

Surrogate Modelling and Uncertainty Quantification

Internship at IBM Research Daresbury

Thanks to...



Małgorzata Zimoń



IBM Research Staff

Sci-Tech Daresbury

- Opened in 1967 as the Daresbury Nuclear laboratory
- Houses > 1,000 people working at ~100 tech companies

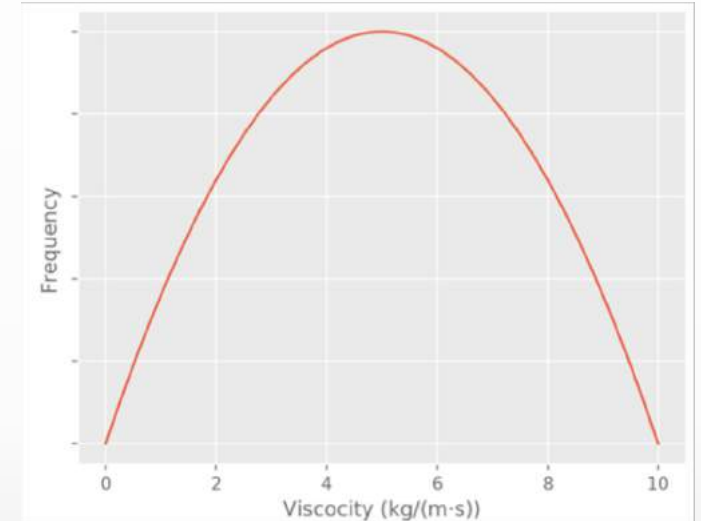
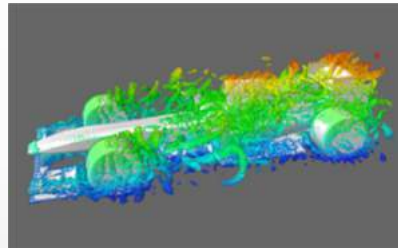
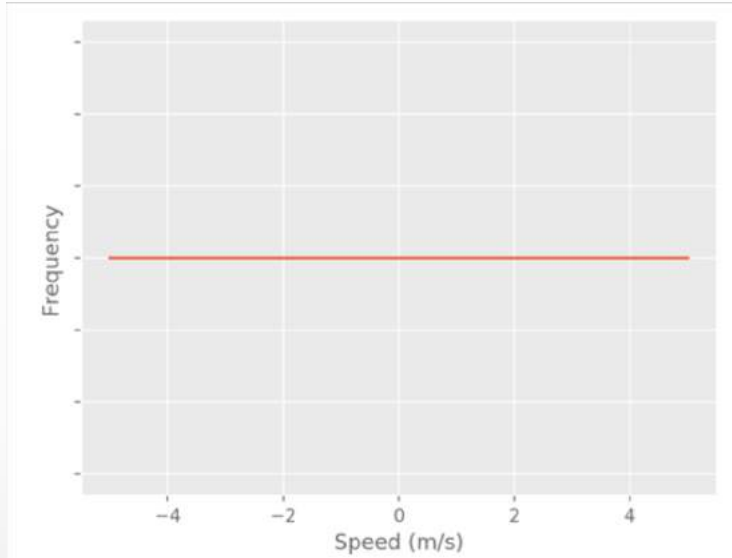


IBM Research UK

- IBM has ~350,000 employees worldwide
- ~60 researchers working for IBM Research UK based Daresbury and Hursley
- Research areas include computational chemistry, life sciences, physical science and engineering, quantum computing



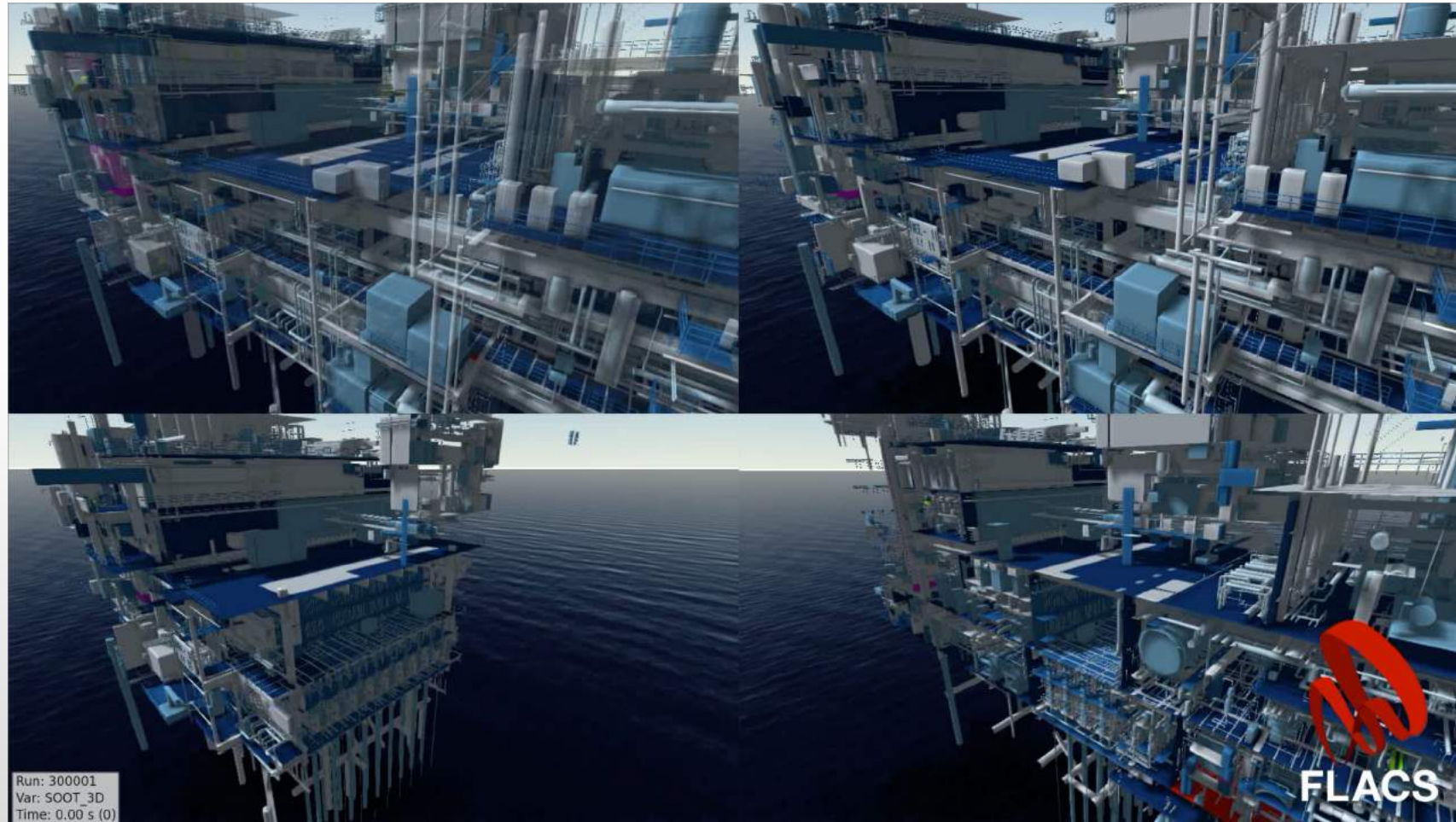
Surrogate Modelling



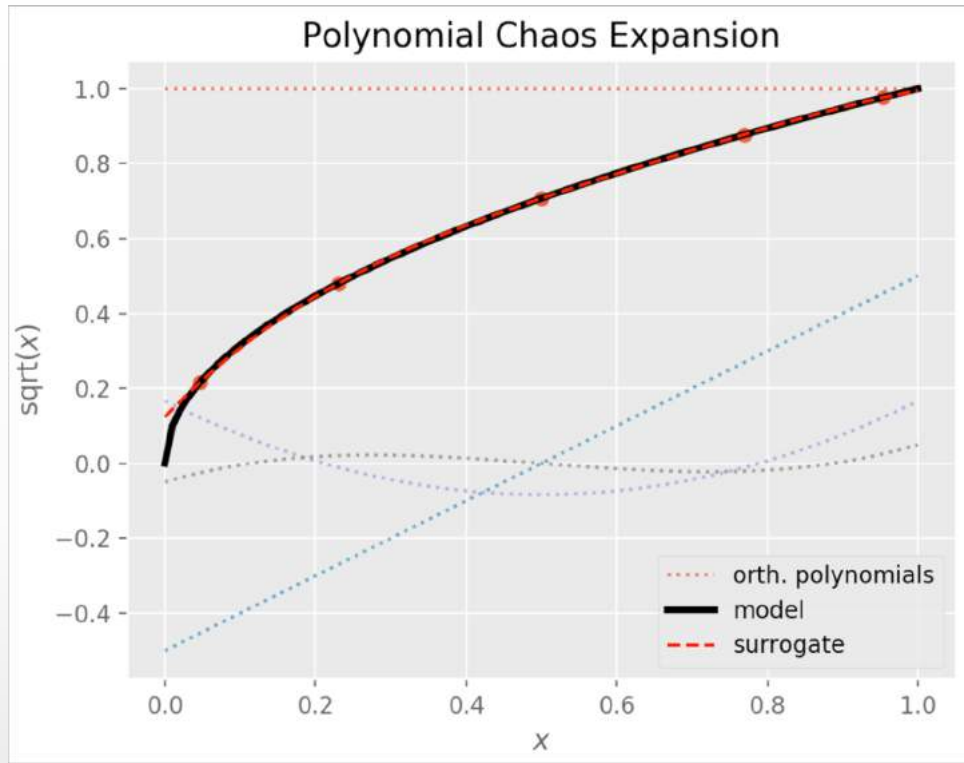
- Computer simulations can be expensive to run, and - to understand the output statistics - require many iterations
- Fast approximations of simulations can be built to quickly construct output statistics, called surrogates

- We focus on two popular non-intrusive surrogate methods - Polynomial Chaos Expansion and Gaussian Process Emulation
- We want to build surrogates non-intrusively, i.e. not knowing anything about the internal processes of the simulation

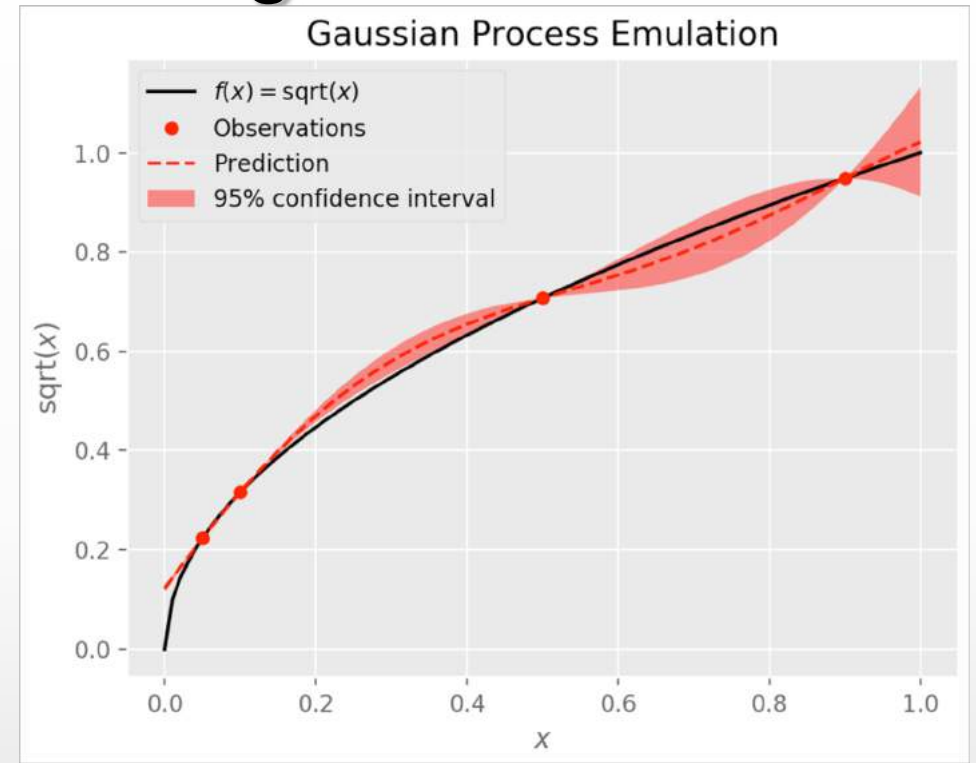
Research Aim: Rare Event Modelling



Surrogate Modelling



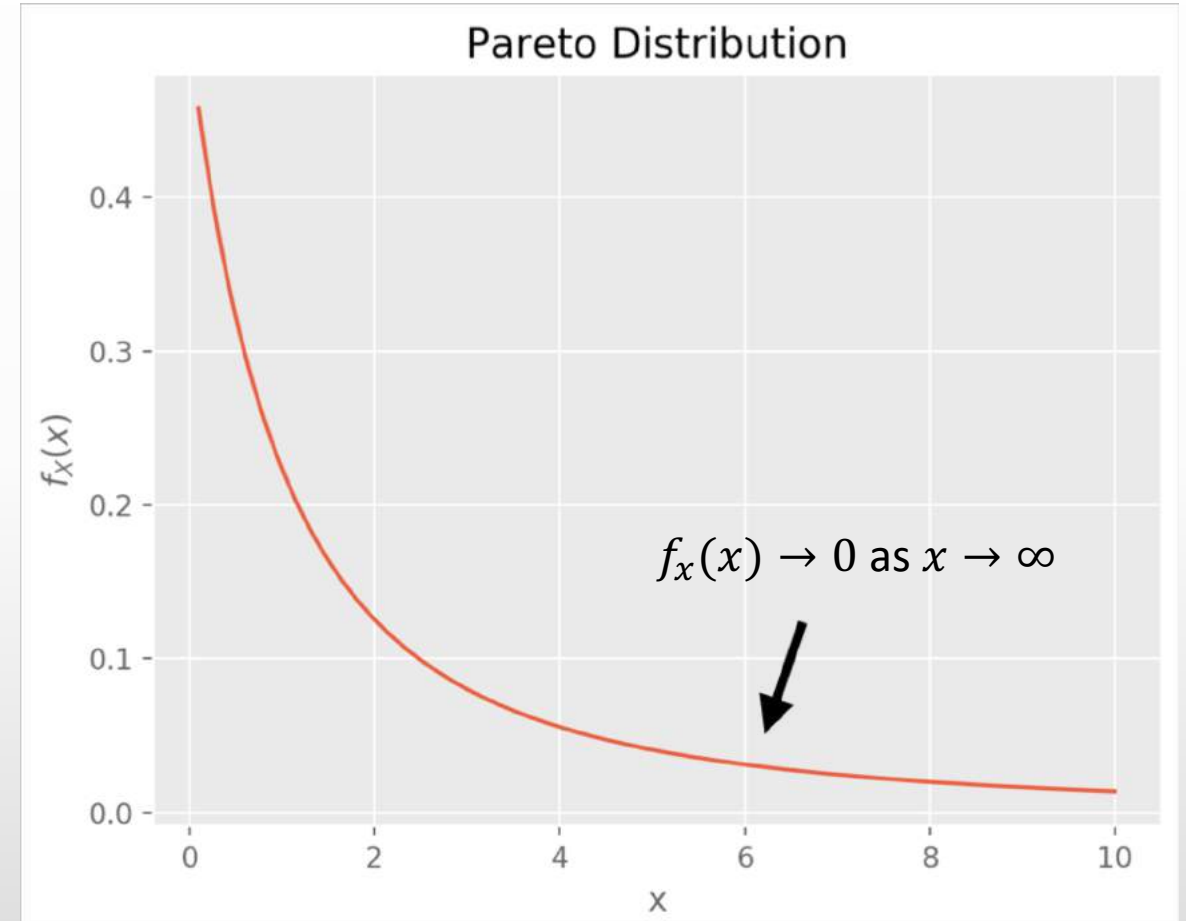
- Expresses the simulator (model) as a sum of weighted orthogonal basis polynomials
- $y(x) \approx \sum_{k=0}^n c_k \psi_k(x)$



- GPE can be used as an interpolation, with uncertainty information provided
- It acts as a joint probability distribution function for simulation outputs $f(x_i)$

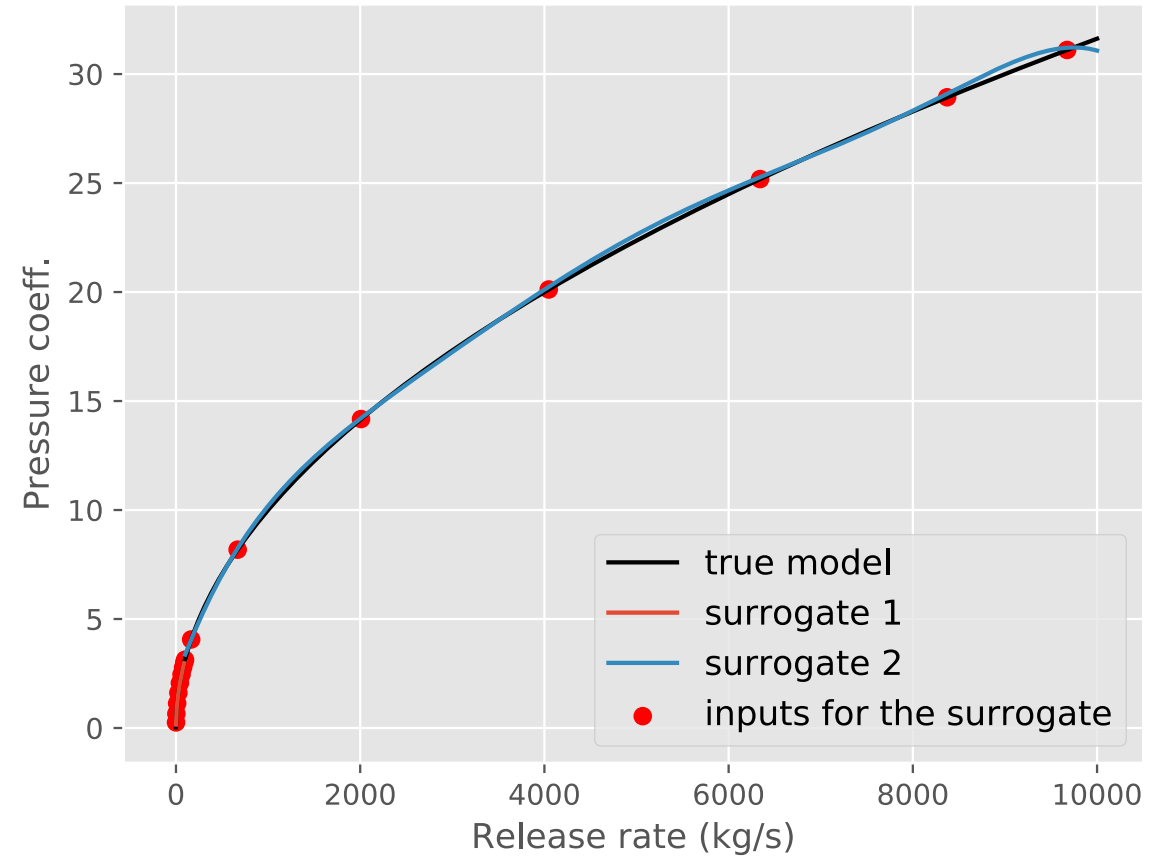
Pareto Distribution

- Considering computational expense and the desire to capture rare events, where should the cutoff value be?
- How accurate can the surrogate methods be in approximating the simulator, and the mean and variance of the outputs?
- There is often a tension between approximating the model across the support and estimating the statistical moments

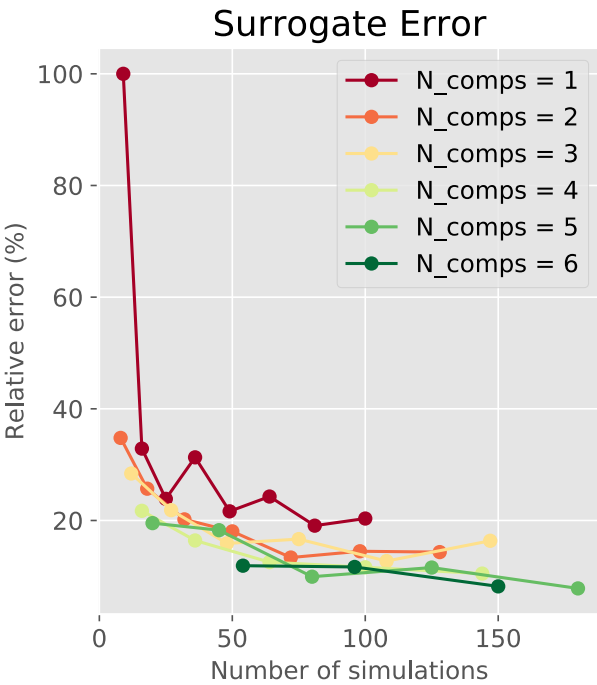
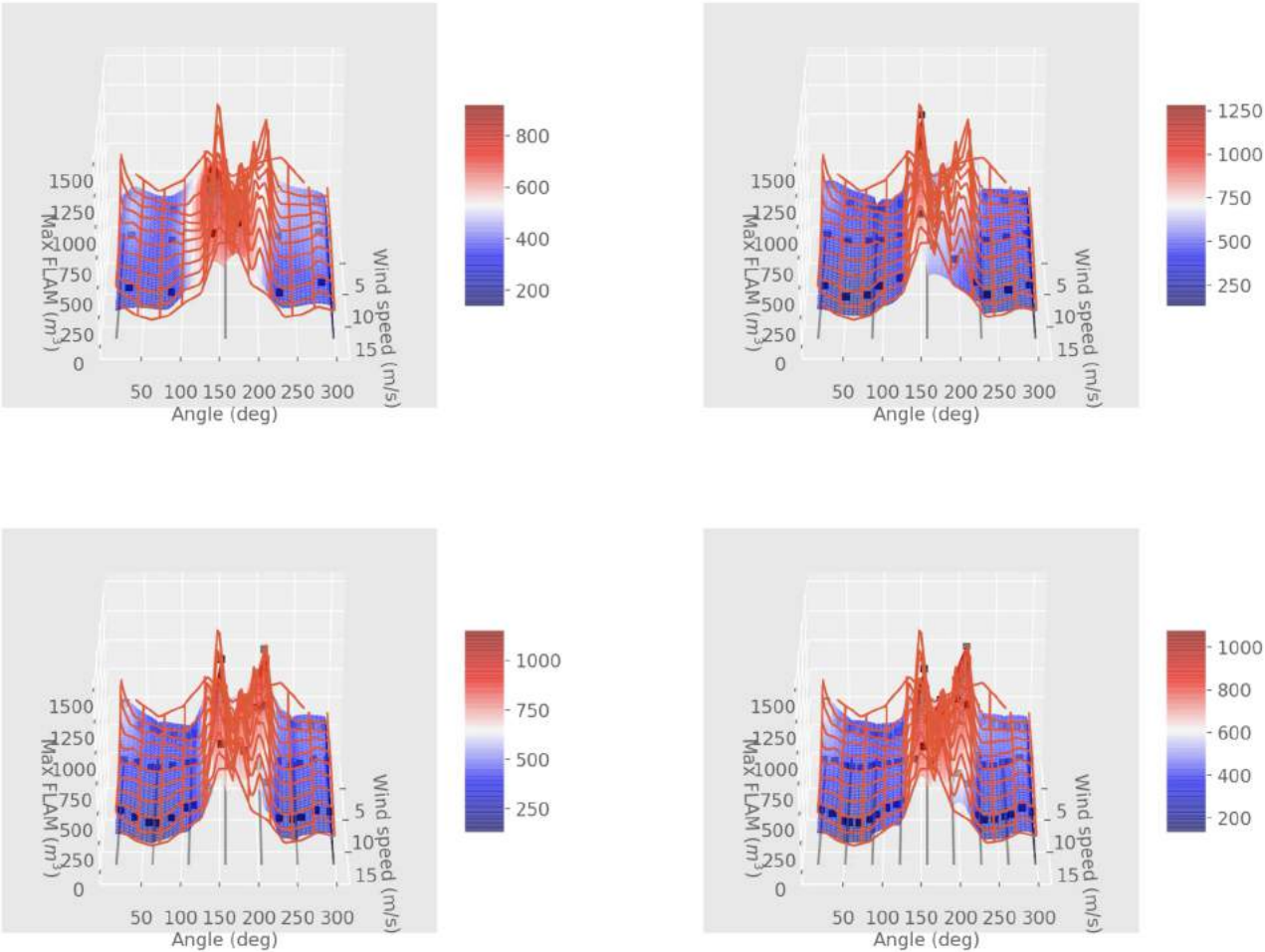


Domain Decomposition

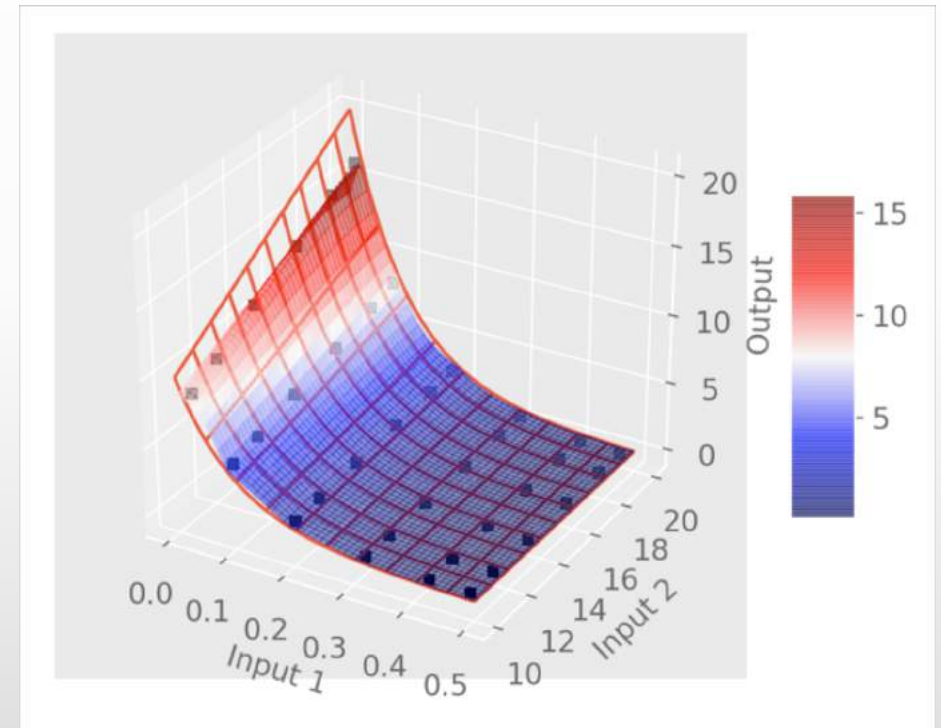
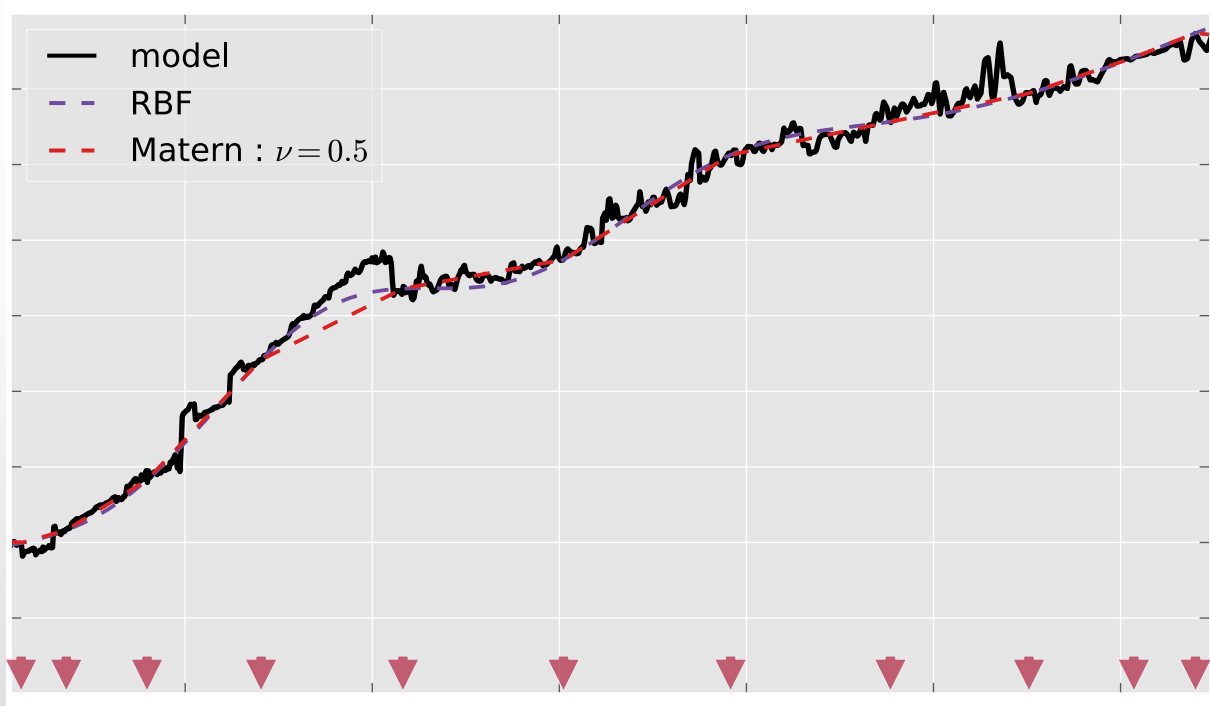
- The resulting training points poorly sampled the region where the bulk of the data lies, resulting in poor statistical estimation
- Domain decomposition provides one way to mitigate against this issue



Multi-Component Surrogate



Visualising and Quantifying Uncertainty for Higher Dimensional Simulations

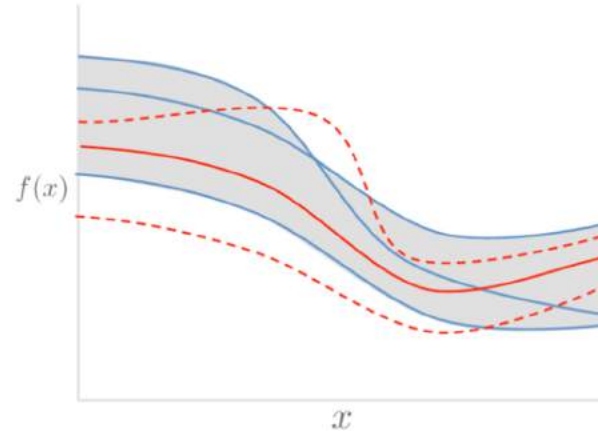


Band Depth and Functional Box Plots

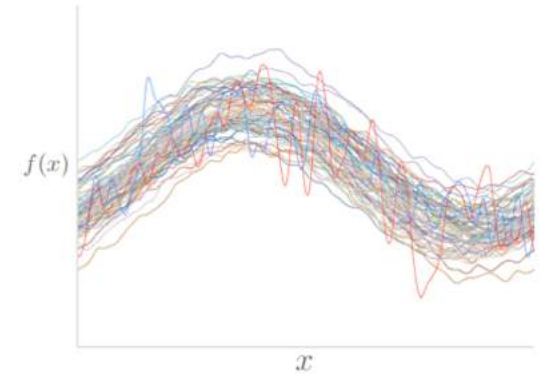
- **Functional boxplots** allow for the quantification and visualisation of uncertainty in the outputs of stochastic, time-series simulations

- $BD^j(g(x)) = P[f(x) \subset B(f_1(x), \dots, f_j(x))]$

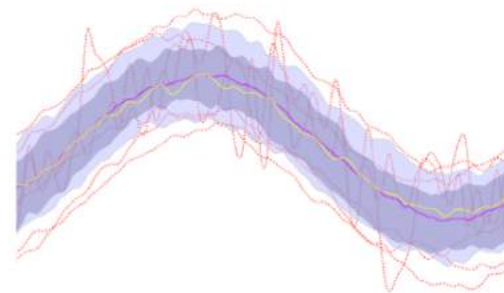
- $BD_J(g(x)) = \sum_{j=2}^J BD^j(g(x))$



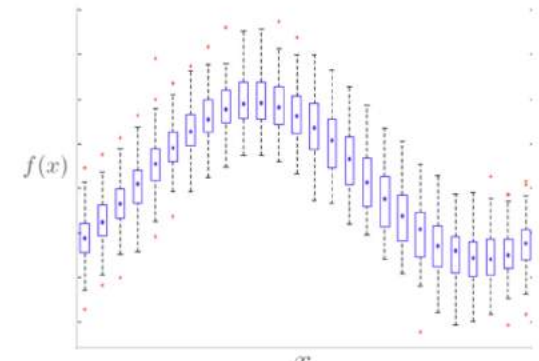
(a)



(b)

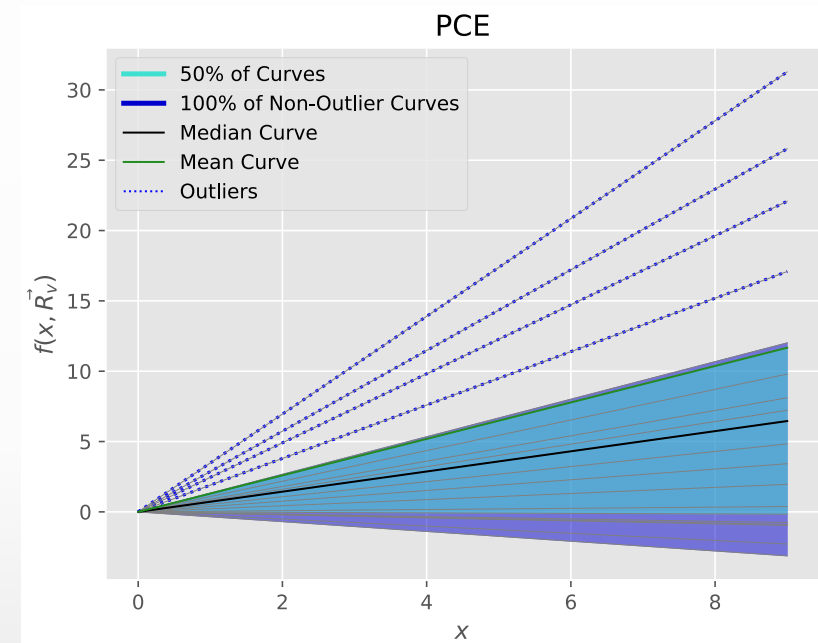
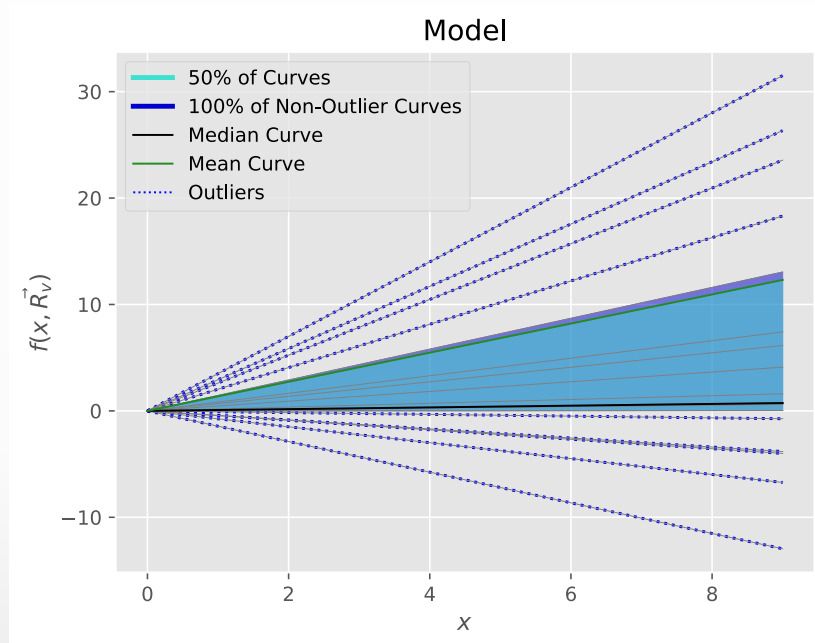


(c)



(d)

Visualising Uncertainty for Higher Dimension Inputs



- This provides a means to assess the performance of a surrogate. Not just important for UQ result, but also different modelling techniques.

Overall Experience

- Nice atmosphere, very 'uncorporate' from an interns persepective
- Range of scientific backgrounds: astronomy, chemistry, computer science etc.
- A lot of academic freedom, but not total
- Good system of accountability
- Journal clubs, seminars, conferences etc. made it feel similar to a university
- Opportunity for future collaboration