Received ; | Revised ; | Accepted

DOI: xxx/xxxx

RESEARCH ARTICLE

Data fusion with Gaussian processes for estimation of environmental hazard events

Xiaoyu Xiong | Benjamin D. Youngman | Theodoros Economou

Department of Mathematics, University of Exeter, Exeter, U.K.

Correspondence
Xiaoyu Xiong, Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, U.K.
Email: x.xiong@exeter.ac.uk

Summary

Environmental hazard events such as extra-tropical cyclones or windstorms that develop in the North Atlantic can cause severe societal damage. Environmental hazard is quantified by the hazard footprint, a spatial area describing potential damage. However, environmental hazards are never directly observed, so estimation of the footprint for any given event is primarily reliant on station observations (e.g., wind speed in the case of a windstorm event) and physical model hindcasts. Both data sources are indirect measurements of the true footprint, and here we present a general statistical framework to combine the two data sources for estimating the underlying footprint. The proposed framework extends current data fusion approaches by allowing structured Gaussian process discrepancy between physical model and the true footprint, while retaining the elegance of how the "change of support" problem is dealt with. Simulation is used to assess the practical feasibility and efficacy of the framework, which is then illustrated using data on windstorm Imogen.

KEYWORDS:
Data integration, change-of-support, Gaussian processes, spatial interpolation, model validation, European windstorm, natural hazards

1 | INTRODUCTION

Environmental hazards, such as flooding, windstorms and tornadoes, can have devastating consequences, such as damage to infrastructure and loss of life. The risk from such events is likely to increase with climate change and is identified as a research topic that needs better quantification for UK climate change adaptation (CCRA 2017). One way to quantify and understand the risk due to such events is to have reliable estimates of the hazard footprint (Hill, Sparks, & Rougier 2013). Hazard footprints are conventionally estimated using measurements from observing stations and/or gridded analyses produced by numerical climate or air pollution simulation models (Chang 2016; Dee et al. 2011; Fuentes & Raftery 2005).

One example of a hazard footprint is given in Figure 1 . This shows observations and the gridded climate model output for windstorm Imogen over France, the details of which are given in Section 5.1. As indicated by the plot, ground monitoring stations lack spatial coverage but can be thought to measure true wind speed fairly accurately. Structured physical model output tend to have complete spatial coverage but can only represent the phenomenon at the model's predetermined resolution and not at smaller scales. In addition, model output tend to exhibit biases compared to the observations.

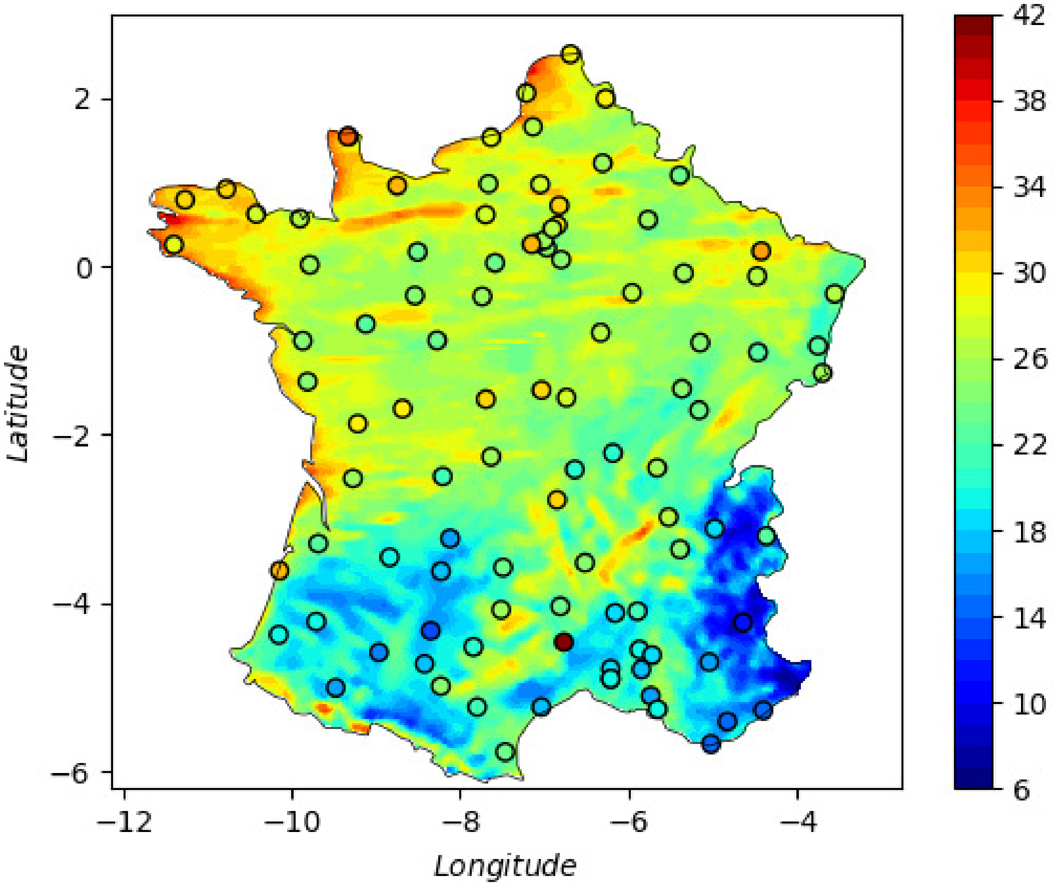


FIGURE 1 Windstorm Imogen footprint over France, according to storm footprints data from Windstorm Information Service (WISC 2019). The circles denote station observations, and are plotted on top of the gridded numerical climate model output.

Knowledge of small scale detail in environmental hazard footprints is important because of the large spatial heterogeneity in vulnerability and exposure. In the UK, for example, insurers require reliable estimates of the windstorm event for individual postcode regions (which may cover just a few houses) in order to be able to calculate losses using portfolio data on building types, values of property, etc. Interpolation at small scale based solely on observational data is often not accurate due to the sparsity of the monitoring stations. It is therefore necessary to utilise additional sources of data, such as physical model output. However, integrating multiple data sources for improved spatial interpolation is a non-trivial task that brings several challenges:

1. All data sources should be thought of as imperfect representations of the truth. Biases can be both systematic and random.
2. The data sources can have different spatial support, e.g. point station measurements and gridded model output. This is often called the "change of support" problem (Cressie 1993).
3. It is important to accurately quantify uncertainty from the different data sources and propagate this to predictions.

Here, we present a general modelling framework that is able to tackle these challenges and provide footprint estimates (predictions) that reliably integrate information across different data sources. With this framework we assume that model output are at sufficiently high resolution to provide a good representation of the true spatial process. Model output at lower resolution are classed as a data source of insufficient value.

The paper is structured as follows: Section 2 presents a brief review of methods of integrating multiple data sources and introduces notation. Section 3 presents a detailed description of the data fusion (DF) framework. Section 4 reports on the experiments and results from simulation experiments. Section 5 presents the application of this to windstorm Imogen. Conclusions and future research directions are discussed in Section 6.

2 | BACKGROUND

Integrating information across a number of data sources is a common problem, and a number of approaches have been proposed to combine data for either calibration of computer code, improved spatial predictions or (physical) model validation. Poole and Raftery (2000) first introduce the Bayesian melding (BM) concept, where "melding" refers to combining information about both inputs and outputs of deterministic simulation models. Kennedy and O'Hagan (2001) have introduced a framework that utilises observations to calibrate inputs of computer code that produces simulations from a physical model. A similar approach has also been presented by Craig, Goldstein, Rougier, and Seheult (2001) for Bayesian estimation for complex systems.

The BM approach of Fuentes and Raftery (2005) utilises air pollutant observations at point level to validate the numerical deterministic simulation models that provide estimates of pollutant concentrations at grid cells over the US. It models the underlying true process $Z(\boldsymbol{s})$ as a Gaussian process (GP) driving both sources of data, where the observational data $Y(\boldsymbol{s})$ is expressed as $Z(\boldsymbol{s})$ plus measurement error $\epsilon(\boldsymbol{s})$, while the numerical model output $X(\boldsymbol{s})$ is related to the true process via a systematic additive bias term $\alpha(\boldsymbol{s})$, a multiplicative bias term $\beta(\boldsymbol{s})$ and an i.i.d. Gaussian error term $\delta(\boldsymbol{s})$:

$$Z(\boldsymbol{s}) \sim GP\big(\mu_Z(\boldsymbol{s}), c_Z(\boldsymbol{s}, \boldsymbol{s}')\big) \tag{1}$$

$$Y(\boldsymbol{s}) = Z(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \tag{2}$$

$$X(\boldsymbol{s}) = \alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})Z(\boldsymbol{s}) + \delta(\boldsymbol{s}) \tag{3}$$

$$\delta(\boldsymbol{s}) \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

where $\mu_Z(\boldsymbol{s})$ and $c_Z(\boldsymbol{s}, \boldsymbol{s}')$ are the mean and covariance functions respectively, and σ is the standard deviation of the model discrepancy. The change of support is solved by integrating the point-level numerical model output $X(\boldsymbol{s})$ over grid cells. So for model output defined on a grid, the contribution to the likelihood of $X(A)$ where A is a grid cell, is based on $\int_A X(\boldsymbol{s})d\boldsymbol{s}$. This assumes that the numerical model output $X(A)$ is interpreted as an aggregation of $X(\boldsymbol{s})$ over the cell. Note also that both data sources are conditional upon the true process, which in principle makes it more straightforward to coherently add more data sources. The BM approach has been further extended to allow for spatio-temporal variability (McMillan, Holland, Morara, & Feng 2010) and multivariate output (Foley & Fuentes 2008).

The method introduced by Paciorek (2012) combines spatial information sources while accounting for systematic errors in proxies and considers different combinations of spatial scale for the true process, discrepancy, and model resolution. While Paciorek (2012) relies on discretisation of the true process, this can in theory be at sufficiently high resolution to give no appreciable difference between the continuous true process and its discrete approximation. In practice, though, this may be computationally prohibitive. BM avoids any discretisation of the true process by virtue of its explicit link between the continuous and aggregated process achieved by integrating over grid cells to obtain covariances. Hence BM needs no consideration of sensitivity to a subjectively chosen fine base grid.

An alternative approach to hierarchical formulations, is to consider a regression setting. Shaddick et al. (2018) introduce a data integration model where different sources of data act as covariates and spatial random coefficients reflect a nested geographical hierarchy to allow the bias (calibration) functions to vary over space. While this approach is flexible, it does not easily allow for the change of support challenge or marginal inference on the true data generating process.

In addition, there are other approaches that consider DF in both spatial and spatial-temporal modelling of data. For example, Keller and Peng (2019) present a framework for evaluating the error in aggregating areal exposure concentrations for air pollution epidemiology. Boaz, Lawson, and Pearce (2019) develop a multivariate fusion framework to deal with air pollution prediction with partial missingness. Wilkie et al. (2019) treat data for fusion as realisations of smooth temporal functions, Gilani, Berrocal, and Batterman (2019) accounts for nonstationarity by incorporating covariates, postulated to drive the nonstationary behaviour, in the covariance function and Ma and Kang (2020) develop a stochastic expectation-maximisation that facilitates the use of large spatio-temporal data sets. Forlani, Bhatt, Cameletti, Krainski, and Blangiardo (2020) demonstrate the added benefit of allowing multiple sources of model output in a framework that combines stochastic partial differential equations and the integrated nested Laplace approximation (Lindgren, Rue, & Lindström 2011).

Among the aforementioned DF approaches, BM seems to be the one that yields simultaneous solutions to the challenges (see Section 1) of bias removal, change of support and uncertainty quantification that arise when attempting to estimate environmental hazard footprints. Note that in BM, the change of support challenge is dealt with in a way that is in line with the interpretation of gridded model output, i.e. that the value in a given (spatio-temporal) grid is an aggregation of the variable in that cell. In the subsequent chapter, we present a generic DF framework that extends the BM approach to allow for a potentially more structured discrepancy term.

3 | ENVIRONMENTAL HAZARD FOOTPRINT MODELLING FRAMEWORK

Brynjarsdóttir and O'Hagan (2014) give a thorough exposition of the importance of modelling discrepancy in uncertainty quantification and has shown that explicitly modelling the discrepancy between simulator outputs and true process using GPs could lead to enormously improved predictions. Thus, we relax the assumption that the discrepancy term $\delta(\boldsymbol{s})$ in (3) is i.i.d. and instead model it as a zero-mean GP:

$$\delta(\boldsymbol{s}) \sim GP\big(0, c_{\delta}(\boldsymbol{s}, \boldsymbol{s}')\big)$$

Recall that this term quantifies the error between the true process and the model output after allowing for structured bias through $\alpha(\boldsymbol{s})$ and $\beta(\boldsymbol{s})$. In the application to environmental hazard footprints, such as windstorm footprints, this discrepancy is likely to be spatially structured. It is true for instance that some climate models tend to get the storm track – the "path" that storms travel over – slightly out of place (Zappa, Shaffrey, Hodges, Sansom, & Stephenson 2013). So while the bias term will capture consistent under- or over-estimation of wind speeds, the discrepancy term can allow, to a certain extent, for spatial misalignment. More generally, the model is able to adapt to any possible form of the spatially structured discrepancy by exploiting the GP nonparametric formulation (Xiong, Šmídl, & Filippone 2017) and hence reduce uncertainties when making predictions.

3.1 | Proposed data fusion framework

Hence, the proposed DF modelling framework for environmental hazard footprints is

$$Z(\boldsymbol{s}) \sim GP\big(\mu(\boldsymbol{s}), c_Z(\boldsymbol{s}, \boldsymbol{s}')\big) \qquad \text{True process} \tag{5}$$

$$Y(\boldsymbol{s}) = Z(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \qquad \text{Data (observations)} \tag{6}$$

$$\epsilon(\boldsymbol{s}) \sim \mathcal{N}(0, \sigma_Y^2) \qquad \text{Measurement error} \tag{7}$$

$$X(\boldsymbol{s}) = \alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})Z(\boldsymbol{s}) + \delta(\boldsymbol{s}) \qquad \text{Numerical model output} \tag{8}$$

$$\delta(\boldsymbol{s}) \sim GP\big(0, c_\delta(\boldsymbol{s}, \boldsymbol{s}')\big) \qquad \text{Discrepancy term} \tag{9}$$

The bias terms ($\alpha(\boldsymbol{s})$, $\beta(\boldsymbol{s})$) in the numerical model output are assumed linear, and are meant to capture spatially varying under- or over-estimation of a hazard event. In the application study of the windstorm footprint (see section 5), $\alpha(\boldsymbol{s})$ is defined as a polynomial function of spatial coordinates and $\beta(\boldsymbol{s})$ is defined as a global scaling factor (constant β) to capture multiplicative bias across all wind speed values. The covariance functions $c_Z(\boldsymbol{s}, \boldsymbol{s}')$ and $c_\delta(\boldsymbol{s}, \boldsymbol{s}')$, which depend on hyper-parameters $\boldsymbol{\psi}_Z$ and $\boldsymbol{\psi}_\delta$, respectively, are key to the performance of the spatial interpolation. If $c_Z(\boldsymbol{s}, \boldsymbol{s}')$ and $c_\delta(\boldsymbol{s}, \boldsymbol{s}')$ depend only on $\| \boldsymbol{s} - \boldsymbol{s'} \|$ then both processes are stationary and well-established covariance functions, such as the powered exponential and Matérn, can be used.

To complete model specification, we need to allow for the fact the data $Y(\boldsymbol{s})$ are point locations while data from the numerical model output are gridded, i.e. $X(A_i)$ where A_i is a region reflecting the size and shape of each grid cell i. Following Fuentes and Raftery (2005), $X(A_i)$ is interpreted as the integration of $X(\boldsymbol{s})$ over A_i and takes the following form

$$X(A_i) = \int_{A_i} X(\boldsymbol{s})d\boldsymbol{s} = \int_{A_i} \alpha(\boldsymbol{s})d\boldsymbol{s} + \int_{A_i} \beta(\boldsymbol{s})Z(\boldsymbol{s})d\boldsymbol{s} + \int_{A_i} \delta(\boldsymbol{s})d\boldsymbol{s}$$

so that $X(\boldsymbol{s})$, for any $\boldsymbol{s} \in A_i$, is unobserved and only its aggregated counterpart $X(A_i)$ is observed.

Let $\boldsymbol{\theta}$ represent all the parameters involved in (5)–(9), namely $\boldsymbol{\psi}_Z$, $\boldsymbol{\psi}_\delta$, σ_Y, and any coefficients in $\alpha(\boldsymbol{s})$ and $\beta(\boldsymbol{s})$. Integrating out the true process $Z(\boldsymbol{s})$ implies that observations $\boldsymbol{Y} = [Y(\boldsymbol{s}_1), ..., Y(\boldsymbol{s}_n)]^T$ and model output $\boldsymbol{X} = [X(A_1), ..., X(A_m)]^T$ jointly follow a multivariate Gaussian distribution

$$p(\boldsymbol{Y}, \boldsymbol{X} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \tilde{\boldsymbol{\mu}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\boldsymbol{Y}} & \Sigma_{\boldsymbol{YX}} \\ \Sigma_{\boldsymbol{XY}} & \Sigma_{\boldsymbol{X}} \end{bmatrix} \right) = \mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{C}) \tag{10}$$

where $\boldsymbol{Y}$ is of size n and $\boldsymbol{X}$ is of size m, with

$$\hat{\boldsymbol{\mu}} = (\mu(\boldsymbol{s}_1), ..., \mu(\boldsymbol{s}_n))^T \quad \text{and} \quad \tilde{\boldsymbol{\mu}} = \Big(\int_{A_1} \big[\alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})\mu(\boldsymbol{s})\big] d\boldsymbol{s}, ..., \int_{A_m} \big[\alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})\mu(\boldsymbol{s})\big] d\boldsymbol{s} \Big)^T.$$

In (10), $\Sigma_{\boldsymbol{Y}}$ is the covariance of $\boldsymbol{Y}$ defined by

$$\text{cov}(Y(\boldsymbol{s}), Y(\boldsymbol{s}')) = c_Z(\boldsymbol{s}, \boldsymbol{s}') + \sigma_Y^2 \mathbb{1}(\boldsymbol{s} = \boldsymbol{s}'). \tag{10.1}$$

Matrix $\Sigma_{\boldsymbol{YX}}$ is the cross-covariance between $\boldsymbol{Y}$ and $\boldsymbol{X}$, with its entries specified by

$$\text{cov}(Y(\boldsymbol{s}_i), X(A_j)) = \frac{1}{|A_j|} \int_{A_j} \beta(\boldsymbol{v}) c_Z(\boldsymbol{s}_i, \boldsymbol{v}) d\boldsymbol{v} \tag{11}$$

where $|A_j|$ is the area of grid cell A_j. Lastly, $\Sigma_{\boldsymbol{X}}$ is the covariance of $\boldsymbol{X}$ with entries determined by

$$\text{cov}(X(A_i), X(A_j)) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} \beta(\boldsymbol{u})\beta(\boldsymbol{v}) c_Z(\boldsymbol{u}, \boldsymbol{v}) d\boldsymbol{u} d\boldsymbol{v} + \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} c_\delta(\boldsymbol{u}, \boldsymbol{v}) d\boldsymbol{u} d\boldsymbol{v}. \tag{12}$$

In practice, we uniformly sample L_i and L_j locations in A_i and A_j respectively, so that (11) is approximated by

$$\widehat{\text{cov}}(Y(\boldsymbol{s}_i), X(A_j)) = \frac{1}{L_j} \sum_{\gamma=1}^{L_j} \beta(\boldsymbol{v}_{j\gamma}) c_Z(\boldsymbol{s}_i, \boldsymbol{v}_{j\gamma})$$

whereas (12) is approximated by

$$\widehat{\text{cov}}(X(A_i), X(A_j)) = \frac{1}{L_i L_j} \sum_{\gamma=1}^{L_i} \sum_{\gamma'=1}^{L_j} \left(\beta(\boldsymbol{u}_{i\gamma}) \beta(\boldsymbol{v}_{j\gamma'}) c_Z(\boldsymbol{u}_{i\gamma}, \boldsymbol{v}_{j\gamma'}) + c_\delta(\boldsymbol{u}_{i\gamma}, \boldsymbol{v}_{j\gamma'}) \right).$$

It should be noted that, particularly in situations where the number of numerical model output $\boldsymbol{X}$ is much larger than that of the observations $\boldsymbol{Y}$, there might be non-identifiability in uniquely characterising $Z(\boldsymbol{s})$ and $\delta(\boldsymbol{s})$. In practice, this can be avoided by constraining the smoothness properties of each GP, i.e. constraining the parameters in $c_Z(\boldsymbol{s}, \boldsymbol{s}')$ and $c_\delta(\boldsymbol{s}, \boldsymbol{s}')$. For environmental hazard footprints, the true process $Z(\boldsymbol{s})$ is assumed smoother than the discrepancy term $\delta(\boldsymbol{s})$ to allow for the fact that the latter is effectively an error term while the former is a physical process. We implement constrained optimisation to achieve this in our simulation study (see Section 4). We find that the results are similar to those achieved by unconstrained optimisation, and hence the simulation study also suggests implementing this constraint is not always necessary in practice.

3.2 | Inference

Let $\boldsymbol{D}$ denote the combined data points of the observations $\boldsymbol{Y}$ and the gridded model output $\boldsymbol{X}$, i.e. $\boldsymbol{D} = [Y(\boldsymbol{s}_1), ..., Y(\boldsymbol{s}_n), X(A_1), ..., X(A_m)]^T$. Then, (10) implies that the likelihood for the data is based on the multivariate Gaussian distribution, so that the log-likelihood is given by

$$\log[p(\boldsymbol{D} \mid \boldsymbol{\theta})] = -\frac{1}{2} \log(|\boldsymbol{C}|) - \frac{1}{2} (\boldsymbol{D} - \boldsymbol{\mu}^+)^T \boldsymbol{C}^{-1} (\boldsymbol{D} - \boldsymbol{\mu}^+) + const. \tag{13}$$

Inference on $\boldsymbol{\theta}$ can then be performed using maximum likelihood, as shown later in Section 5.

Note however that the primary goal here is the estimation or rather prediction of the true environmental hazard footprint $Z(\boldsymbol{s})$, given all the data. Then suppose that a prediction for the true footprint is sought for an arbitrary in-sample or out-of-sample location $\boldsymbol{s}$. The covariance matrix between $Z(\boldsymbol{s})$ and $\boldsymbol{D}$ is given by

$$\boldsymbol{k}^*(\boldsymbol{s}) = [c_Z(\boldsymbol{s}, \boldsymbol{s}_1), ..., c_Z(\boldsymbol{s}, \boldsymbol{s}_n), \text{cov}(Z(\boldsymbol{s}), X(A_1)), ..., \text{cov}(Z(\boldsymbol{s}), X(A_m))]^T \tag{13.1}$$

where n is the number of observations, m is the number of grid cells ($A_i, i = 1, ..., m$) relating to model output and $\text{cov}(\cdot, \cdot)$ is defined by (10.1) and (11). Letting $k(\boldsymbol{s}) = c_Z(\boldsymbol{s}, \boldsymbol{s})$, the Gaussian identity implies that the joint distribution of the combined data points $\boldsymbol{D}$ and the true footprint at $\boldsymbol{s}$ is

$$\begin{bmatrix} \boldsymbol{D} \\ Z(\boldsymbol{s}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^+ \\ \mu(\boldsymbol{s}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{C} & \boldsymbol{k}^*(\boldsymbol{s}) \\ \boldsymbol{k}^*(\boldsymbol{s})^T & k(\boldsymbol{s}) \end{bmatrix} \right)$$

where

$$\begin{aligned} Z(\boldsymbol{s}) \mid \boldsymbol{D}, \boldsymbol{\theta} &\sim \mathcal{N}(\hat{\mu}(\boldsymbol{s}), \sigma^2(\boldsymbol{s})) \\ \hat{\mu}(\boldsymbol{s}) &= \mu(\boldsymbol{s}) + \boldsymbol{k}^*(\boldsymbol{s})^T \boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{\mu}^+) \\ \sigma^2(\boldsymbol{s}) &= k(\boldsymbol{s}) - \boldsymbol{k}^*(\boldsymbol{s})^T \boldsymbol{C}^{-1} \boldsymbol{k}^*(\boldsymbol{s}) \end{aligned} \tag{14}$$

Thus, predictions are achieved by plugging in the estimate of $\boldsymbol{\theta}$ into (14).

3.3 | Model checking

In this section, we present a suite of model checking diagnostics, motivated by Bastos and O'Hagan (2009) who discuss diagnostics for GP emulators. One important aspect of these diagnostics is that unlike conventional verification and validation methods, they take into account the correlation structure in the predictions. This is particularly important for ensuring GP assumptions are valid.

Let $\boldsymbol{S}^* = (\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_v)$ denote a set of locations where observation data are not available, $\boldsymbol{\mu}^* = (\mu(\boldsymbol{s}_1), \mu(\boldsymbol{s}_2), ..., \mu(\boldsymbol{s}_v))$ denote the mean values of the true process at these locations and $\boldsymbol{Y}^* = (Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2), ..., Y(\boldsymbol{s}_v))$ denote the corresponding predicted values (truth plus the measurement error). The predictive distribution of $\boldsymbol{Y}^*$ given the data is

$$\boldsymbol{Y}^* \mid \boldsymbol{D}, \boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^*, \boldsymbol{V}^*)$$

with

$$\hat{\boldsymbol{\mu}}^* = \boldsymbol{\mu}^* + \boldsymbol{K}^{*T} \boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{\mu}^+)$$

$$\boldsymbol{V}^* = \boldsymbol{\Sigma}_{\boldsymbol{Y}^*} - \boldsymbol{K}^{*T} \boldsymbol{C}^{-1} \boldsymbol{K}^*$$

where each entry of $\boldsymbol{\Sigma}_{\boldsymbol{Y}^*}$ is specified by $c_Z(\boldsymbol{s}_i, \boldsymbol{s}_j) + \sigma_Y^2 \mathbb{1}(\boldsymbol{s}_i = \boldsymbol{s}_j)$ and each column of $\boldsymbol{K}^*$ is $\boldsymbol{k}^*(\boldsymbol{s}_i)$ from (13.1) for $i = 1, ..., v$.

Then by definition, the diagonals of $\boldsymbol{V}^*$ are the marginal predictive variances of $\boldsymbol{Y}^*$ (denoted by $\mathrm{Var}(\boldsymbol{Y}^*)$). Let $\tilde{\boldsymbol{Y}}$ denote the real observations at locations $\boldsymbol{S}^*$ that are left out for validation. The first diagnostic relates to the standardised individual prediction errors for the validation data $\tilde{\boldsymbol{Y}}$, and is defined as

$$D_I(\tilde{\boldsymbol{Y}}) = \frac{\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\mu}}^*}{\sqrt{\mathrm{Var}(\boldsymbol{Y}^*)}}$$

A plot of $\boldsymbol{D}_I(\tilde{\boldsymbol{Y}})$ against the predictions ($\hat{\boldsymbol{\mu}}^*$) or the validation data index provides a graphical diagnostic for the model. The distributional assumptions of the model are valid if $\boldsymbol{D}_I(\tilde{\boldsymbol{Y}})$ fluctuate around 0 with a constant variance and display no particular patterns. Large individual errors indicate an underestimation of the variance, whereas small individual errors indicate an overestimation of the variance. It may also suggest inappropriate assumptions of stationarity. In particular, the plot of $\boldsymbol{D}_I(\tilde{\boldsymbol{Y}})$ against the predictions ($\hat{\boldsymbol{\mu}}^*$) may suggest a problem in the mean function if the errors are systematically positive (or negative) for some particular ranges of the predictions.

Since by definition, $\boldsymbol{D}_I(\tilde{\boldsymbol{Y}})$ are potentially correlated, the next diagnostic produces uncorrelated prediction errors defined by

$$D_G(\tilde{\boldsymbol{Y}}) = \boldsymbol{G}^{-1}(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\mu}}^*)$$

where $\boldsymbol{G}$ is the standard deviation matrix which satisfies $\boldsymbol{V}^{\star} = \boldsymbol{G}\boldsymbol{G}^T$. Following Bastos and O'Hagan (2009), we use the pivoted Cholesky decomposition to obtain $\boldsymbol{G}$ where the data are permuted so that the first data point is the one with the largest predictive variance, the second data point is the one with the largest predictive variance conditioned on the first one, and so on. The pivoted Cholesky decomposition returns a permutation matrix $\boldsymbol{P}$ and a lower triangular matrix $\boldsymbol{L}$ such that $\boldsymbol{P}^T\boldsymbol{V}^*\boldsymbol{P} = \boldsymbol{L}\boldsymbol{L}^T$, leading to $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{L}$. We call the corresponding errors pivoted Cholesky errors and denote them by $\boldsymbol{D}_{PC}(\tilde{\boldsymbol{Y}})$. The main advantage of $\boldsymbol{D}_{PC}(\tilde{\boldsymbol{Y}})$ is that, in the case where the covariance function is the Gaussian covariance function, i.e. $c(\boldsymbol{s}, \boldsymbol{s}') = \sigma^2 \exp\left\{-(\| \mathbf{s} - \mathbf{s}' \| / \phi)^2\right\}$, it could diagnose both the estimation of the marginal variance σ^2 and the estimation of the correlation length-scale ϕ. Unusually large or small errors in the first part of the sequence indicate poor estimation of the marginal variance σ^2, whereas large or small errors in the latter part of the sequence suggest poor estimation of the correlation length-scale ϕ or a necessity to reconstruct the correlation structure.

In the subsequent section, we present a simulation study of windstorm events to investigate the implementation and performance of the proposed framework. This is followed by an application of this to data relating to storm Imogen.

TABLE 1 **Algorithm for simulating data.**

1. Generate $Z(\boldsymbol{s}_i)(i = 1, ..., 10^6)$ from $Z(\boldsymbol{s}) \sim \text{GP}(0, \eta(\boldsymbol{s}, \boldsymbol{s}'))$ where $\eta(\boldsymbol{s}, \boldsymbol{s}'))$ is specified by (16), $\boldsymbol{s}_i$ are ordered as a 1000×1000 grid and are within the area of France (denoted by $\mathcal{S}$), i.e. $\text{lon}(\boldsymbol{s}) \in (-12.0, -3.0), \text{lat}(\boldsymbol{s}) \in (-6.3, 3.0)$.
2. Simulate observations $Y(\boldsymbol{s}_i)(i = 1, ..., n_Y)$ where n_Y is set to 200 in this case. First randomly sample $Z(\boldsymbol{s}_i)$ of size n_Y. Then $Y(\boldsymbol{s}_i)$ is simulated according to the equation $Y(\boldsymbol{s}_i) = Z(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \epsilon(\boldsymbol{s}_i) \sim \mathcal{N}(0, 0.25)$.
3. Generate the model output $X(A_i)(i = 1, ..., n_X)$ where n_X is set to 625 in this case, A_i are ordered as a 25×25 block and the average of $Z(\boldsymbol{s})$ within each A_i is taken to be $X(A_i)$.

Notes: In the above step (3), the number of spatial locations within each block A_i could be any arbitrary number ν. In this simulation, ν is set to 100.

4 | SIMULATION STUDY

4.1 | Algorithm for data simulation

In this simulation study, we assume a zero-mean GP for $Z(\boldsymbol{s})$, i.e. $\boldsymbol{\mu}(\boldsymbol{s})$ in (5) is $\boldsymbol{0}$. In both the simulation and application study in this work, the covariance functions in (5) and (9) are both Gaussian covariance functions defined as $c_Z(\boldsymbol{s}, \boldsymbol{s}') = \sigma_Z^2 \exp\left\{-(\| \mathbf{s} - \mathbf{s}' \| / \phi_Z)^2\right\}$ and $c_\delta(\boldsymbol{s}, \boldsymbol{s}') = \sigma_\delta^2 \exp\left\{-(\| \mathbf{s} - \mathbf{s}' \| / \phi_\delta)^2\right\}$. Thus, we are assuming that both $Z(\boldsymbol{s})$ and $\delta(\boldsymbol{s})$ are isotropic and relatively smooth. For $Z(\boldsymbol{s})$, σ_Z^2 represents the marginal variance of the unobserved true process and the correlation length-scale ϕ_Z accounts for the magnitude of dependence between $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s}')$, with a smaller value of ϕ_Z indicating stronger dependence.

The additive bias $\alpha(\boldsymbol{s})$ is modelled as a polynomial of longitude and latitude at location $\boldsymbol{s}$ (denoted by $\text{lon}(\boldsymbol{s})$ and $\text{lat}(\boldsymbol{s})$, respectively) with the form:

$$\alpha(\boldsymbol{s}) = \alpha_0 + \alpha_1 \text{lon}(\boldsymbol{s}) + \alpha_2 \text{lat}(\boldsymbol{s}) + \alpha_3 \text{lon}(\boldsymbol{s}) \text{lat}(\boldsymbol{s}) + \alpha_4 \text{lon}(\boldsymbol{s})^2 + \alpha_5 \text{lat}(\boldsymbol{s})^2. \tag{15}$$

In order to simulate data that better represent real windstorm footprints, $Z(\boldsymbol{s})$ is generated using the following powered exponential covariance function

$$\eta(\boldsymbol{s}, \boldsymbol{s}') = \tau \exp\left\{- \| \mathbf{s} - \mathbf{s}' \|^\alpha / \varphi^2\right\} \tag{16}$$

where τ is set to 1, φ is set to 1.5 and α is set to 1.8. It looks quite skewed compared to data simulated using standard Gaussian kernel and hence the model's ability to capture non-Gaussian data can be tested.

The algorithm for simulating the true windstorm process $Z(\boldsymbol{s})$, the observations $Y(\boldsymbol{s})$ and the areal model output $X(A_i)$ (A_i denotes the ith area) is given in Table 1 .

Figure 2 (a) presents an example of the simulated true process $Z(\boldsymbol{s})$ of a windstorm event. Figure 2 (b) shows the associated $X(A_i)$ which represent the areal climate model output, while Figure 2 (d) shows a plot of 200 $Z(\boldsymbol{s})$ values against corresponding

$Y(\boldsymbol{s})$ values as well as 400 $Z(\boldsymbol{s})$ values against corresponding $X(A_i)$ values. Averaging $Z(s)$ to get $X(A_i)$ has clearly induced bias. Finally, Figure 2 (c) presents predictions $Z(\boldsymbol{s})|Y(s), X(A_i)$ from the model fitted to the data presented in Figure 2 (d). The plot indicates that the model has captured the true process $Z(s)$ (Figure 2 (a)) reasonably well.

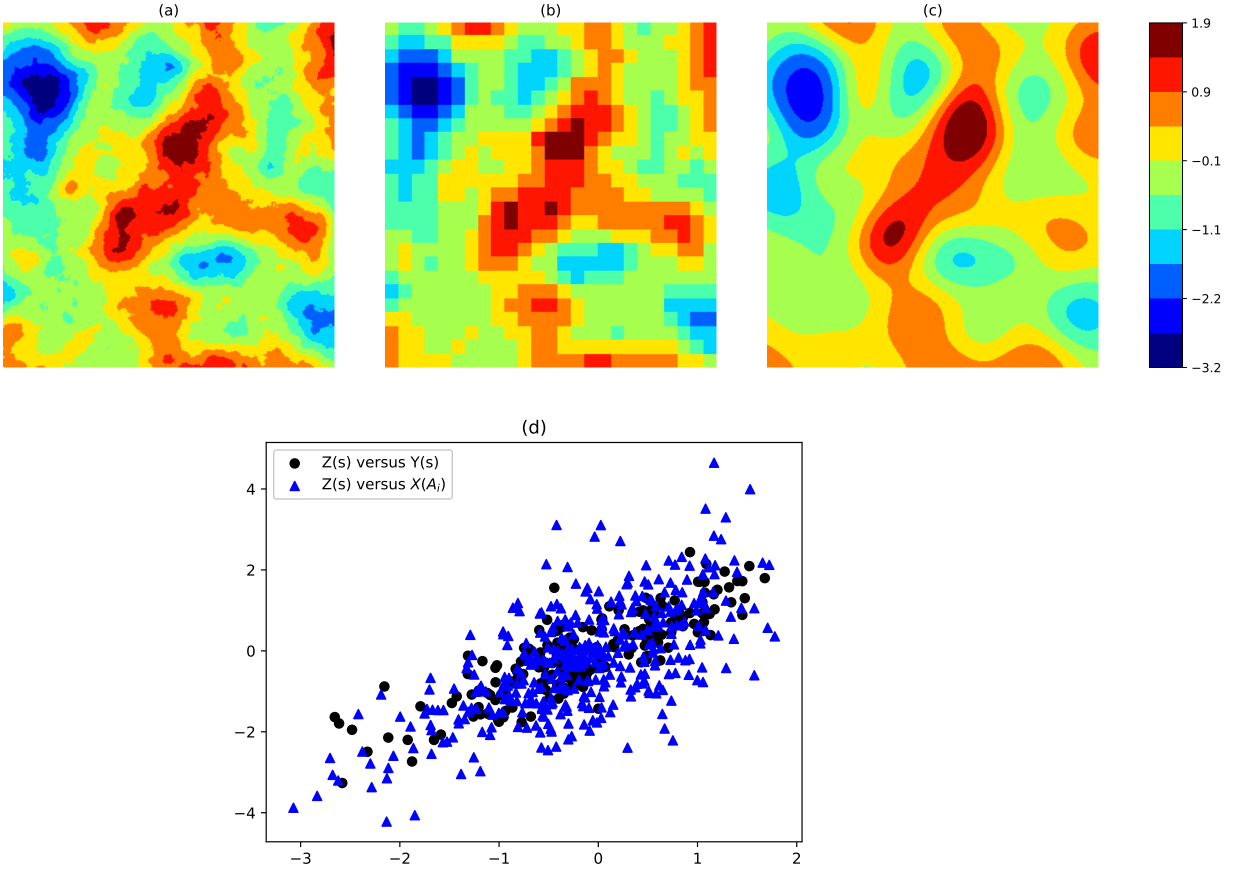
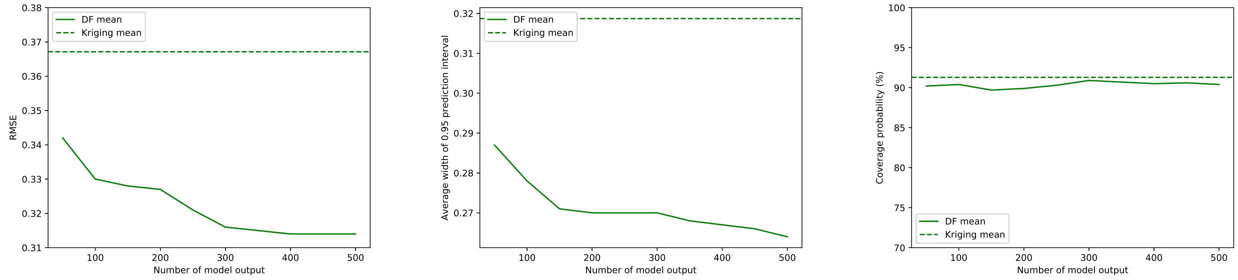


FIGURE 2 Illustration of simulated $Z(\boldsymbol{s}), Y(\boldsymbol{s})$ and $X(A_i)$ and predicted $Z(\boldsymbol{s})$. (a) Simulated $Z(\boldsymbol{s})$ over a 1000 × 1000 grid. (b) Simulated $X(A_i)$ over a 25 × 25 block. (c) Predicted $Z(\boldsymbol{s})$ over a 1000 × 1000 grid. (d) Simulated Z(s) versus $Y(s)$ of size 200, $X(A_i)$ of size 400.

4.2 | Results

Thirty such simulations (see the algorithm presented in Table 1) were performed to represent 30 different windstorm events. In order to test the impact of the number of model output on the DF model, for each simulation, we fix the number of observations n_Y to be 200. Then we start by randomly sampling 50 areal model output out of the total 625 ones, then increasingly add 50 different areal model output until we reach the maximum number of model output, which is set to 500 for computational convenience. These varying numbers of areal model output are denoted by $\boldsymbol{X}_{50}, \boldsymbol{X}_{100}, ..., \boldsymbol{X}_{500}$. Accordingly for each simulation there are 10 data combinations $(\boldsymbol{Y}, \boldsymbol{X}_{50})$, $(\boldsymbol{Y}, \boldsymbol{X}_{100})$, ..., $(\boldsymbol{Y}, \boldsymbol{X}_{500})$. Then the DF model is fitted to the 10 data combinations and an out-of-sample prediction is performed over the 10^6 grid cells where we have the simulated truth $Z(s)$. The prediction accuracy is measured in terms of the root-mean-square error (RMSE), the average width of prediction interval (AWPI) and the coverage probability. We also fit a Kriging model, basically defined by (5)–(6), only to the observations $\boldsymbol{Y}$. This is to ensure that using the DF framework to include biased realisations of the true process as extra information actually makes a difference in the predictions. The comparison of the out-of-sample prediction results between the two models are displayed in Figure 3 where the RMSE, AWPI and coverage probability are averages over the 30 simulations. The results suggest that by combining the model output with observations, the DF model achieves a lower RMSE and AWPI while maintaining a similar coverage probability when compared to the Kriging model, which only utilises the observational data. In addition, the impact of the number of model output is apparent: the higher the number of model output, the lower the RMSE and AWPI. We also perform another 30 simulations where the number of observations and areal model output is set to 200 and 400 respectively, and implement an out-of-sample prediction at the 10^6 grid cells. Figure 4 presents the out-of-sample RMSE, AWPI and coverage probability over the 30 simulations at the grid cells for both models. The left-hand column is for the Kriging model, whereas the right-hand column is for the DF model. As can be seen, compared to the Kriging model at the majority of the grid cells, the DF model exhibits a much lower RMSE and AWPI while maintaining a competitive coverage probability. The results from the simulation study suggest the presented DF model in Section 3 is able to more accurately quantify the uncertainty at high spatial resolution when predicting the true footprint.



(a) RMSE

(b) AWPI

(c) Coverage probability

FIGURE 3 The out-of-sample RMSE, AWPI and coverage probability over 30 simulations for Kriging and the DF approach.

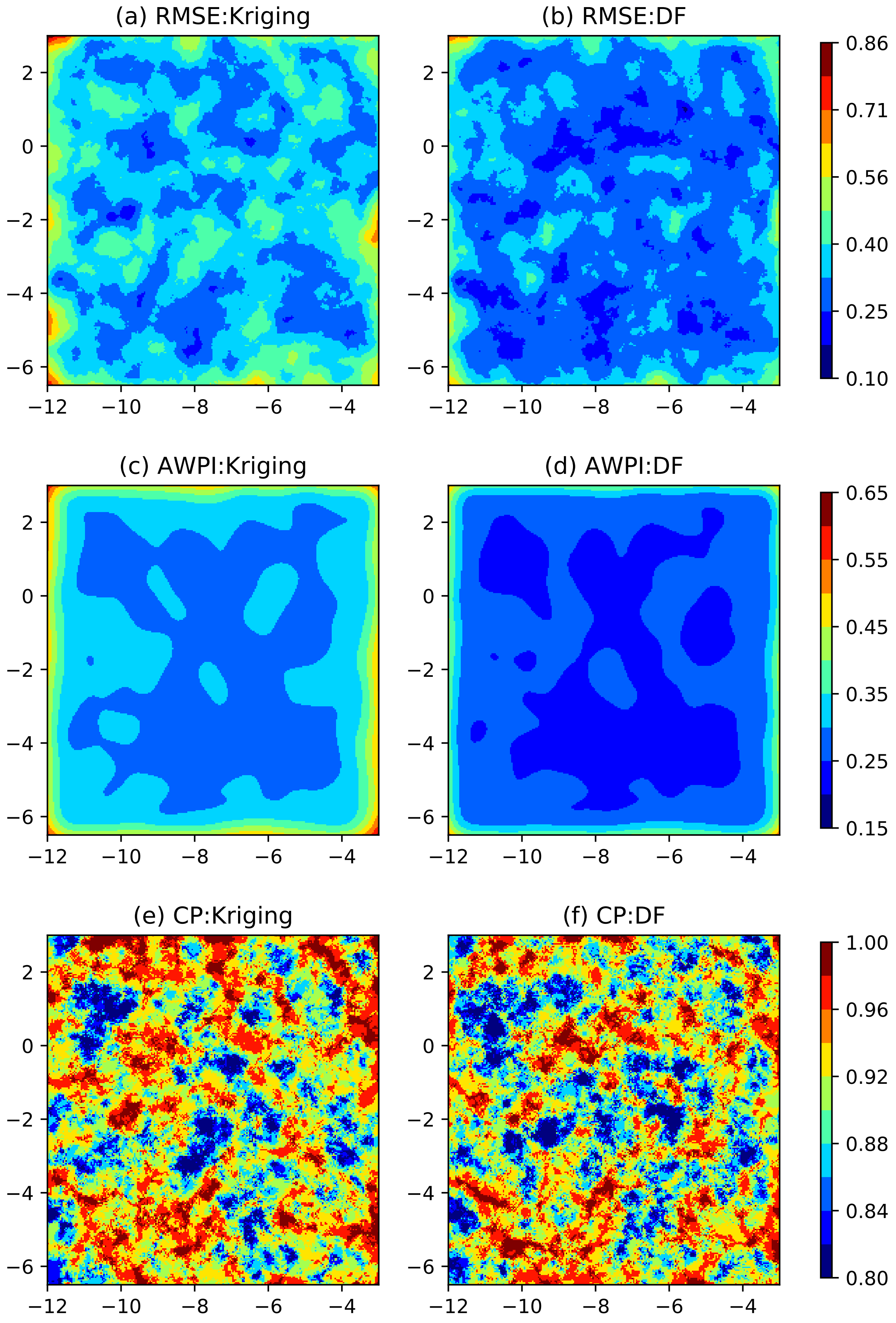


FIGURE 4 Out-of-sample RMSE, AWPI and coverage probability (CP) for Kriging and the DF approach at the 10^6 grid cells.

5 | APPLICATION TO STORM IMOGEN

This section presents the application of the DF model presented in Section 3 to windstorm data.

5.1 | Data

The windstorm hazard footprint is usually defined as a spatial map of the maximum 3-second gust speed over a 72-hour period centered at its occurrence time (Haylock 2011), and is something used as a proxy for property damage and insured loss (Klawa & Ulbrich 2003). We study European windstorm Imogen within the spatial domain of France. This avoids the added complexity of the distribution of gust speed being different over land compared to sea. It occurred between 06/02/2016 - 09/02/2016 and caused severe damage by bringing trees down, causing power cuts and producing large coastal waves. Figure 5 (a) reveals the simulated storm footprint obtained by a dynamic down-scaling of the ERA Interim re-analysis using the Met Office Unified Model(MetUM) (Davies et al. 2005; Dee et al. 2011; Roberts et al. 2014). While the ERA-Interim re-analysis assimilates observational data, here it only drives the MetUM at the boundary of the original footprint domain (which captures most of Europe). Hence we assume that no observational data are used twice. Figure 5 (a) has been superimposed by circles the 128 station observations that we use. The figure reveals a discrepancy of gust speed intensity between the simulated ones and the station observations. To fit the DF model, we randomly sample 500 simulated gust speeds across France, referred to as MetUM model output throughout the rest of this work. We also randomly sample 28 points from the 128 observations as out-of-sample test points, leaving 100 observations as training points. In this application study, to account for the effect of topology on the prediction, we assume the mean surface of the true footprint is determined by altitude, while the remaining spatial structure is captured by the covariance structure $c_Z(\boldsymbol{s}, \boldsymbol{s}')$. Consequently, $\mu(\boldsymbol{s})$ in (5) takes the following form:

$$\mu(\boldsymbol{s}) = \alpha_0 + \alpha_1 \mathrm{alt}(\boldsymbol{s}) + \alpha_2 \mathrm{alt}(\boldsymbol{s})^2 \tag{17}$$

where $\mathrm{alt}(\boldsymbol{s})$ is the altitude at the location $\boldsymbol{s}$ and $\alpha_0, \alpha_1, \alpha_2$ are the polynomial coefficients to be estimated.

5.2 | Results

We first predict the gust speeds using the DF model at the 500 locations where we have MetUM model output and plot them against the station observations which is displayed in Figure 5 (b). Compared to Figure 5 (a) which shows the MetUM model output against the station observations, it is evident that the discrepancy between predicted gust speeds and observations is lower at the majority of locations which suggests the proposed DF model is able to effectively produce improved predictions and could work as a way of validating numerical simulator outputs.

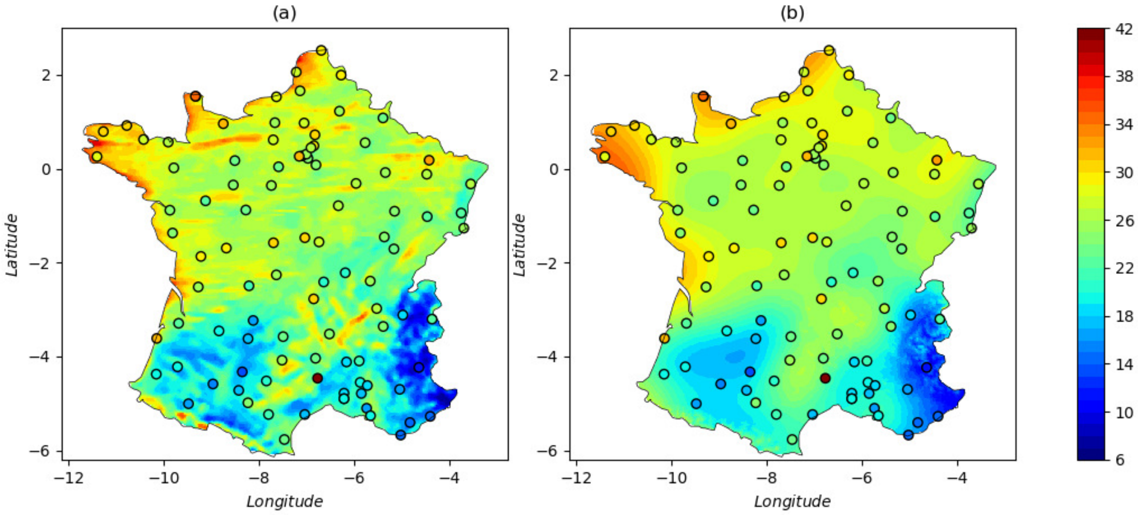


FIGURE 5 Plot of wind gust speed observations at weather stations (represented by coloured circles) and MetUM model output. (a) Station observations and MetUM model output used in the DF model. (b) Station observations and predicted model output.

Figure 6 (a) presents the predictions versus the observations for the 28 out-of-sample test points. The solid line indicates the case where the predictions exactly match the observations. The black circles represent observations versus the predicted mean, with the dotted lines indicating the 95% confidence interval for the predictions. Apart from one extreme value that is underestimated, the predictions for the rest of 27 points are within the 95% confidence interval. Figure 6 (b) presents a spatial map of the standardised residuals where the largest one is marked in red. Figure 7 is a colour map showing the values of the 100 observations (denoted by circles in (a)), 500 MetUM model output (denoted by circles in (b)) and the test point with the largest residual (represented by the square). It is obvious that the value of the red square (test point) seems to be much higher than the observations and the MetUM model output in the nearby area. This suggests that the DF model can be used to highlight outliers, such as potentially erroneous observations. The underestimation of the test point also implies future work could improve on the model's capability of capturing extreme values.

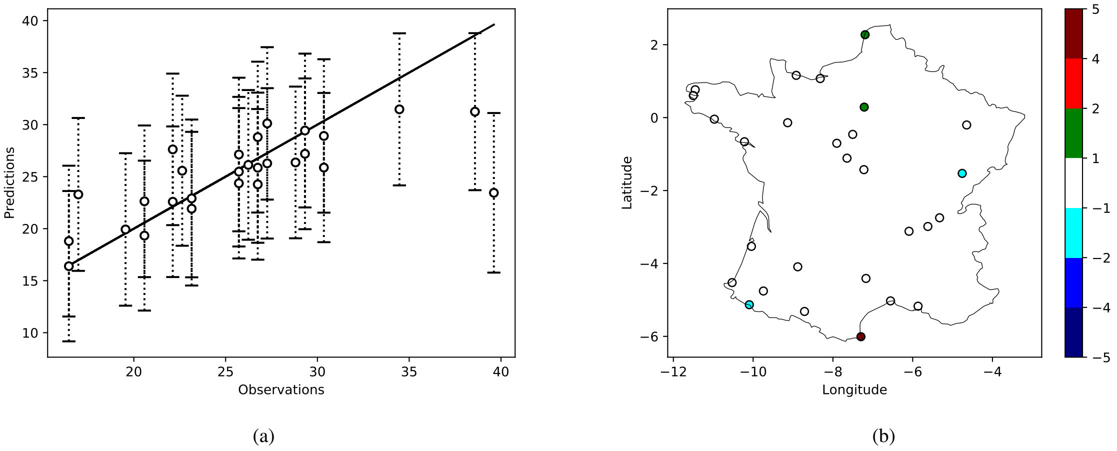


FIGURE 6 Plot of out-of-sample prediction. (a) Out-of-sample predictions versus observations. (b) Standardised residuals plotted on the spatial domain of France.

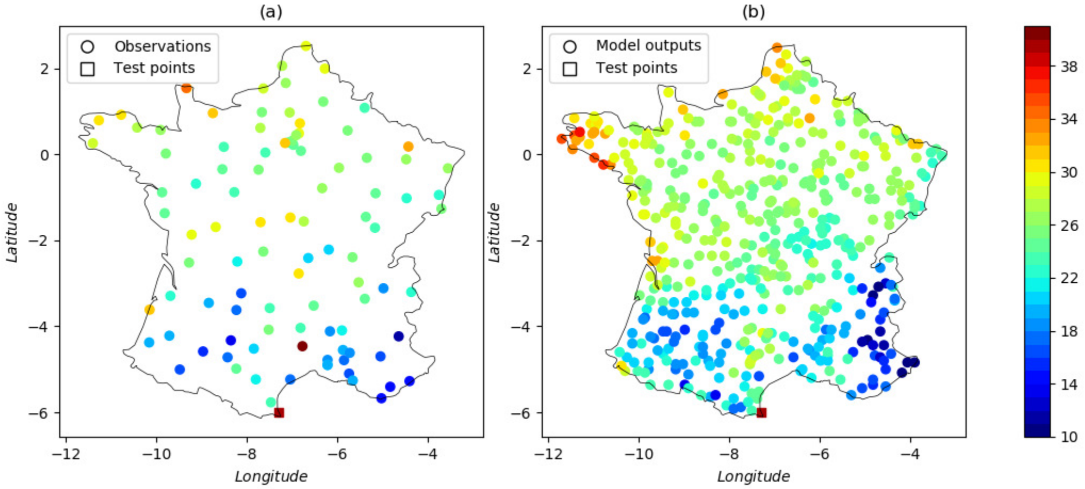
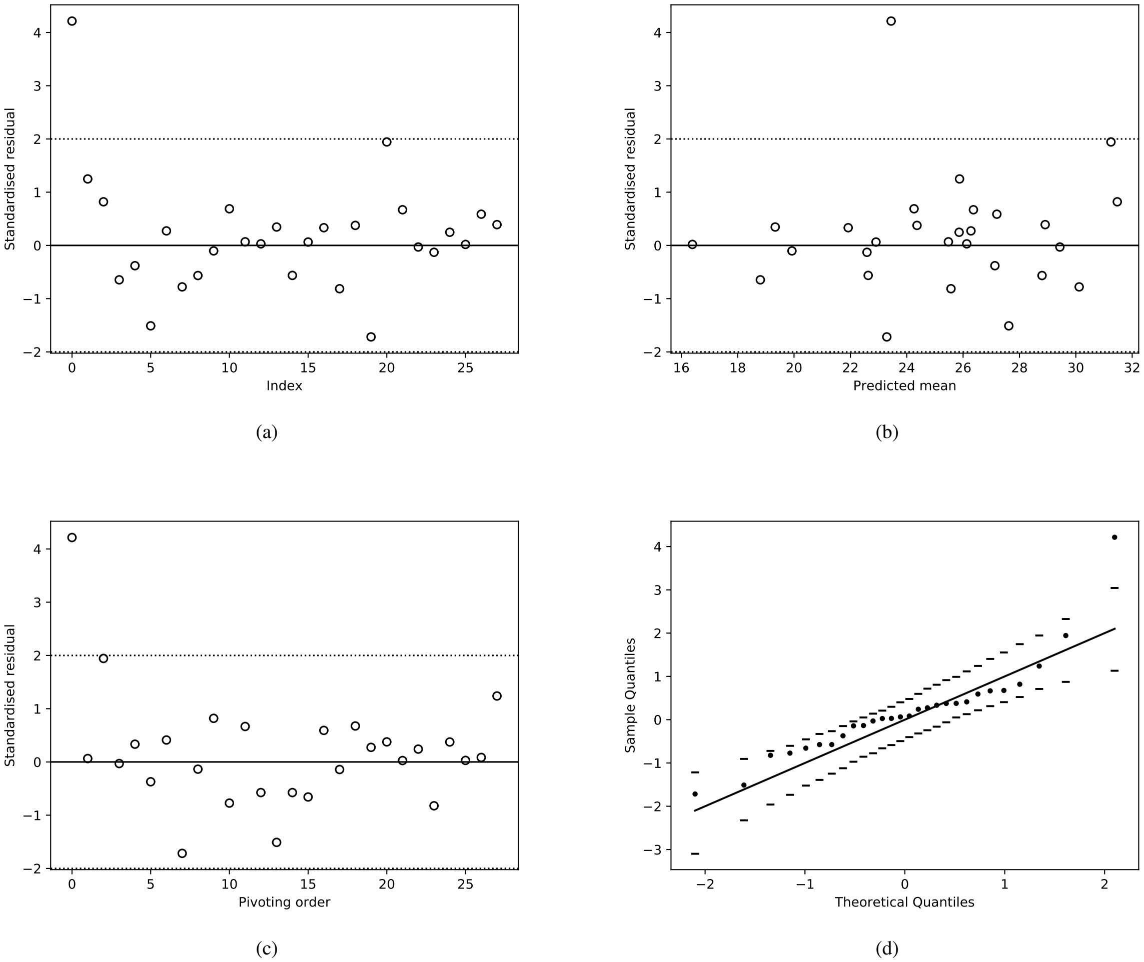


FIGURE 7 Plots of observations, MetUM model output used in the DF model and the test point with the largest residual. (a) Observations and the test point with the largest residual. (b) MetUM model output and the test point with the largest residual.

Figure 8 presents some graphical diagnoses for the out-of-sample prediction where 28 observations were left out for validation. Figure 8 (a)&(b) presents the standardised individual prediction errors ($D_I(\tilde{\boldsymbol{Y}})$) against the validation data index and the predicted mean respectively. Figure 8 (c) presents the pivoted Cholesky errors ($D_{PC}(\tilde{\boldsymbol{Y}})$) against the pivoting order. No obvious pattern can be seen from these graphical diagnoses except for one largest error, which justifies the GP assumption of the DF model.

Figure 8 (d) presents the QQ-plot of the pivoted Cholesky errors. The solid line is the 45 degrees line where the sample quantiles exactly match the theoretical ones. Let $p(\boldsymbol{\theta})$ denote the approximate multivariate normal distribution of the parameters where the mean is the maximum likelihood estimate (MLE) and the covariance is the negated inverse Hessian matrix evaluated at the MLE. The solid points are the sample quantiles of the pivoted Cholesky errors (computed at the mean of $p(\boldsymbol{\theta})$) against the theoretical quantiles. The slope of the solid points is slightly lower than 1, indicating a possible slight overestimation of the predictive variability. In order to take into account the parameter uncertainty and sample variability, we randomly sample 1000 samples of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$, and for each of the 1000 samples, we compute the pivoted Cholesky errors. Thus we compute the 95% confidence interval for the QQ-plot which is designated as error bars in the plot. As can be seen, only one out-of-sample validation point is outside the 95% confidence interval, which doesn't give reason to doubt the GP assumption of the model. The GP model assumption is further justified by the diagnostics for the in-sample predictions; see Supplementary Material.



(a)

(b)

(c)

(d)

FIGURE 8 Graphical diagnostics for the out-of-sample prediction for the observed data. (a) Standardised individual prediction errors $D_I(\tilde{Y})$ versus the validation data index. (b) Standardised individual prediction errors $D_I(\tilde{Y})$ versus the predictions. (c) Pivoted Cholesky errors $D_{PC}(\tilde{Y})$ versus the pivoting order. (d) QQ-plot of the pivoted Cholesky errors.

6 | SUMMARY

In this paper we have presented a generic DF framework for modelling environmental hazard events. The motivation is that for the hazard events, such as windstorms, hail storms and tornadoes, it is usually difficult to obtain observational data at high geographical resolution and hence utilising other sources of data (physical model output in particular) plays an important role in the assessment of those events. This DF framework combines observations from monitoring stations at point level and gridded numerical simulation model output for improved predictions at high spatial resolution. It extends the BM method by utilising a GP to flexibly model the discrepancy structure between the true footprint and physical model output. The effectiveness of the DF framework has been demonstrated in the simulation study as well as in the application to European windstorm data by providing reliable out-of-sample estimates. It has proved to be able to significantly improve the prediction accuracy by lowering

RMSE and AWPI, while maintaining a competitive coverage probability when compared to a Kriging model based solely on observational data in the simulation study. When applied to the Imogen windstorm data, DF provides reliable estimates at locations where there are no observational data with a prediction accuracy of 96%. The assumption of the DF framework has also been validated through several model checking diagnostics. In addition, the uncertainty of the parameters are fully taken account of by maximising the joint likelihood and estimating the parameters simultaneously. The maximum likelihood estimation is objective, robust and relatively fast, making the DF approach applicable to large scale hazard footprints.

Despite providing a generic effective statistical framework for estimating environmental hazard events by combining multiple data sources, there are a number of interesting directions for future work. First, we use a Gaussian covariance function, which implies that GP sample paths are infinitely differentiable. Such an assumption might be inappropriate for some processes, such as high-resolution rainfall measurements derived from radar data, which tend to have much greater short scale variability and hence the power exponential or Matérn covariance function could be considered for such processes. Second, the isotropic Gaussian covariance function assumes the same correlation scale along the longitude and latitude. However, this might not be the case for some processes where the distance along the longitude and latitude accounts for different magnitude of dependence between $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s}')$. Therefore, exploring an anisotropic covariance function where the length-scale for the longitude and latitude are different would be another direction of future work. The model of (6) implies a Gaussian marginal distribution of wind gust speeds. While our study supports this being a reasonable model for gust speeds within a single windstorm event, a more flexible marginal model such as the transformed GP where a Box-Cox transformation of the response is performed (Diggle & Ribeiro 2007) may be needed for other processes. In addition, we have sub-sampled up to 500 model output out of the total available 27,185 model output for computational feasibility. However, our DF framework would produce more reliable and accurate estimates if we could use all the model output, hence a natural extension of the framework is to incorporate methods such as Banerjee (2017); Wang et al. (2019) for efficient inference in high-dimensional cases. Finally, as the DF framework is in principle able to generalise to other data sources we could extend the model framework to incorporate more than two data sources, such as including the data collected from the amateur meteorologists.

7 | ACKNOWLEDGEMENTS

7.1 | Funding

We gratefully acknowledge support from the National Environment Research Council (NERC) UK 2017-2021 fund (reference number NE/P017436/1).

7.2 | Availability of data and material

The data that support the findings of this study are openly available at `https://wisc.climate.copernicus.eu/wisc/#/help/products#stormtrack_download`, WISC (2019).

References

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, *12*(2), 583–614.

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

Boaz, R., Lawson, A., & Pearce, J. (2019). Multivariate air pollution prediction modeling with partial missingness. *Environmetrics*, *30*(7), e2592.

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*(11), 114007.

CCRA. (2017). *UK Climate Change Risk Assessment.* Available from: `https://www.gov.uk/government/publications/uk-climate-change-risk-assessment-2017`[last accessed April 2019].

Chang, H. (2016). Data assimilation for environmental pollution fields. *Handbook of Spatial Epidemiology*, 289–302.

Craig, P. S., Goldstein, M., Rougier, J. C., & Seheult, A. H. (2001). Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, *96*(454), 717–729.

Cressie, N. (1993). *Statistics for spatial data, 2nd ed.* New York: John Wiley & Sons.

Davies, T., Cullen, M. J., Malcolm, A. J., Mawson, M., Staniforth, A., White, A., & Wood, N. (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., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., … Vitart, F. (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., & Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer: New York.

Foley, K. M., & Fuentes, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane surface wind prediction. *Journal of agricultural, biological, and environmental statistics*, *13*(1), 37–59.

Forlani, C., Bhatt, S., Cameletti, M., Krainski, E., & Blangiardo, M. (2020). A joint Bayesian space–time model to integrate spatially misaligned air pollution data in R-INLA. *Environmetrics*, e2644.

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

Gilani, O., Berrocal, V., & Batterman, S. A. (2019). Nonstationary spatiotemporal Bayesian data fusion for pollutants in the near-road environment. *Environmetrics*, *30*(7), e2581.

Haylock, M. (2011). European extra-tropical storm damage risk from a multi-model ensemble of dynamically-downscaled global climate models. *Natural Hazards and Earth System Sciences*, *11*(10), 2847–2857.

Hill, L., Sparks, R., & Rougier, J. (2013). Risk assessment and uncertainty in natural hazards. *Risk and Uncertainty Assessment for Natural Hazards, edited by: Rougier, JC, Sparks, RS J., and Hill, LJ*, 1–18.

Keller, J. P., & Peng, R. D. (2019). Error in estimating area-level air pollution exposures for epidemiology. *Environmetrics*, *30*(8), e2573.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464.

Klawa, M., & Ulbrich, U. (2003). A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Natural Hazards and Earth System Science*, *3*(6), 725–732.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498.

Ma, P., & Kang, E. L. (2020). Spatio-temporal data fusion for massive sea surface temperature data from MODIS and AMSR-E instruments. *Environmetrics*, *31*(2), e2594.

McMillan, N. J., Holland, D. M., Morara, M., & Feng, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics: The official journal of the International Environmetrics Society*, *21*(1), 48–65.

Paciorek, C. J. (2012). Combining spatial information sources while accounting for systematic errors in proxies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(3), 429–451.

Poole, D., & Raftery, A. E. (2000). Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, *95*(452), 1244–1255.

Roberts, J., Champion, A., Dawkins, L., Hodges, K., Shaffrey, L., Stephenson, D., … Youngman, B. (2014). The XWS open access catalogue of extreme European windstorms from 1979 to 2012. *Natural Hazards and Earth System Sciences*, *14*, 2487–2501.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., … Prüss-Ustün, A. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(1), 231–253.

Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., & Wilson, A. G. (2019). Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems* (pp. 14648–14659).

Wilkie, C. J., Miller, C. A., Scott, E. M., O'Donnell, R. A., Hunter, P. D., Spyrakos, E., & Tyler, A. N. (2019). Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, *30*(3), e2549.

WISC. (2019). *WISC products.* Available from: `https://wisc.climate.copernicus.eu/wisc/#/help/products#stormtrack_download`[last accessed April 2019].

Xiong, X., Šmídl, V., & Filippone, M. (2017). Adaptive multiple importance sampling for Gaussian processes. *Journal of Statistical Computation and Simulation*, *87*(8), 1644–1665.

Zappa, G., Shaffrey, L. C., Hodges, K. I., Sansom, P. G., & Stephenson, D. B. (2013). A multimodel assessment of future projections of North Atlantic and European extratropical cyclones in the CMIP5 climate models. *Journal of Climate*, *26*(16), 5846–5862.

Supplementary Material

We also perform an in-sample prediction at the locations where we have the 100 observations to fit the DF model. The predictive distribution $p(\boldsymbol{Y} \mid \boldsymbol{X})$ in this case is given by:

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{X}}, \Sigma_{\boldsymbol{Y}|\boldsymbol{X}})$$

where

$$\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{X}} = \boldsymbol{\mu} + \Sigma_{\boldsymbol{YX}}\Sigma_{\boldsymbol{X}}^{-1}(\boldsymbol{X} - \boldsymbol{\alpha})$$

$$\Sigma_{\boldsymbol{Y}|\boldsymbol{X}} = \Sigma_{\boldsymbol{Y}} - \Sigma_{\boldsymbol{YX}}\Sigma_{\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{XY}}$$

Figure 9 presents the graphical diagnostics for the in-sample prediction as shown in Figure 8 . Out of the 100 in-sample test points, we only observe three abnormal ones in these plots which justifies the model assumption that the conditional distribution of $\boldsymbol{Y}$ (observations) given $\boldsymbol{X}$ (model output) is also a GP.

Similarly, we conduct an in-sample prediction test at the locations where we have the 500 MetUM model output. The predictive distribution $p(\boldsymbol{X} \mid \boldsymbol{Y})$ in this case takes the form:

$$p(\boldsymbol{X} \mid \boldsymbol{Y}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Y}}, \Sigma_{\boldsymbol{X}|\boldsymbol{Y}})$$

where

$$\boldsymbol{\mu}_{\boldsymbol{X}|\boldsymbol{Y}} = \boldsymbol{\alpha} + \Sigma_{\boldsymbol{XY}}\Sigma_{Y}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

$$\Sigma_{\boldsymbol{X}|\boldsymbol{Y}} = \Sigma_{X} - \Sigma_{\boldsymbol{XY}}\Sigma_{Y}^{-1}\Sigma_{\boldsymbol{YX}}$$

The graphical diagnoses are given in Figure 10 . In this case, we observe a number of abnormal points in these plots. However, in Figure 10 (d) which is the QQ-plot of the pivoted Cholesky errors, only four points are outside the 95% confidence interval in the lower tail which relates to small values of MetUM model output. We do not investigate further for this case, as we are in general more interested in estimates of environmental hazard events of high intensities. However, further investigations could be made in the future by changing the mean function (15) or use a different covariance structure for $c_Z(\boldsymbol{s}, \boldsymbol{s}')$ and $c_\delta(\boldsymbol{s}, \boldsymbol{s}')$.

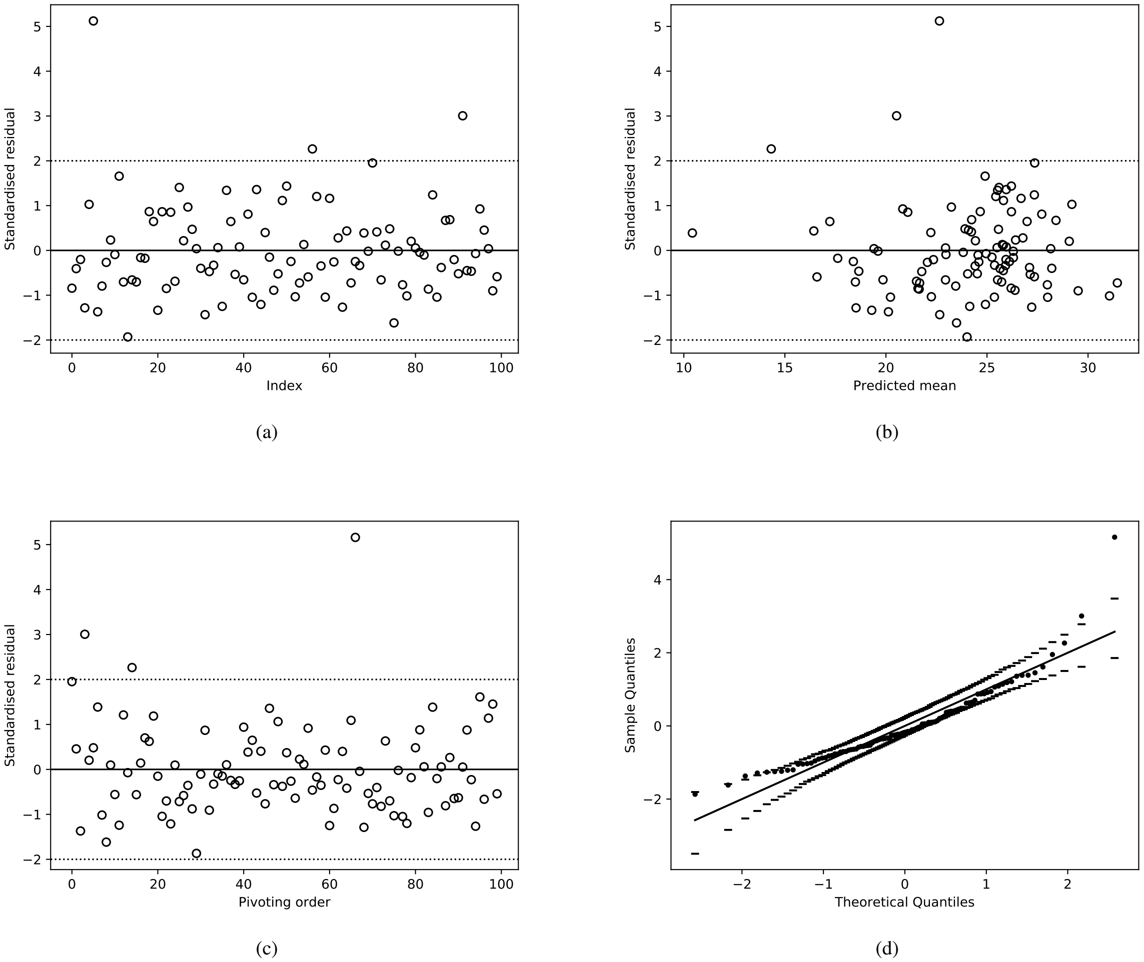
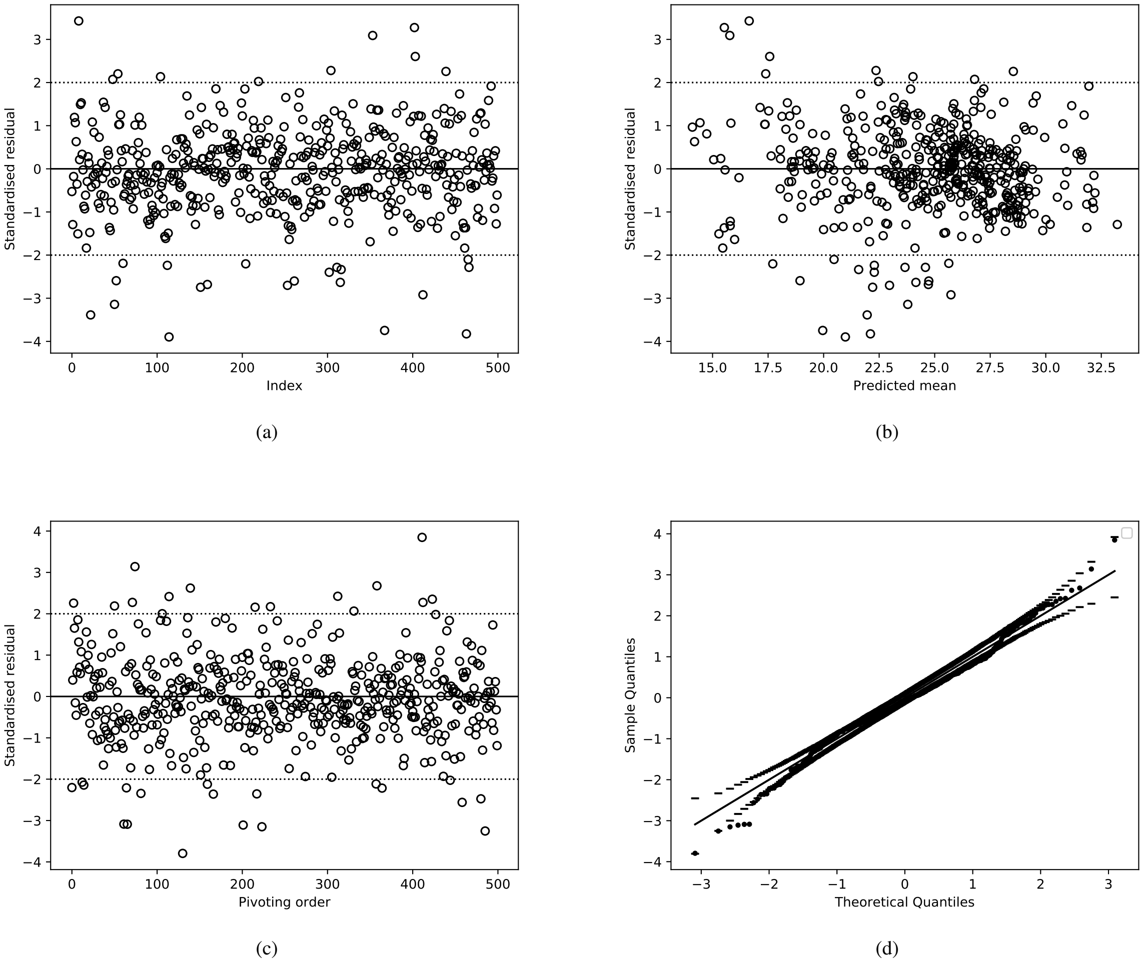


FIGURE 9 Graphical diagnostics for the in-sample prediction for the observed data. (a) Standardised individual prediction errors versus the validation data index. (b) Standardised individual prediction errors versus the predictions. (c) Pivoted Cholesky errors versus the pivoting order (d) QQ-plot of the pivoted Cholesky errors.



(a)

(b)

(c)

(d)

FIGURE 10 Graphical diagnostics for in-sample prediction for the MetUM model output. (a) Standardised individual prediction errors versus the validation data index. (b) Standardised individual prediction errors versus the predictions. (c) Pivoted Cholesky errors versus the pivoting order. (d) QQ-plot of the pivoted Cholesky errors.

□