



Data fusion with Gaussian processes for estimation of environmental hazard events

Journal:	<i>Environmetrics</i>
Manuscript ID:	env-19-0118
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	03-Sep-2019
Complete List of Authors:	Xiong, Xiaoyu; University of Exeter College of Engineering Mathematics and Physical Sciences Youngman, Benjamin; University of Exeter College of Engineering Mathematics and Physical Sciences Economou, Theodoros; University of Exeter College of Engineering Mathematics and Physical Sciences
Keywords:	Data integration, data simulation, Gaussian processes, spatial interpolation, model validation, European windstorm, natural hazard
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
GPs_for_DF.tex wileyNJD-APA.bib WileyNJD-APA.bst WileyNJD-v2.cls NJDaapcite.sty	

SCHOLARONE™
Manuscripts

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 hazard 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 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, data simulation, Gaussian processes, spatial interpolation, model validation, European windstorm, natural hazards

1 | INTRODUCTION

Environmental hazards, such as air pollution, extreme wind speeds and rainfall/flooding caused by extra-tropical cyclones or windstorms, can have devastating socioeconomic impacts such as damage to infrastructures 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). The 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 outputs for windstorm Imogen over France, the details of which are introduced 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 model outputs tend to have complete spatial coverage but can only represent true values at the model's predetermined resolution and not at smaller scales.

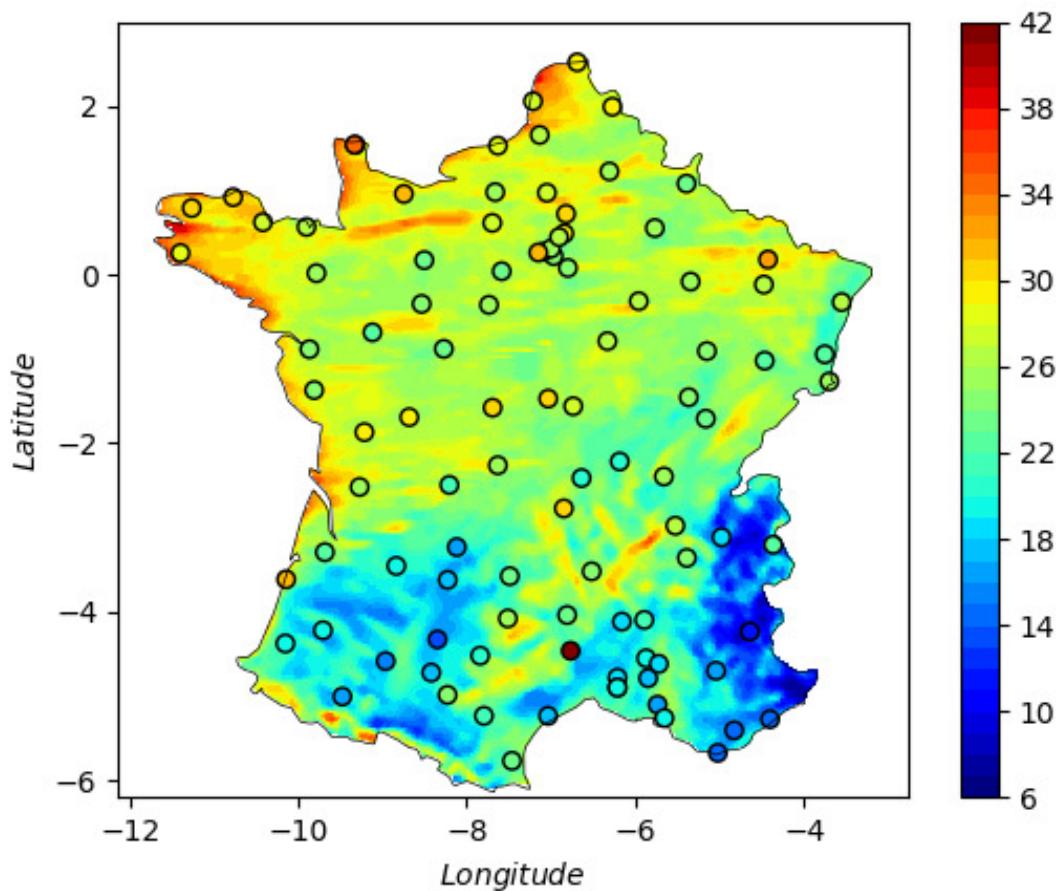


FIGURE 1 Windstorm Imogen footprint over France, according to storm footprints data from Windstorm Information Service (WISC 2019). The circles denote station observations, and are plotted on top of the gridded numerical climate model outputs.

Knowledge of small scale detail in environmental hazard footprints is important because of the large spatial heterogeneity in vulnerability and exposure. In the United Kingdom, 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 and such an endeavour entails 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 all available data sources.

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 while 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. 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. To describe this framework in the context of our application to environmental hazard footprints, let s denote a spatial location at which both an observation (from a monitoring station) and prediction from a physical model is potentially available. The framework can then be formulated as

$$Y(s) = Z(s) + \epsilon(s) \quad (1)$$

$$Z(s) = \beta X(s, \vartheta) + \delta(s) \quad (2)$$

where $Y(s)$ denotes the observation, ϑ are the calibration parameters, $Z(s)$ denotes the true process, $\epsilon(s)$ is the observation error, β is an unknown regression (bias) parameter, $X(s, \vartheta)$ denotes the output of a physical model, and $\delta(s)$ represents the

discrepancy between the physical model output and the real process. The physical model output $X(.,.)$ and the discrepancy term $\delta(.)$ are modelled as two independent Gaussian processes (GPs):

$$X(.,.) \sim GP(\mu_X(.,.), c_X(.,.)) \quad (3)$$

$$\delta(.) \sim GP(\mu_\delta(.), c_\delta(.,.)) \quad (4)$$

where $\mu_X(.,.)$ and $\mu_\delta(.)$ are the mean functions and $c_X(.,.)$ and $c_\delta(.,.)$ are the covariance functions, specified by hyper-parameters ψ_X, ψ_δ respectively. Then observations $\mathbf{Y} = [Y(s_1), \dots, Y(s_n)]^T$ and physical model outputs $\mathbf{X} = [X(s_1, \vartheta_1), \dots, X(s_n, \vartheta_n)]^T$ at the set of points comprising known variable inputs and calibration inputs are jointly modelled as a multivariate normal distribution. The hyper-parameters are estimated in a two-stage process: ψ_δ is obtained in the second stage by fixing the estimates of the ψ_X obtained in the first stage, where estimation only uses the model outputs \mathbf{X} whilst ignoring the observational data \mathbf{Y} that also contains information about ψ_X . Consequently, the uncertainty of the model parameters is not fully taken into account. Note that in this framework the change of support problem is not necessarily addressed. Also, the model output $X(s, \vartheta)$ is treated as fixed so it is not obvious how this can be extended to further data sources.

Fuentes and Raftery (2005) have developed a Bayesian melding (BM) approach that 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 true underlying process $Z(s)$ as a GP driving both sources of data, where the observational data $Y(s)$ is expressed as $Z(s)$ plus measurement error $\epsilon(s)$, while the numerical model output $X(s)$ is related to the true process via a systematic additive bias term $\alpha(s)$, a multiplicative bias term $\beta(s)$ and an i.i.d. Gaussian error term $\delta(s)$:

$$Z(s) \sim GP(\mu_Z(s), c_Z(s, s')) \quad (5)$$

$$Y(s) = Z(s) + \epsilon(s) \quad (6)$$

$$X(s) = \alpha(s) + \beta(s)Z(s) + \delta(s) \quad (7)$$

$$\delta(s) \sim \mathcal{N}(0, \sigma^2) \quad (8)$$

where $\mu_Z(s)$ and $c_Z(s, s')$ 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 $X(s)$ over grid cells (sub-regions). 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 $\int_A X(s) ds$. This assumes that the numerical model output $X(A)$ is interpreted as an aggregation of $X(s)$ 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).

An alternative approach to hierarchical formulations, is to consider a regression setting. Shaddick et al. (2018) introduces 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.

Among the aforementioned DF approaches, the BM method 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. 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

The work by Brynjarsdóttir and O'Hagan (2014) gives 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. As such, we relax the assumption that the discrepancy term $\delta(s)$ in (7) is i.i.d. and instead model it as a zero-mean GP:

$$\delta(s) \sim GP(0, c_\delta(s, s')) \quad (9)$$

Recall that this term quantifies the error between the true process and the model output after allowing for structured bias through $\alpha(s)$ and $\beta(s)$. 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 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

As such, the proposed DF modelling framework for environmental hazard footprints is:

$$Z(s) \sim GP(\mu(s), c_Z(s, s')) \quad \text{True process} \quad (10)$$

$$Y(s) = Z(s) + \epsilon(s) \quad \text{Data (observations)} \quad (11)$$

$$\epsilon(s) \sim \mathcal{N}(0, \sigma_Y^2) \quad \text{Measurement error} \quad (12)$$

$$X(s) = \alpha(s) + \beta(s)Z(s) + \delta(s) \quad \text{Numerical model output} \quad (13)$$

$$\delta(s) \sim GP(0, c_\delta(s, s')) \quad \text{Discrepancy term} \quad (14)$$

The bias terms ($\alpha(s)$, $\beta(s)$) 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), $\alpha(s)$ is defined as a polynomial function of spatial coordinates and $\beta(s)$ is defined as a global scaling factor (constant β) to capture multiplicative bias across all wind speed values. The covariance functions $c_Z(s, s')$ and $c_\delta(s, s')$ specified by the hyper-parameters ψ_Z and ψ_δ respectively are key to the performance of the spatial interpolation. If $c_Z(s, s')$ and $c_\delta(s, s')$ depend only on $\|s - s'\|$ then both processes are stationary such that 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 $Y(s)$ are point locations while data from the numerical model output is 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 $X(s)$ over A_i and takes the following form:

$$X(A_i) = \int_{A_i} X(s) ds = \int_{A_i} \alpha(s) ds + \int_{A_i} \beta(s)Z(s)ds + \int_{A_i} \delta(s)ds \quad (15)$$

As such, $X(s)$, for any $s \in A_i$, in (13) is unobserved and the only data on it are aggregated in the form of $X(A_i)$.

Let θ represent all the parameters involved in (10)–(14), namely ψ_Z , ψ_δ , σ_Y , and any coefficients in $\alpha(s)$ and $\beta(s)$. Integrating out the true process $Z(s)$ implies that observations $\mathbf{Y} = [Y(s_1), \dots, Y(s_n)]^T$ and model output $\mathbf{X} = [X(A_1), \dots, X(A_m)]^T$ jointly follow a multivariate Gaussian distribution:

$$p(\mathbf{Y}, \mathbf{X} | \theta) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \Sigma_Y & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_X \end{bmatrix}\right) = \mathcal{N}(\boldsymbol{\mu}^+, \mathbf{C}) \quad (16)$$

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

$$\boldsymbol{\mu} = (\mu(s_1), \dots, \mu(s_n))^T \quad \text{and} \quad \boldsymbol{\alpha} = \left(\int_{A_1} \alpha(s) ds, \dots, \int_{A_m} \alpha(s) ds \right)^T$$

In (16), Σ_Y is the covariance of \mathbf{Y} defined by

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = c_Z(\mathbf{s}, \mathbf{s}') + \sigma_Y^2 \mathbf{1}(\mathbf{s} = \mathbf{s}'). \quad (17)$$

Matrix Σ_{YX} is the cross-covariance between \mathbf{Y} and \mathbf{X} , with its entries specified by:

$$\text{cov}(Y(\mathbf{s}_i), X(A_j)) = \frac{1}{|A_j|} \int_{A_j} \beta(\mathbf{v}) c_Z(\mathbf{s}_i, \mathbf{v}) d\mathbf{v} \quad (18)$$

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

$$\text{cov}(X(A_i), X(A_j)) = \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} \beta(\mathbf{u}) \beta(\mathbf{v}) c_Z(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} + \frac{1}{|A_i||A_j|} \int_{A_i} \int_{A_j} c_\delta(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}. \quad (19)$$

In practice, we uniformly sample L_i and L_j locations in A_i and A_j respectively, so that (18) is approximated by:

$$\widehat{\text{cov}}(Y(\mathbf{s}_i), X(A_j)) = \frac{1}{L_j} \sum_{\gamma} \beta(\mathbf{v}_{j\gamma}) c_Z(\mathbf{s}_i, \mathbf{v}_{j\gamma}) \quad (20)$$

whereas (19) is approximated by:

$$\widehat{\text{cov}}(X(A_i), X(A_j)) = \frac{1}{L_i L_j} \sum_{\gamma} \sum_{\gamma'} (\beta(\mathbf{u}_{i\gamma}) \beta(\mathbf{v}_{j\gamma'}) c_Z(\mathbf{u}_{i\gamma}, \mathbf{v}_{j\gamma'}) + c_\delta(\mathbf{u}_{i\gamma}, \mathbf{v}_{j\gamma'})). \quad (21)$$

It should be noted that, particularly in situations where the number of numerical model outputs \mathbf{X} is much larger than that of the observations \mathbf{Y} , there might be non-identifiability in uniquely characterising the two GPs - $Z(\mathbf{s})$ and $\delta(\mathbf{s})$. In practice, this can be avoided by constraining the smoothness properties of each GP, i.e., constraining the parameters in $c_Z(\mathbf{s}, \mathbf{s}')$ and $c_\delta(\mathbf{s}, \mathbf{s}')$. For environmental hazard footprints, the true process $Z(\mathbf{s})$ is assumed smoother than the discrepancy term $\delta(\mathbf{s})$ to allow for the fact that the latter is effectively an error term while the former is a physical process.

3.2 | Inference

Let \mathbf{D} denote the combined data points of the observations \mathbf{Y} and the gridded model outputs \mathbf{X} , i.e., $\mathbf{D} = [Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n), X(A_1), \dots, X(A_m)]^T$. Then, (16) implies that the likelihood for the data is based on the multivariate Gaussian distribution, so that the log-likelihood is given by:

$$\log[p(\mathbf{D} | \boldsymbol{\theta})] = -\frac{1}{2} \log(|\mathbf{C}|) - \frac{1}{2} (\mathbf{D} - \boldsymbol{\mu}^+)^T \mathbf{C}^{-1} (\mathbf{D} - \boldsymbol{\mu}^+) + \text{const.} \quad (22)$$

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 $Z(\mathbf{s})$, given all the data. Suppose then we would like the true footprint at some in-sample/out-of-sample location \mathbf{s} . The covariance

matrix between $Z(s)$ and \mathbf{D} is given by:

$$\mathbf{k}^*(s) = [c_Z(s, s_1), \dots, c_Z(s, s_n), \text{cov}(Z(s), X(A_1)), \dots, \text{cov}(Z(s), X(A_m))]^T \quad (23)$$

where n is the number of observations, m is the number of grid cells ($A_i, i = 1, \dots, m$) relating to model outputs and $\text{cov}(\cdot, \cdot)$ is defined by (18). Letting $k(s) = c_Z(s, s)$, the Gaussian identity implies that the joint distribution of the combined data points \mathbf{D} and the true footprint at s is:

$$\begin{bmatrix} \mathbf{D} \\ Z(s) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}^+ \\ \mu(s) \end{bmatrix}, \begin{bmatrix} \mathbf{C} & \mathbf{k}^*(s) \\ \mathbf{k}^*(s)^T & k(s) \end{bmatrix}\right) \quad (24)$$

where

$$Z(s) | \mathbf{D}, \theta \sim \mathcal{N}(\hat{\mu}(s), \sigma^2(s)) \quad (25)$$

$$\hat{\mu}(s) = \mu(s) + \mathbf{k}^*(s)^T \mathbf{C}^{-1} (\mathbf{D} - \boldsymbol{\mu}^+) \quad (26)$$

$$\sigma^2(s) = k(s) - \mathbf{k}^*(s)^T \mathbf{C}^{-1} \mathbf{k}^*(s) \quad (27)$$

Thus, predictions are achieved by plugging in the estimate of θ into (25).

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 $\mathbf{S}^* = (s_1, s_2, \dots, s_v)$ denote a set of locations where observation data are not available, $\boldsymbol{\mu}^* = (\mu(s_1), \mu(s_2), \dots, \mu(s_v))$ denote the mean values of the true process at these locations, $\mathbf{Y}^* = (Y(s_1), Y(s_2), \dots, Y(s_v))$ denote the corresponding predicted values (truth plus the measurement error). The predictive distribution of \mathbf{Y}^* given the data is:

$$\mathbf{Y}^* | \mathbf{D}, \theta \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^*, \mathbf{V}^*) \quad (28)$$

with

$$\begin{aligned} \hat{\boldsymbol{\mu}}^* &= \boldsymbol{\mu}^* + \mathbf{K}^{*T} \mathbf{C}^{-1} (\mathbf{D} - \boldsymbol{\mu}^+) \\ \mathbf{V}^* &= \boldsymbol{\Sigma}_{\mathbf{Y}^*} - \mathbf{K}^{*T} \mathbf{C}^{-1} \mathbf{K}^* \end{aligned} \quad (29)$$

where each entry of $\boldsymbol{\Sigma}_{\mathbf{Y}^*}$ is specified by $c_Z(s_i, s_j) + \sigma_Y^2 \mathbf{1}(s_i = s_j)$ and each column of \mathbf{K}^* is $\mathbf{k}^*(s_i)$ defined in (23) for $i = 1, \dots, v$.

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

prediction errors for the validation data $\tilde{\mathbf{Y}}$, and is defined as:

$$D_I(\tilde{\mathbf{Y}}) = \frac{\tilde{\mathbf{Y}} - \hat{\boldsymbol{\mu}}^*}{\sqrt{\text{Var}(\mathbf{Y}^*)}} \quad (30)$$

A plot of $D_I(\tilde{\mathbf{Y}})$ 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 $D_I(\tilde{\mathbf{Y}})$ 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 that stationarity, when assumed, is inappropriate. In particular, the plot of $D_I(\tilde{\mathbf{Y}})$ 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, $D_I(\tilde{\mathbf{Y}})$ are correlated, the next diagnostic produces uncorrelated prediction errors defined by:

$$D_G(\tilde{\mathbf{Y}}) = \mathbf{G}^{-1}(\tilde{\mathbf{Y}} - \hat{\boldsymbol{\mu}}^*) \quad (31)$$

where \mathbf{G} is the standard deviation matrix which satisfies $\mathbf{V}^* = \mathbf{G}\mathbf{G}^T$. Following Bastos and O'Hagan (2009), we use the pivoted Cholesky decomposition to obtain \mathbf{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 \mathbf{P} and a lower triangular matrix \mathbf{L} such that $\mathbf{P}^T \mathbf{V}^* \mathbf{P} = \mathbf{L} \mathbf{L}^T$, leading to $\mathbf{G} = \mathbf{P} \mathbf{L}$. We call the corresponding errors pivoted Cholesky errors and denote them by $D_{PC}(\tilde{\mathbf{Y}})$. The main advantage of $D_{PC}(\tilde{\mathbf{Y}})$ is that, in the case where the covariance function is the Gaussian covariance function, i.e., $c(\mathbf{s}, \mathbf{s}') = \sigma \exp \{-(\|\mathbf{s} - \mathbf{s}'\|/\phi)^2\}$, it could diagnose both the estimation of the marginal variance σ 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 , 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.

4 | SIMULATION STUDY

4.1 | Algorithm for data simulation

In both the simulation and application study in this work, the covariance functions $c_Z(\mathbf{s}, \mathbf{s}')$ and $c_\delta(\mathbf{s}, \mathbf{s}')$ are both Gaussian covariance functions defined by $\sigma_Z \exp \{-(\|\mathbf{s} - \mathbf{s}'\|/\phi_Z)^2\}$ and $\sigma_\delta \exp \{-(\|\mathbf{s} - \mathbf{s}'\|/\phi_\delta)^2\}$ respectively. As such, we are assuming that both $Z(\mathbf{s})$ and $\delta(\mathbf{s})$ are isotropic and relatively smooth.

TABLE 1 Algorithm for simulating data.

1. Generate $Z(s_i)(i = 1, \dots, 10^6)$ from $Z(s) \sim GP(0, c_Z(s, s'))$ where s_i are ordered as a 1000×1000 grid and are within the area of France (denoted by S), i.e., $s_{lon} \in (-12.0, 3.0)$, $s_{lat} \in (-6.3, 3.0)$.
2. Simulate observations $Y(s_i)(i = 1, \dots, n_Y)$ where n_Y is set to 200 in this case. First randomly sample $Z(s_i)$ of size n_Y and transform their probability integral with a gamma distribution $Gamma(5.40, 0.21)$. The transformed ones are denoted by $\tilde{Z}(s_i)$. Then $Y(s_i)$ is simulated according to the equation $Y(s_i) = \tilde{Z}(s_i) + \epsilon(s_i)$, $\epsilon(s) \sim \mathcal{N}(0, 0.25)$.
3. Generate the model outputs $X(A_i)(i = 1, \dots, n_X)$ where n_X is set to 625 in this case, A_i are ordered as a 25×25 block and the average of $Z(s)$ 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 v . In this simulation, v is set to 100.

For $Z(s)$, σ_Z represents the marginal variance of the unobserved true process and the correlation length-scale ϕ_Z accounts for the magnitude of dependence between $Z(s)$ and $Z(s')$, with a smaller value of ϕ_Z indicating stronger dependence. To generate $Z(s)$, we set σ_Z to 1 and ϕ_Z to 0.2. A length-scale of 0.2 specifies a wiggly surface and it is chosen to reflect the large variability of that real windstorm footprint data tends to exhibit.

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

$$\alpha(s) = \alpha_0 + \alpha_1 s_{lon} + \alpha_2 s_{lat} + \alpha_3 s_{lon}s_{lat} + \alpha_4 s_{lon}^2 + \alpha_5 s_{lat}^2. \quad (32)$$

To also reflect the fact that wind speed values will not necessarily follow a normal distribution, a gamma probability integral transformation is performed on the generated $Z(s)$ to test the model's ability to capture non-Gaussian data. The parameters of the gamma distribution are estimated using wind speed observations relating to storm Imogen. The algorithm for simulating the true windstorm process $Z(s)$, the observations $Y(s)$ and the areal model outputs $X(A_i)$ (A_i denotes the i th area) is given in Table 1 .

Figure 2 (a) presents an example of the simulated true process $Z(s)$ of a windstorm event. Figure 2 (b) shows the associated $X(A_i)$ which represents the areal climate model output, while Figure 2 (d) shows a plot of 200 $Z(s)$ values against corresponding $Y(s)$ values as well as 300 $Z(s)$ 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 $Z(s)|Y(s), X(A_i)$ 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.

4.2 | Results

We perform 100 such simulations (see the algorithm presented in Table 1) to represent 100 different windstorm events. In order to test the impact of the number of model outputs on the DF model, for each simulation, we fix the number of observations n_Y

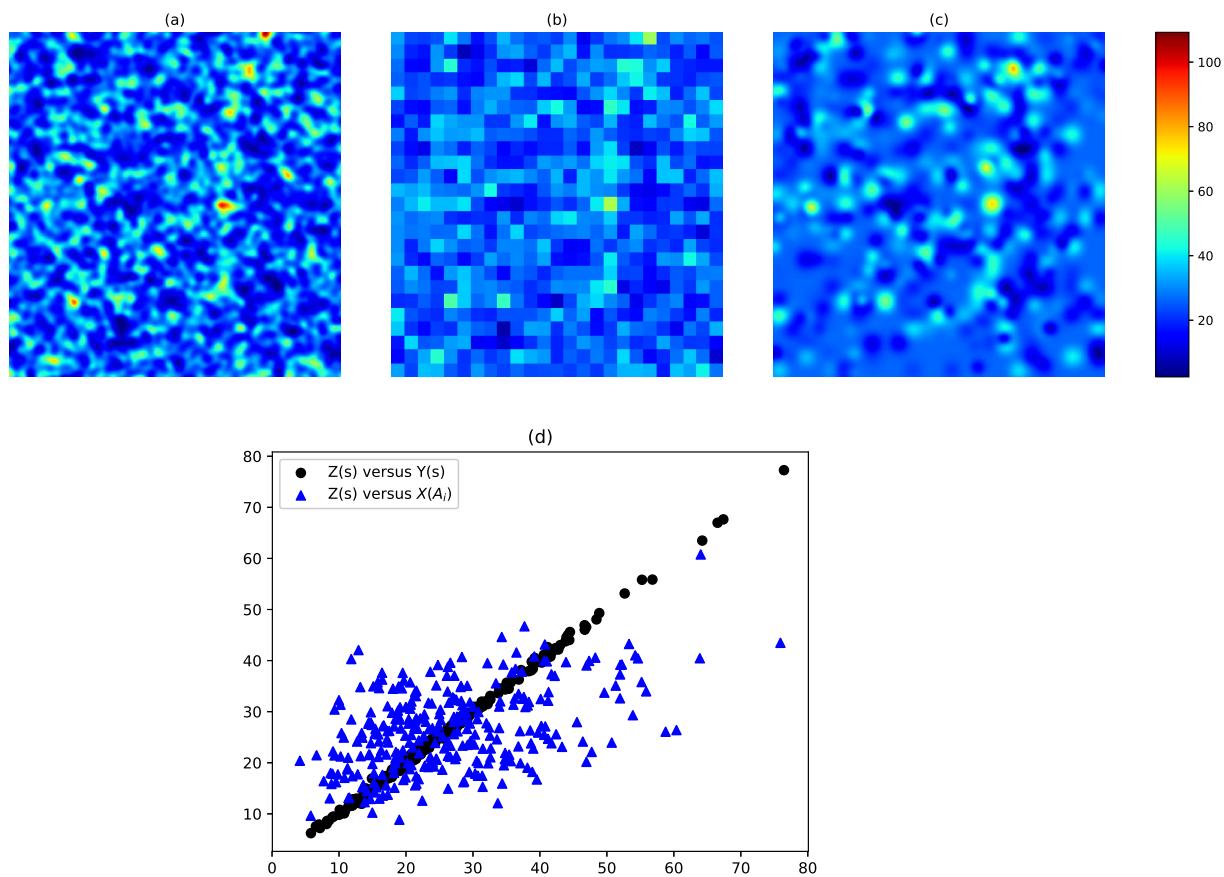


FIGURE 2 Illustration of simulated $Z(s)$, $Y(s)$ and $X(A_i)$ and predicted $Z(s)$. (a) Simulated $Z(s)$ over a 1000×1000 grid. (b) Simulated $X(A_i)$ over a 25×25 block. (c) Predicted $Z(s)$ over a 1000×1000 grid. (d) Simulated $Z(s)$ versus $Y(s)$ of size 200, $X(A_i)$ of size 300.

to be 200, start by randomly sampling 50 areal model outputs out of the total 625 ones, then increasingly add 50 different areal model outputs until we reach the maximum number of model outputs which is set to 500 for computational convenience. We denote these different number of areal model outputs by \mathbf{X}_{50} , \mathbf{X}_{100} , ..., \mathbf{X}_{500} . Accordingly for each simulation we have 10 data combinations $(\mathbf{Y}, \mathbf{X}_{50})$, $(\mathbf{Y}, \mathbf{X}_{100})$, ..., $(\mathbf{Y}, \mathbf{X}_{500})$. We then fit the DF model to the 10 data combinations and perform an out-of-sample prediction 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 confidence interval (AWCI) and the coverage probability. We also fit a Kriging model only to the observations \mathbf{Y} , basically defined by (10)–(11). 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, AWCI and coverage probability are averages over the 100 simulations. The results suggest that by combining the model outputs with observations, the DF model achieves a lower RMSE and AWCI while maintaining a similar coverage probability when compared to the Kriging model that only utilises the observational data. In addition, the impact of the number of model outputs is apparent: the higher

the number of model outputs, the lower the RMSE and AWCI. We also perform another 50 simulations where the number of observations and areal model outputs is set to 200 and 300 respectively, and implement an out-of-sample prediction at the 10^6 grid cells. Figure 4 presents the out-of-sample RMSE, AWCI and coverage probability over the 50 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, at the majority of the grid cells, the RMSE, AWCI of the DF model is much lower than those of the Kriging model whereas the coverage probability of the former is similar to that of the latter. 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 a windstorm event's true footprint.

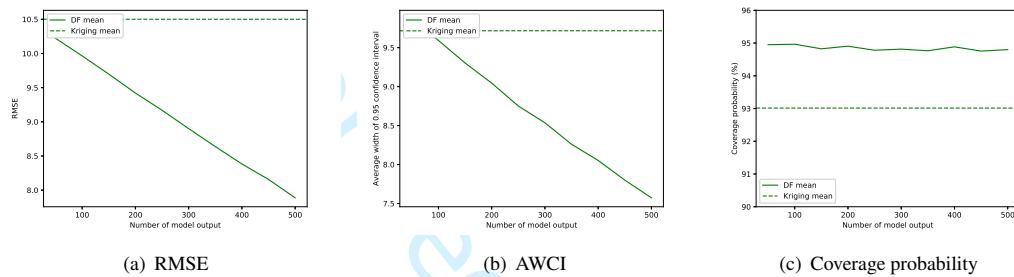


FIGURE 3 The out-of-sample RMSE, AWCI and coverage probability over 100 simulations for Kriging and the DF approach.

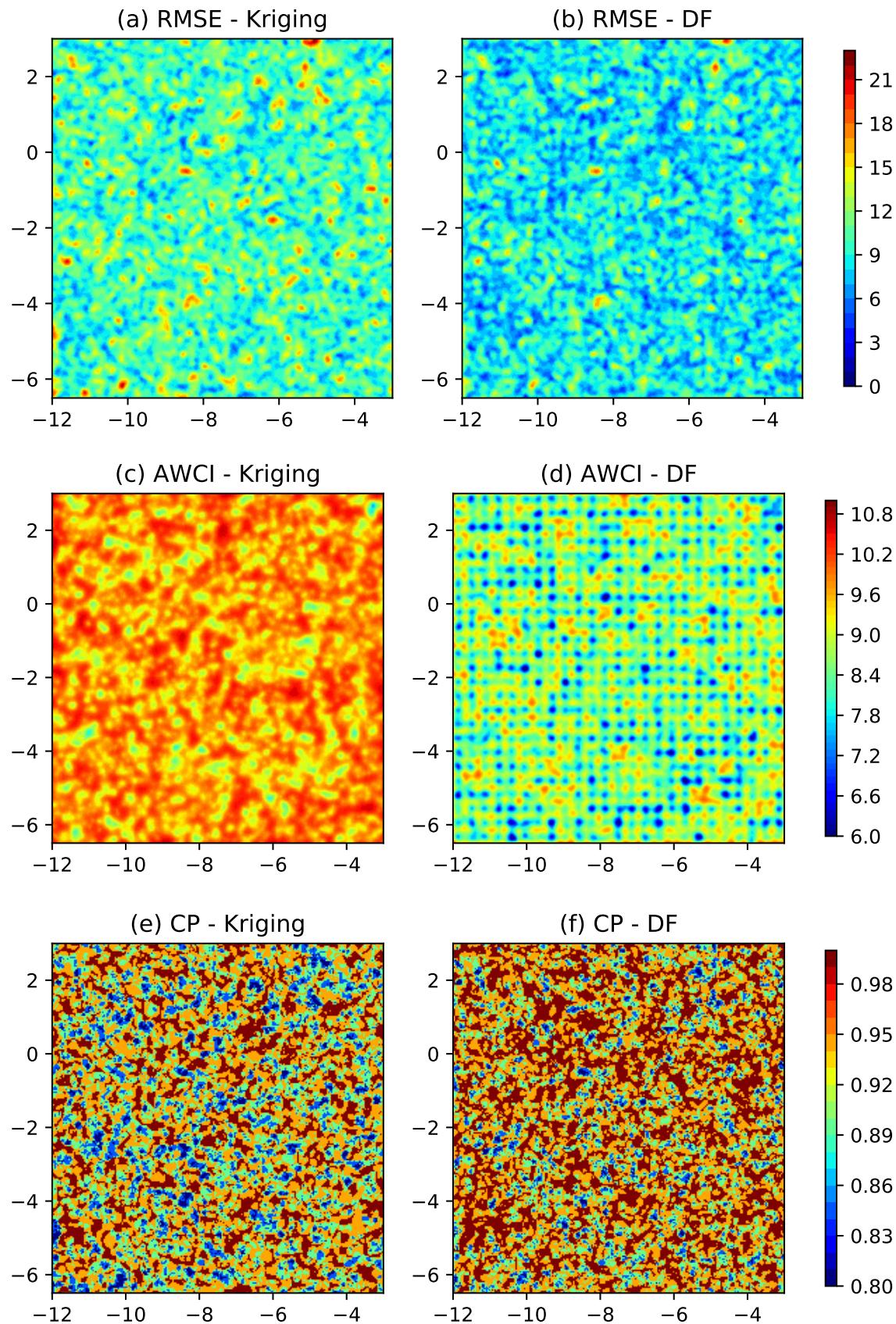


FIGURE 4 Out-of-sample RMSE, AWCI and coverage probability (CP) for Kriging and the DF approach at the 10^6 grid cells .

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 generally defined as a spatial map of the maximum 3-second gust speed over a 72-hour period centered at its occurrence time (Haylock 2011), which can be used to predict property damage and insured loss (Klawa & Ulbrich 2003). We study the European windstorm named Imogen which occurred in 2016 within the spatial domain of France. Figure 5 (a) shows 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), on top of which are the 128 station observations denoted by circles. 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 outputs 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.

5.2 | Results

We first predict the gust speeds using the DF model at the 500 locations where we have MetUM model outputs and plot them against the station observations which is displayed in Figure 5 (b). Compared to Figure 5 (a) which shows the MetUM model outputs 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.

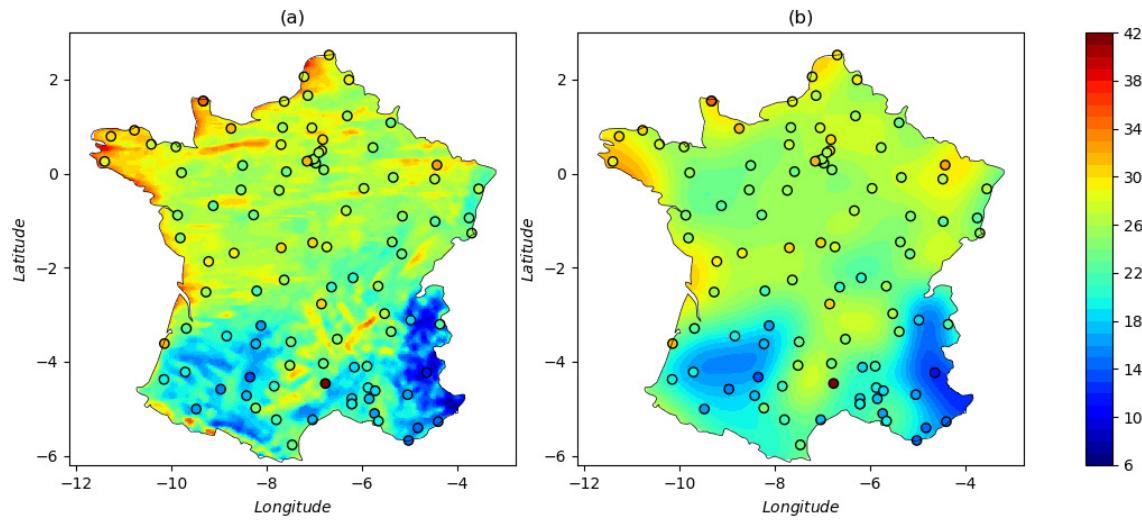


FIGURE 5 Plot of wind gust speed observations at weather stations (represented by coloured circles) and MetUM model outputs. (a) Station observations and MetUM model outputs to fit the DF model. (b) Station observations and predicted model outputs.

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 the two extreme values that are underestimated and one point that are overestimated, the predictions for the rest of 25 points are within the 95% confidence interval. Figure 6 (b) presents a spatial map of the standardised residuals where the three largest ones are marked in red and blue. Figure 7 (a) displays the observations (denoted by circles) and the three test points with the largest residuals (represented by squares). The values of the two red squares (test points) seem to be much higher than the observations in the nearby area, the underestimation of which suggests future work could improve on the model's capability of capturing extreme values. Figure 7 (b) shows the 500 MetUM model outputs (denoted by circles) and the three test points with the largest residuals (represented by squares). The overestimation of one point (denoted by the blue square in Figure 7 (a)&(b)) could be explained by the fact it situates at the boundary of France where much more uncertainty is expected for the model due to lack of data in the area.

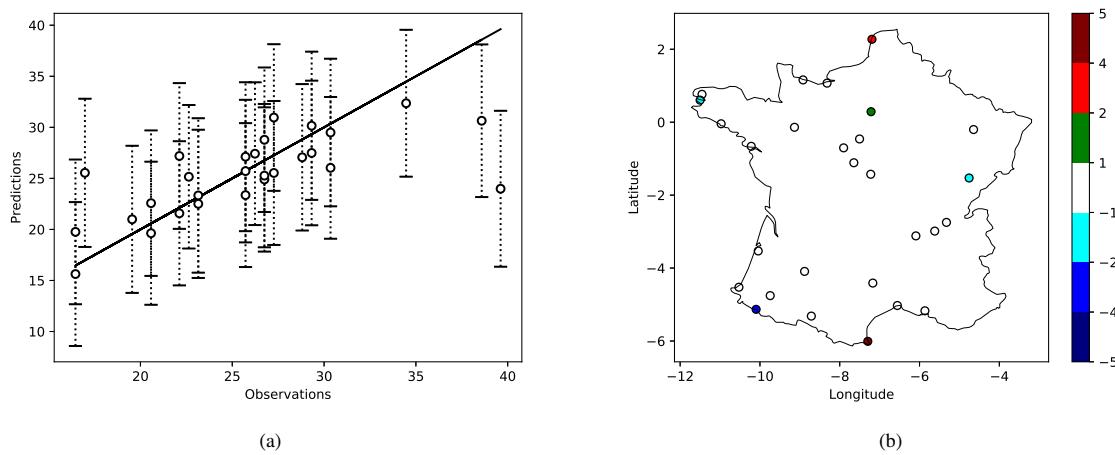


FIGURE 6 Plot of out-of-sample prediction. (a) Out-of-sample predictions versus observations. (b) Standardised residuals plotted on the spatial domain of France.

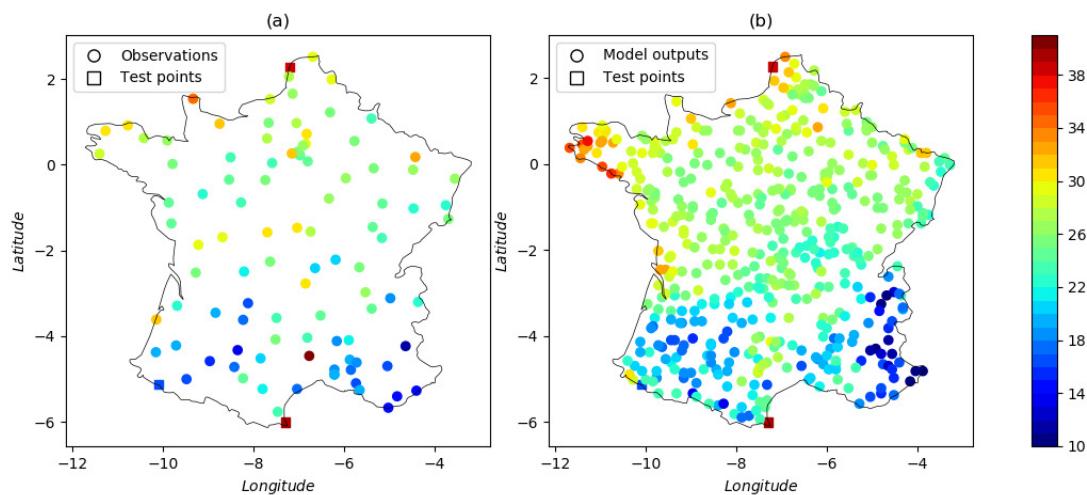


FIGURE 7 Plots of observations, MetUM model outputs for fitting the model and 3 test points with the largest residuals. (a) Observations and 3 test points with the largest residuals. (b) MetUM model outputs and 3 test points with the largest residuals.

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 ($D_I(\tilde{Y})$) against the validation data index and the predicted mean respectively. Figure 8 (c) presents the pivoted Cholesky errors ($D_{PC}(\tilde{Y})$) against the pivoting order. No obvious pattern can be seen from these graphical diagnoses except for three largest errors, which justifies the GP assumption of the DF model. The three largest errors actually correspond to the three points locating at the boundary of France (see Figure 6 (b)), where the model exhibits more uncertainty about the predictions due to insufficient observations around the area.

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 $p(\theta)$ denote the approximate multivariate normal distribution of the parameters where the mean is the maximum likelihood estimate (MLE) and the covariance is the minus of the inverse of Hessian matrix obtained from the MLE. The solid points are the sample quantiles of the pivoted Cholesky errors (computed at the mean of $p(\theta)$) 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 θ from $p(\theta)$, 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 justifies the GP assumption of the model.

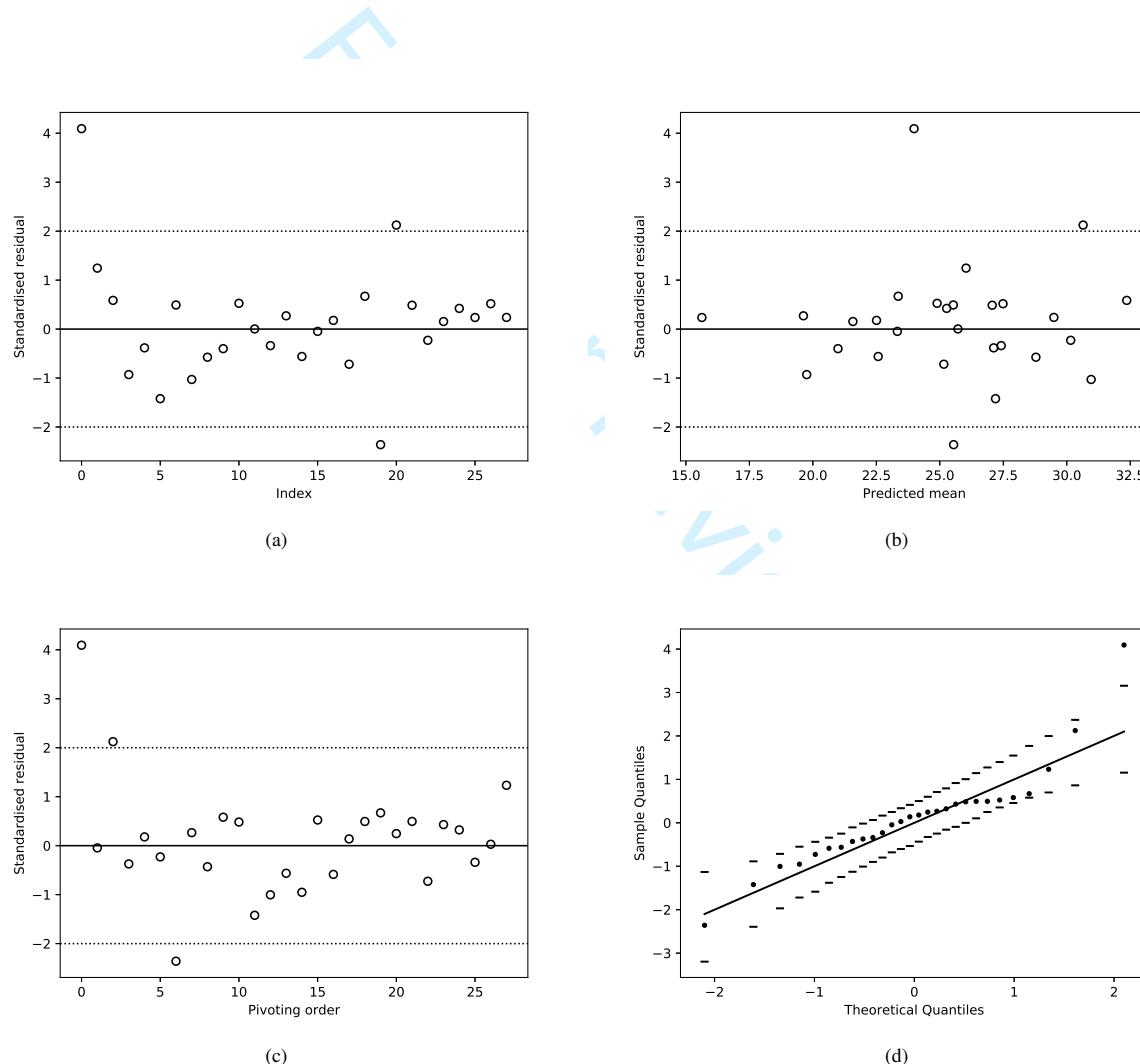


FIGURE 8 Graphical diagnostics for the out-of-sample prediction for the observed data. (a) Standardised individual prediction errors $D_I(\tilde{Y})$ versus the validation data index. (b) Standardised individual prediction errors $D_I(\tilde{Y})$ versus the predictions. (c) Pivoted Cholesky errors $D_{PC}(\tilde{Y})$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

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

$$p(\mathbf{Y} | \mathbf{X}) \sim \mathcal{N}(\boldsymbol{\mu}_{Y|X}, \Sigma_{Y|X}) \quad (33)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{Y|X} &= \boldsymbol{\mu} + \Sigma_{YX} \Sigma_X^{-1} (\mathbf{X} - \boldsymbol{\alpha}) \\ \Sigma_{Y|X} &= \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} \end{aligned} \quad (34)$$

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

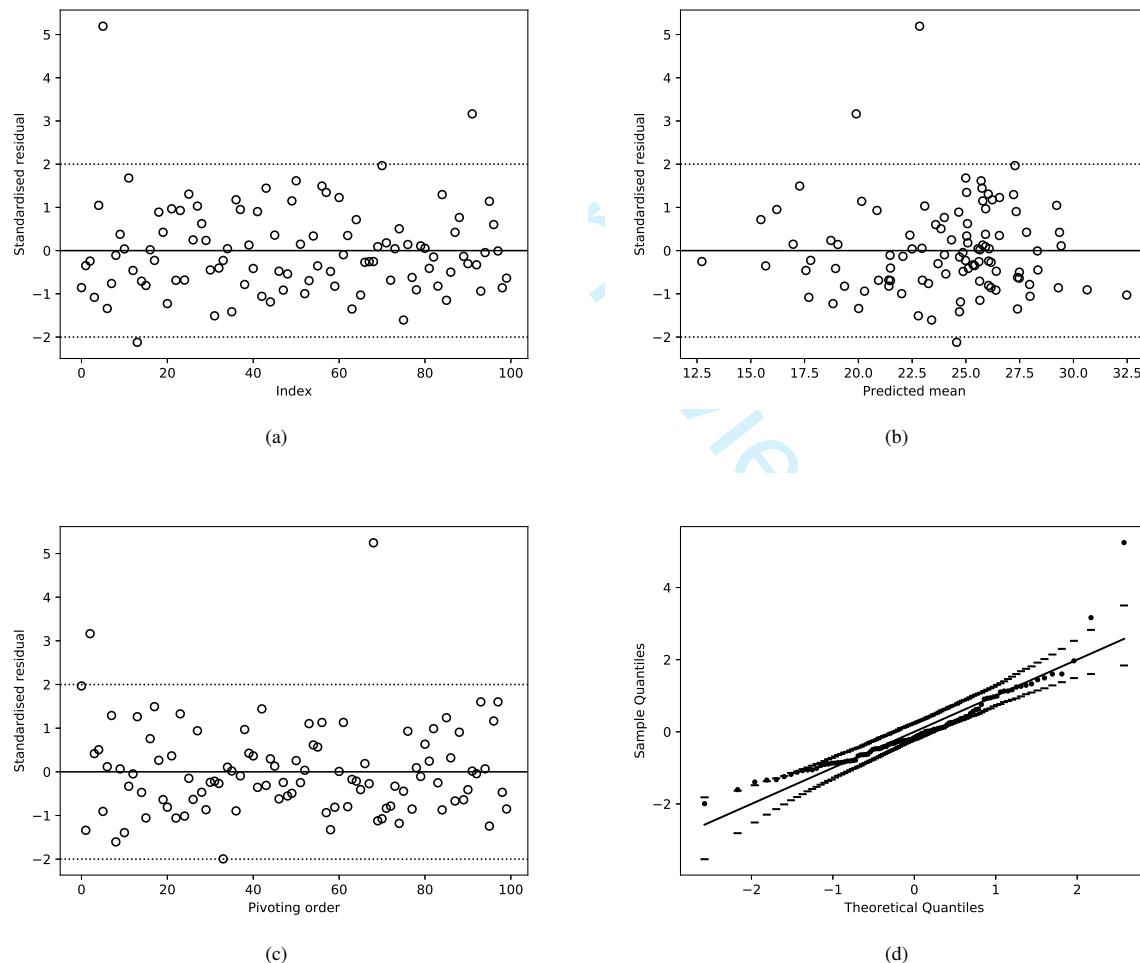


FIGURE 9 Graphical diagnostics for the in-sample prediction for the observed data. (a) Standardised individual prediction errors versus the validation data index. (b) Standardised individual prediction errors versus the predictions. (c) Pivoted Cholesky errors versus the pivoting order (d) QQ-plot of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

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

$$p(\mathbf{X} | \mathbf{Y}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}|\mathbf{Y}}, \Sigma_{\mathbf{X}|\mathbf{Y}}) \quad (35)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{X}|\mathbf{Y}} &= \boldsymbol{\alpha} + \Sigma_{\mathbf{XY}} \Sigma_Y^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \\ \Sigma_{\mathbf{X}|\mathbf{Y}} &= \Sigma_X - \Sigma_{\mathbf{XY}} \Sigma_Y^{-1} \Sigma_{YX} \end{aligned} \quad (36)$$

The graphical diagnoses are given in Figure 10 . In this case, we observe a number of abnormal points in these plots. In particular, in Figure 10 (d) which is the QQ-plot of the pivoted Cholesky errors, 14 points are outside the 95% confidence interval in the lower tail which relates to small values of MetUM model outputs. 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 (32) or use a different covariance structure for $c_Z(s, s')$ and $c_\delta(s, s')$.

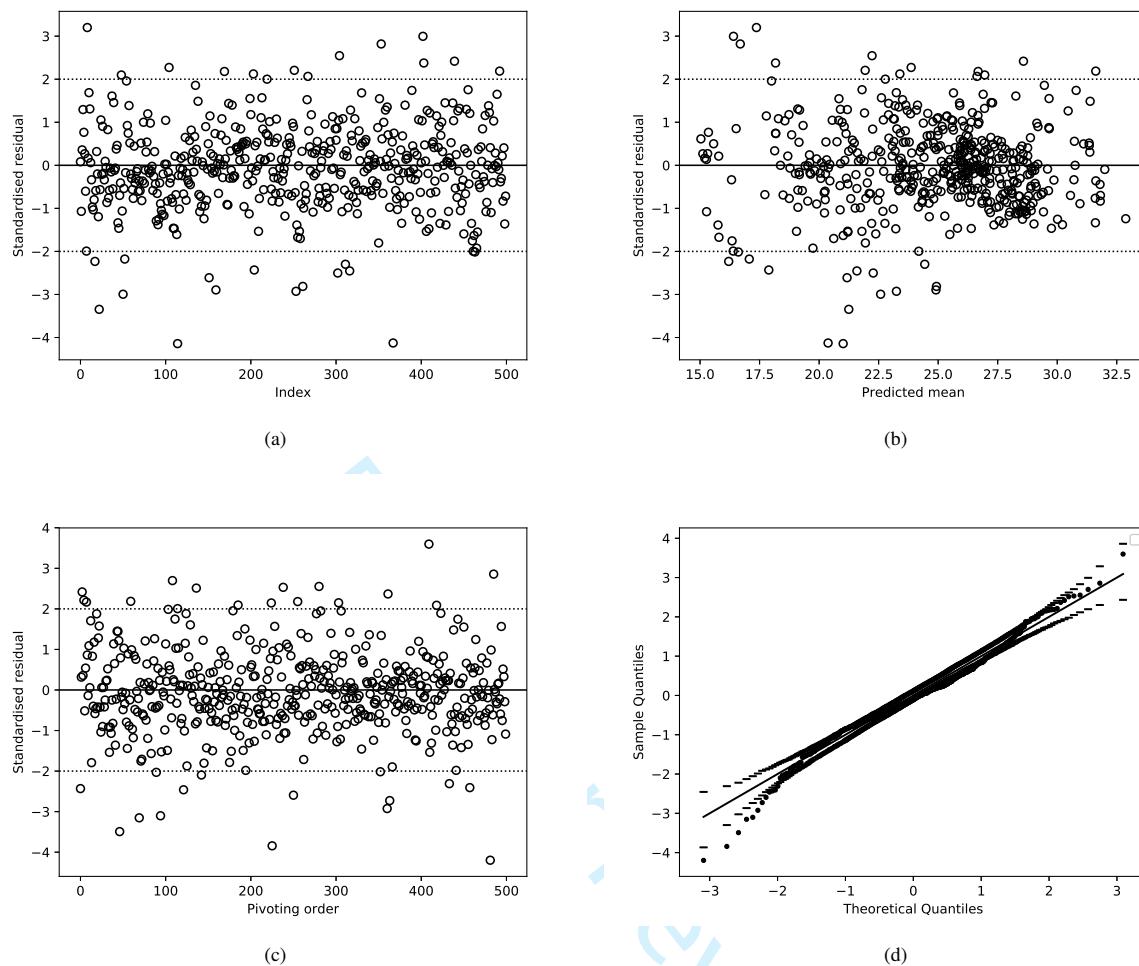


FIGURE 10 Graphical diagnostics for in-sample prediction for the MetUM model outputs. (a) Standardised individual prediction errors versus the validation data index. (b) Standardised individual prediction errors versus the predictions. (c) Pivoted Cholesky errors versus the pivoting order. (d) QQ-plot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

6 | CONCLUSION

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, hails and tornadoes, it is usually difficult to obtain observational data at high geographical resolution and hence utilising other sources of data (physical model outputs 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 outputs for improved predictions at high spatial resolution. It extends the BM method by utilising a GP to flexibly model the discrepancy structure between the observational data and the numerical model outputs. 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

providing a lower RMSE and AWCI while maintaining a similar coverage probability when compared to the Kriging model based solely on observational data in the simulation study. When applied to the Imogen windstorm data, it provides reliable estimates at locations where there is no observational data with a prediction accuracy of 93%. 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 the sample paths of the GP are infinitely differentiable. Such an assumption might be inappropriate for some process such as the case of high-resolution rainfall measurements derived from radar data which tend to have much greater short scale variability and hence the power exponential, or the Matérn covariance function could be considered for such processes. Second, the 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 $Z(s)$ and $Z(s')$. Therefore, exploring a different length-scale for the longitude and latitude in the covariance function would be another direction of future work. The model of (10) 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 P 2007) may be needed for other processes. In addition, we have sub-sampled up to 500 model outputs out of the total available 27,185 model outputs for computational feasibility. However, it would produce more reliable and accurate estimates if we could use all the model outputs, hence a natural extension of the framework is to incorporate methods such as Banerjee (2017) for efficient inference in high-dimensional cases. Last but not least, 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.

7.2 | Availability of data and material

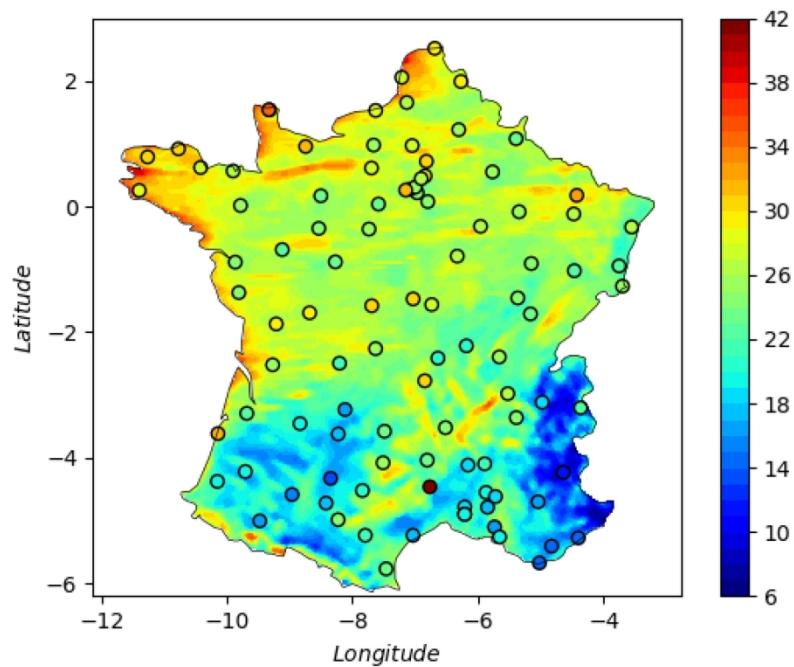
The data that support the findings of this study are openly available at https://wisc.climate.copernicus.eu/wisc/#/help/products#stormtrack_download, WISC (2019).

References

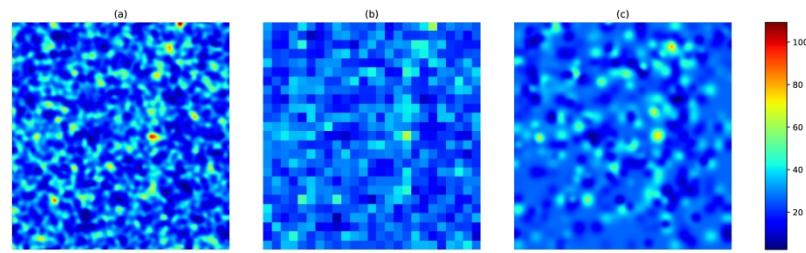
- Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian analysis*, 12(2), 583.
- Bastos, L. S., & O'Hagan, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics*, 51(4), 425–438.
- Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance 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., Berrisford, P., Poli, P., Kobayashi, S., ... others (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, R. P. (2007). *Model-based geostatistics*. Springer: New York.
- 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.
- 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.
- 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.
- 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.
- 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., Donkelaar, A., Burnett, R., ... others (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.
- 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.

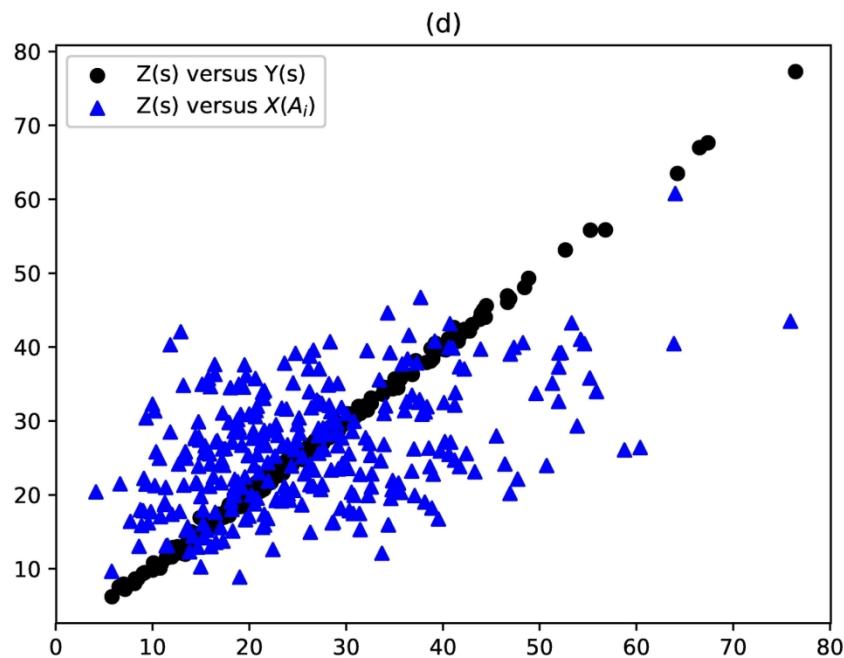




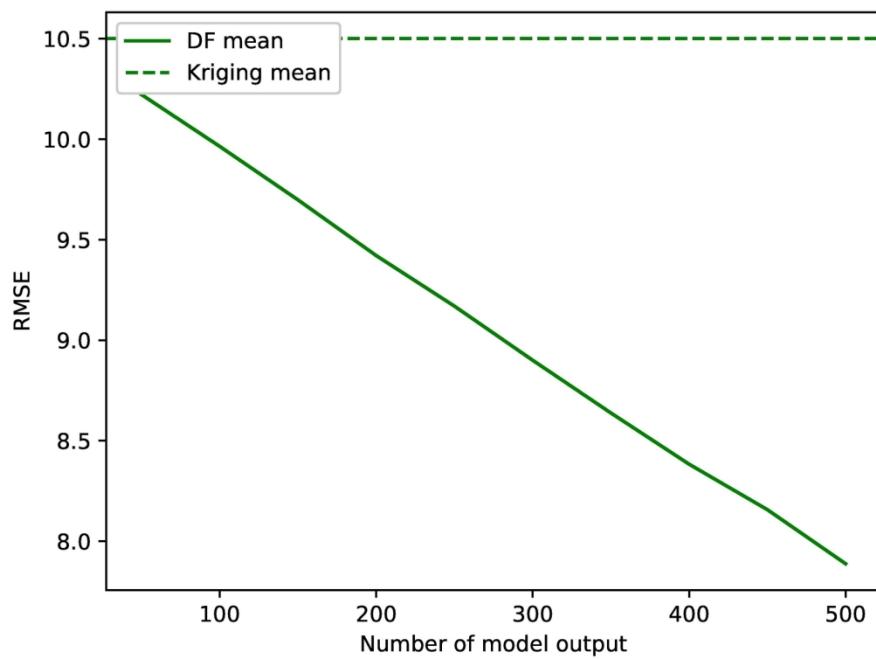
162x121mm (300 x 300 DPI)



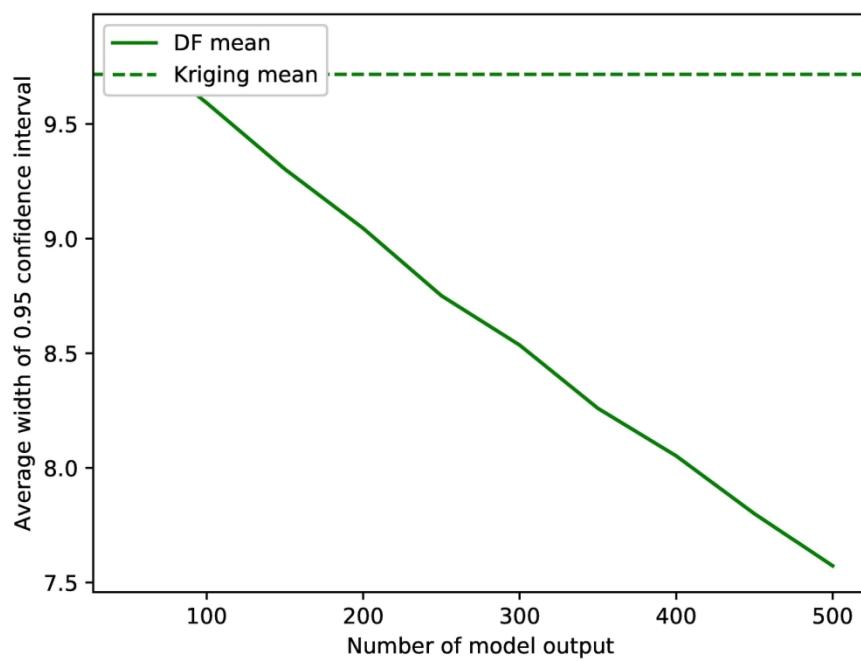
431x127mm (300 x 300 DPI)



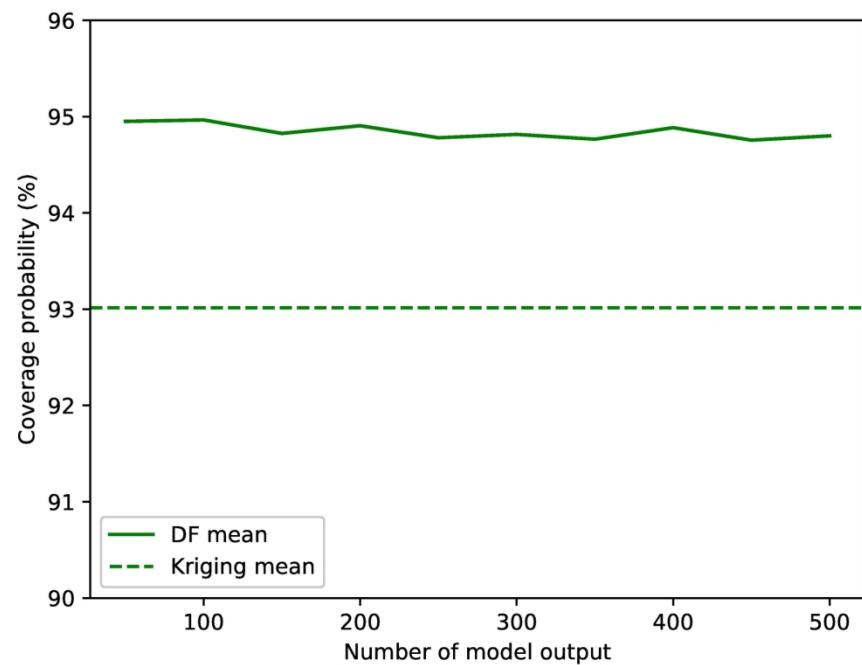
162x121mm (300 x 300 DPI)



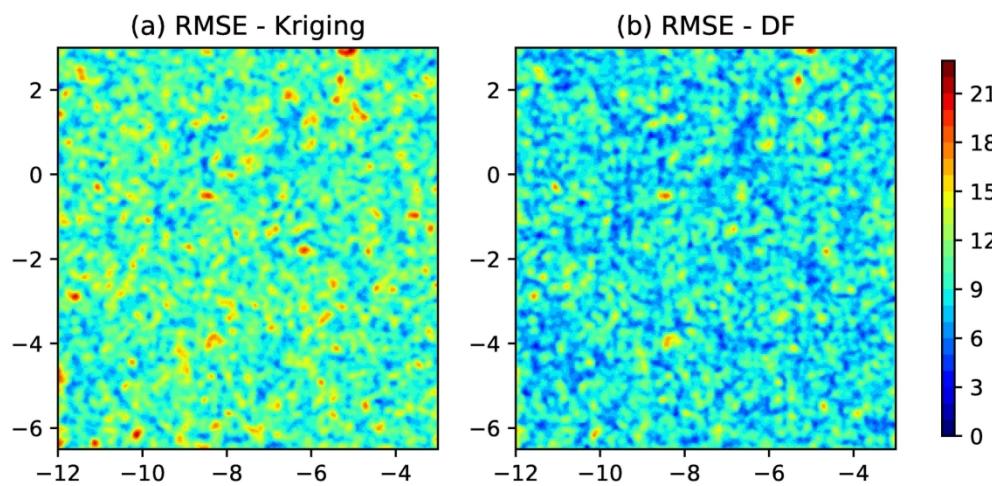
162x121mm (300 x 300 DPI)



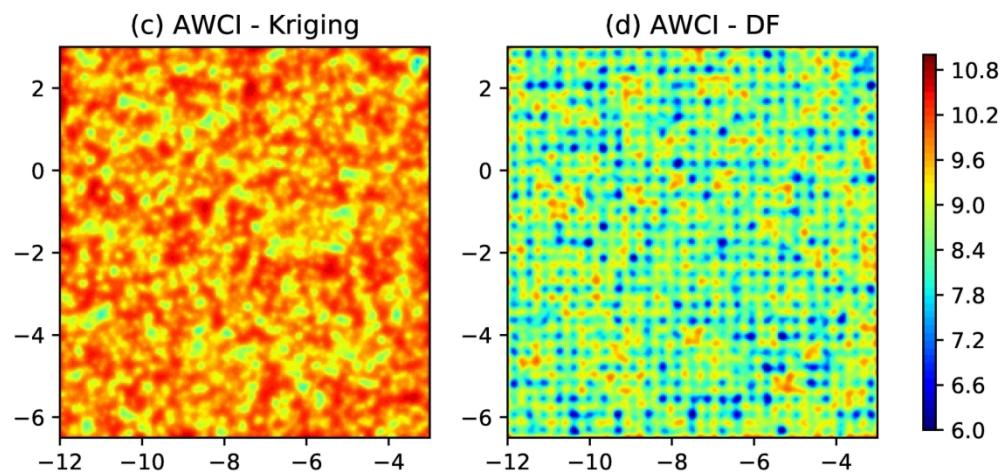
162x121mm (300 x 300 DPI)



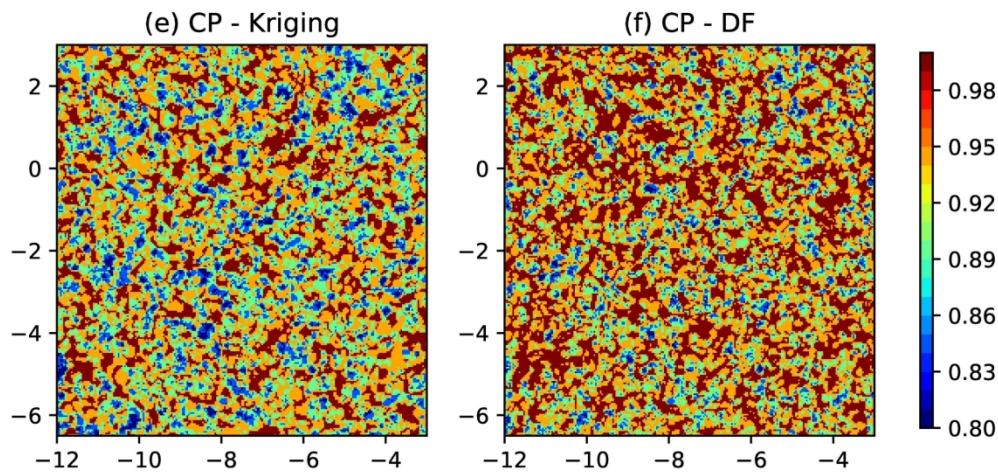
162x121mm (300 x 300 DPI)



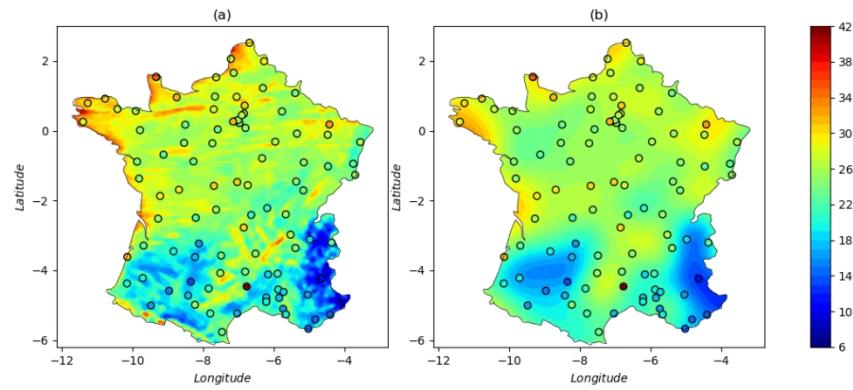
162x121mm (300 x 300 DPI)



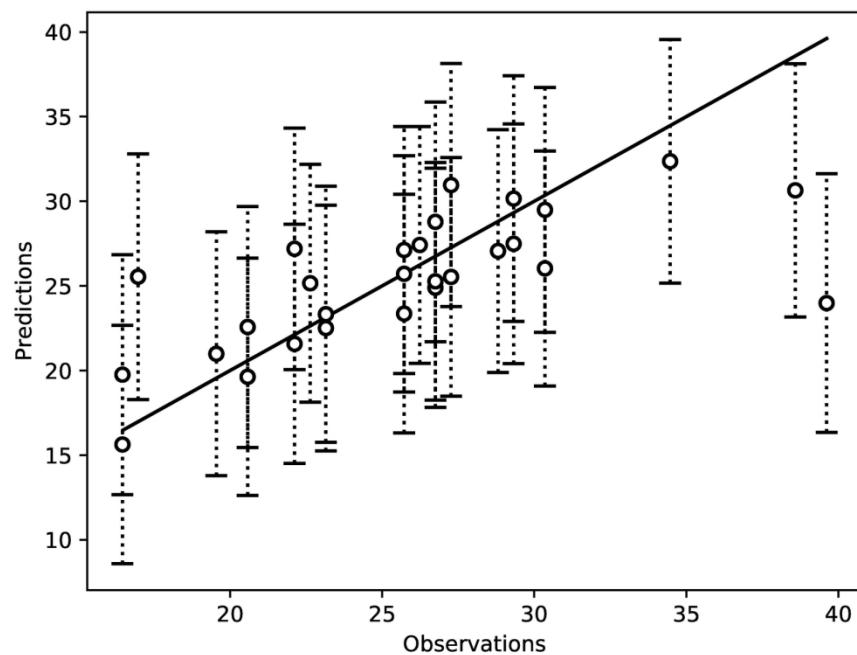
162x121mm (300 x 300 DPI)



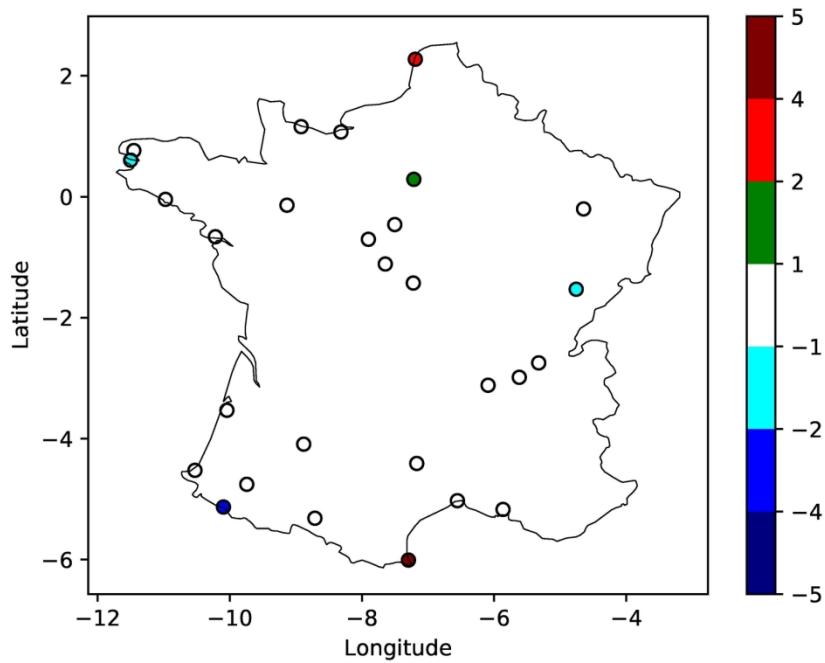
162x121mm (300 x 300 DPI)



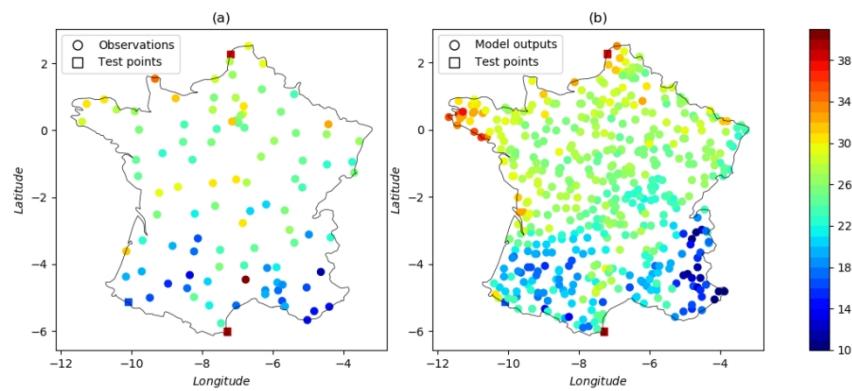
304x127mm (300 x 300 DPI)



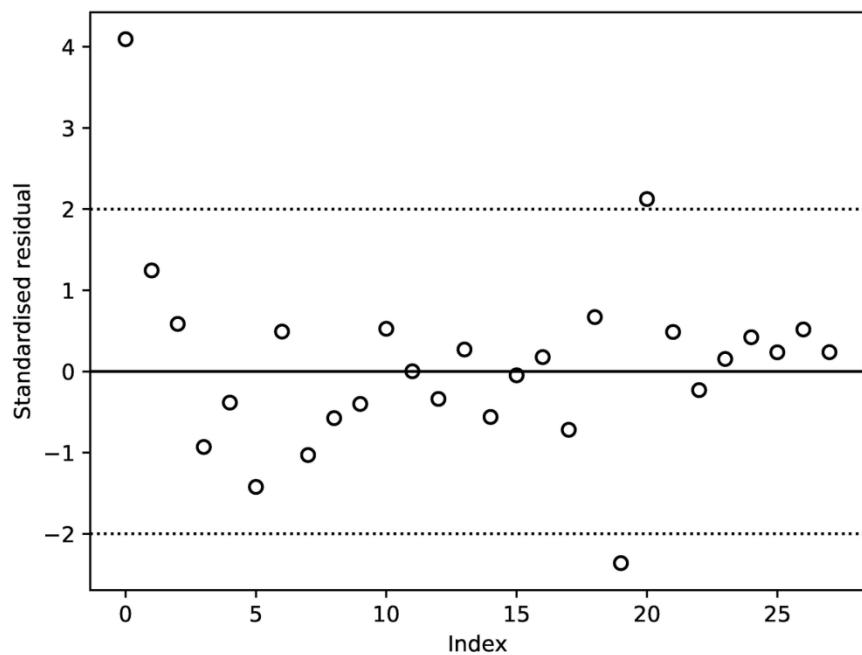
162x121mm (300 x 300 DPI)



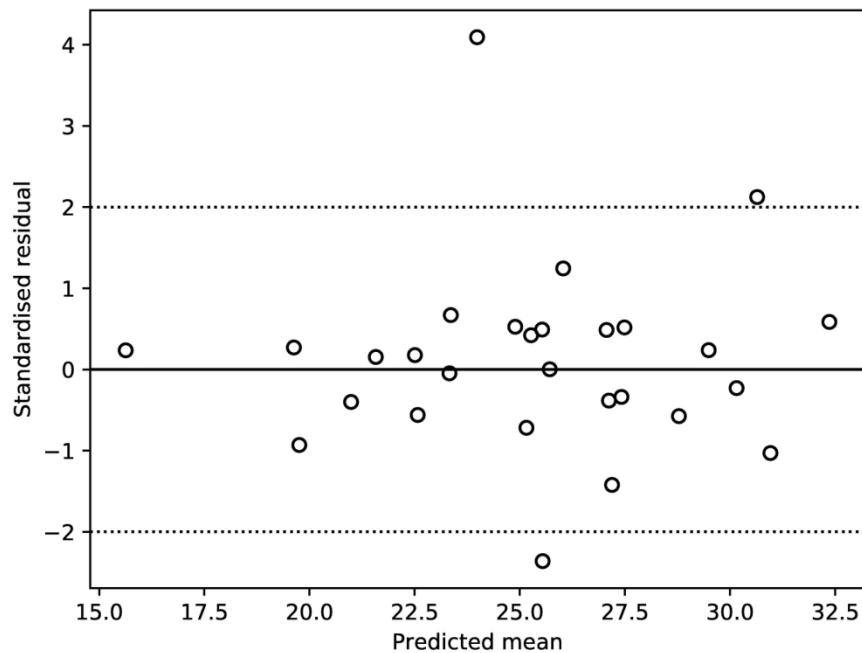
162x121mm (300 x 300 DPI)



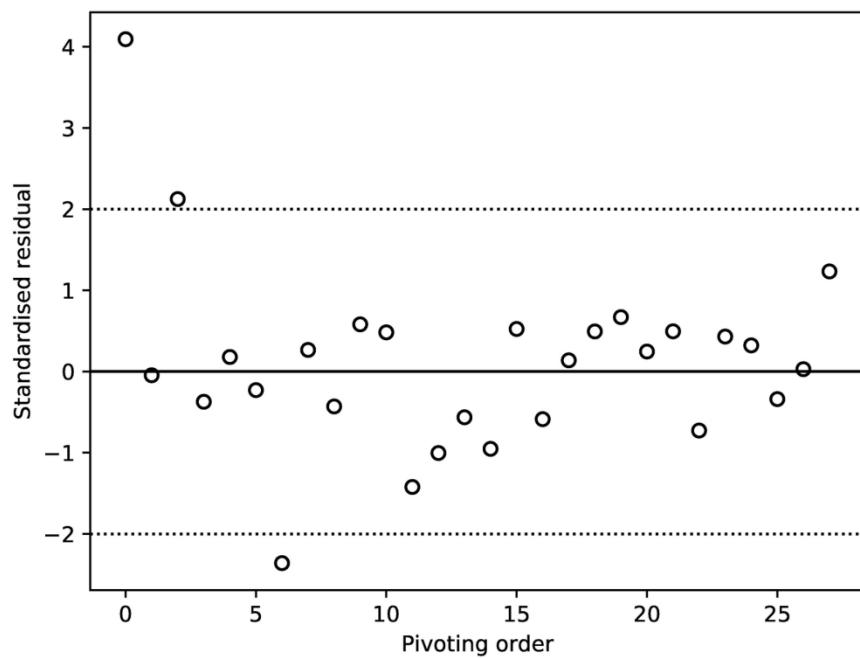
304x127mm (300 x 300 DPI)



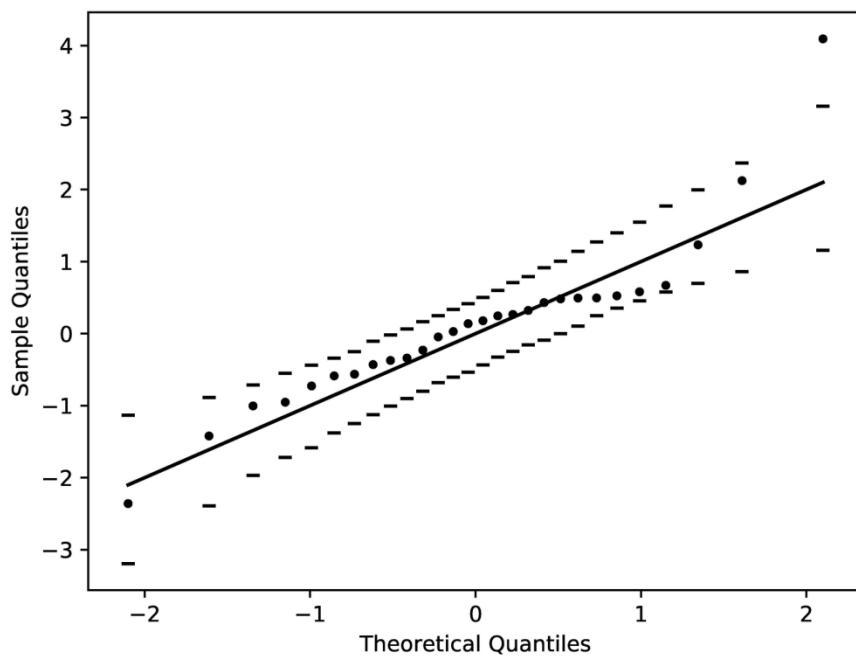
162x121mm (300 x 300 DPI)



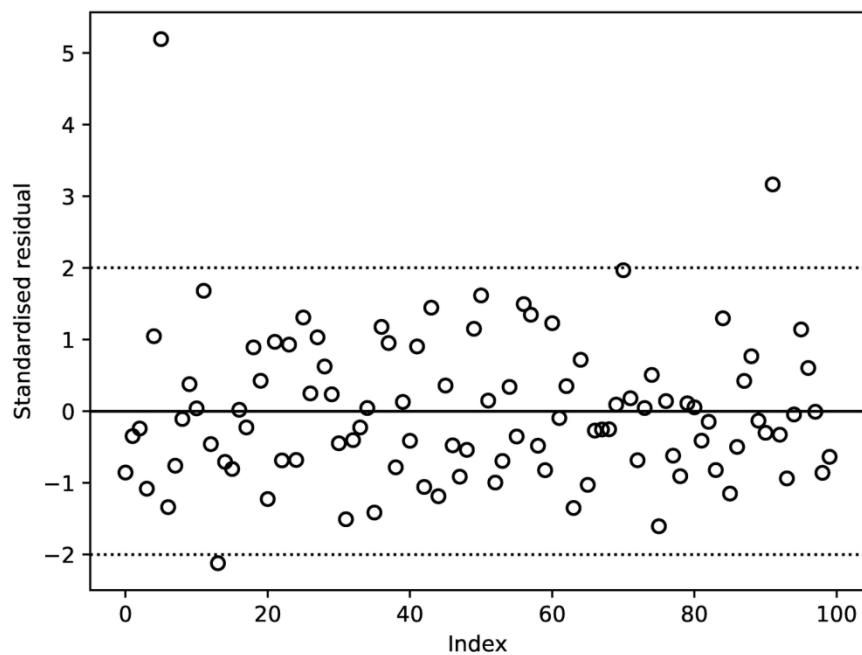
162x121mm (300 x 300 DPI)



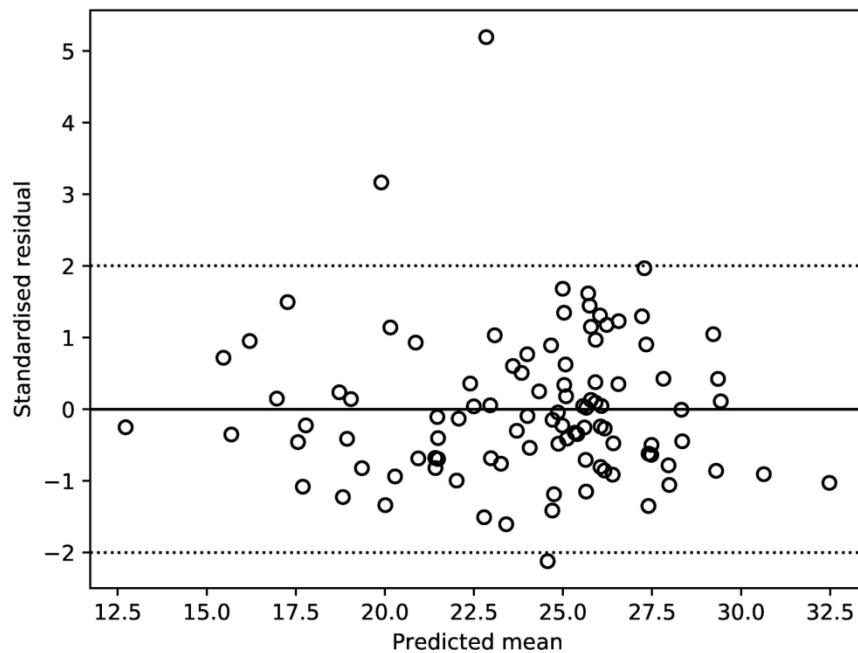
162x121mm (300 x 300 DPI)



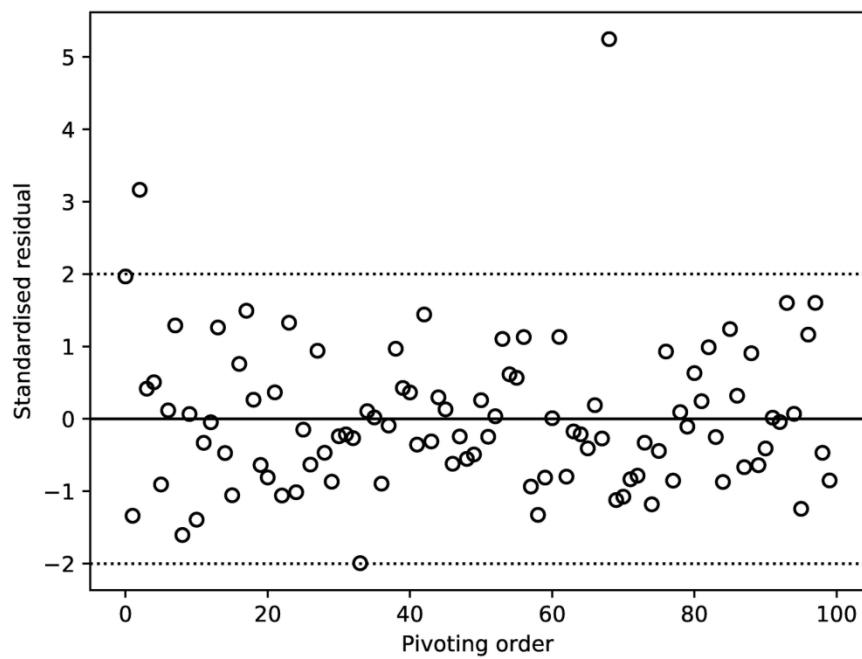
162x121mm (300 x 300 DPI)



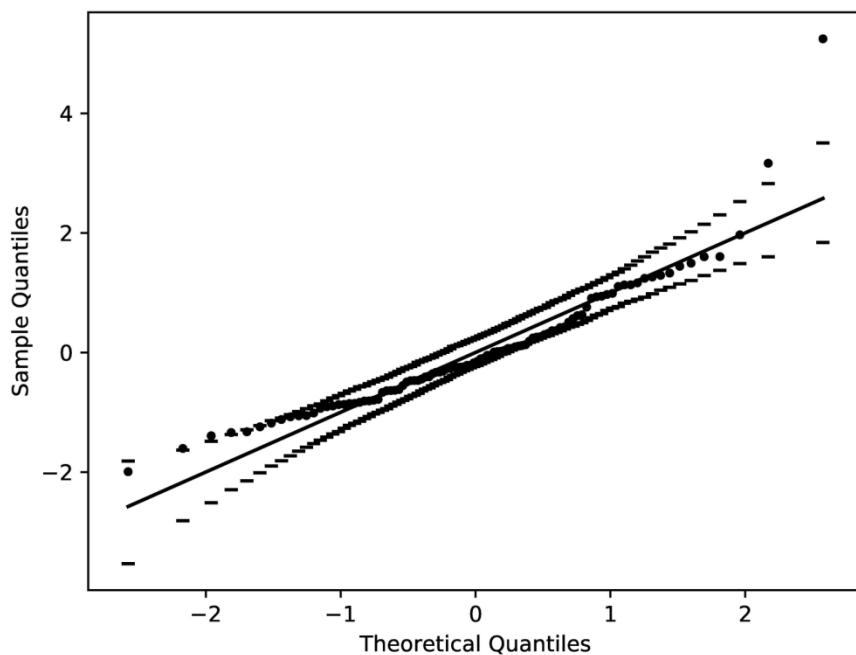
162x121mm (300 x 300 DPI)



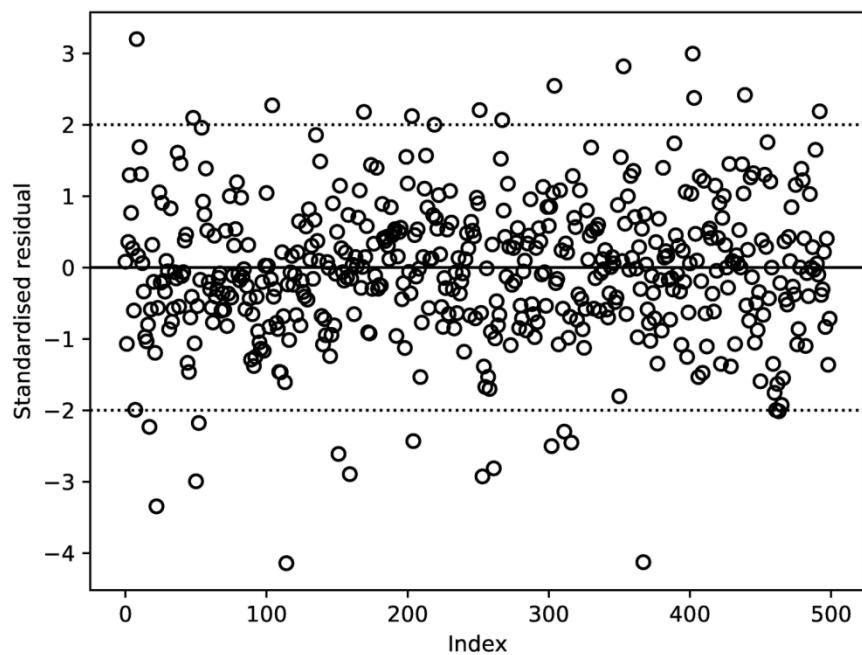
162x121mm (300 x 300 DPI)



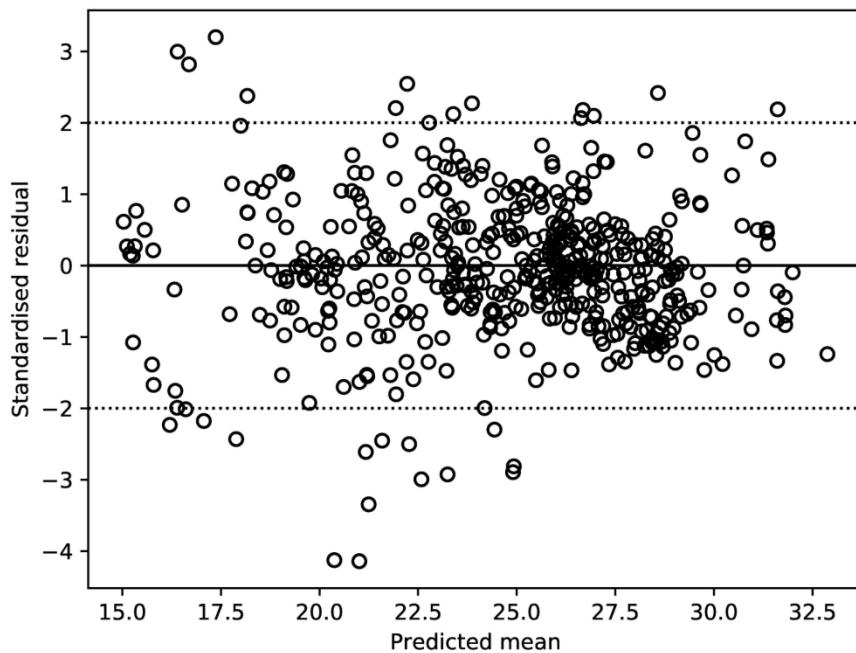
162x121mm (300 x 300 DPI)



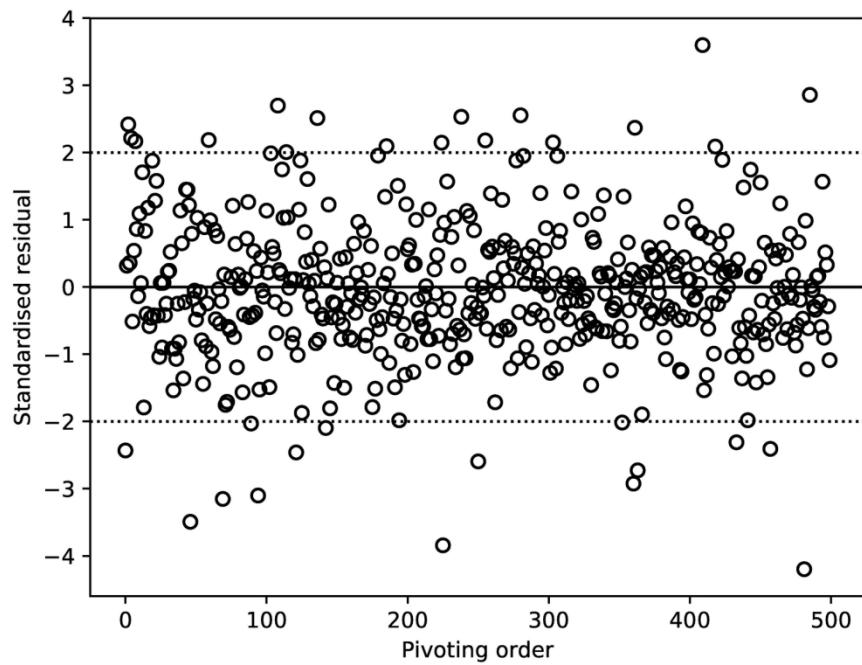
162x121mm (300 x 300 DPI)



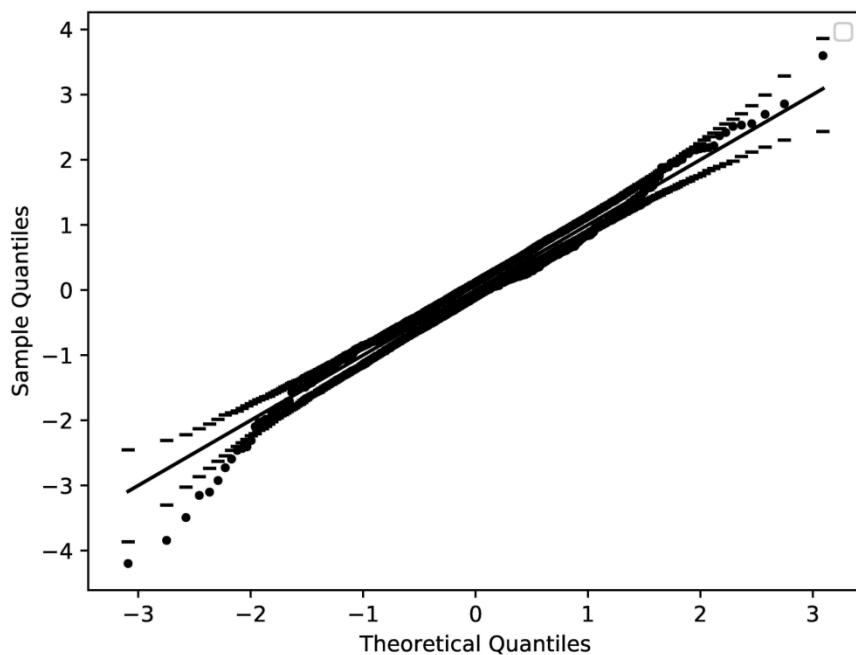
162x121mm (300 x 300 DPI)



162x121mm (300 x 300 DPI)



162x121mm (300 x 300 DPI)



162x121mm (300 x 300 DPI)