Data fusion with Gaussian processes for estimation of windstorm events

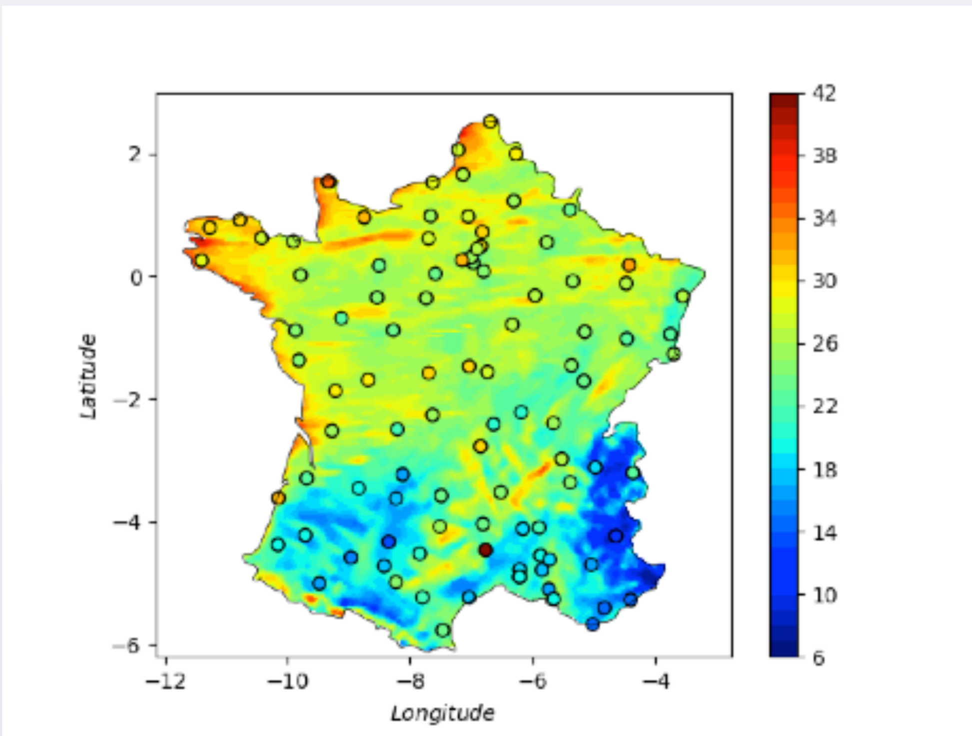
Xiaoyu Xiong, Benjamin D. Youngman, Theodoros Economou

University of Exeter email: x.xiong@exeter.ac.uk, b.youngman@exeter.ac.uk
t.economou@exeter.ac.uk

Motivation

- The windstorm footprint, a spatial area describing quantified wind gust speeds, is commonly used for risk estimation
- Windstorm footprints are conventionally estimated using measurements from observing stations and/or gridded analyses produced by numerical climate simulation models
- 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

Windstorm Imogen footprint over France



- Knowledge of small scale detail in windstorm footprints is important because of the large spatial heterogeneity in vulnerability and exposure
- Interpolation at small scale based solely on sparse observational data is often not accurate, hence integrating additional data sources for improved spatial interpolation is necessary but entails several challenges:
  - All data sources should be thought of as imperfect representations of the truth. Biases can be both systematic and random
  - The data sources can have different spatial support
  - It is important to accurately quantify uncertainty from the different data sources and propagate this to predictions
- 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

Data fusion (DF) with Gaussian processes (GPs)

$Z(s) \sim GP(\mu(s), c_z(s, s'))$	True process	(1)
$Y(s) = Z(s) + \varepsilon(s)$	Data (observations)	(2)
$\varepsilon(s) \sim \mathcal{N}(0, \sigma_Y{}^2)$	Measurement error	(3)
$X(s) = \alpha(s) + \beta(s)Z(s) + \delta(s)$	Numerical model output	(4)
$\delta(s) \sim GP(\mu(s), c_\delta(s, s'))$	Discrepancy term	(5)

Main novelty of the model:

While $\alpha(s)$, $\beta(s)$ will capture consistent under- or over-estimation of wind speeds, the GP nonparametric formulation of $\delta(s)$ can allow for any possible form of spatially structured discrepancy.

Change of support:

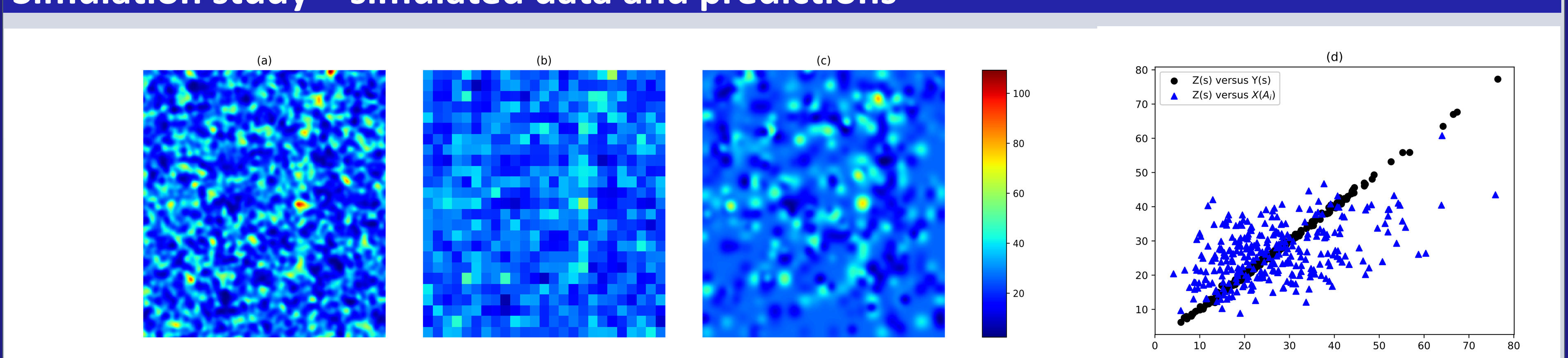
$$X(A_i) = \int_0^{A_i} X(s)\,ds = \int_0^{A_i} \alpha(s)\,ds + \int_0^{A_i} \beta(s)Z(s)\,ds + \int_0^{A_i} \delta(s)\,ds$$

The joint distribution:

$p(\boldsymbol{Y},\boldsymbol{X}|\boldsymbol{\theta}) \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\alpha}\end{pmatrix}, \begin{pmatrix}\Sigma_Y & \Sigma_{YX}\\ \Sigma_{XY} & \Sigma_X\end{pmatrix}\right)$ where $\boldsymbol{Y}$ is of size n and $\boldsymbol{X}$ is of size m with

$$\boldsymbol{\mu}=(\mu(s_1),\dots,\mu(s_n))^T \quad \boldsymbol{\alpha} = \left(\int_0^{A_1}\alpha(s)ds,\dots,\int_0^{A_m}\alpha(s)ds\right)^T$$

Simulation study – simulated data and predictions

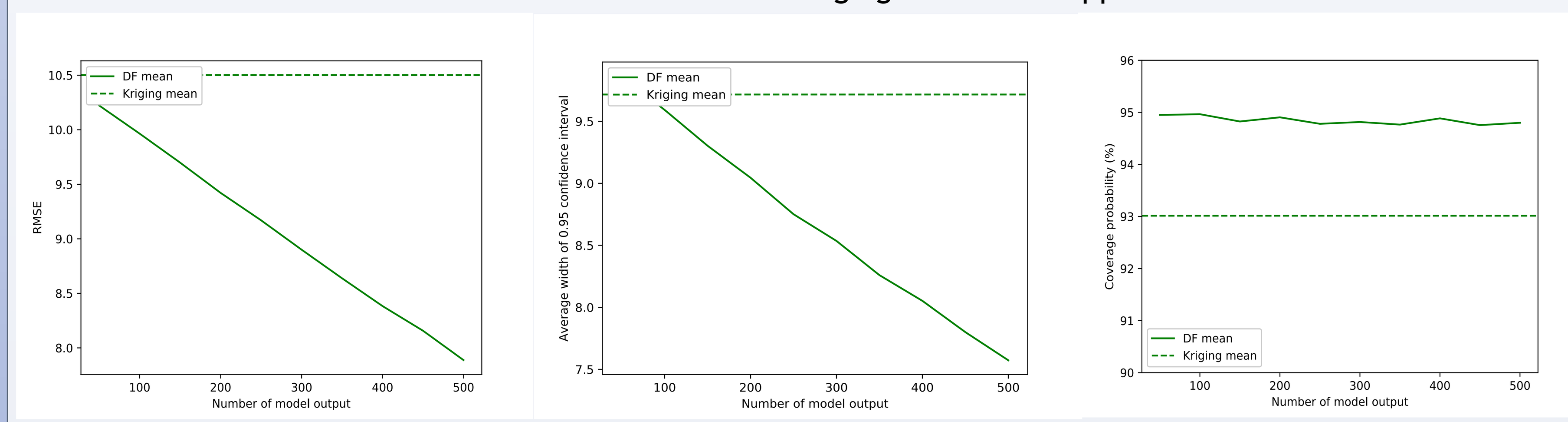


(a) Simulated Z(s) over a 1000 × 1000 grid (b) Simulated $X(A_i)$ over a 25 × 25 block
(c) Predicted Z(s) (d) Simulated Z(s) versus Y(s) and $X(A_i)$

- Figure (d) shows averaging Z(s) to get $X(A_i)$ has clearly induced bias
- Figure (c) indicates that the model has captured the true process Z(s) reasonably well

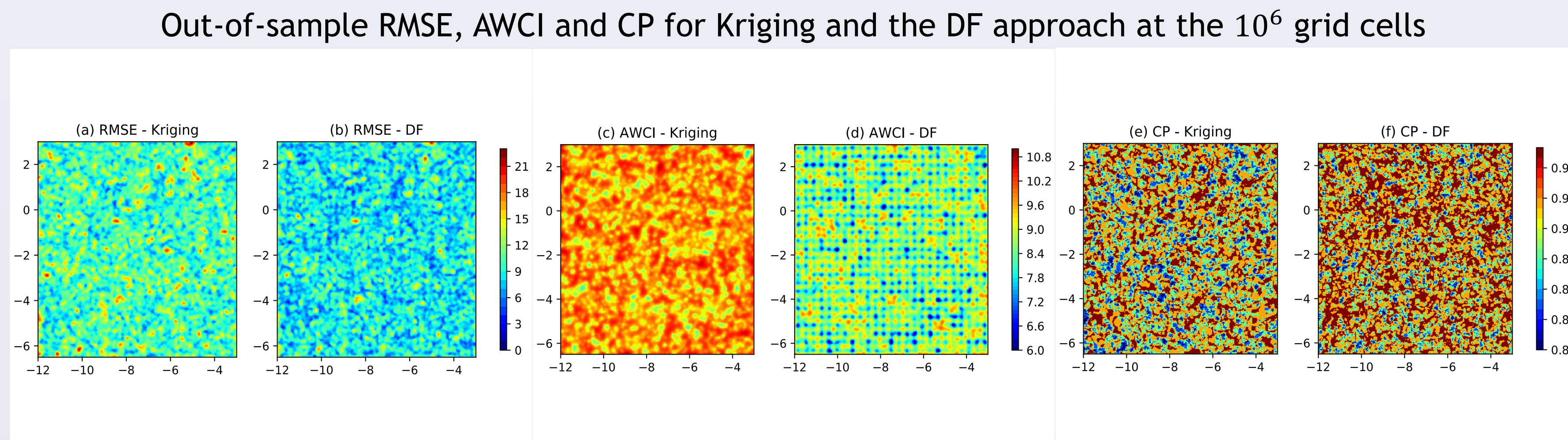
Simulation study – out-of-sample predictions

Out-of-sample RMSE, average width of 95% confidence interval and coverage probability over 100 simulations for Kriging and the DF approach



- Metrics: RMSE (root mean square error), average width of 95% confidence interval (AWCI), coverage probability (CP); the Kriging model is described by equations (1) and (2)
- DF model achieves a lower RMSE and AWCI while maintaining a similar CP
- The higher the number of model outputs, the lower the RMSE and AWCI

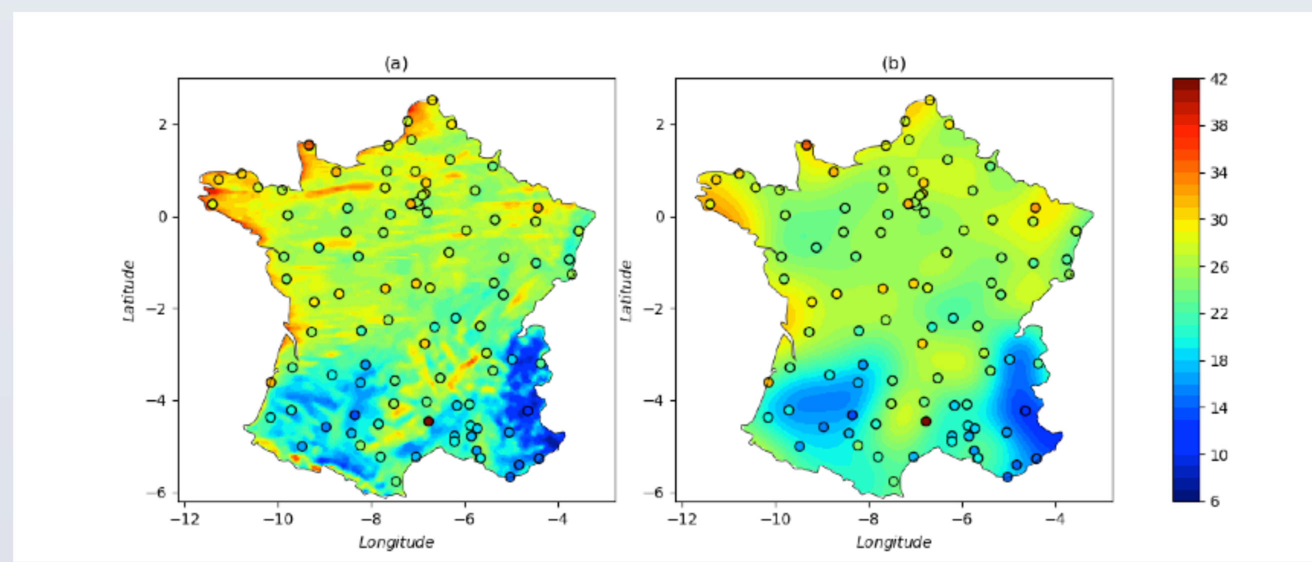
Out-of-sample RMSE, AWCI and CP for Kriging and the DF approach at the 10^6 grid cells



- 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
- DF model is able to more accurately quantify the uncertainty at high spatial resolution

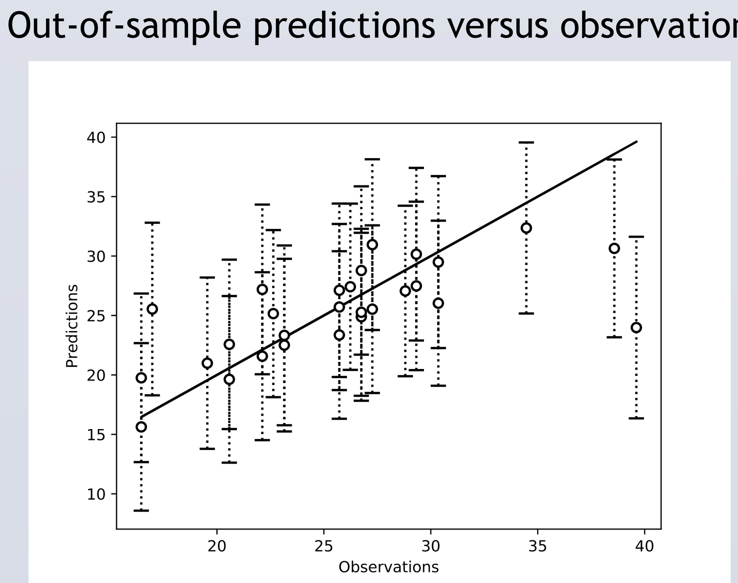
Application to Imogen windstorm data

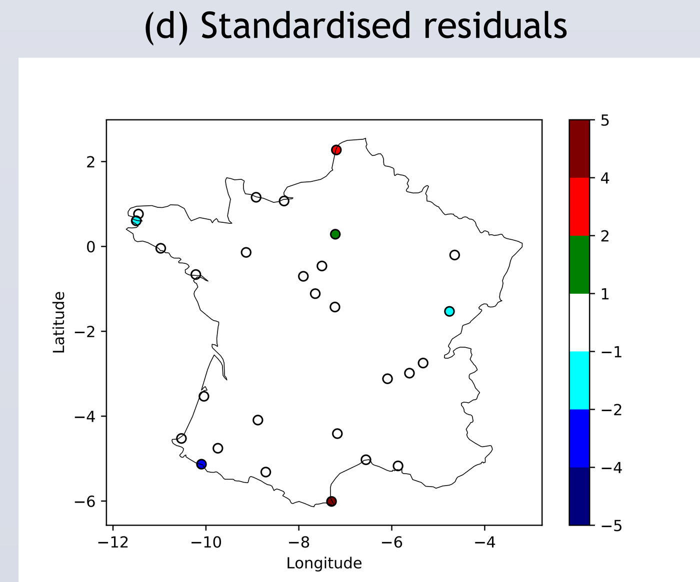
(a) Observations and model outputs to fit the DF model (b) Observations and predicted model outputs



- Improved spatial interpolation at small scales
- Could work as a way of validating numerical simulator outputs

(c) Out-of-sample predictions versus observations (d) Standardised residuals





- High prediction accuracy of 93%
- Two abnormal ones locates at the boundary where higher uncertainty is expected

Conclusions

- We present a generic DF framework that utilises 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 at high spatial resolution
- The DF framework is in principle able to generalise to other data sources
- Current work attempts to exploit sparse GPs to deal with millions of data points

References

[1] Kennedy, M. C., and O'Hagan, A. (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(3), 425-464.

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

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

[4] Xiong, X., Smídl, V., and Filippone, M. (2017). Adaptive multiple importance sampling for Gaussian processes. Journal of Statistical Computation and Simulation, 87(8), 1644-1665.