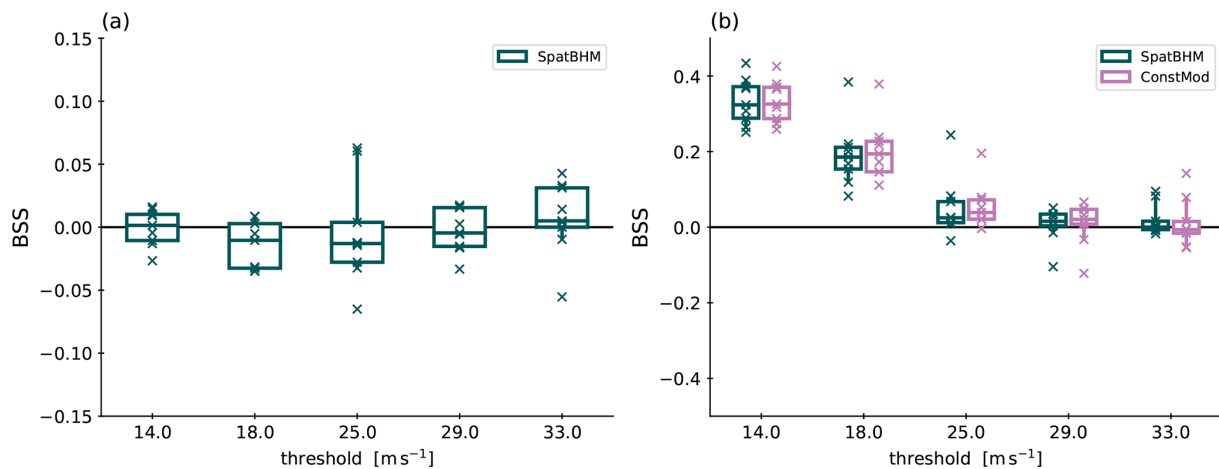


Figure 7. (a) BSS of SpatBHM against ConstMod, aggregated over all stations. Each boxplot comprises 9 values, one for each year included in the evaluation data set. The skill score values are calculated from the mean score at all stations throughout each year. (b) As in (a) but for SpatBHM and ConstMod against climatology.

(Lerch et al., 2017). The top 25 %-wind gust events are selected from the whole data set. The majority of these events is observed at coastal locations and mountain tops. As the stratification significantly reduces the data at some stations, the score values are averaged over all locations. We stratify once based on the predictions from SpatBHM and once based on the predictions from ConstMod. For the estimation of uncertainties, score values are calculated for each year of the verification data set, resulting in a sample of 9 independent score estimations. The sample size is insufficient to yield statistically robust results, but it provides an indication of the potential outcome. The resulting skill score values for SpatBHM and ConstMod are depicted in Fig. 8.

In contrast to the threshold exceedance probabilities, the prediction quantiles in Fig. 8a almost consistently exhibit higher skill for SpatBHM than for ConstMod, across all evaluated quantiles. The spatial model exhibits considerably higher quantile skill for the top 25 % gust events than for the complete data set. This skill improvement is found for both types of stratification, so it does not depend on the model used for the stratification, although there are minor differences. Similar to the separate evaluation for all locations, the spatially aggregated skill compared to ConstMod increases with higher quantile. Moreover, SpatBHM consistently outperforms ConstMod, independent of which of the two models was used to select the top 25 % of predic-

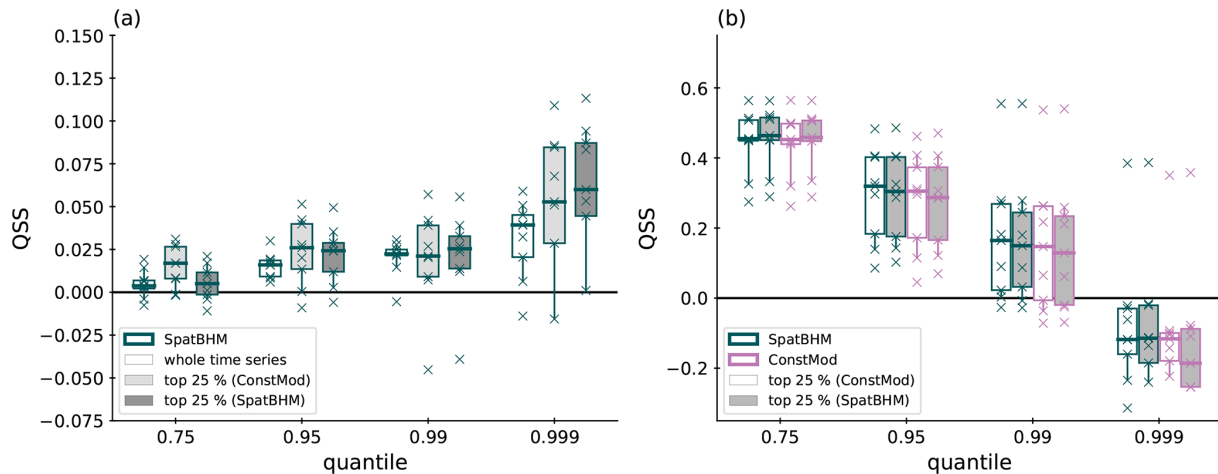


Figure 8. (a) QSS for SpatBHM against ConstMod. Each boxplot comprises one value for each of the 9 years included in the evaluation data set. The skill score values are aggregated over all 109 investigated locations. The white shaded boxes represent the average annual skill evaluated on the whole time series. The light and dark gray shaded boxes represent the skill, when only using the top 25 %, stratified on the highest respective quantile forecasts selected from ConstMod and SpatBHM. (b) Same for SpatBHM against climatology in comparison to ConstMod. The time series for verification are stratified based on the predictions from ConstMod (white) and SpatBHM (gray).

tions. Therefore, the spatial modeling approach can be considered to particularly improve the prediction of rare or extreme events in terms of prediction quantiles. However, in this context it should be noted, that the overall skill of both models compared to the climatology decreases with higher quantile (Fig. 8b).

For a better understanding of the origin for the improved skill, Fig. 9 shows the score decomposition from Eq. (15) for BS and QS. The inter-model differences of MCB and RES for the prediction of threshold exceedance probabilities (cf. Fig. 9c and d) are small. MCB is better in SpatBHM for the higher thresholds, whereas RES is better in ConstMod for the higher thresholds. However, the inter-model differences are more distinctive for the prediction quantiles in Fig. 9a and b. MCB is higher for SpatBHM and lower for ConstMod, pointing to poorer calibration in the spatial approach. The 0.999-quantile is an exception in that MCB is similar for both models and slightly better for SpatBHM for both the complete and reduced data sets. A possible reason for the inferior calibration of SpatBHM compared to ConstMod is the spatial interpolation procedure. For locations with wind gust characteristics similar to the spatial mean, ConstMod already has high skill and the spatial approach would not be required. If a station exhibits distinct wind gust characteristics from its surrounding stations, the kriging process results in a value too close to nearby observations and introduces a bias. This bias can only partially be removed by the elevation offset in Eq. (8). Consequently, SpatBHM only exhibits better calibration at locations, where the wind gust characteristics diverge from the spatial mean, yet align with the surrounding stations. Therefore, the calibration of SpatBHM is reduced compared to ConstMod. RES, shown in Fig. 9b and d, is very

similar for both models. RES is slightly better for SpatBHM, most visibly for the extreme quantiles. As RES values are an order of magnitude larger than MCB, the resolution part of the score is what defines the local skill. Therefore we conclude that SpatBHM outperforms ConstBHM not due to better calibration but rather due to a better resolution.

In summary, our results can be regarded as a proof of concept, that the spatial hierarchical approach can contribute to an improved wind gust post-processing from reanalysis data, as it provides on average more spatially homogeneous estimates of the observed wind gusts. Our stratified analysis showed that the increase in skill is especially noteworthy for higher prediction quantiles and in higher wind conditions, which points to a better representation of the variability of the wind gusts. Nevertheless, these results do not generalize to locations with low average windiness with low LocMod skill against climatology. Consequently, the spatial probabilistic post-processing provides high skill for coastal stations and some mountain tops, and less skill for valley locations. The high skill at coastal stations, however, suggests that the spatial model successfully detects and corrects systematic regional gust behavior deviations.

5.4 Quality of the spatial parameter prediction

To assess the added value of SpatBHM over ConstMod with regard to the spatial prediction of the model parameters, we calculate the CRPS for the marginal posterior predictive distributions of the regression coefficients μ^i and ζ^i . As observations, the parameter estimations from the LocMod-fits are used (Sect. 5.1). In the case of ConstMod, the posterior predictive distribution is obtained from the MCMC samples (Sect. 4.2), while we use the posterior predictive distri-

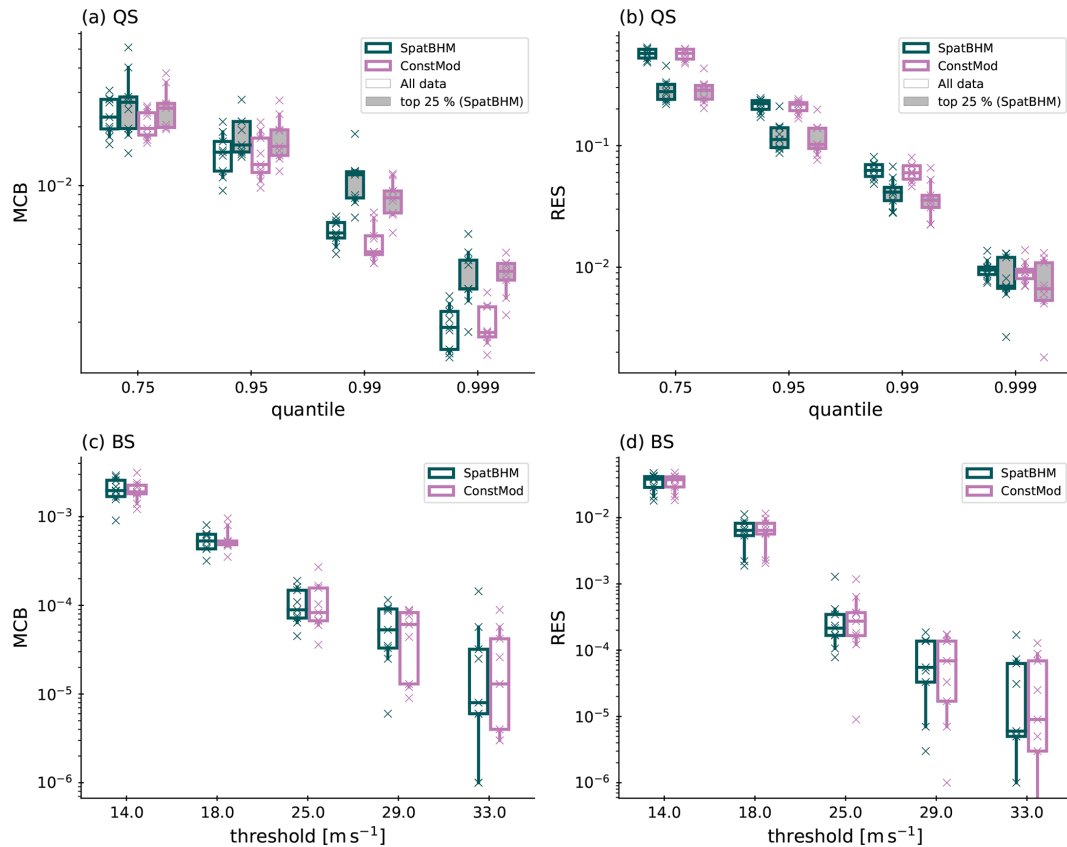


Figure 9. (a) QS miscalibration and (b) QS resolution for SpatBHM and ConstMod in comparison for evaluation on the complete data set and only the top 25 % of data. Results for stratification are based on SpatBHM. Specifications of box plots as in Fig. 8. (c) BS miscalibration and (d) BS resolution for SpatBHM (green) and ConstMod (pink). The score decomposition is calculated over the full time series without stratification.

bution obtained from the spatial interpolation for SpatBHM (Sect. 4.3.2). As LocMod is similarly estimated by MCMC, we calculate CRPS once for each element of the LocMod posterior samples, in order to obtain an estimate of the uncertainty of CRPS at each station. Both the posterior distribution of the constant coefficients in ConstMod and SpatBHM and the spatially interpolated distribution from SpatBHM are assumed as Gaussian, so that we evaluate CRPS following Eq. (21). The spatially constant coefficients in SpatBHM were sampled from their posterior distribution for this part of the evaluation, as are their counterparts in LocMod.

The resulting score values are displayed in Fig. 10 for the six regression coefficients. As μ^z was not estimated locally, the altitude predictor contribution $\mu^3 \Delta z_i$ in ConstMod and SpatBHM is added to the intercept for the location parameter, i.e. μ^0 , for each station. The CRPS values show a general improvement of the parameter prediction in SpatBHM compared to ConstMod. The superiority of SpatBHM is most pronounced for μ^1 and μ^2 and less for the scale coefficients. For the ζ^2 there are hardly any differences. From these values, we calculated the skill scores via Eq. (18), shown in Fig. 11a. The CRPS values translate into skill values of up

to 50 % for μ^2 of SpatBHM over ConstMod. As μ^2 is one of the GRFs, this clearly points to the added value of the spatial approach.

Following from this, we evaluate the multivariate posterior distribution of the regression coefficients using energy score (Eq. 22). As can be seen in Fig. 11b, ES shows similarly high skill values of SpatBHM against ConstMod, likewise found around 50 %. In the multivariate case, the skill is rather homogeneous over all stations, pointing to a clear superiority of SpatBHM with regard to the spatial prediction of the parameters.

5.5 Post-processing time series

The objective of the spatial hierarchical extreme value model is to provide an enhanced description of surface wind gusts in reanalysis by providing a probabilistic diagnostic for wind gusts based on a statistical post-processing of predictor variables. SpatBHM stands to the task quite well as illustrated by Fig. 12, depicting time series of postprocessed wind gust from SpatBHM at three distinct locations for September and October 2002. The observations consistently fall within the

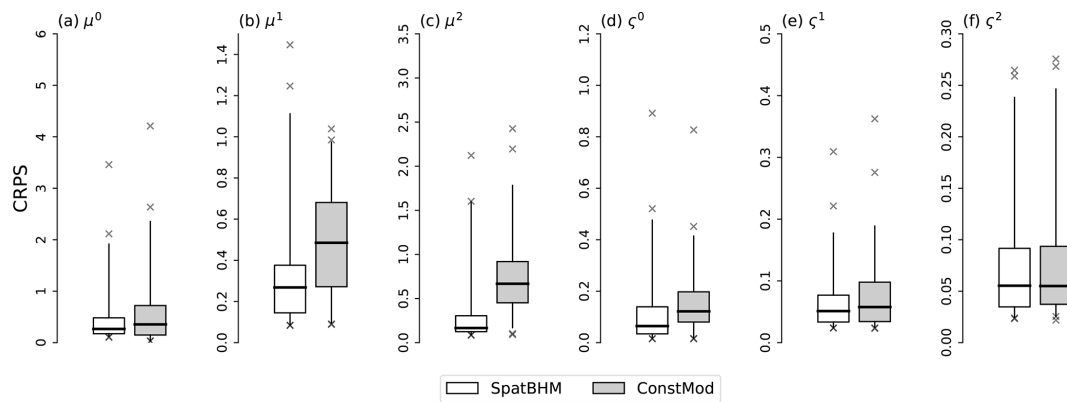


Figure 10. CRPS for the regression coefficients evaluated against the parameter estimations from LocMod fit, trained only on the local data. Score values are shown for (a) μ^0 including the altitude predictor and μ^z , (b) μ^1 , (c) μ^2 , (d) ζ^0 , (e) ζ^1 and (f) ζ^2 . Boxplot specifications as in Fig. 4.

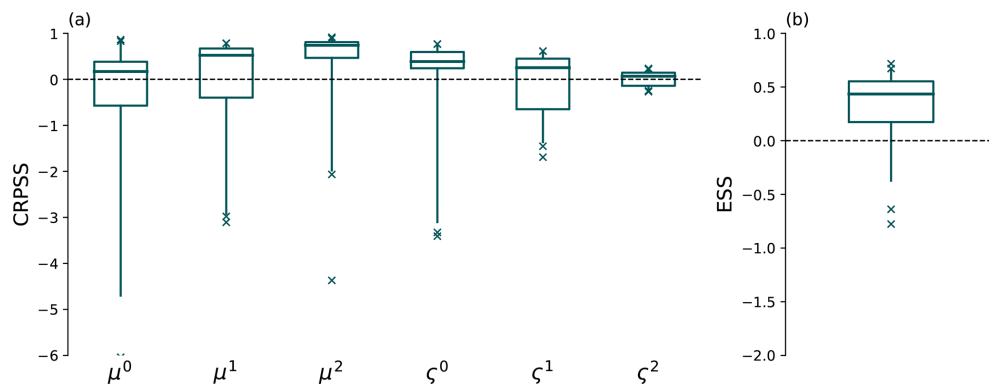


Figure 11. (a) CRPSS skill of SpatBHM compared to ConstMod for the regression coefficients. (b) Energy skill score of SpatBHM compared to ConstMod for the multivariate estimation of the regression coefficients. The single parameters are weighted with their spatial variance before multivariate aggregation. Boxplot specifications as in Fig. 4.

displayed distributional range. Moreover, at the first station, the median of the predicted distributions is visibly corrected towards the corresponding observations, indicating that the statistical model improves the calibration of the reanalysis. The third location in Fig. 12c is used to contrast the results with a location with no skill.

Figure 12a depicts the postprocessed wind gusts at Arkona, a coastal station in northeastern Germany. The post-processing is capable of capturing the transient weather patterns, which is especially evident from 21 September to 7 October. The skill score values against climatology are quite high with values larger than 50 %, and the observations typically lie inside the 98 % confidence interval of the predicted distributions. The forecasts are well calibrated at this location, as can be understood from the probability integral transform (PIT) histogram (Dawid, 1984) in Fig. 13a for the same station. In general, the histogram looks homogeneous with an indication of a too high dispersion of the forecast distribution.

For central German locations, SpatBHM shows a similarly strong performance, as illustrated by the example of Frankfurt in Fig. 12b. In most cases, the observed wind gusts are found close to the median of the predicted distribution. However, $V_{\max}$ and the observations display usually similar values at this location, so that the surface maximum wind speed in the reanalysis can already be regarded as a good prediction for the observation. At this location, the PIT-histogram (Fig. 13b) provides interesting added information on the gust prediction. Apparently, there is a consistent negative bias, combined with a too large variance of the forecast distribution. This suggests that the high skill is achieved by a bias-variance trade-off, where a larger variance compensates the negative bias for the quantiles and thresholds of interest.

The spatial post-processing encounters challenges in complex terrain with low wind speed, as illustrated by the example of Garmisch-Partenkirchen in Fig. 12c. This station is the location with the lowest average wind speed among the selected stations and even LocMod has no skill. At this location, SpatBHM corrects the observations systematically to

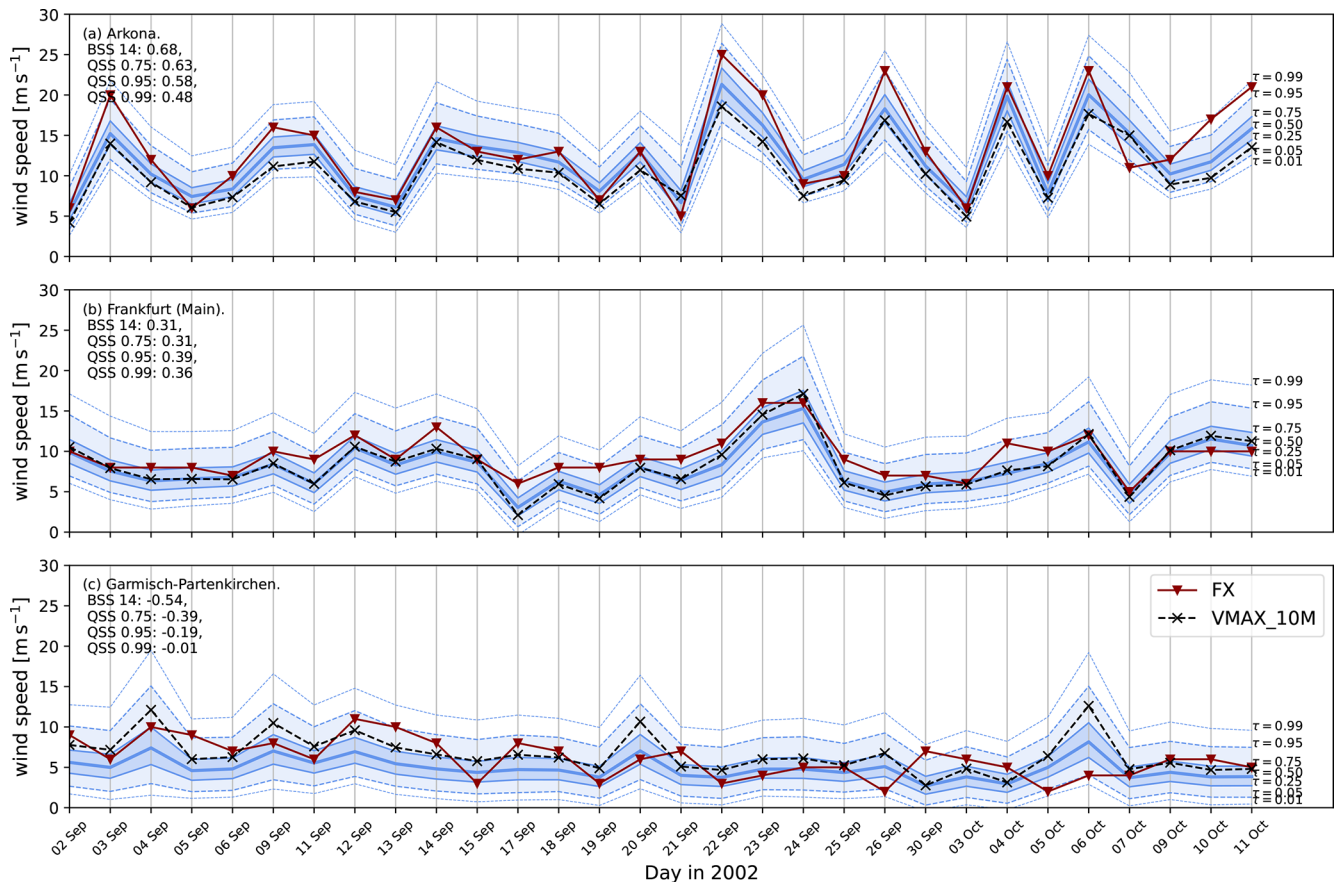


Figure 12. Predictive distributions generated by SpatBHM in 2002 for (a) Arkona, (b) Frankfurt (Main) and (c) Garmisch-Partenkirchen. The station observations FX are marked by the red line, while $V_{\max}$ is marked in black. The blue shading denotes the various quantiles of the predictive distribution, as given by the annotations. The displayed skill score values are calculated against climatology. It should be noted that the time series are incomplete due to missing observations. The predictions are cross-validated, i.e. each station was removed from the data prior to model training and prediction.

lower values, which leads to a pronounced negative bias. Accordingly, the PIT-histogram in Fig. 13c shows a similarly strong negative bias. Different to Fig. 13b, there is no visible excess of dispersion of the forecast. Therefore, the bias-variance trade-off does not succeed at this location. However, we still observe a slight improvement in skill with respect to ConstMod. The low skill at this location is due to the lack of explanatory power of the predictor variables at this particular location, as shown by the non-systematic bias and a low correlation between the reanalysis and the observations.

5.6 Post-processing on a spatial grid

Post-processing gridded data sets using SpatBHM involves the application of the spatial interpolation procedure to high dimensional data. In theory, this is readily doable because of the use of GRFs in the model formulation. However, the interpolation process involves the inversion of the covariance matrix, which is computationally expensive in high dimensions. Therefore, we advise to split the target region in subre-

gions and to draw iteratively from the GRF. We start by simulating a horizontal slice of the data set in the south and iteratively simulate horizontal slices further northward. In each iteration, the probability for the next horizontal slice is conditioned on the results of the previous draw for the adjoining slice in addition to conditioning on the fitted values at the training locations. This iterative drawing process is repeated, until the complete field has been simulated. In the absence of a unique station altitude for each grid cell, we set the Δz predictor to zero. Thereby, we enable a statistical downscaling of the gust model to the subgrid-scale in the event that the altitude predictor exists.

Figure 14a provides an example for the result of this iterative interpolation, obtained for μ^0 , for a grid of 121 times 161 cells over Germany. Each simulated horizontal slice contains 5 rows of grid cells, resulting in 605 simultaneously drawn representations in each iteration. The mean of μ^0 is a smooth spatial field with some visible minima and maxima, not all of which are easily explainable by the geography. However, some geographical features, such as the Hartz

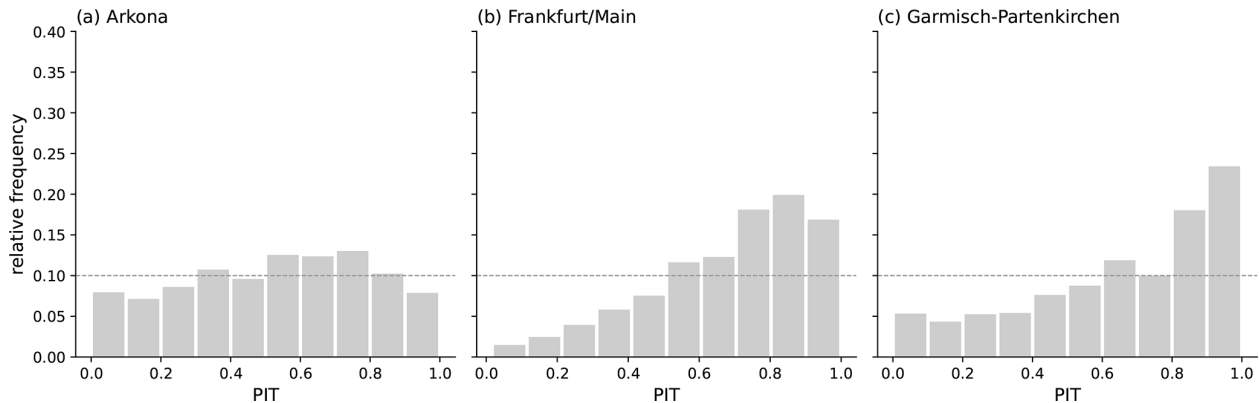


Figure 13. PIT histograms for the three stations from Fig. 12. PIT-values are obtained by evaluating the predicted empirical gust distribution function at the observed values.

Mountains in the center and the Cologne Lowland in the west are discernible. The parameter predictions are lower over the ocean and the adjoining coastlines. Areas that are sufficiently far away from any of the training locations, are estimated close to the spatial field mean. Owing to the small range parameter estimates (cf. Table C2), the process is still considerably rough, as it represents a local station correction to the spatial mean and SpatBHM does not include a nugget effect. A nugget effect could account for outlying stations, but we did not include it as it decreased the predictive performance in preliminary tests. Figure 14b shows the standard deviation of the predicted μ^0 -field, obtained from 100 draws of the same spatial field. The variability of the interpolated field is lower near the training locations and increases with distance. Mountain stations, such as Zugspitze, Brocken and Feldberg are not as distinctly visible as non-mountain stations, as the elevation offset (Eq. 9) successfully segregates them from their immediate surroundings. Thereby, their influence on the direct neighborhood is reduced, albeit they remain inside the training data set.

Following the interpolation of the GRFs, it is possible to estimate spatial fields of the Gumbel parameters of the wind gust distribution. Figure 15a and b depict such spatial fields for the location and the scale parameter, including the predictor variables. The shown area is a portion of the German North Sea coastline. The fields are obtained from SpatBHM by retrieving one representation of the interpolated spatial fields for the coefficients μ^0 and μ^2 and evaluating Eq. (23) for the predictor variables on 15 September 2010, while setting $\Delta z_i \equiv 0$. There is a strong contrast between the sea and land surfaces due to the contribution of the covariate variables. Both location and scale parameter are estimated higher over land than over the ocean. For the scale parameter, this contrast is even more pronounced. Apart from the land-sea-difference, the fields look reasonably smooth. Although the higher parameter predictions over land appear counterintuitive at first, they are plausible considering that SpatBHM does not predict the gust speed, but $FX - V_m$.

Adding V_m to the expectation value of the SpatBHM prediction results in overall higher postprocessed gust predictions over the ocean than over land as shown, in Fig. 15c. Concluding from this, wind gusts possess considerably less variability over the ocean and are found closer to the mean wind values. Therefore, the model parameters are estimated lower. Figure 15d shows the difference between the $V_{\max}$ field from COSMO-REA6 and the expectation value of the postprocessed field from SpatBHM. In most areas, the correction applied by SpatBHM to the expectation is negative, i.e. reducing the mean wind prediction. Only two regions display a positive correction, namely the northern depicted part of the ocean and two intricate bays (Jade bight, Weser estuary), which are standing out. This difference compared to the surrounding area comes from the interpolation of land-area location parameters, that are higher than location parameters over sea. Simultaneously, V_m is higher over the bay areas due to reduced roughness. The combination of these two factors leads to the distinctly positive correction. However, as we do not have any verifying observations over the ocean we refrain from making any statements about the accuracy of this correction. Nevertheless, including the land-sea-mask as spatial predictor can be a useful addition for SpatBHM in the future.

6 Conclusions

In this study, we present a spatial probabilistic post-processing approach (SpatBHM) for wind gusts from re-analysis models, with the COSMO-REA6 model serving as a case study. The probabilistic post-processing is based on Bayesian hierarchical modeling and extreme value statistics. Wind gust observations are modeled as a non-stationary Gumbel distribution, with location and scale parameters varying both spatially and temporally. The temporal dependency is modeled as a linear model on spatio-temporal predictor variables. The spatial dependency is included by letting the regression coefficients vary in space following a Gaussian random field with an isotropic covariance structure.

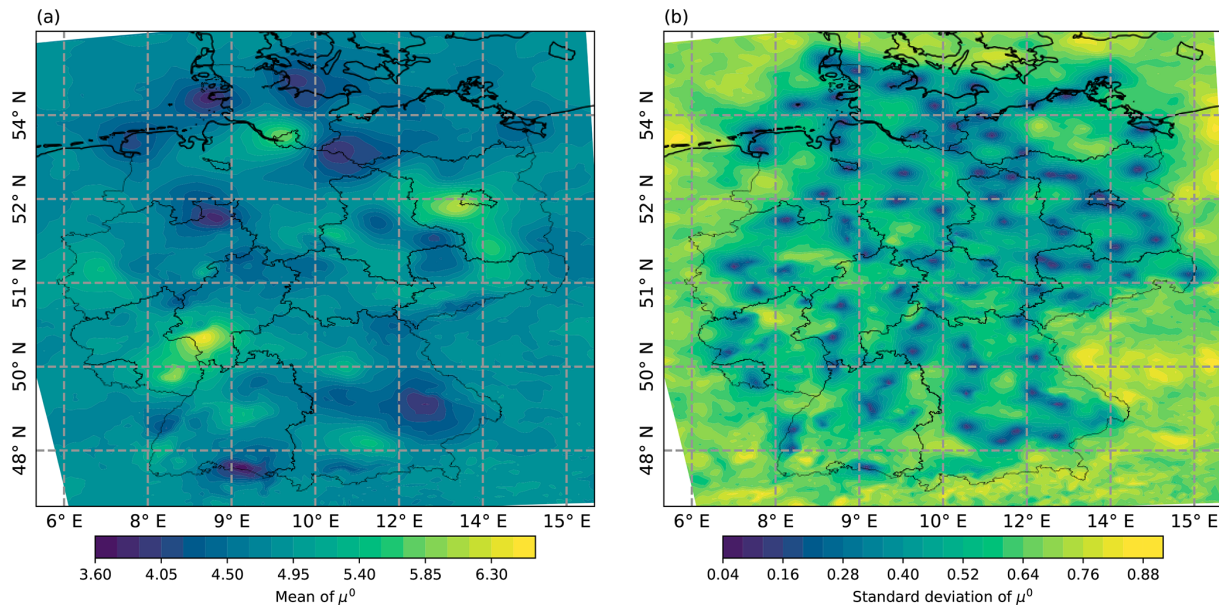


Figure 14. Posterior distribution of μ^0 after spatial interpolation. Left: expectation for the coefficient μ^0 , interpolated from SpatBHM. Right: The standard deviation of the same field based on 100 draws from the same model.

We compared the SpatBHM against a model with spatially constant coefficients (ConstMod) using proper scoring rules for quantiles and threshold exceedance probabilities and their decomposition to assess miscalibration, resolution, and uncertainty.

Although we used a computationally expensive MCMC-algorithm for model fitting, the sampling process can be adapted for operational use. Pystan (3.9.0) with Python (3.9.18) was run sequentially in a JupyterHub service running Debian GNU/Linux 12 (x86_64), requiring about one hour of sampling time for the final version of SpatBHM. Training can be sped up by about 25 % by selecting a more informative prior for the range parameters. The spatial interpolation to large data sets is a computational bottleneck because the parameter fields must be interpolated to the entire domain for all elements of the posterior sample. Thus, the entire process must be repeated N times. For this analysis, the process was carried in R (4.2.2) on a computer with an AMD Ryzen 2400G processor, running at 3.6 GHz using 29 GB of RAM, and an openSUSE Leap 15.5 system (x86_64). The process described in Sect. 5.6 with $N = 100$ consumed about 3 h of computation time, so the complete analysis can be reproduced on a standard desktop (≥ 8 cores, 16 GB RAM). For operational purposes, the spatial interpolation can also be run in parallel.

We have reached a proof of concept in favor of the spatial modeling approach, indicating that SpatBHM possesses higher skill than ConstMod for all investigated features of the predictive distribution, although we find that ConstMod already exhibits high skill against the climatology. While SpatBHM does not show much improvement for the pre-

diction of threshold exceedance probabilities, it clearly outperforms ConstMod for prediction quantiles by 2 %–5 % in terms of median skill, depending on the exact quantile under investigation. Moreover, SpatBHM demonstrates particular skill in predicting high quantiles, due to a better representation of the variability of wind gusts. This is evident more when focusing the evaluation on top quartile of the predictions. As a further advantage, the GRFs in SpatBHM can be readily interpolated to large data sets by an iterative simulation procedure. Therefore, the model facilitates simple post-processing including gridded data sets.

However, the skill of all models is low at locations with little wind. Provided that a contribution of the station altitude difference is included into the estimation of the spatial covariance structure, SpatBHM can enhance the predictive skill at these locations to some degree. The elevation offset stabilizes the estimation of the range parameter of the GRF and results in a more accurate and better calibrated representation of the spatial structure. However, even the spatial model falls short of attaining the same level of calibration as the simple climatology in these low wind conditions. An examination of the score decomposition in these cases revealed that the challenges observed at these locations are not solely caused by poor calibration, but also by a lack of model resolution. This is an indication of a lack of explanatory power of the predictor variables in low-wind conditions.

In conclusion, wind gust post-processing remains a challenging topic due to the localized nature, short duration, and inherent extremity of wind gusts. Further research is required to account for an annual or diurnal cycle and to assess the models' performance in different boundary layer regimes.

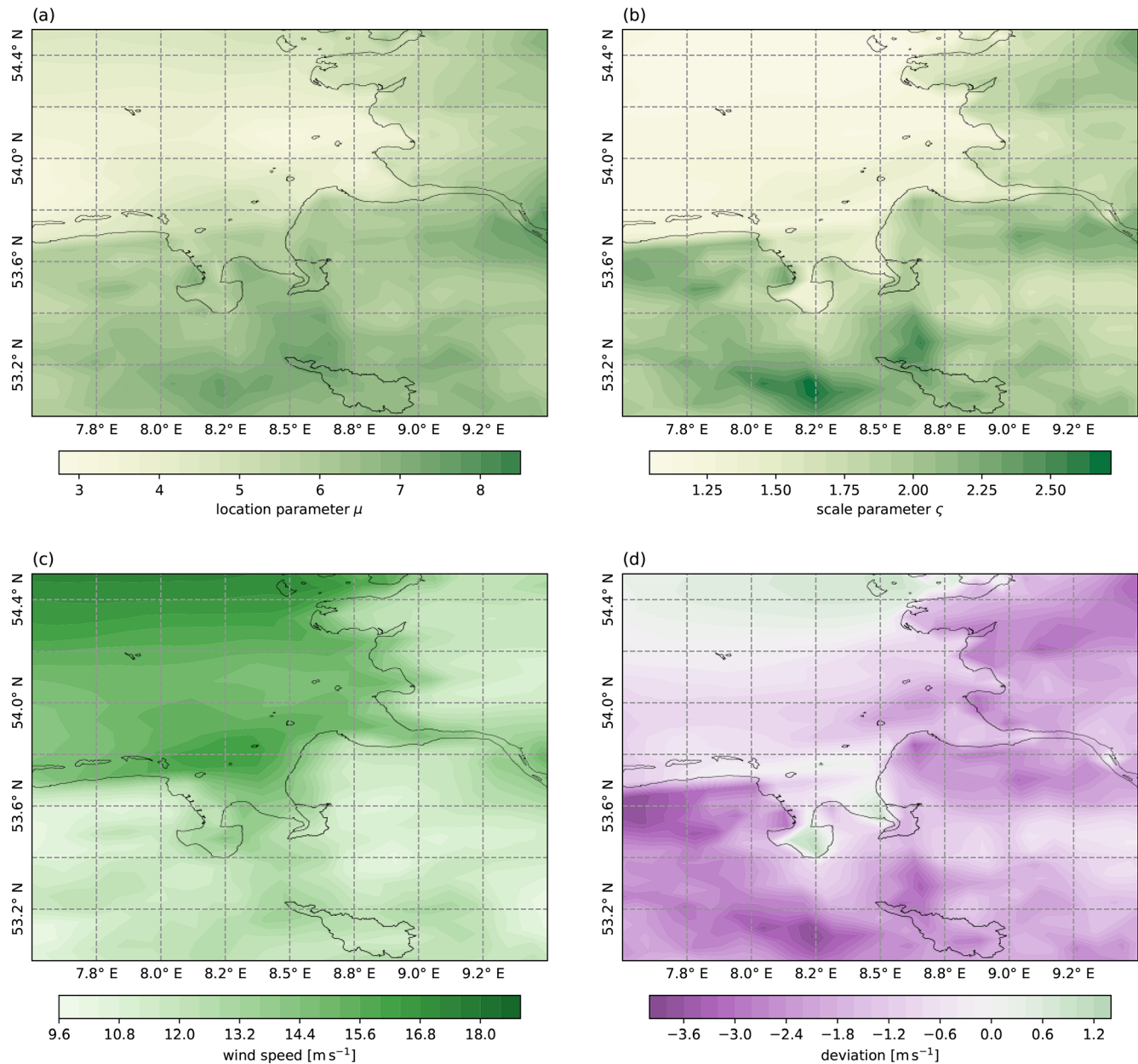


Figure 15. Spatial fields of (a) the location parameter and (b) the scale parameter on 15 September 2010, predicted by SpatBHM over the German North Sea area. (c) Postprocessed expectation value for FX from SpatBHM. (d) Difference between postprocessed expectation value and the $V_{\max}$ predictor field. Green shading indicates a positive deviation of SpatBHM and purple indicates a negative deviation.

One way to include the annual and diurnal cycles is to use harmonic functions with suitable frequencies as additional predictors. Further testing is required to determine whether their effects are best modeled as spatially constant or as GRFs. Currently, the model code allows the flexible inclusion of new predictors by adding columns to the predictor matrix and specifying whether their coefficient is modeled as constant or as a GRF. The spatial interpolation process remains unchanged, but the prediction code requires updates to include the harmonic function predictors. In summary, the

annual and diurnal cycles can be easily included, but their inclusion requires further model checking and comparison, as well as a more extensive data preprocessing. Finally, developing a spatial prediction procedure, that is capable of accounting for existing gust observations at specific locations, namely, conditionally sampling from the complete gust model, can be an interesting follow up to this work. As the spatial post-processing approach demonstrates added skill to existing linear approaches, it can contribute to an improved representation of wind gust characteristics in reanalysis.

Appendix A: List of symbols

y_{it}	observation at location $\mathbf{r}_i, i = 1, \dots, q$ and time $t_k, k = 1, \dots, n$
μ_{it}, ς_{it}	GEV parameters at location $\mathbf{r}_i, i = 1, \dots, q$ and time $t_k, k = 1, \dots, n$
$x_{j,ik}^\mu / x_{j,ik}^\varsigma$	set of covariates $j = 1, \dots, m_\mu$ for $\mu_{it} / j = 1, \dots, m_\varsigma$ for ς_{it}
$\mathbf{x}_{it}^\mu / \mathbf{x}_{it}^\varsigma$	vector of covariates $\mathbf{x}_{it}^\mu = (x_{0,ik}^\mu, \dots, x_{m_\mu,ik}^\mu) / \mathbf{x}_{it}^\varsigma = (x_{0,ik}^\varsigma, \dots, x_{m_\varsigma,ik}^\varsigma)$
μ_i^j / ς_i^j	time constant regression coefficients for covariate $x_{j,ik}^\mu / x_{j,ik}^\varsigma$
$\boldsymbol{\beta}_i^\mu / \boldsymbol{\beta}_i^\varsigma$	vector with regression coefficients $\boldsymbol{\beta}_i^\mu = (\mu_i^0, \dots, \mu_i^{m_\mu}) / \boldsymbol{\beta}_i^\varsigma = (\varsigma_i^0, \dots, \varsigma_i^{m_\varsigma})$
$\boldsymbol{\mu}^j / \boldsymbol{\varsigma}^j$	$\boldsymbol{\mu}^j = (\mu_1^j, \dots, \mu_q^j) / \boldsymbol{\varsigma}^j = (\varsigma_1^j, \dots, \varsigma_q^j)$
$\boldsymbol{\alpha}_{\mu^j}, \boldsymbol{\Sigma}_{\mu^j} / \boldsymbol{\alpha}_{\varsigma^j}, \boldsymbol{\Sigma}_{\varsigma^j}$	parameters of multivariate Gaussian distribution for regression coefficients $\boldsymbol{\mu}^j / \boldsymbol{\varsigma}^j$
$\alpha_{\mu^j}, \sigma_{\mu^j}, \rho_{\mu^j} / \alpha_{\varsigma^j}, \sigma_{\varsigma^j}, \rho_{\varsigma^j}$	parameters of homogeneous GRFs for regression coefficients $\mu^j(\mathbf{r}) / \varsigma^j(\mathbf{r})$
$\gamma_{\mu^j}, \tau_{\mu^j}^2, \delta_{\mu^j}, \epsilon_{\mu^j} \zeta_{\mu^j}, \eta_{\mu^j}$	parameters of priors for random field parameters of $\mu^j(\mathbf{r})$
$\gamma_{\varsigma^j}, \tau_{\varsigma^j}^2, \delta_{\varsigma^j}, \epsilon_{\varsigma^j} \zeta_{\varsigma^j}, \eta_{\varsigma^j}$	parameters of priors for random field parameters of $\varsigma^j(\mathbf{r})$
f_z	scaling factor for elevation offset

Appendix B: Predictor selection for linear models

The selection of the optimal predictor variables for the spatial hierarchical post-processing model was performed on ConstMod, which also serves as baseline model during evaluation in Sect. 5.1. We tested and evaluated a variety combinations of predictors and predictands. For ConstMod, all regression coefficients $\boldsymbol{\mu}^j$ and $\boldsymbol{\varsigma}^j$ are spatially constant. All model versions are trained as outlined in Sect. 5.1. Table B1 provides an overview over the all trained versions, including the selection of predictand and predictor variables for location and scale, as well as the median score values obtained at the 109 verification locations. The predictor variables follow the nomenclature as outlined in Sect. 5.1. The locally estimated model local model (LocMod), uses the same formulation as ConstMod 4 without the altitude predictor Δz . By estimating ConstMod locally for each station, the contribution of Δz is directly added to the intercept μ^0 . LocMod is trained on the local station data only, so it learns the local wind gust characteristics. For purposes of comparison, an overview of all ConstMod versions, LocMod and the climatology are presented in Table B1 together with the median score values for assessment of the predictive skill. Figure B1 shows the cross-validated skill scores for ConstMod compared to the local climatology for the 14 and 18 ms^{-1} threshold and the 0.75, 0.95, 0.99 and 0.999-quantiles for the different versions of ConstMod, obtained at all investigated locations. All versions of ConstMod demonstrate a high level of skill against the local climatology. The skill is large over all investigated distributional features with average skill score values between 10 % and 45 %.

The model version exhibiting the least skill is ConstMod 1, which only uses $V_{\max}$ as predictor. The

incorporation of the altitude predictor Δz (ConstMod 2) does not improve the skill for non-mountain stations, but it is visible that the quantile predictions at some locations are improved compared to the simplest model version. The first noticeable improvement in skill is observed when the predictand is modified to $\text{FX} - V_m$ (ConstMod 3). This finding is consistent with the findings of Friederichs et al. (2009). The improvement in skill is particularly evident for the 14 ms^{-1} threshold and for higher quantiles. The optimal spatially constant linear model was found to be the full version (ConstMod 4), as specified in Eq. (23). Therefore, we base all spatial models on ConstMod 4. The final fitted values of the parameters for the various ConstMod versions are presented in Table B2.

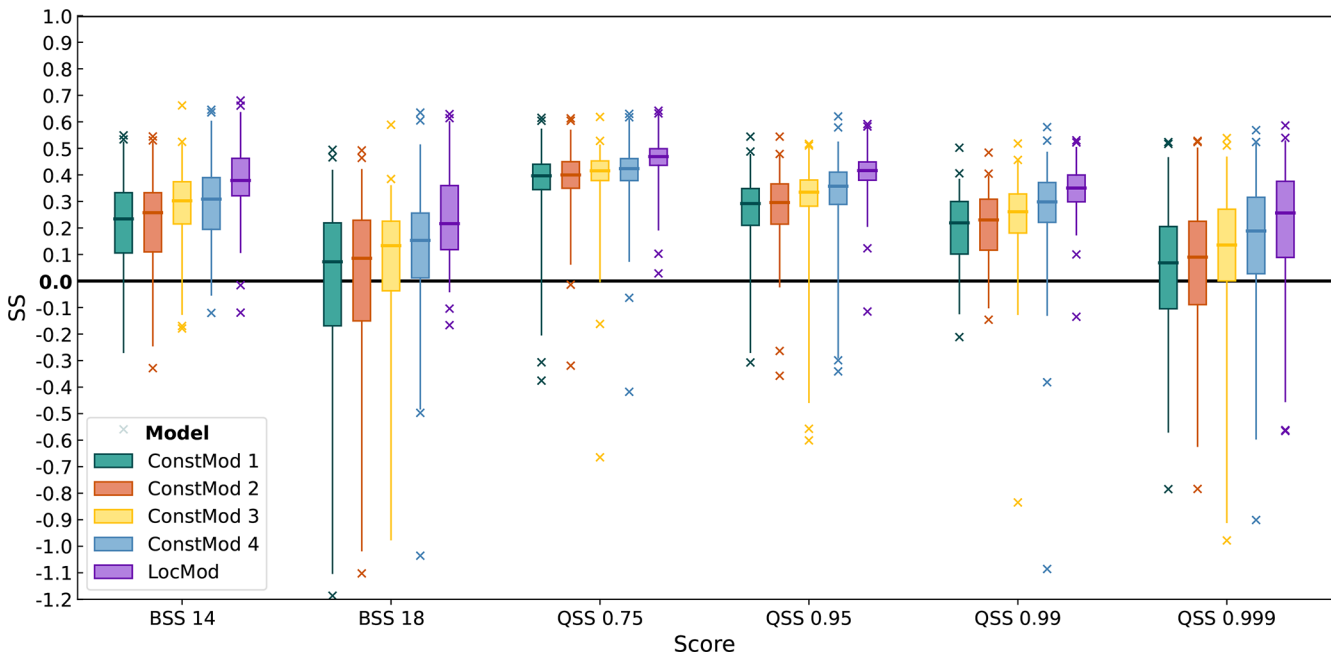


Figure B1. Cross-validated skill scores against climatology of various versions of ConstMod and LocMod (colored boxes) for exceedance probabilities of the 14 m s^{-1} (BSS 14) and 18 m s^{-1} (BSS 18) threshold, and the 0.75 to 0.999-quantiles (QSS 0.75 to QSS 0.999). Each box plot contains skill scores at 109 locations. Boxes represent the interquartile range and whiskers extend to the 0.01 and 0.99-quantiles. Bold lines mark the median scores. Model definitions and median scores are given in Table B1.

Table B1. Model versions and median score values for threshold excess probabilities and quantiles. For climatology, no predictors are used and parameters are estimated at each station, respectively. ConstMod 1–4 are estimated at all 109 stations simultaneously. LocMod refers to a local version of ConstMod 4 without Δz trained at each location separately.

Model	Predictand	Predictors μ	Predictors ζ	BS 14	BS 18	QS 0.75	QS 0.95	QS 0.99	QS 0.999
Climatology	FX	none	none	0.0637	0.0113	1.105	0.414	0.120	0.0163
ConstMod 1	FX	$V_{\max}$	$V_{\max}$	0.0479	0.0113	0.666	0.297	0.094	0.0144
ConstMod 2	FX	$V_{\max}, \Delta z$	$V_{\max}$	0.0477	0.0113	0.659	0.295	0.093	0.0142
ConstMod 3	$\text{FX} - V_{\text{m}}$	$V_{\max}, \Delta z$	$V_{\max}$	0.0443	0.0103	0.639	0.271	0.086	0.0134
ConstMod 4	$\text{FX} - V_{\text{m}}$	$V_{\max}, V_{\text{m}}, \Delta z$	$V_{\max}, V_{\text{m}}$	0.0433	0.0103	0.636	0.265	0.084	0.0127
LocMod	$\text{FX} - V_{\text{m}}$	$V_{\max}, V_{\text{m}}$	$V_{\max}, V_{\text{m}}$	0.0397	0.0092	0.597	0.244	0.077	0.012

Table B2. Parameter estimations for all ConstMod version together with their 99 % confidence intervals, assuming normal marginal posterior distributions. The values are obtained from training the model on the complete training data set.

Parameter	ConstMod 1	ConstMod 2	ConstMod 3	ConstMod 4
μ^0	8.334 ± 0.016	8.288 ± 0.016	4.681 ± 0.014	4.735 ± 0.014
μ^1	2.457 ± 0.016	2.445 ± 0.016	1.118 ± 0.014	1.843 ± 0.027
μ^2	—	—	—	-0.890 ± 0.025
μ^z	—	0.152 ± 0.007	0.221 ± 0.007	0.226 ± 0.008
ζ^0	0.680 ± 0.005	0.673 ± 0.005	0.556 ± 0.005	0.535 ± 0.005
ζ^1	0.330 ± 0.005	0.332 ± 0.005	0.218 ± 0.005	0.403 ± 0.009
ζ^2	—	—	—	-0.239 ± 0.008

Appendix C: Selection of spatial fields for SpatBHM

Table C1. SpatBHM versions with model name, spatially variable regression coefficients and cross-validated median score values obtained from the 109 stations. In each model, the remaining regression coefficients are assumed to be spatially constant.

Model	Spatial parameters	BS 14	BS 18	QS 0.75	QS 0.95	QS 0.99	QS 0.999
SpatBHM 1	μ^0	0.0427	0.01	0.633	0.261	0.082	0.0124
SpatBHM 2a	$\mu^0, \mu^1(V_{\max})$	0.0434	0.0101	0.6308	0.260	0.081	0.0121
SpatBHM 2b	$\mu^0, \mu^2(V_m)$	0.0435	0.01	0.6314	0.2598	0.0805	0.0121
SpatBHM 2c	μ^0, σ^0	0.0427	0.0101	0.6293	0.2628	0.0833	0.0129
SpatBHM 3	$\mu^0, \mu^1(V_{\max}), \mu^2(V_m)$	0.0436	0.01	0.6331	0.2615	0.0806	0.0122

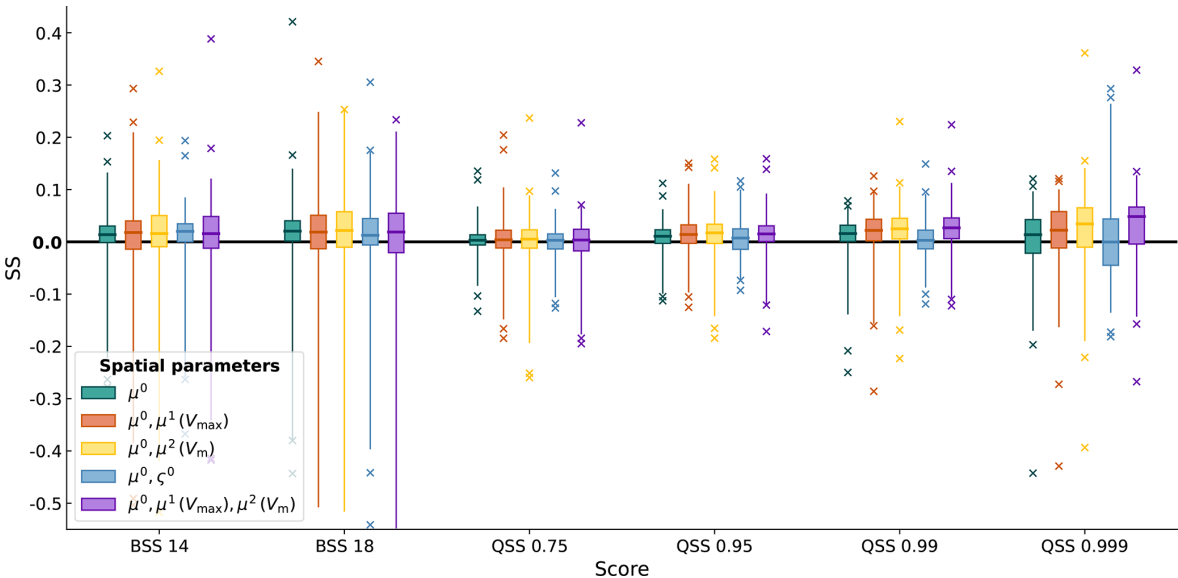


Figure C1. As Fig. 4 but for cross-validated skill scores of SpatBHM against ConstMod. Model definitions and median scores are given in Table C1.

Subsequent to the implementation of the elevation offset, SpatBHM is constructed in a gradual manner, commencing from ConstMod 4 and permitting an increasing number of model parameters to vary in space. This approach permits the assessment of the added value of individual spatial parameters. Firstly, only the intercept of the location parameter, μ^0 , is made variable in space, while the remaining parameters are kept spatially constant. Subsequently, the number of spatially varying parameters is increased through the testing of different combinations of spatial and non-spatial parameters, as illustrated in Table C1. The parameter μ^z (i.e. the regression coefficient for Δz) is spatially constant in all model versions, as it controls a spatial covariate assuming a fixed value at each location. The fitted values of the parameters are shown in Table C2. The median score values for SpatBHM are displayed in Table C1. All spatial models demonstrate a comparable level of skill with respect to climatology as ConstMod across the entire range of distribution characteristics investigated.

The median skill score with respect to ConstMod (Fig. C1) is non-negative for all model versions and for all evaluated thresholds and quantiles. With regard to the threshold excess probabilities, there are not many skill differences among the models which indicates that the gain in benefit from selecting a more complex model is minimal. For higher quantiles, the skill improvement is higher than for the central region of the predictive distribution, indicating superior performance of SpatBHM for extreme events.

A spatially variable location parameter leads to a significant improvement in the predictive skill of threshold exceedance probabilities and high quantiles. The inclusion of a spatially variable regression coefficient for the location parameter leads to an improvement in the predictive skill of threshold exceedance probabilities and particularly improves the skill for the extreme quantiles. A spatially variable regression coefficient for V_m (μ^2) provides more skill for the upper quantiles than a spatially variable regression coefficient for

$V_{\max}(\mu^1)$. A spatially variable scale parameter affects different prediction quantities in contradictory ways, rendering its overall effect ambiguous. It improves the skill for the exceedance probabilities of lower thresholds and reduces the dispersion of the Brier score, but also reduces the skill for all shown predicted quantiles compared to the other spatial model versions. This is particularly evident for the extreme 0.99 and 0.999 quantiles, where the median skill score compared to ConstMod is approximately zero. Whether a spatially variable scale parameter should be selected therefore depends on the desired predicted property.

The addition of a third spatially variable regression coefficient leads to a slight improvement in skill compared to a model with only two spatial parameters. However, this improvement is not as significant and the results for different parts of the distribution are inconsistent. The third spatial parameter improves the extreme quantiles 0.99 and 0.999, but does not provide much more information for the prediction of threshold exceedances. Further versions of SpatBHM with varying regression coefficients for the scale parameter (not shown) do not show significantly different results from those described above and are less skillful. We therefore proceed with SpatBHM 2b, while further refinements remain possible.

Table C2. Parameter estimations for SpatBHM within their 99 % confidence intervals, assuming normal marginal posterior distributions. The values are obtained from training the model on the complete training data set and using Eq. (9) as distance metric. Cells with only one number for the spatial parameters indicate spatially constant parameters.

Parameter	SpatBHM 1	SpatBHM 2a	SpatBHM 2b	SpatBHM 2c	SpatBHM 3	
μ^0	α_{μ^0}	4.78 ± 0.2	4.78 ± 0.24	4.80 ± 0.27	4.78 ± 0.22	4.82 ± 0.23
	σ_{μ^0}	0.56 ± 0.11	0.70 ± 0.16	0.69 ± 0.14	0.61 ± 0.13	0.67 ± 0.14
	ρ_{μ^0}	32 ± 17.5	37 ± 18.1	38 ± 22.2	32 ± 18.3	30 ± 18.7
μ^1	α_{μ^1}		1.56 ± 0.17			1.54 ± 0.20
	σ_{μ^1}	1.63 ± 0.03	0.4 ± 0.10	1.47 ± 0.03	1.61 ± 0.03	0.35 ± 0.11
	ρ_{μ^1}		45 ± 26.8			76 ± 39.4
μ^2	α_{μ^2}			-0.30 ± 0.28		-0.34 ± 0.34
	σ_{μ^2}	-0.57 ± 0.03	-0.37 ± 0.04	0.60 ± 0.16	-0.52 ± 0.03	0.63 ± 0.22
	ρ_{μ^2}			62 ± 38.4		68 ± 37.3
μ^z	0.23 ± 0.08	0.26 ± 0.10	0.32 ± 0.10	0.31 ± 0.08	0.33 ± 0.09	
ς^0	α_{ς^0}				0.49 ± 0.08	
	σ_{ς^0}	0.491 ± 0.005	0.473 ± 0.005	0.468 ± 0.005	0.20 ± 0.05	0.464 ± 0.005
	ρ_{ς^0}				35 ± 25.7	
ς^1	0.373 ± 0.009	0.365 ± 0.009	0.348 ± 0.009	0.40 ± 0.01	0.347 ± 0.009	
ς^2	-0.224 ± 0.008	-0.240 ± 0.008	-0.227 ± 0.008	-0.26 ± 0.01	-0.227 ± 0.008	
f_z	116 ± 73.8	157 ± 69.6	184 ± 81.2	133 ± 68.9	194 ± 78.8	

Appendix D: GEV shape of wind gust distributions

The GEV shape parameter ξ has a great effect on the distribution. E.g., if $\xi < 0$, the distribution becomes a Weibull-Type distribution with an upper bound on the supported values, while the Gumbel-type distribution for $\xi = 0$ is unbounded. As ξ is difficult to infer stably from the data, we decided to set it to a fixed and spatially constant value. As we are hesitant to exclude potential extreme events, we would like to intentionally omit the Weibull-type distribution as a valid option, although it may theoretically occur.

To determine an appropriate choice for the constant ξ , we fitted non-stationary GEV-models at each station to obtain local estimates of ξ for the gust model. We calculated the model for 1000 bootstrap iterations in order to obtain 95 %-credibility intervals for the parameter estimation. If the credibility intervals include 0, we conclude there is no signal for $\xi \geq 0$ at the station and the Gumbel-assumption is justified. If the credibility intervals are entirely positive, we say wind gusts at the station follow a Fréchet-type distribution. Finally, if the entirety of the credibility interval is below 0, we say that wind gusts follow a Weibull-type distribution, which we would like to exclude. Therefore, if all stations are either Gumbel or Weibull, we regard the assumption that $\xi \equiv 0$ as justified for SpatBHM.

We found only one station with a Fréchet-type distribution, but the 95 %-credibility interval is entirely $\xi < 0.1$, so that the value is small. The majority with 84 out of 109 stations show a Weibull-Type distribution with mean shape values between $-0.3 < \xi < 0$. The remaining stations have no clear signal with respect to ξ . We could not detect a spatial pattern for ξ . Some stations with low local GEV shape values coincide with locations where our spatial model exhibits limited skill. However, there are also stations with near-zero local GEV shape values, where SpatBHM still performs poorly. Conversely, several stations with a significantly negative ξ are well captured, suggesting that factors other than the local shape parameter influence model performance. Hence, assuming $\xi = 0$ for SpatBHM appears justified or at least not contradicted by the data.

Appendix E: Positive definiteness of the covariance matrix

Applying the Matérn-3/2 covariance function is known to be not positive definite on the sphere, when used with the great-circle distance (Gneiting, 2013). This results from the restriction of a function defined on $\mathbb{R}^+$ to a finite domain. However, for SpatBHM we are only modeling a limited portion on the sphere with maximum distances of $\mathcal{O}(10^3)$ km, so that the investigated area is almost flat. Nevertheless, we like to work on the real distances, which is why we selected the great-circle distance.

In earlier stages, we experimented with an exponential covariance kernel, which is known to be positive definite on the

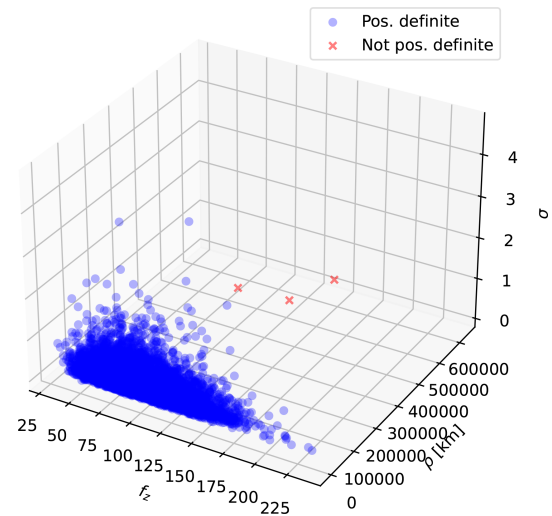


Figure E1. Positive definiteness of the covariance matrix depending on the covariance parameters.

sphere. The exponential kernel is equivalent to the Matérn-1/2-Kernel and therefore the roughest member of the family. However, due to the existence of very short distances between some of the training locations, the exponential kernel produced numerical instabilities. Therefore, we selected a function with a higher smoothness, thereby sacrificing the guarantee of a positive definite matrix.

We ran sensitivity tests to ensure that the covariance matrices in SpatBHM remain positive definite in practice for the limited modeling window over central Europe. To this end we calculated the covariance matrix using the coordinates of the training locations and a wide variety of possible values for f_z , ρ and σ , by sampling from their respective prior distributions (Table 1) 10 000 times. For each set of parameters, we determined whether the resulting covariance matrix remains positive definite.

As shown in Fig. E1, from all 10 000 parameter draws, only 3 combinations resulted in a non-positive definite covariance matrix. These draws had in common that the range parameter was very large at values $\rho > 10^5$ km, while the values of the altitude scaling factor f_z and the sill σ did not visibly affect the outcome. The largest range parameters occurring in the posterior samples from SpatBHM are of $\mathcal{O}(10^3)$ km, so we conclude that the assumption that the covariance matrices in SpatBHM remain positive definite in practice, is justified.

Additionally, we conducted a sensitivity test of SpatBHM using an even smoother Matérn-5/2 covariance function. While the smoother covariance function had little effect on visible roughness of the μ^0 -field and the model's predictive performance, its numerical stability compared to the exponential covariance function supports the use of a higher smoothness assumption for our modeling area.

Code and data availability. The SYNOP observations and COSMO-REA6 reanalysis data used in this study are publicly accessible through DWD (German weather service) via an open data server (<https://opendata.dwd.de>, last access: 31 October 2023). The model training was performed using Stan (<https://mc-stan.org>), using the Python interface pystan. The spatial interpolation and prediction were implemented using GNU-licensed free software from the R Project for Statistical Computing (<http://www.r-project.org>, last access: 12 May 2025). The Stan model code is provided along a Jupyter Notebook with model training examples, the preprocessed training and evaluation data sets and the R-Code for the spatial interpolation and the prediction are available from Zenodo at <https://doi.org/10.5281/zenodo.15437958> (Ertz and Friederichs, 2025). The same repository contains two recreation Jupyter Notebooks for test experiments regarding the shape parameter of the GEV distribution and the positive definiteness of the covariance matrix, that have been suggested by the associate editor and one anonymous reviewer.

Author contributions. PF is responsible for conceptualization of this work, acquisition of funding and editing of the manuscript. PE is responsible for conducting the investigation including data processing, model training and prediction, verification and visualization, and for writing the original draft of the manuscript.

Competing interests. The contact author has declared that neither of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We are grateful to Prof. Dr. Andreas Hense, University of Bonn, for fruitful discussions and constructive suggestions. Finally, we want to thank Sebastian Buschow for reviewing the manuscript and valuable feedback on the accessibility of the accompanying code.

Financial support. This research has been supported by the Bundesministerium für Bildung und Forschung (grant nos. 01LP2327F and 01LP2323A) and the Bundesministerium für Verkehr und Digitale Infrastruktur (grant no. 4823DWDP5A).

Review statement. This paper was edited by Likun Zhang and reviewed by two anonymous referees.

References

- Apputhurai, P. and Stephenson, A. G.: Spatiotemporal hierarchical modelling of extreme precipitation in Western Australia using anisotropic Gaussian random fields, *Environ. Ecol. Stat.*, 20, 667–677, <https://doi.org/10.1007/s10651-013-0240-9>, 2013.
- Atger, F.: Estimation of the reliability of ensemble-based probabilistic forecasts, *Q. J. Roy. Meteor. Soc.*, 130, 627–646, <https://doi.org/10.1256/qj.03.23>, 2004.
- Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E.: An empirical distribution function for sampling with incomplete information, *Ann. Math. Stat.*, 26, 641–647, <https://doi.org/10.1214/aoms/1177728423>, 1955.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E.: Hierarchical modeling and analysis for spatial data, Chapman and Hall/CRC, New York, ISBN 9780429205231, <https://doi.org/10.1201/9780203487808>, 2003.
- Baran, S. and Lakatos, M.: Clustering-based spatial interpolation of parametric postprocessing models, *Weather Forecast.*, 39, 1591–1604, <https://doi.org/10.1175/WAF-D-24-0016.1>, 2024.
- Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J.: Statistics of extremes: Theory and applications, Wiley, Chichester, ISBN 9780471976479, 2004.
- Bentzien, S. and Friederichs, P.: Decomposition and graphical portrayal of the quantile score, *Q. J. Roy. Meteor. Soc.*, 140, 1924–1934, <https://doi.org/10.1002/qj.2284>, 2014.
- Best, M. J. and Chakravarti, N.: Active set algorithms for isotonic regression; A unifying framework, *Math. Program.*, 47, 425–439, <https://doi.org/10.1007/BF01580873>, 1990.
- Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S.: Towards a high-resolution regional reanalysis for the European CORDEX domain, *Q. J. Roy. Meteor. Soc.*, 141, 1–15, <https://doi.org/10.1002/qj.2486>, 2015.
- Bradbury, W. M. S., Deaves, D. M., Hunt, J. C. R., Kershaw, R., Nakamura, K., Hardman, M. E., and Bearman, P. W.: The importance of convective gusts, *Meteorol. Appl.*, 1, 365–378, <https://doi.org/10.1002/met.5060010407>, 1994.
- Brasseur, O.: Development and application of a physical approach to estimating wind gusts, *Mon. Weather Rev.*, 129, 5–25, [https://doi.org/10.1175/1520-0493\(2001\)129<0005:DAAOAP>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0005:DAAOAP>2.0.CO;2), 2001.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, 78, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2), 1950.
- Bröcker, J.: Reliability, sufficiency, and the decomposition of proper scores, *Q. J. Roy. Meteor. Soc.*, 135, 1512–1519, <https://doi.org/10.1002/qj.456>, 2009.
- Climate Data Center: Observations Germany, https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/ (last access: 31 October 2023), 2023.
- Coles, S.: An introduction to statistical modeling of extreme values, Springer Series in Statistics, Springer-Verlag, London, ISBN 9781447136750, <https://doi.org/10.1007/978-1-4471-3675-0>, 2001.
- Cooley, D., Nychka, D., and Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels, *J. Am. Stat. Assoc.*, 102, 824–840, <https://doi.org/10.1198/016214506000000780>, 2007.

- Davison, A. C. and Gholamrezaee, M. M.: Geostatistics of extremes, *Proc. Roy. Soc. A.*, 468, 581–608, <https://doi.org/10.1098/RSPA.2011.0412>, 2012.
- Davison, A. C., Padoan, S., and Ribatet, M.: Statistical modeling of spatial extremes, *Stat. Sci.*, 27, 161–186, <https://doi.org/10.1214/11-STS376>, 2012.
- Dawid, A. P.: Present position and potential developments: Some personal views: Statistical Theory: The prequential approach, *J. R. Stat. Soc. Series A (General)*, 147, 278, <https://doi.org/10.2307/2981683>, 1984.
- de Leeuw, J., Hornik, K., and Mair, P.: Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and Active Set Methods, *J. Stat. Softw.*, 32, 1–24, <https://doi.org/10.18637/jss.v032.i05>, 2009.
- Deutscher Wetterdienst: Warnkriterien, https://www.dwd.de/DE/wetter/warnungen_aktuell/kriterien/warnkriterien.html?nn=607268, (last access: 14 February 2024), 2024.
- Dimitriadis, T., Gneiting, T., and Jordan, A. I.: Stable reliability diagrams for probabilistic classifiers, *P. Natl. Acad. Sci. USA*, 118, e2016191118, <https://doi.org/10.1073/pnas.2016191118>, 2021.
- Dyrddal, A. V., Lenkoski, A., Thorarindottir, T. L., and Stordal, F.: Bayesian hierarchical modeling of extreme hourly precipitation in Norway, *Environmetrics*, 26, 89–106, <https://doi.org/10.1002/env.2301>, 2015.
- Ertz, P. and Friederichs, P.: Wind gust postprocessing from COSMO-REA6 with a spatial Bayesian hierarchical extreme value model, Zenodo [code], <https://doi.org/10.5281/zenodo.15437958>, 2025.
- Ferro, C. A. T.: Fair scores for ensemble forecasts: Fair scores for ensemble forecasts, *Q. J. Roy. Meteor. Soc.*, 140, 1917–1923, <https://doi.org/10.1002/qj.2270>, 2014.
- Friederichs, P. and Hense, A.: Statistical downscaling of extreme precipitation events using censored quantile regression, *Mon. Weather Rev.*, 135, 2365–2378, <https://doi.org/10.1175/MWR3403.1>, 2007.
- Friederichs, P., Goeber, M., Bentzien, S., Lenz, A., and Krampitz, R.: A probabilistic analysis of wind gusts using extreme value statistics, *Meteorol. Z.*, 18, 615–629, <https://doi.org/10.1127/0941-2948/2009/0413>, 2009.
- Friederichs, P., Wahl, S., and Buschow, S.: Postprocessing for extreme events, in: *Statistical postprocessing of ensemble forecasts*, edited by: Vannitsem, S., Wilks, D. S., and Messner, J. W., Elsevier, 127–154, <https://doi.org/10.1016/B978-0-12-812372-0.00005-4>, 2018.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B.: *Bayesian data analysis*, Texts in statistical science series, 3rd edn., CRC Press, Taylor and Francis Group, Boca Raton London New York, ISBN 9781439840955, 2014.
- Gilks, R. W., Richardson, S., and Spiegelhalter, D. J. (Eds.): *Markov Chain Monte Carlo in practice*, Chapman & Hall/CRC interdisciplinary statistics series, Chapman and Hall/CRC, Boca Raton, FL, ISBN 9781482214970, 1995.
- Giorgi, F., Jones, C., and Asrar, G. R.: Addressing climate information needs at the regional level: the CORDEX framework, *WMO Bull.*, 58, 175–183, 2009.
- Gneiting, T.: Strictly and non-strictly positive definite functions on spheres, *Bernoulli*, 19, <https://doi.org/10.3150/12-BEJSP06>, 2013.
- Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, *J. Amer. Statist. Assoc.*, 102, 359–378, <https://doi.org/10.1198/016214506000001437>, 2007.
- Gneiting, T. and Ranjan, R.: Comparing density forecasts using threshold- and quantile-weighted scoring rules, *J. B. Econ. Stat.*, 29, 411–422, <https://doi.org/10.1198/jbes.2010.08110>, 2011.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, 133, 1098–1118, <https://doi.org/10.1175/MWR2904.1>, 2005.
- Gneiting, T., Kleiber, W., and Schlather, M.: Matérn cross-covariance functions for multivariate random fields, *J. Amer. Stat. Assoc.*, 105, 1167–1177, <https://doi.org/10.1198/jasa.2010.tm09420>, 2012.
- Gneiting, T., Wolfram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., and Schienle, M.: Model diagnostics and forecast evaluation for quantiles, *Ann. Rev. Stat. Appl.*, 10, 597–621, <https://doi.org/10.1146/annurev-statistics-032921-020240>, 2023.
- Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, 15, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.
- Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively setting path length in Hamiltonian Monte Carlo, *J. Mach. Learn. Res.*, 15, 1593–1623, <http://jmlr.org/papers/v15/hoffman14a.html> (last access: 4 January 2024), 2014.
- Hyndman, R. and Fan, Y.: Sample quantiles in statistical packages, *Am. Stat.*, 50, 361–365, <https://doi.org/10.1080/00031305.1996.10473566>, 1996.
- Jordan, A. I., Mühlemann, A., and Ziegel, J. F.: Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals, *Ann. Inst. Stat. Math.*, 74, 489–514, <https://doi.org/10.1007/s10463-021-00808-0>, 2022.
- Koenker, R.: *Quantile regression*, vol. 38 of *Econometric Society Monographs*, Cambridge University Press, ISBN 9780511754098, 2005.
- Koenker, R. and Machado, J. A. F.: Goodness of fit and related inference processes for quantile regression, *J. Amer. Stat. Assoc.*, 94, 1296–1310, <https://doi.org/10.1080/01621459.1999.10473882>, 1999.
- Krüger, F., Lerch, S., Thorarindottir, T., and Gneiting, T.: Predictive inference based on Markov Chain Monte Carlo output, *Int. Stat. Rev.*, 89, 274–301, <https://doi.org/10.1111/insr.12405>, 2021.
- Lerch, S., Thorarindottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster’s Dilemma: Extreme events and forecast evaluation, *Stat. Sci.*, 32, 106–127, <https://doi.org/10.1214/16-STS588>, 2017.
- Lindgren, F., Lindström, J., and Rue, H.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *J. R. Stat. Soc.: Series B*, 73, 423–498, <https://doi.org/10.1111/j.1467-9868.2011.00777.x>, 2011.
- Miles, R. E.: The complete amalgamation into blocks, by weighted means, of a finite set of real numbers, *Biometrika*, 46, 317–327, <https://doi.org/10.2307/2333529>, 1959.
- Möller, A., Thorarindottir, T. L., Lenkoski, A., and Gneiting, T.: Spatially adaptive, Bayesian estimation for

- probabilistic temperature forecasts, arXiv [preprint], <https://doi.org/10.48550/arXiv.1507.05066>, 2016.
- Murphy, A. H.: A new vector partition of the probability dcore, *J. Appl. Meteor.*, 12, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2), 1973.
- Nakamura, K., Kershaw, R., and Gait, N.: Prediction of near-surface gusts generated by deep convection, *Meteorol. Appl.*, 3, 157–167, <https://doi.org/10.1002/met.5060030206>, 1996.
- Oesting, M., Schlather, M., and Friederichs, P.: Statistical post-processing of forecasts for extremes using bivariate brown-resnick processes with an application to wind gusts, *Extremes*, 24 pp., <https://doi.org/10.1007/s10687-016-0277-x>, 2016.
- Perrin, O., Rootzén, H., and Taesler, R.: A discussion of statistical methods used to estimate extreme wind speeds, *Theor. Appl. Climatol.*, 85, 203–215, <https://doi.org/10.1007/s00704-005-0187-3>, 2006.
- R Development Core Team: R: A language and environment for statistical computing, Version 4.2.3, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org> (last access: 12 May 2025), ISBN 3-900051-07-0, 2023.
- Rasmussen, C. E. and Williams, C. K. I.: Gaussian processes for machine learning, Adaptive computation and machine learning, 3rd edn., MIT Press, Cambridge, Mass, ISBN 9780262182539, 2008.
- Riddell, A., Hartikainen, A., and Carter, M.: pystan (3.9.0), PyPI, <https://pypi.org/project/pystan> (last access: 1 March 2024), 2021.
- Rockel, B., Will, A., and Hense, A.: Special issue: regional climate modelling with COSMO-CLM(CCLM), *Meteorol. Z.*, 347–348, <https://doi.org/10.1127/0941-2948/2008/0309>, 2008.
- Rue, H., Martino, S., and Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations, *J. R. Stat. Soc. B*, 71, 319–392, <https://doi.org/10.1111/j.1467-9868.2008.00700.x>, 2009.
- Sang, H. and Gelfand, A. E.: Continuous spatial process models for spatial extreme values, *J. Agric. Biol. Envir. Stat.*, 15, 49–65, <https://doi.org/10.1007/s13253-009-0010-1>, 2010.
- Schulz, J.-P.: Revision of the turbulent gust diagnostics in the COSMO-model, *COSMO News Letter*, 8, 17–22, 2008.
- Schulz, J.-P. and Heise, E.: A new scheme for diagnosing near-surface convective gusts, *COSMO News Letter*, 3, E221–225, 2003.
- Stan Development Team: Stan: A probabilistic programming language, Stan Development Team, version 2.33, <https://mc-stan.org/> (last access: 1 March 2024), 2022.
- Stein, M. L.: Interpolation of spatial data: some theory for kriging, Springer series in statistics, Springer, New York, ISBN 9780387986296, 1999.
- Stephenson, A. G., Lehmann, E. A., and Phatak, A.: A max-stable process model for rainfall extremes at different accumulation durations, *Weather Clim. Extrem.*, <https://doi.org/10.1016/j.wace.2016.07.002>, 2016.
- Walshaw, D.: Getting the most from your extreme wind data: A step by step guide, *J. Res. Natl. Inst. Stan.*, 99, 399–411, <https://doi.org/10.6028/jres.099.038>, 1994.
- Walshaw, D. and Anderson, C. W.: A model for extreme wind gusts, *J. R. Stat. Soc. C*, 49, 499–508, <https://doi.org/10.1111/1467-9876.00208>, 2000.
- Wessel, J. B., Ferro, C. A. T., Evans, G. R., and Kwasniok, F.: Improving probabilistic forecasts of extreme wind speeds by training statistical post-processing models with weighted scoring rules, *Mon. Weather Rev.*, <https://doi.org/10.1175/MWR-D-24-0151.1>, 2025.
- Wilks, D. S.: Statistical methods in the atmospheric sciences, Elsevier, Amsterdam, Netherlands; Cambridge, MA, 4. edn., ISBN 9780128158234, 2019.
- Wolffram, D.: Replication package for “Model Diagnostics and Forecast Evaluation for Quantiles”, GitHub [code], <https://github.com/dwolffram/replication-ARSIA2023> (last access: 27 November 2024), 2022.