Spatial inference for hazard event intensities using imperfect observation and simulation data

Benjamin D. Youngman[*][†]and David B. Stephenson[*]

June 3, 2019

Abstract

Intensities of natural hazard events are often represented by maps. The data on which these maps are based, however, are subject to error. We propose a statistical framework for estimating the true—or actual—intensities of hazard events that explicitly accounts for data error and can accommodate various sources of data. Perhaps most commonly used are observations and simulations from complex computer models. We combine these through a Gaussian process, which gives an analytical expression for the joint distribution of a hazard event's actual intensities, given data. This formulation allows us to intuitively and explicitly represent various forms of discrepancy, such as measurement error and the difference in spatial structure between simulated and actual values.

We study windstorm Kyrill and its *footprint*, which maps its maximum wind gust speeds over a 72-hour period. Simulated data at 25km and 4km resolution are used, which are generated by a downscaled version of the Met Office's Unified model. We verify the proposed framework's performance by reliably statistically downscaling the 25km footprint to 4km resolution, and then proceed to estimate Kyrill's actual footprint using the 25km footprint and observations from stations across Europe. This actual footprint estimate has realistic spatial structure and gust speeds much closer to the observations than originally simulated. Results for all of a further 47 European windstorms support these findings.

Keywords: Gaussian process; geostatistics; natural hazard; European windstorm; latent process; wind gusts.

[*] Department of Mathematics, University of Exeter, UK.

[†] B.D. Youngman, Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QF, UK. E-mail: b.youngman@exeter.ac.uk

1 Introduction

Devastating natural hazard events, such as earthquakes, hurricanes, floods and tsunamis, tend to coincide with extreme environmental events, which seldom occur. However, detailed quantitative knowledge of such events is vital for disaster prevention and relief. Such knowledge is often gained by studying past events. For example, intensity of rainfall over a region during a past event—perhaps from satellite imagery—may be informative for flooding. Natural hazard event intensities are usually quantified by multiple sources of data. While different sources will have their own merits, none can ever perfectly represent the *actual* intensities of a hazard event—i.e. reality or the 'unobserved truth' (Fuentes et al., 2003). We propose a statistical framework for inferring actual intensities of natural hazard events from multiple data sources. This can benefit various industries, such as catastrophe modelling (described in Grossi et al. (2005)), where natural hazard data, such as intensity at a set of locations, are combined with property data, such as value and susceptibility to damage, to produce estimates of loss to property.

We focus on using output from a computer simulation model and observation data from stations to infer actual intensities. These are two sources of data commonly available for natural hazard events. While the long-standing use of observation data offers trustworthiness, their accuracy can vary considerably between natural hazard types and also between measurement methods. Locations of observations are also finite, and usually coincide with an irregularly-spread set of stations, which may not be located at a point of interest. A computer simulation model, such as a variant of a climate model, will typically offer spatially complete data. Some discrepancy between its output and actual intensities will always exist, but this may be reduced by improving the calibration and/or increasing the resolution of the model. This work will explicitly account for deficiencies in data sources, which is necessary for reliable inferences about reality to be made.

The proposed framework uses a Gaussian process to represent actual intensities of a natural hazard event over space, given the data. This formulation allows the spatial discrepancy between simulated and actual intensities to be readily incorporated through spatial covariance functions, which are often used in geostatistics; see Diggle and Ribeiro (2007). We additionally draw on the statistical emulation literature (Oakley, 2011) by using checks derived for Gaussian process emulators (Bastos and O'Hagan, 2009) to verify the appropriateness of model assumptions. Our reliance on Gaussian processes means that the joint distribution of actual intensities is fully and analytically characterised.

Estimating actual intensities of natural hazard events over regions is closely related to the work of Guttorp and Walden (1987) and Fuentes et al. (2003). In the latter observations on particulate matter and simulated data are related through a Gaussian model to the 'unobserved truth', whose posterior distribution can be sampled (Fuentes and Raftery, 2005). Our approach is related to Fuentes et al. (2003) by explicitly accounting for sources of discrepancy in model output and observations and incorporating these through Gaussian processes. We differ by offering an alternative treatment of aggregation in the model output, which leads to an off-the-shelf approach to inference for actual intensities of hazard events. Part of the Extreme Wind Storms (XWS) catalogue (Roberts et al., 2014) had the same aim: to combine observations and simulated footprints to give 'recalibrated footprints', which were estimated as linear functions of simulated footprints using a mixed effects model. Our work may be seen as an extension that derives actual intensities using a variant of kriging, which relaxes the assumption of linearity between actual and simulated values. Kriging was used to derive the E-OBS gridded datasets (Haylock et al., 2008; van den Besselaar et al., 2011) from observations. Unlike the E-OBS data, we also consider simulated data.

Section 2 of this paper gives details of the proposed framework for estimating actual intensities of natural hazard events. Sections 3 presents two analyses of the footprint of windstorm Kyrill, i.e. analyses of a map of its maximum gust speeds for a 72-hour period over Europe. Specifically, Section 3.2 gives a proof-of-concept study in which its 25km footprint is downscaled to 4km resolution and then Section 3.3 gives estimates of its actual footprint. Section 4 summarises the proposed framework.

2 Statistical framework

This section proposes a statistical framework for estimating actual intensities of natural hazard events over regions, using observations and computer simulation model data.

2.1 Hazard event model outline

Let $Z(s)$ denote the unobservable actual value of a natural hazard event at location s in region R and $Y(s)$ denote an observation on $Z(s)$. We assume that $Y(s) = Z(s) + \epsilon(s)$, where $\epsilon(s)$ are independent Gaussian measurement errors, so that

$$Y(s) \,|\, Z(s) \sim N\big(Z(s), \sigma_Y^2\big). \tag{1}$$

Let $x(s)$ denote a simulated estimate of the hazard event at location s. A key aspect of this framework, which differs Fuentes et al. (2003), is to assume that simulator output, $x(s)$, forms a covariate, and is therefore fixed. To estimate $Z(s)$, $x(s)$ is required for any location s. As $x(s)$ is typically represented on a grid, if a location s does not exactly correspond to a grid cell centroid, interpolation may be used. We assume that $Z(s)$ and $x(s)$ are related by

$$Z(s) = m\big(x(s)\big) + \varepsilon(s), \tag{2}$$

where m and ε are systematic and random discrepancy terms, respectively. The discrepancy captures various deficiencies in the simulator, in particular its inability to perfectly represent reality and its representation of a continuous process at finite resolution. This aggregation effect is mimicked in Fuentes et al. (2003) by averaging higher-resolution simulations over grid cells, whereas here a discrepancy term is introduced. We assume a Gaussian process (GP) model for $Z(s)$ given its simulated counterpart, $x(s)$, so that

$$Z(s) \,|\, x(s) \sim GP\big(m\big(x(s)\big), \sigma_X^2 c_X(\,,\,)\big), \tag{3}$$

where $m(\,)$ and $\sigma_X^2 c_X(\,,\,)$ are its mean and covariance functions, respectively. The covariance function can allow some smooth spatial variation in how actual and simulated intensities differ.

Relation (3) represents $[Z(s) \,|\, X(s),\, \Theta]$, where $[\,]$ denotes 'distribution of' and Θ represents an arbitrary parameter set. Guttorp and Walden (1987) propose a similar decomposition to equation (2) for $[X(s) \,|\, Z(s),\, \Theta]$. Adopting the uniform prior $[Z(s) \,|\, \Theta] \propto 1$ unites our approach with that of Guttorp and Walden (1987) for symmetric discrepancy terms, as then $[X(s) \,|\, Z(s),\, \Theta] \propto [Z(s) \,| X(s),\, \Theta]$. For simulation models that act as 'smoothers', i.e. fail to capture small-scale processes, the conditioning direction of Relation (3) offers the interpretation that $Z(s)$ is a function of $X(s)$ plus some noise, as in Equation (2). Part of rationale behind this is that we believe adjusting $x(s)$ through m and ε will provide sufficient information about $Z(s)$. This, however, is likely to require that $x(s)$ corresponds to a simulator of moderate resolution or higher, or one in which coupling $x(s)$ with additional covariates in m, such as elevation or entities related to sub-grid-scale information, provides sufficient information about $Z(s)$. Furthermore, the conditioning direction of Relation (2) leads to the marginal model of Relation (4), the form of which matches downscaling approaches classed as 'regression methods' in Wilby and Wigley (1997).

2.2 Inference

2.2.1 Parameter estimation

Relations (1) and (3) in Section 2.1 imply the marginal model

$$Y(s) \mid x(s), \sigma^2, \boldsymbol{\beta}, \boldsymbol{\theta} \sim GP\big(m\big(x(s)\big), \sigma^2 c(\,,\,)\big), \tag{4}$$

where $m(\,)$ is as in Relation (3) and $\sigma^2 c(\,,\,) = \sigma_Y^2 c_Y(\,,\,) + \sigma_X^2 c_X(\,,\,)$, as Relation (1) may be written as a GP with covariance function $\sigma_Y^2 c_Y(\,,\,)$. For tractability, suppose that $m(x) = \mathbf{h}^T(x)\boldsymbol{\beta}$, where $\mathbf{h}(\,)$ comprises q basis functions (e.g. $\mathbf{h}(x) = (1, x)^T$) and $\boldsymbol{\beta}$ comprises q regression coefficients. Depending on the forms chosen for the correlation functions, not all their parameters, collectively denoted $\boldsymbol{\theta}$, may be identifiable without prior knowledge, in particular if both $c_X(\,,\,)$ and $c_Y(\,,\,)$ contain nugget terms. We address this in Section 3 by specifying the measurement error. Relation (4) allows us to directly establish the relationship between the observations and simulator output and in turn perform inference. A possible drawback to this tractability is that only observation locations are used to infer $Z(s)$, whereas in Fuentes et al. (2003) all simulator output locations are used. Careful consideration must be given as to whether observation locations are sufficient for inferring $Z(s)$ for any s of interest.

Let $\mathbf{y} = \big(y(s_1), \ldots, y(s_n)\big)'$ denote observations on a hazard event at locations $s_1, \ldots, s_n$, and let $\mathbf{x} = \big(x(s_1), \ldots, x(s_n)\big)'$ denote corresponding simulator output. Construct $n \times q$ matrix $\mathbf{H}$ with ith row $\mathbf{h}^T\big(x(s_i)\big)$ for $i = 1, \ldots, n$, and the $n \times n$ matrix $\mathbf{A}(\boldsymbol{\theta})$ with (i, j)th element $c(s_i, s_j)$. The restricted log-likelihood, obtained by integrating over a uniform prior for $\boldsymbol{\beta}$ (Harville, 1974), is given by

$$\ell_R(\boldsymbol{\theta}) = -\frac{n-q}{2}\big(\log(2\pi) + \log\hat{\sigma}^2\big) - \frac{1}{2}|\mathbf{A}(\boldsymbol{\theta})| - \frac{1}{2}|\mathbf{H}^T\{\mathbf{A}(\boldsymbol{\theta})\}^{-1}\mathbf{H}|, \tag{5}$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\{\mathbf{A}(\boldsymbol{\theta})\}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\{\mathbf{A}(\boldsymbol{\theta})\}^{-1}\mathbf{y}, \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{n-q}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}})^T\{\mathbf{A}(\boldsymbol{\theta})\}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}). \tag{7}$$

We choose $\boldsymbol{\theta}$ to maximise Equation (5).

2.2.2 Actual process estimation

We can use an estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, and the assumptions made in Section 2.1 to infer $Z(s)$ for the hazard event for any $s \in R$. Noting that

$$\begin{pmatrix} \mathbf{Y} \\ Z(s) \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mathbf{H}\hat{\boldsymbol{\beta}} \\ \mathbf{h}^T\big(x(s)\big)\hat{\boldsymbol{\beta}} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2\mathbf{A}(\hat{\boldsymbol{\theta}}) & \hat{\sigma}_X^2\mathbf{t}(s) \\ \hat{\sigma}_X^2\mathbf{t}^T(s) & \hat{\sigma}_X^2 c_X(s,s) \end{pmatrix}\right), \tag{8}$$

where $\mathbf{t}^T(s) = \big(c_X(s_1,s),\ldots,c_X(s_n,s)\big)$, it follows that

$$Z(s)\,|\,\mathbf{Y} = \mathbf{y} \sim GP\big(m^*\big(x(s)\big), c^*(\ ,\ )\big), \tag{9}$$

where

$$m^*\big(x(s)\big) = \mathbf{h}^T\big(x(s)\big)\hat{\boldsymbol{\beta}} + \hat{\sigma}_X^2\mathbf{t}^T(s)\{\hat{\sigma}^2\mathbf{A}(\hat{\boldsymbol{\theta}})\}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}), \tag{10}$$

$$c^*(s,s') = \hat{\sigma}_X^2\big[c_X(s,s') - \hat{\sigma}_X^2\mathbf{t}^T(s)\{\hat{\sigma}^2\mathbf{A}(\hat{\boldsymbol{\theta}})\}^{-1}\mathbf{t}(s')\big]. \tag{11}$$

Implementation of equations (9)–(11) may be simplified by noting that

$$\{\hat{\sigma}^2\mathbf{A}(\hat{\boldsymbol{\theta}})\}^{-1} = \{\hat{\Sigma}_X(\hat{\boldsymbol{\theta}}) + \hat{\sigma}_Y^2\mathbf{I}_n\}^{-1} \tag{12}$$

$$= \hat{\sigma}_Y^{-2}[\hat{\Sigma}_X(\hat{\boldsymbol{\theta}})]^{-1}\{[\hat{\Sigma}_X(\hat{\boldsymbol{\theta}})]^{-1} + \hat{\sigma}_Y^{-2}\mathbf{I}_n\}^{-1}, \tag{13}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and the (i,j)th elements of $\hat{\Sigma}_X(\hat{\boldsymbol{\theta}})$ are given by $\hat{\sigma}_X c_X(s_i, s_j)$.

2.3 Model checking

Diagnostics originally designed for statistical emulators offer robust methods for assessing the assumptions of Section 2.1. These are summarised here; for fuller details see Bastos and O'Hagan (2009). For a hazard event consider $\tilde{n}$ validation points $\tilde{\mathbf{Y}} = \big(Y(\tilde{s}_1), \ldots, Y(\tilde{s}_{\tilde{n}})\big)^T$ at locations $\tilde{s}_1, \ldots, \tilde{s}_{\tilde{n}}$ with simulator output $x(\tilde{s}_1), \ldots, x(\tilde{s}_{\tilde{n}})$. Define matrix $\tilde{\mathbf{V}}$ to have (i,j)th element $\hat{\sigma}^2 c^\dagger(\tilde{s}_i, \tilde{s}_j)$, for $i, j = 1, \ldots, \tilde{n}$ and the vector $\tilde{\mathbf{m}} = \big(m^\dagger\big(x(\tilde{s}_1)\big), \ldots, m^\dagger\big(x(\tilde{s}_{\tilde{n}})\big)\big)^T$, where

$$
\begin{aligned}
m^\dagger\big(x(s)\big) &= \mathbf{h}^T\big(x(s)\big)\hat{\boldsymbol{\beta}} - \mathbf{t}^T(s)\{A(\hat{\boldsymbol{\theta}})\}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}),\\
c^\dagger(s, s') &= c(s, s') - \mathbf{t}^T(s)\{A(\hat{\boldsymbol{\theta}})\}^{-1}\mathbf{t}(s').
\end{aligned}
$$

These alternative GP mean and covariance functions, in comparison to equations (10) and (11), account for the fact that the observations that form the validation data are subject to measurement error, unlike the actual values. It follows from model (4) that $\tilde{\mathbf{Y}} \sim MVN_{\tilde{n}}(\tilde{\boldsymbol{m}}, \tilde{\mathbf{V}})$.

Now define $\tilde{e}(s_i) = \big(Y(\tilde{s}_i) - m^\dagger\big(x(\tilde{s}_i)\big)\big)/\sqrt{\hat{\sigma}^2 c^\dagger(s_i, s_i)}$ for $i = 1, \ldots, \tilde{n}$. Test 1 is based on the approximate distributional result that $\tilde{e}(s_i) \sim N(0,1)$; thus large $\tilde{e}(s_i)$ relative to the standard Gaussian distribution highlight model inadequacy. In the present case, plotting $\tilde{e}(s_i)$ against i or spatially against $\tilde{s}_i$ may help show regions in which the predictions are poor. Now consider the pivoted Cholesky decomposition of $\tilde{\mathbf{V}}$ so that permutation matrix $\mathbf{P}$ and upper triangular matrix $\mathbf{U}$ satisfy $\mathbf{P}^T\tilde{\mathbf{V}}\mathbf{P} = \mathbf{U}^T\mathbf{U}$. Define $\mathbf{G} = \mathbf{P}\mathbf{U}^T$. Elements of the vector $\mathbf{e}^{PC} = \mathbf{G}^{-1}(\tilde{\mathbf{Y}} - \tilde{\mathbf{m}})$, where $\mathbf{e}^{PC} = \big(e_1^{PC}, \ldots, e_{\tilde{n}}^{PC}\big)^T$, are independent and approximately satisfy $e_i^{PC} \sim N(0,1)$. Test 2 is a quantile-quantile plot of these errors and Test 3 plots e_i^{PC} against i. Model inadequacy is indicated in Test 2 with points deviating from the line with zero intercept and unit slope, and in Test 3 with large absolute or non-random values. Test 4 uses the Mahalanobis distance $D_{MH} = (\mathbf{e}^{PC})^T\tilde{\mathbf{e}}^{PC}$, for which $D_{MH} \sim F_{\tilde{n}, n-q}$ has the F-Snedecor distribution with degrees of freedom $\tilde{n}$ and $n - q$. Large D_{MH} relative to $F_{\tilde{n}, n-q}$ indicates model inadequacy.

3 Extreme European windstorm footprints

3.1 Data

We study windstorm Kyrill, which caused the largest insured loss of any European windstorm since 2000 ($6.7bn, indexed to 2012 prices) . Windstorm Kyrill reached its peak intensity on 18th January 2007. We define a windstorm by a 72-hour period centred on the time of its peak intensity. Figure 1 shows two footprints for windstorm Kyrill, which are defined as maxima of maximum 3 second wind gusts at 10 metres above ground. The footprints in Figure 1 have approximate resolutions of 25km and 4km (Roberts et al., 2014; WISC, 2017). Both are obtained from a dynamic downscaling of the ERA Interim reanalysis using the Met Office Unified Model (Dee et al., 2011; Davies et al., 2005). Figure 1 also shows gust speed observations, maximised over the 72-hour period defining windstorm Kyrill, for 752 stations, which are obtained from the NOAA Global Summary of the Day (GSOD) database[1]. The observations show the highest gust speeds to occur in similar regions to the footprints, although a less smooth structure in spatial variation is evident. The observations will be used in Section 3.3 to derive Kyrill's actual footprint using the methodology of Section 2. Figure 1 also shows a comparison of the observed and simulated gust speeds, and reveals a discrepancy between gust speeds from all three data sources. (Bi-linear interpolation is used to match simulated gust speeds to station locations throughout this section.) Figure 1 shows that simulated gust speeds tend to be lower than those measured and also that speeds simulated from the 25km model tend to be lower than from the 4km model, especially when high gust speeds are simulated. This suggestion of bias is also evident in Figure 1 from the footprints. The main motivation for estimating actual intensities of windstorms is to overcome such biases. This is particularly important for understanding risk, as underestimation of gust speeds can lead to underestimation of losses, or to buildings being designed to insufficient criteria.

[1] `http://www7.ncdc.noaa.gov/CDO/cdoselect.cmd`

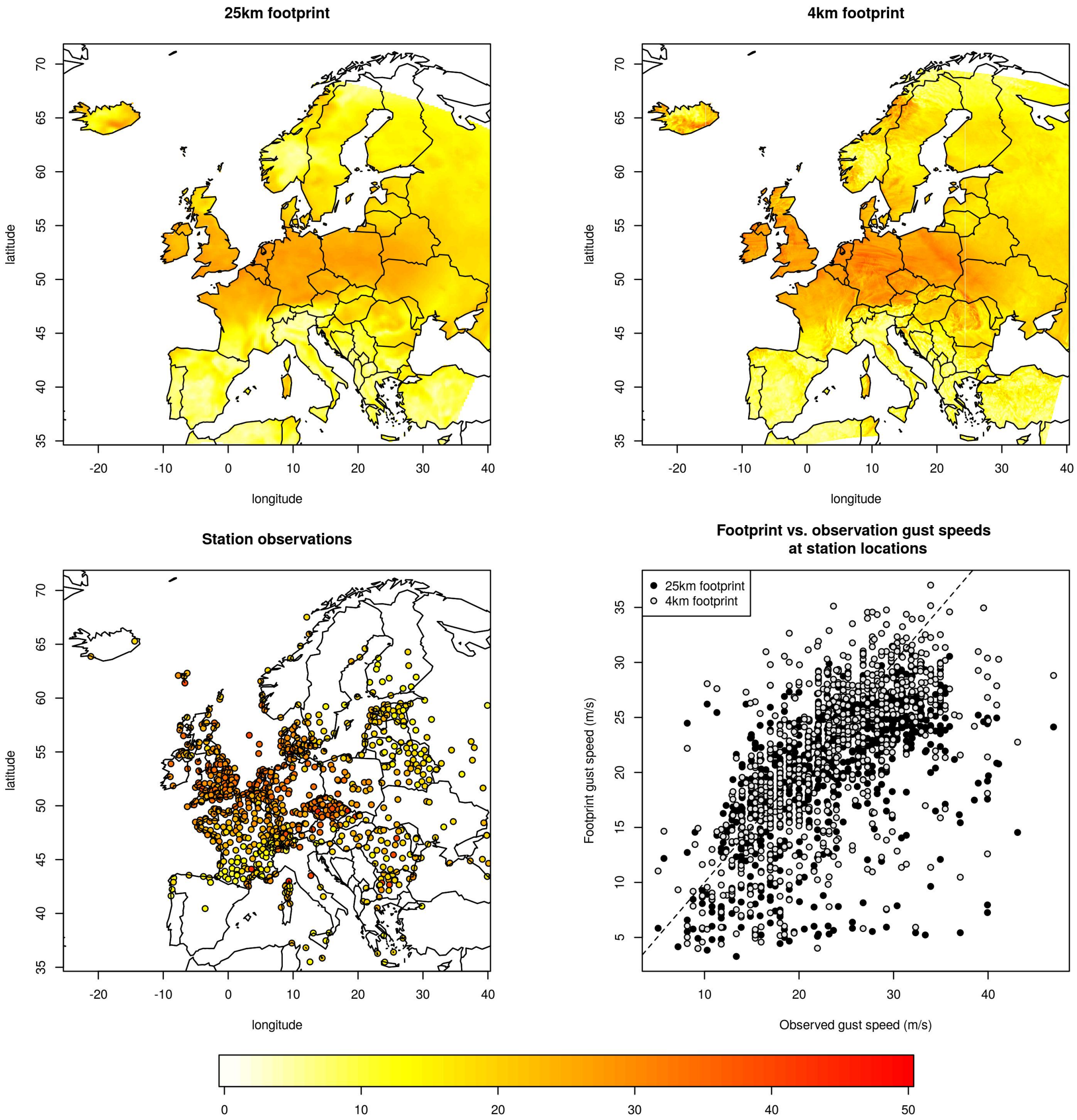


Figure 1: Maximum wind gust speeds for windstorm Kyrill in ms^{-1}. Footprints simulated at 25km (upper left) and 4km (upper right) resolutions. Station-based observed wind gust speeds (lower left) and footprint versus observed wind gust speeds (lower right) at station locations. (Footprint speeds are bi-linearly interpolated to coincide with station locations.)

3.2 Proof-of-concept: downscaling from 25km to 4km

To indicate whether actual footprints may be derived from the simulated 25km footprint, an intermediate example in which the 25km footprint is downscaled to 4km is first considered. In terms of Relation (4), $Y(s)$ and $x(s)$ are 4km and 25km wind gust speeds at location s, respectively. The mean function of Relation (4) is chosen as

$$m(s) = g_1\big(x(s)\big) + g_2\big(elev^{\dagger}(s)\big) + g_3\big(orog^{\dagger}(s)\big), \tag{14}$$

where g_d, $d = 1, 2, 3$, are cubic regression splines of basis dimension five, $elev^{\dagger}(s) = \sqrt{elev(s)}$, where $elev(s)$ is the elevation in metres of location s, and $orog^{\dagger}(s) = \log(1 + orog(s))$, where $orog(s)$ measures the variation in the orography near location s. Splines are used to accommodate a fairly broad range of relationships, e.g. in the mean discrepancy between $Y(s)$ and $x(s)$. A relatively low basis dimensions of five ensures that overly wiggly relationships are avoided. Knots for splines are placed at evenly distributed quantiles of $x(s)$, $elev^{\dagger}(s)$ and $orog^{\dagger}(s)$. If greater generality is required, higher basis dimensions can be chosen and smoothing parameters introduced to optimise wiggliness, which is common in generalised additive models; see Wood (2017).

The covariance function for the GP of Relation (4) is defined for a transformed space. Let $s^{\dagger} = (s_{lon}, s_{lat})^T$ be a location defined by geographical coordinates. Then its corresponding location s in transformed space is given by $s = (s_1, s_2)^T = (Rs^{\dagger})\Phi$, where

$$R = \begin{pmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} 1/\phi_{lon} & 0 \\ 0 & 1/\phi_{lat} \end{pmatrix}, \tag{15}$$

which corresponds to a clockwise rotation through ω followed by a scaling of ϕ_{lon} and ϕ_{lat} relative to the x and y axes, respectively. This transformation reflects that windstorms may follow tracks in directions not necessarily parallel to the longitude or latitude axes; these direction tends to coincide with any directionality in spatial discrepancy, i.e. in $c(\, , \,)$. The correlation function, $c(\, , \,)$ in Relation (4), is chosen as

$$c(s, s') = \begin{cases} 1 & \text{if } \; s = s', \\ (1-\tau)\exp(-||s - s'||^{\delta}) & \text{otherwise,} \end{cases} \tag{16}$$

where $0 \le \tau \le 1$, $0 < \delta \le 2$ and τ is a nugget effect. A powered exponential form for

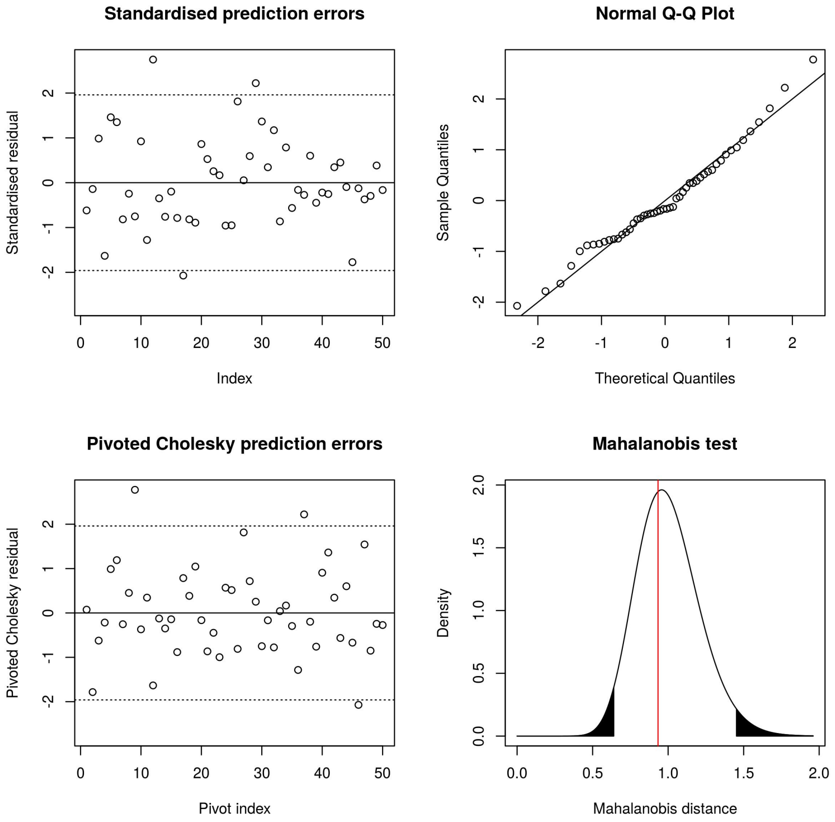


Figure 2: Diagnostic checks for downscaling model. See §2.3 for definitions of standardised prediction, pivoted Cholesky errors and test details.

$c(,)$ offers similar flexibility to a Matérn form, but greater analytical tractability when numerically maximising restricted log-likelihood (5).

Prior to assessing the model's performance, the validity of assumptions made is tested. Figure 2 shows fours diagnostics plots proposed in Bastos and O'Hagan (2009), which are described in Section 2.3 and based on withholding $\tilde{n} = 50$ stations' data for validation, which leaves 702 stations for model fitting. The plots indicate adequacy of the GP's assumptions as the 4km validation data are consistent with their predictive distribution. Specifically, the standardised residuals and pivoted Cholesky errors match the standard Gaussian distribution, and the Mahalanobis distance its $F_{\tilde{n},n-q}$ distribution. More generally the diagnostics offer good support that the form of GP of Relation (4) and its specifications in equations (14)–(16) capture the discrepancy between the 25km and 4km footprints.

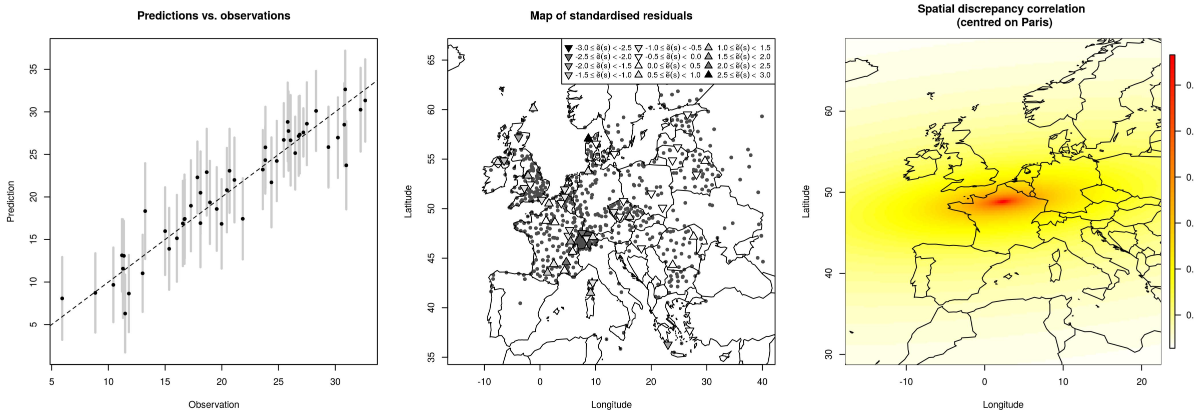


Figure 3: Further diagnostic checks for downscaling model. Predicted wind gusts (●) with 95% uncertainty bounds (grey) versus observations (left); mapped standardised residuals (centre); visual representation of correlation function $c(s, s')$ centred on Paris (right).

Summaries of fit more relevant to the application are shown in Figure 3. The plot of observed versus predicted wind gusts shows good agreement between the two, once uncertainties are taken into account. In particular there are no obvious signs of systematic bias. The correlation map in Figure 3 represents $c(s, s')$, which shows a fairly quick decay in spatial discrepancy; i.e. any difference between the 4km and 25km footprints tends to be fairly localised.

The downscaling model is represented in Figure 4, which shows the mean estimate of the downscaled 25km footprint and its standard error. Estimates assume that $\sigma_Y = 0$ms^{-1}. Visual agreement between the 4km and downscaled 25km footprints is clearly good. The standard error estimates highlight greater precision near to where 4km data have been used. The increase in standard errors away from the 4km data locations is fairly rapid, as reflected in the correlation function shown in Figure 3. The standardised residual errors of Figure 4 are perhaps most informative, showing errors to be concentrated at zero. Greatest absolute errors tend to occur by the boundary of the study region, i.e. where station data become scarce. This finding, coupled with Figure 4, highlights that, although standard errors tend to grow as stations become more distant, the method might be seen as an approach to interpolation, i.e. relied upon within the range of data, which here corresponds to in the vicinity of stations.

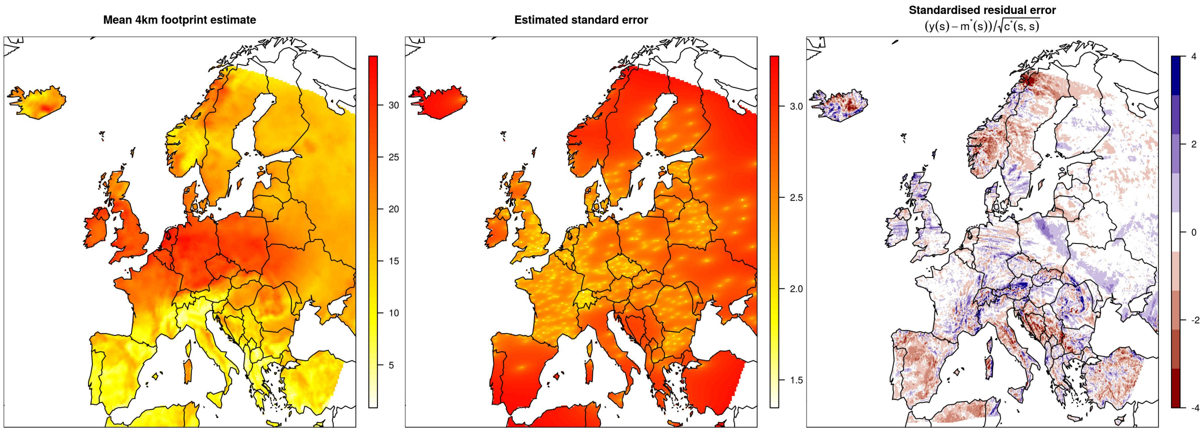

Mean 4km footprint estimate
30
25
20
15
10
5
Estimated standard error
3.0
2.5
2.0
1.5
Standardised residual error
$(y(s) - m^*(s))/\sqrt{c^*(s, s)}$
4
2
0
-2
-4

Figure 4: Downscaling model estimates. Mean 4km footprint estimate (left) and its standard error (center) based on downscaling of 25km footprint. Standardised residuals for 4km versus downscaled 25km footprints (right).

A final simple summary of the downscaling is achieved by comparing the root mean squared error (RMSE) between the 4km footprint, the 25km footprint, and its estimate downscaled from the 25km footprint, which reduces from 3.15ms^{-1} to 2.70ms^{-1}. Figure 8 of Appendix A shows RMSE values obtained similarly for 47 of the 52 footprints in the XWS catalogue (which are those that coincide with WISC footprints, excluding the November 1981 storm, which had only 40 available observations). For every storm a reduction in RMSE from downscaling the 25km footprint, as opposed to using its raw values, can be seen. Furthermore an average RMSE reduction of approximately 24% offers more robust support of the model's performance.

3.3 Actual footprint estimation

The ultimate aim of this work is to estimate the *actual* footprint for windstorm Kyrill, which, given the downscaling of Section 3.2, corresponds to an estimate at infinitely high resolution, or effectively 0km. This is achieved by relating the 25km footprint to the observations, which are both shown in Figure 1. Given that Figure 1 shows a tendency for the 25km footprint's gust speeds to be lower than those observed, we may anticipate that estimating actual intensities for windstorm Kyrill will raise some of its higher gust speeds, which coincide with those that cause greatest damage. For Relation (1) we assume that $\sigma_Y = 2\text{ms}^{\text{-1}}$, i.e. that measured gust speeds deviate from reality on average by $2\text{ms}^{\text{-1}}$. This is intended to be a fairly conservative choice, which is partially informed by $\hat{\tau}\hat{\sigma}^2$ placing an upper bound on σ_Y^2. Remaining model specifications are the same as those of Section 3.2.

The model of Relation (4) is again fitted to 702 stations' data, leaving $\tilde{n} = 50$ validation data points. Figure 5 shows diagnostics plots for the 50 validation data points based on Section 2.3. Again the diagnostic plots indicate good agreement between assumptions in the GP model and predictions of observations from the model, which support the GP providing an adequate representation of the actual intensities, given the simulated and observation data. The further diagnostic plots of Figure 6 offer further support of the model's assumption in the context of mapping actual wind gust speeds, as good agreement between predictions and observations given uncertainties is seen, and systematic patterns in standard errors do not appear evident. Larger prediction uncertainty in the actual gusts speeds, compared to those downscaled to 4km in Section 3.2, is primarily a consequence of the increase to $\sigma_Y^2 = 2\text{ms}^{\text{-1}}$ to allow for measurement error in the observations. The correlation map of Figure 6, which represents $c(s, s')$, shows the estimated spatial discrepancy to extend further than in Section 3.2. This indicates discrepancy between the simulated and actual footprint to be less localised than between the 4km and 25km footprints.

Figure 7 shows the results of estimating actual intensities for Kyrill's footprint. The mean actual footprint estimate qualitatively resembles that simulated at 25km resolution. The plot of their difference, however, reveals increases in gust speeds over central England and Ireland. Additional increases, for example over Iberia and the Balkans, coincide with large standard error estimates; these occur where observations are distant, which effectively reduces $m^*(s)$ to $m(s)$. This is further discussed in Section 4.

Finally, we repeat and similar analysis for actual values for the additional 47 European windstorms described in Section 3.2. The analysis differs as now comparison is made to observations at validation stations, as opposed to over an entire footprint. For most storms, 50 validation stations are used, unless this corresponds to more than 10% of the stations for which data are available; then the latter number is used. RMSE plots that compare 25km footprints and actual footprint estimates to observations at the validation stations are shown in Figure 8 of Appendix A. This, coupled with an average decrease in RMSE of approximately 35% over all storms, offers further support for the gain in estimating the actual footprint using the methods of Section 2, in comparison to using the original footprints.

4 Discussion

We have presented a spatial framework for estimating actual intensities of natural hazard events from observations and simulated data. The framework is based on assuming that, given simulated and observed data, actual intensities may be represented by a Gaussian process. This formulation offers various benefits: actual intensities can exhibit some degree of smoothness over space if geostatistical covariance functions are adopted; discrepancies between data and actual intensities can often be specified intuitively, such as the formulation for measurement error in sections 2.1 and 3.3; and the joint distribution of actual intensities over a region is fully characterised. The framework is also an off-the-shelf approach to inference for actual intensities. It is implemented in the `R` package `recalibrate`, which is provided as Supplementary Material (R Core Team, 2017).

The proposed framework has relied upon Gaussian assumptions. Although these are adequate for the windstorm models of sections 3.2 and 3.3, their suitability for other natural hazard phenomena or other sources of data is not guaranteed. For example, given simulated values, a natural hazard's actual intensities could be heavy tailed. A possible solution to overcome this is to consider transformed Gaussian processes (Diggle and Ribeiro, 2007, Section 3.8), whereby data undergo a transformation prior to fitting Gaussian processes. The Box-Cox transformation is a natural candidate; whether it is applied, or its power transform parameter, can differ between data sources. Such transforms may reduce some intuitiveness of the measurement error, σ_Y; in particular, constant σ_Y may no longer be suitable. A greater challenge may arise if bias in observations is expected, as constraints may be needed to ensure identifiability between any specified bias and $m(s)$. This could be achieved through prior specifications, if a Bayesian approach to inference is adopted.

We have suggested in Section 2 that, because it does not explicitly account for aggregation that leads to model output, the proposed framework is best suited to models of moderate resolution or higher, unless additional information able to disaggregate the output is available. This is in contrast to Fuentes et al. (2003) in which averaging over cells is used to mimic effects of aggregation. By allowing for discrepancy, our approach does not ignore aggregation effects, and gives estimates that are valid at any resolution, given accompanying uncertainty estimates. We see allowing for discrepancy, as opposed to mimicking effects of aggregation, as a fair trade-off given the tractability that it brings, e.g. by allowing a REML approach to inference.

Estimates of actual intensities for windstorm Kyrill show that uncertainties away from observation locations can be relatively large. This indicates that observations are likely to be required close to any location of interest; otherwise the mean actual intensity estimate, $m^*(s)$, reverts to $m(s)$. Setting $m(s) = x(s)$ will cause $m^*(s)$ to revert to $x(s)$, which may be benefit some problems, and if observations are plentiful may have little effect on regions of interest. The flexibility of this may be improved by partitioning $c_X(\, , \,)$: for example, $c_X(s, s', x(s), x(s')) = c_{X,1}(s, s')c_{X,2}(x(s), x(s'))$ separates the discrepancy between the actual and simulated intensities into a spatial component and a component related to the intensity itself, which would essentially govern the overall bias in the simulated intensity (which is governed by $g_1(x(s))$ in sections 3.2 and 3.3). For the modelling of windstorms, further improvement may be gained by partitioning the spatial component of the discrepancy to reflect that windstorms tend to decelerate over land, due to drag, and then accelerate over sea, where drag reduces; see, for example, Wallace and Hobbs (2006, Chapter 7). Finally, although the transformation of $s^* \mapsto s$ used in sections 3.2 and 3.3 allows anisotropy in the spatial discrepancy, it does not allow for misalignment, e.g. if the highest gust speeds were in slightly the wrong place. Different transformations can accommodate this feature, some of the most general being those based on spatial deformations; see Sampson and Guttorp (1992) and Schmidt and O'Hagan (2003).

Acknowledgements

We thank Phil Sansom for helpful discussion and the Willis Research Network for supporting this work.

References

Bastos, L. S. and A. O'Hagan (2009). Diagnostics for Gaussian process emulators. *Technometrics 51*(4), 425–438.

Davies, T., M. J. P. Cullen, A. J. Malcolm, M. H. Mawson, A. Staniforth, A. A. White, and N. Wood (2005). A new dynamical core for the met office's global and regional modelling of the atmosphere. *Quarterly Journal of the Royal Meteorological Society 131*(608), 1759–1782.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society 137*(656), 553–597.

Diggle, P. J. and P. J. Ribeiro (2007). *Model-based Geostatistics*. Springer series in statistics. Springer.

Fuentes, M., P. Guttorp, and P. G. Challenor (2003). Statistical assessment of numerical models. *International Statistical Review 71*(2), 201–221.

Fuentes, M. and A. E. Raftery (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics 61*(1), 36–45.

Grossi, P., H. Kunreuther, and C. C. Patel (2005). *Catastrophe Modeling: A New Approach to Managing Risk*. Catastrophe Modeling. Springer.

Guttorp, P. and A. Walden (1987). On the evaluation of geophysical models. *Geophysical Journal International 91*(1), 201–210.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika 61*(2), 383–385.

Haylock, M. R., N. Hofstra, A. M. G. K. Tank, E. J. Klok, P. D. Jones, and M. New (2008). A European daily high-resolution gridded dataset of surface temperature and precipitation. *J. Geophys. Res, (Atmospheres) 113*.

Oakley, J. E. (2011). Modelling with deterministic computer models. In M. Christie, A. Cliffe, P. Dawid, and S. Senn (Eds.), *Simplicity, Complexity and Modelling*, pp. 51–67. John Wiley & Sons, Ltd.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Roberts, J. F., A. J. Champion, L. C. Dawkins, K. I. Hodges, L. C. Shaffrey, D. B. Stephenson, M. A. Stringer, H. E. Thornton, and B. D. Youngman (2014). The XWS open access catalogue of extreme European windstorms from 1979–2012. *Natural Hazards and Earth System Sciences Discussions 2*(3), 2011–2048.

Sampson, P. D. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association 87*(417), 108–119.

Schmidt, A. M. and A. O'Hagan (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(3), 743–758.

van den Besselaar, E. J. M., M. R. Haylock, G. van der Schrier, and A. M. G. Klein Tank (2011). A European daily high-resolution observational gridded data set of sea level pressure. *J. Geophys. Res. 116*.

Wallace, J. M. and P. V. Hobbs (2006). *Atmospheric Science: An Introductory Survey.* International Geophysics. Elsevier Science.

Wilby, R. and T. Wigley (1997). Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography 21*(4), 530–548.

WISC (2017). Windstorm Information Service. `https://wisc.climate.copernicus.eu/wisc/#/`. Accessed: 18/10/2017.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition.* CRC Press.

A RMSE values for additional European windstorms

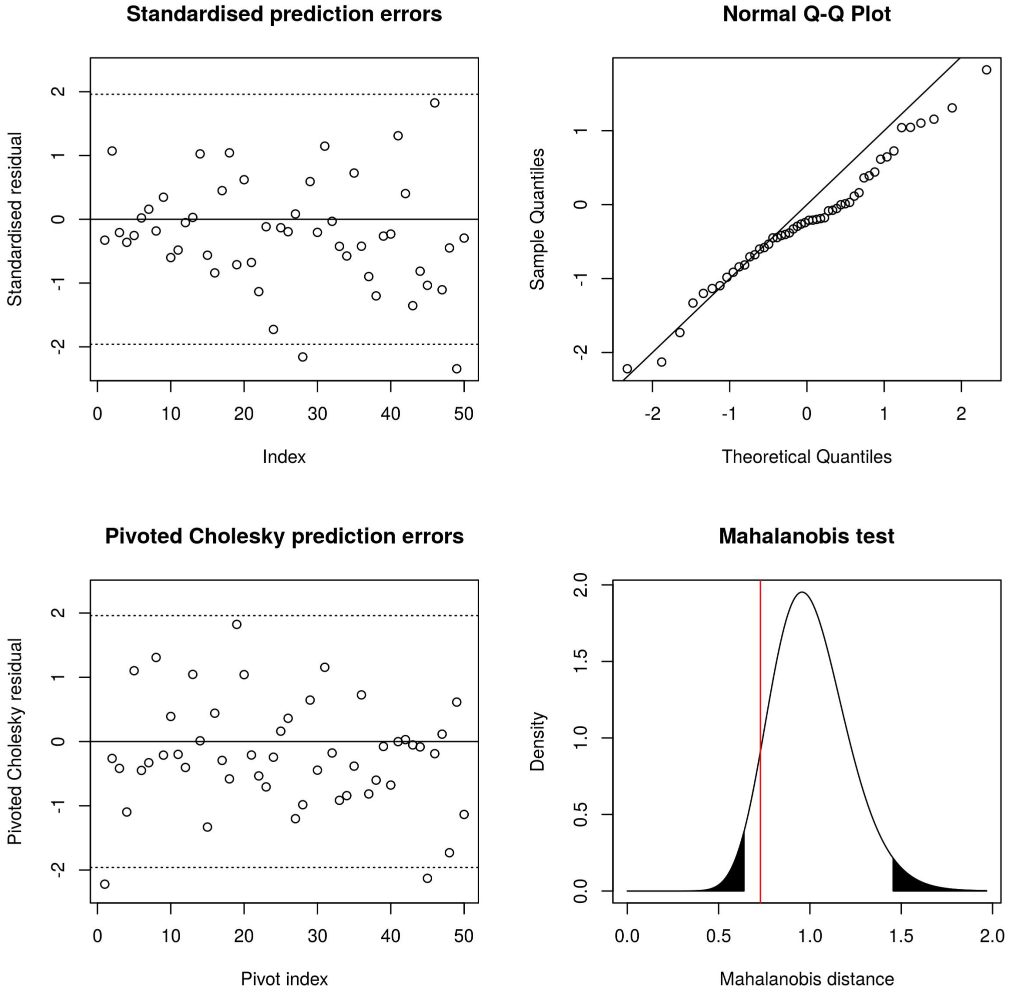


Figure 5: Diagnostic checks for actual footprint estimation model. See Figure 2 for plot descriptions.

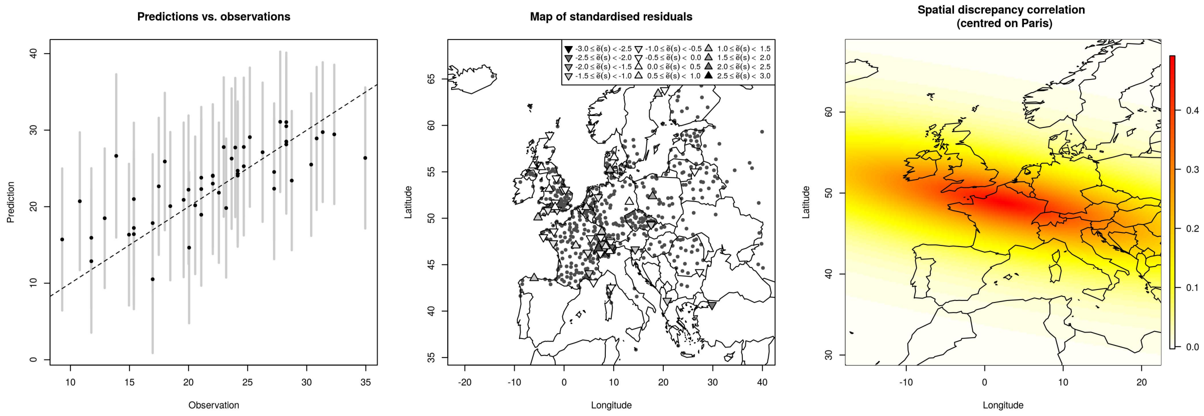
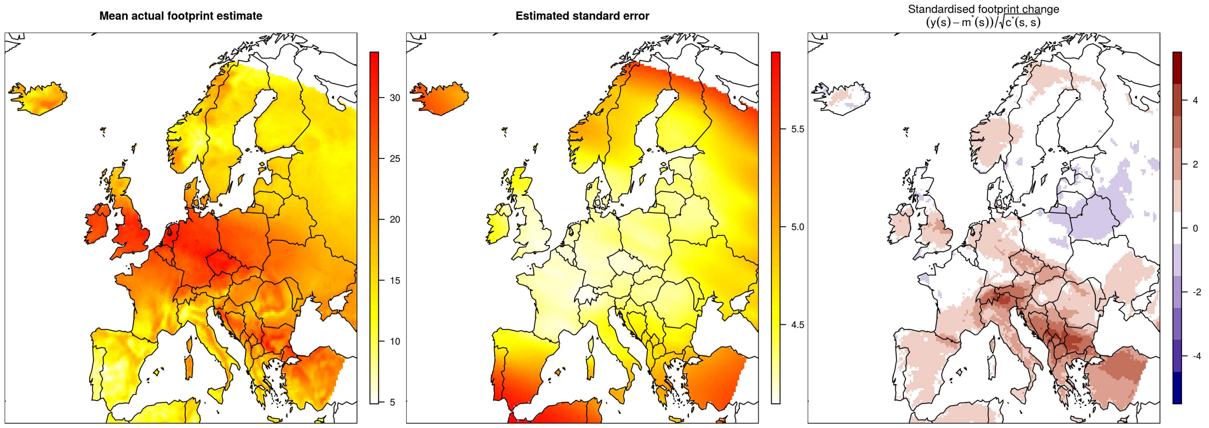


Figure 6: Further diagnostic checks for actual footprint estimation model. See Figure 3 for plot descriptions.



Mean actual footprint estimate
5
10
15
20
25
30
Estimated standard error
4.5
5.0
5.5
Standardised footprint change
$(y(s) - m^*(s))/\sqrt{c^*(s, s)}$
-4
-2
0
2
4

Figure 7: Actual footprint estimation model. Actual footprint estimate (left) and its standard error (center) based on 25km footprint. Standardised differences between estimated actual and 25km footprints (right).

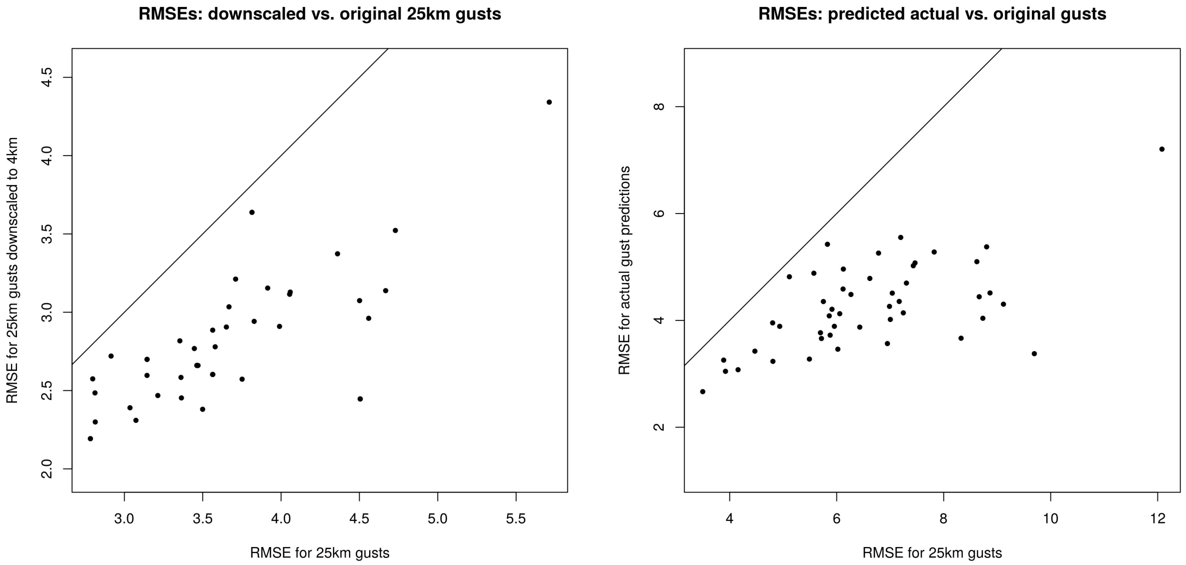


Figure 8: RMSE values comparing the 25km footprint to the 4km footprint with the 25km footprint downscaled to 4km across the entire domain (left). RMSE values comparing the 25km footprint and actual gust predictions to observations for validation stations (right). The line $y = x$ is superimposed.