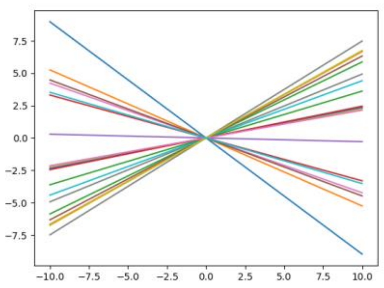
Can geologists and statisticians communicate? Generative models as a tool supporting communication, elicitation and learning.

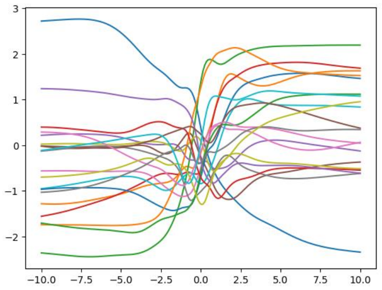
Dan Cornford, IGI Ltd

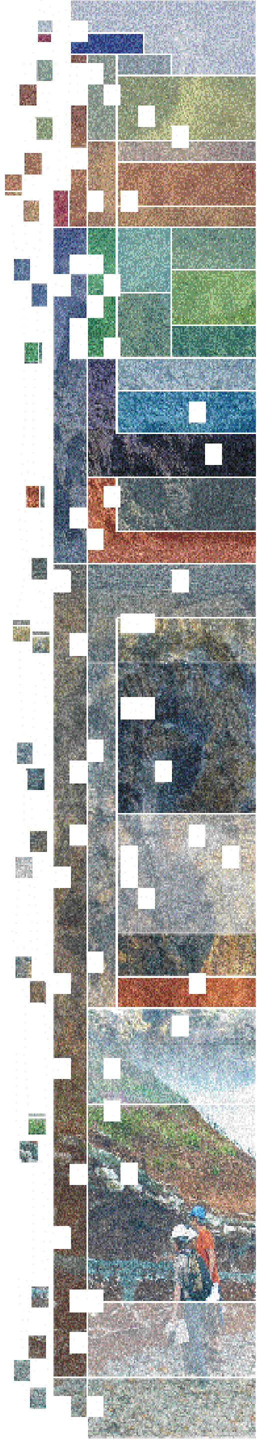
FORCE meeting on

Making good decisions under subsurface uncertainty: How difficult can it be?

6-7 Feb 2024







What's coming up

- Context
  - admitting to my own biases
- Modelling and the role of models
  - what is a model, and what properties do models have
- Inference
  - how do we learn about models from data
- Specification of priors
  - elicitation and thinking about models generatively
- Choosing appropriate models
  - should your data define your models?
- Practical advice
  - some tips if you want to approach your problems in a Bayesian manner

Context

What are my assumptions (biases)

- Why care about uncertainty?
  - typically, a decision problem drives everything
    - should we do A, B or C?
  - informed by an estimation problem and a loss function → expected loss
    - how much oil, gas, can we produce / find? What about GORs? Where is the fluid?
  - but we rarely know everything!
- Uncertainty is subjective
  - I know different things to you
    - so this means my uncertainty can be very different to yours
  - reality is not random, it just is
    - but it is imperfectly known
  - (Bayesian) probability provides a consistent framework for representing uncertainty theoretically
    - the Bayesian part is more about updating beliefs
- A model is a tool to help us understand a (decision or estimation) problem
  - physically motivated, e.g. conservation equations → partial differential equations + empirical ‘closures’
  - data driven, e.g. observations → statistical and ML models

all models are wrong, some are useful

- We rarely have real problems where we know nothing before measuring

the model represents our prior assumptions

What is a model

Are physical and statistical models different?

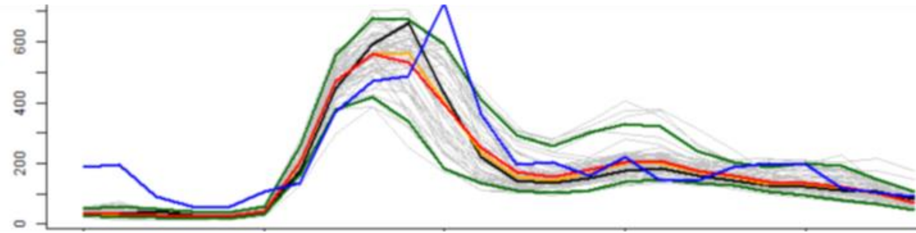
- In essence a model imposes constraints on the solution space of a given problem
    - helpful to sketch ideas in 1D, but generalises to any number of dimensions

    - consider a very general equation $y = f(x) + e$ where x is the ‘input’ and y is the ‘response’, and e is the ‘noise’
        - differential equation e.g. $dy/dt = a*d^2y/dx^2$ $[y_{t+1} = f(y_t) + e]$
        - linear in parameters regression: $y = m*r(x) + c$

    - both types impose a constraint over the admissible solutions
        - both are in essence based on ‘smoothness’ or ‘conservation’ assumptions
        - both have parameters (state) which must be estimated
    - the noise term is important too – “model error” and “observation error”

- The above conclusions generalise to all physical and statistical models
    - most physical models are **dynamic**, so relate more directly to spatio-temporal statistical models

- Given that all models are wrong, we need to talk about uncertainty
    - and all observations are wrong too…
    - fitting / training / calibrating models … all in essence inference in a probabilistic setting
        - maximum likelihood, Bayesian, Kalman filter, …



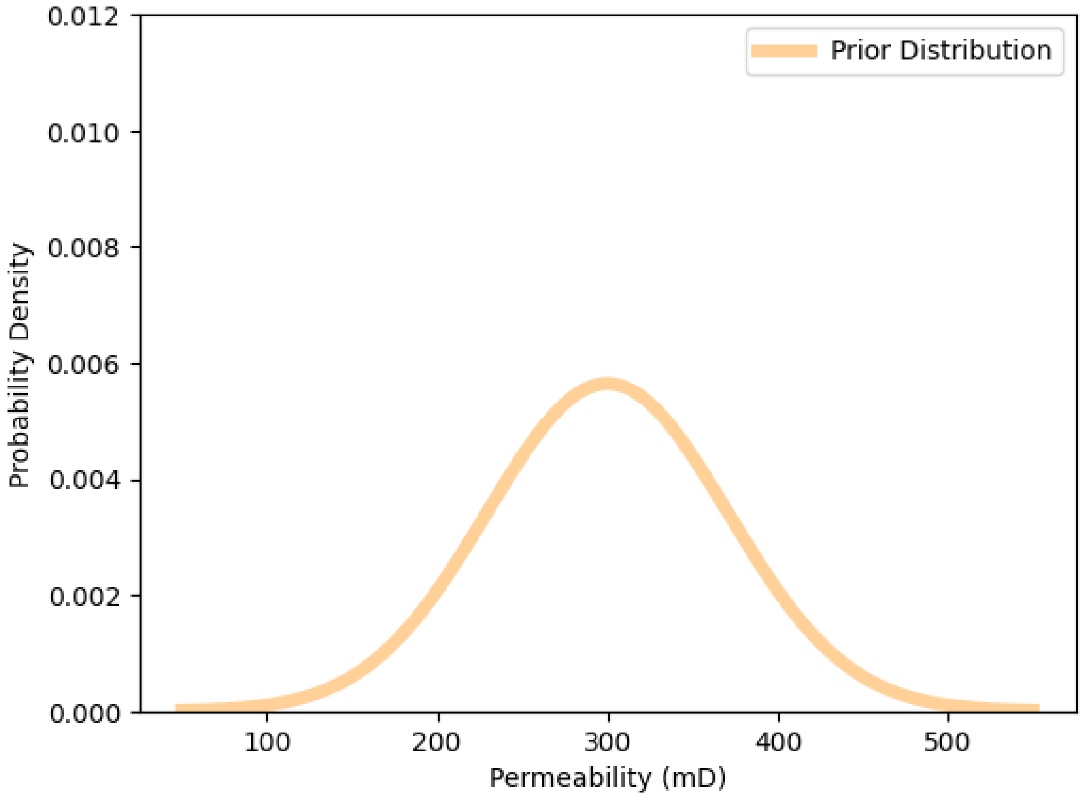
Fitting models to data

How do we use observations to constrain our judgements

- Before we see any observations, our model is an expression of our prior beliefs
  - think about models as being generative allows us to reason about our priors

- Start with something super simple
  - a 1D example thinking about permeability in a section of a reservoir – this is a truly trivial model!

We could start thinking about our prior … what might the permeability be in the absence of any data?



This requires elicitation

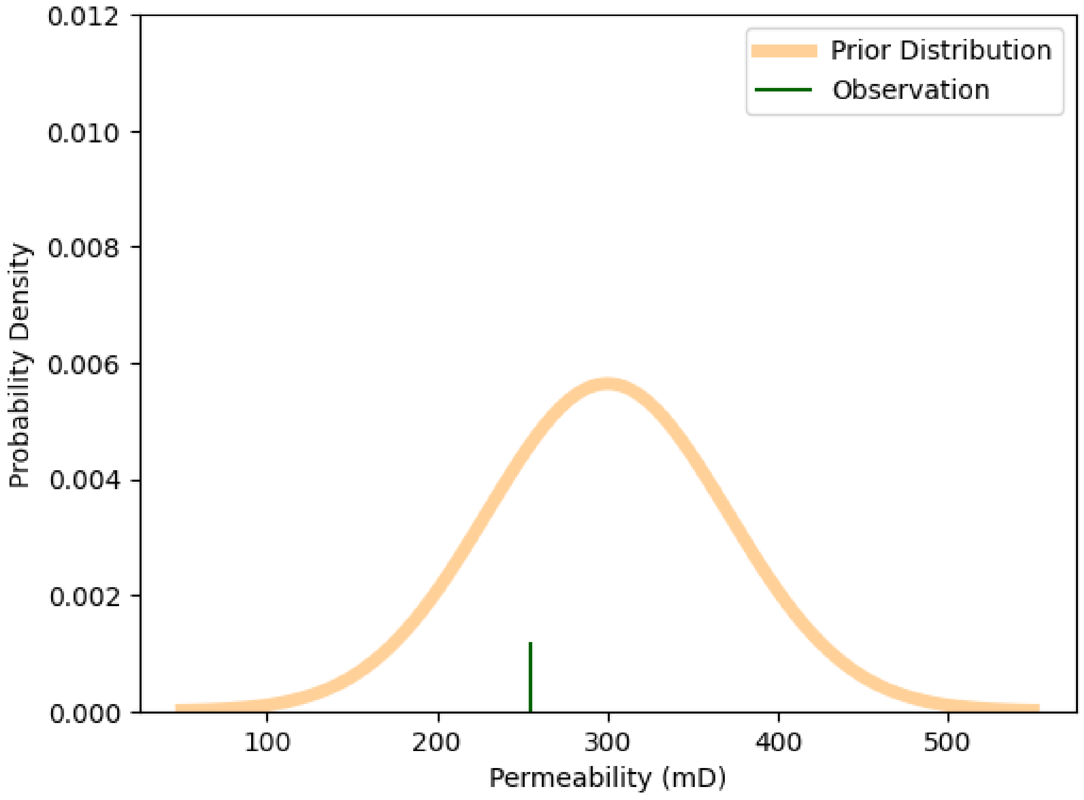
Fitting models to data

How do we use observations to constrain our judgements

- Before we see any observations, our model is an expression of our prior beliefs
  - think about models as being generative allows us to reason about our priors

- Start with something super simple
  - a 1D example thinking about permeability in a section of a reservoir

Now let's assume we can make an observation …



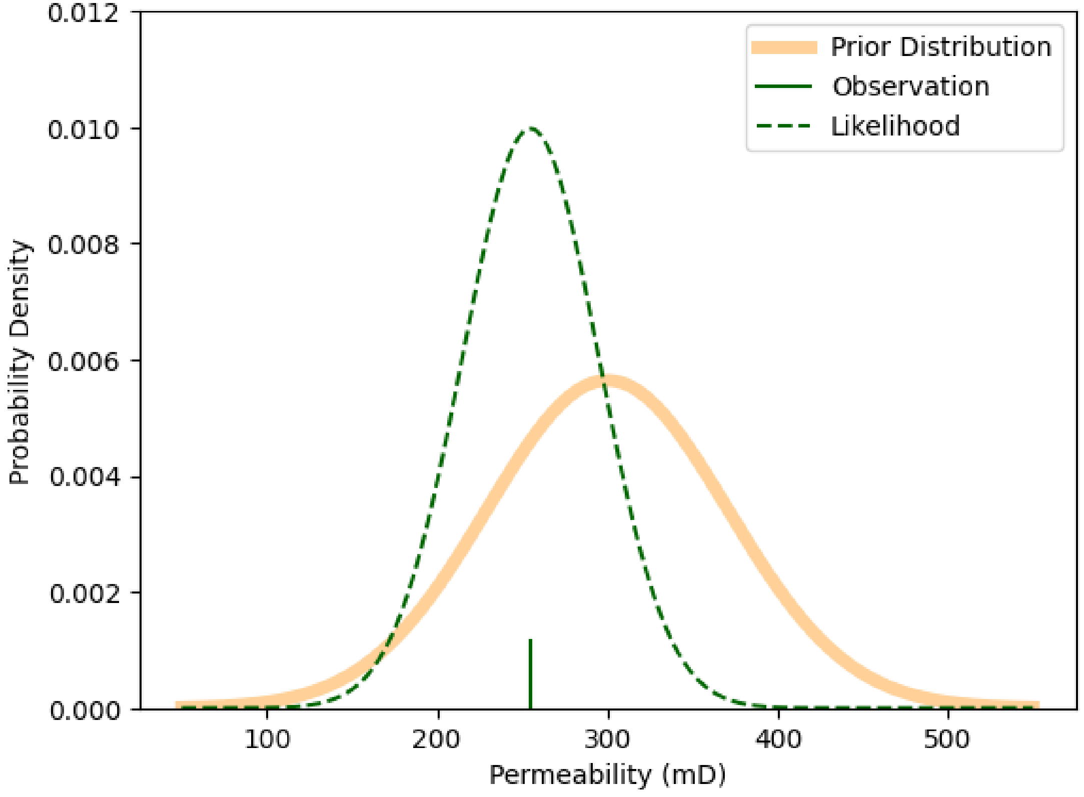
Fitting models to data

How do we use observations to constrain our judgements

- Before we see any observations, our model is an expression of our prior beliefs
  - think about models as being generative allows us to reason about our priors

- Start with something super simple
  - a 1D example thinking about permeability in a section of a reservoir

But remember the observation is likely to have some noise…



This is our likelihood function, with the noise representing all aspects of the observation errors:

measurement error,
representativity error,
observation process error

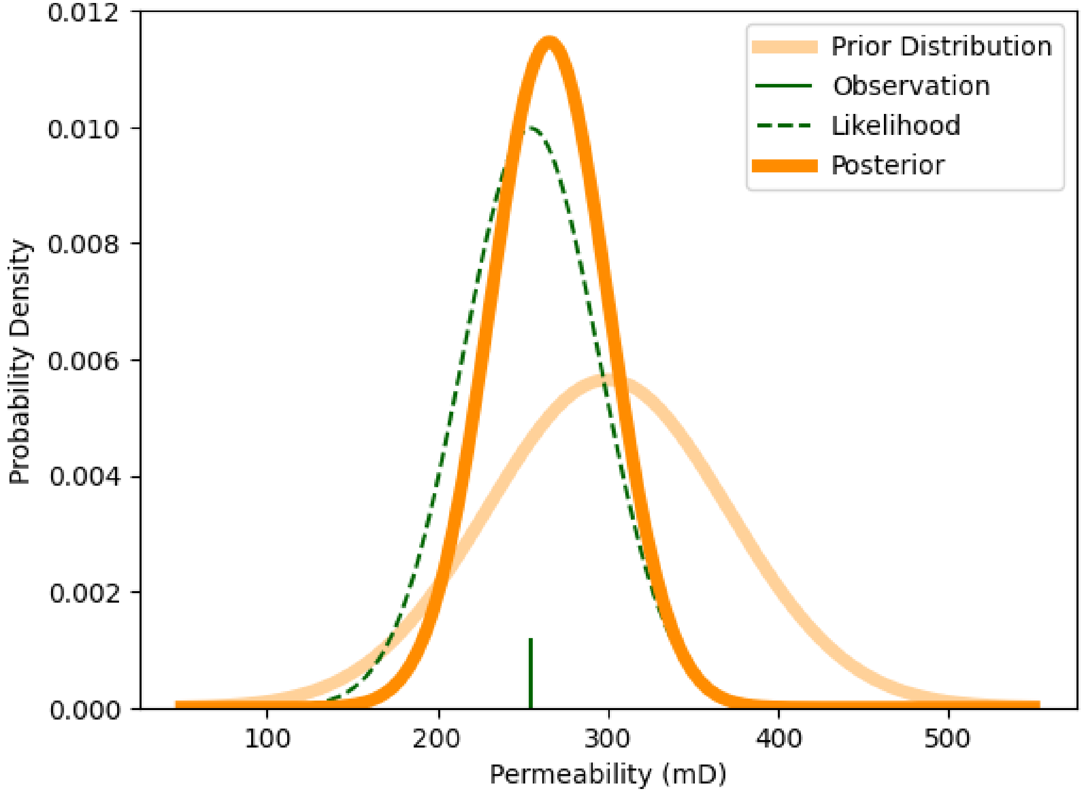
Fitting models to data

How do we use observations to constrain our judgements

- Before we see any observations, our model is an expression of our prior beliefs
  - think about models as being generative allows us to reason about our priors

- Start with something super simple
  - a 1D example thinking about permeability in a section of a reservoir

We can now update our beliefs about the permeability given the observation to provide a posterior estimate



This is Bayesian inference … but really you can just use the definition of conditional probability for this.

The thing that really makes it Bayesian is the interpretation of the distributions!

What is elicitation?

The art of specifying priors

- This talk is not about elicitation, but we need to consider some of the challenges

- Elicitation of (expert) uncertainty is well studied, especially for univariate problems
  - e.g. the SHELF tool from Jeremy Oakley and Tony O'Hagan:

Multiple experts

Eliciting individual distributions from multiple experts. Includes methods for mathematical aggregation using linear pooling.

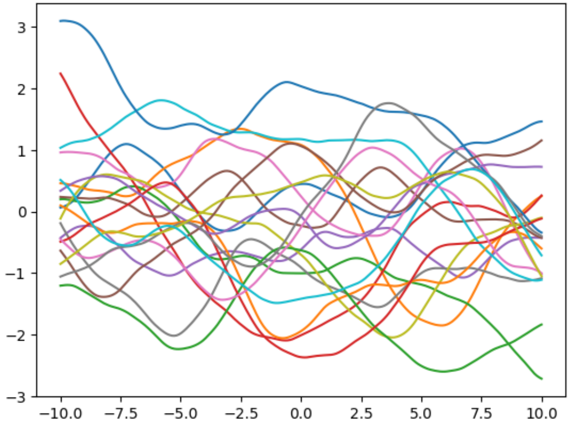
Access multiple experts online

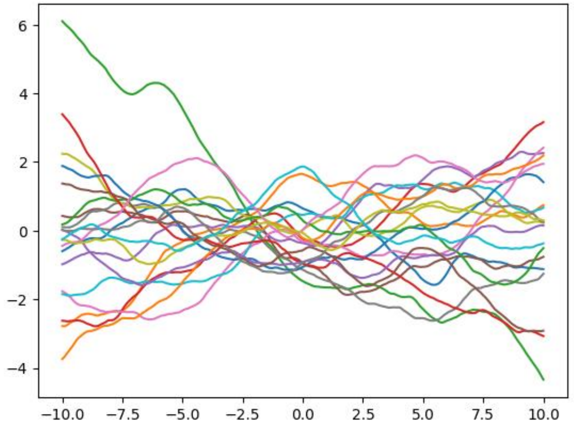
  - in oil and gas, Martin Neumaier et al have done nice work with their 'dancing distributions' ArianeLogiX
- However there remain many challenges
  - handling multivariate problems, e.g. spatial fields, complex models, correlated variables
  - this is where thinking generatively really helps

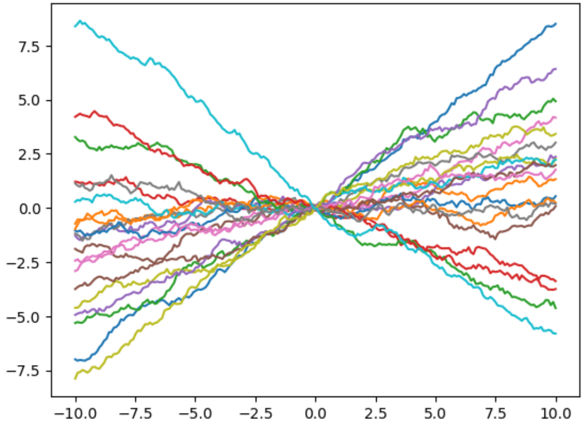
Thinking generatively

How to understand your models

- A model is a mapping from inputs x, to outputs y: y=f(x;w) + e
  - it will typically depend on parameters w
    - regression coefficients, diffusion parameters, forcing / boundary values, etc
      - without going into detail, I'd also include hyper-parameters here …
- Imagine we want to 'think about what f()' looks like before we see data
  - this requires us to simulate (Monte Carlo, if you like) from the possible f's that correspond to our beliefs







- Thinking about models generatively helps you understand what they can do, what their assumptions are and whether they are relevant to what you are trying to do
  - it can be a challenge to visualise in many dimensions, but is worth doing where possible
  - you still need to elicit your beliefs about the uncertain quantities

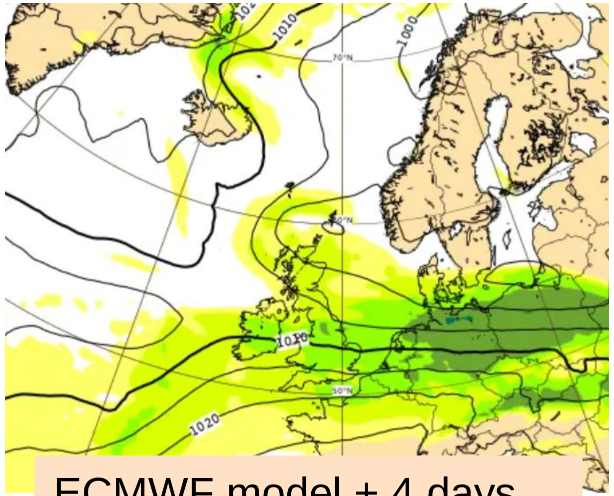
Model complexity and learning

How can you use data to reduce uncertainty

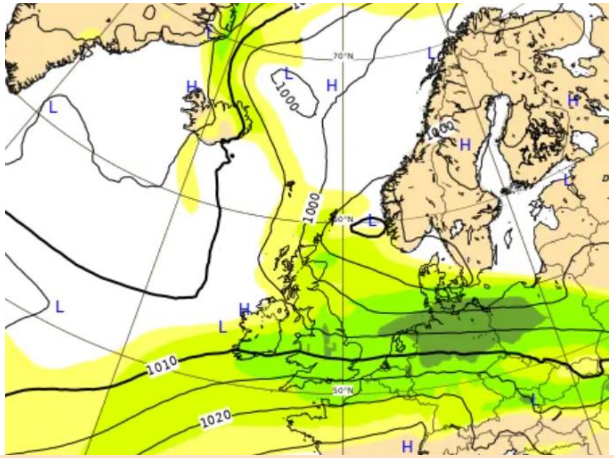
- Beware the power modeller
  - excited by the complexity of their model, often use terms like:
    - full physics
    - high resolution
- These are not intrinsically bad things but ask yourself two questions
  - what do I actually care about?
  - does my knowledge and observations support this complexity?

Digital ~~twins~~ cousins remain popular concepts

- Imagine weather forecasting … if I had to decide whether to go fishing in my boat next weekend …
  - in general I would use a physics driven model …
  - and if I wanted to know whether it will rain here in the next 30 minutes I'd use a data driven model …



ECMWF model + 4 days



GraphCast ML model + 4 days

… but recently ML based models have been shown to perform as well as, if not better than, physics-based models, at least on some key metrics

From: https://charts.ecmwf.int/

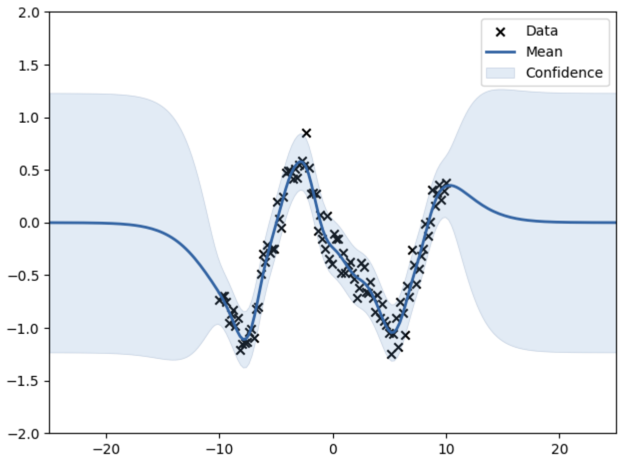
Big data?

When do your (model) assumptions really matter?

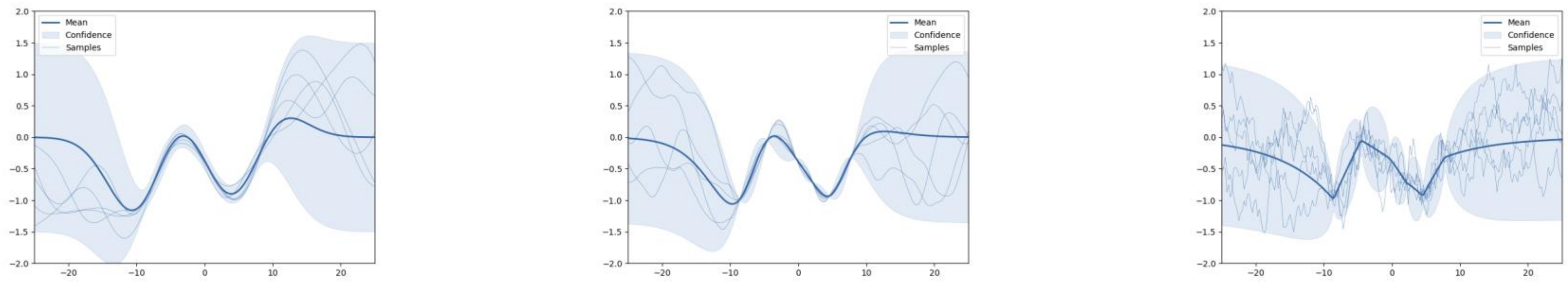
- We’ve seen lots of news around large language models … ChatGPT etc …
  - deep learning has been popular in the ML community for some years
  - but … I’d argue that is not relevant to a lot of problems we face

- If you are Amazon, Microsoft, or Google you probably have big data …
  - but big is relative to the problem space, and the data information content
  - we are considering the subsurface with large spatial (x,y,z) variability and very few direct measurements
    - using remote sensing, e.g. seismic, can help, but has its own challenges – that’s another talk!

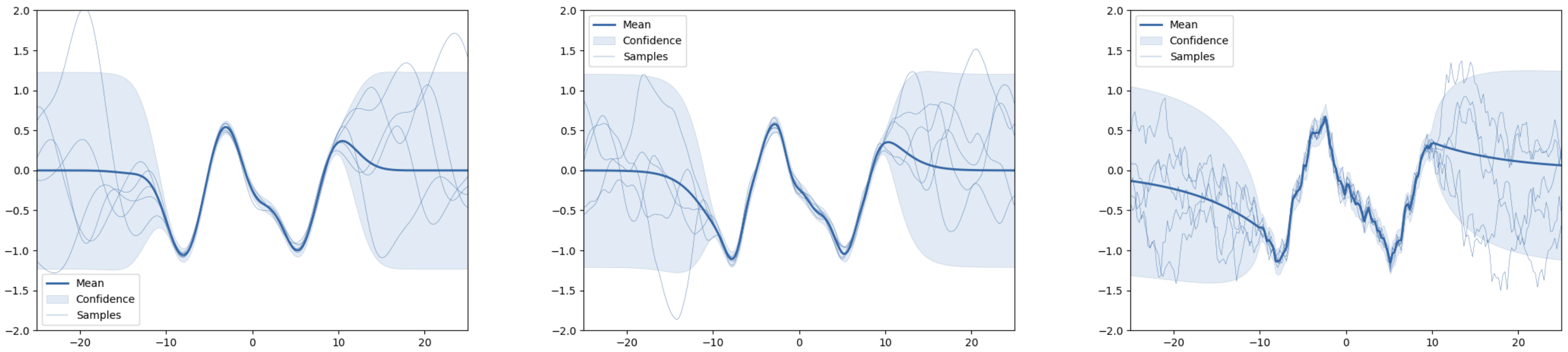
Interpolation is less dependent on the prior, but this is not true for extrapolation



Small data



Big data

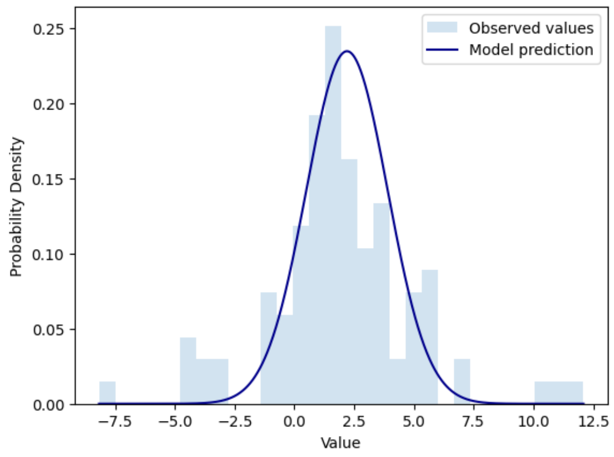


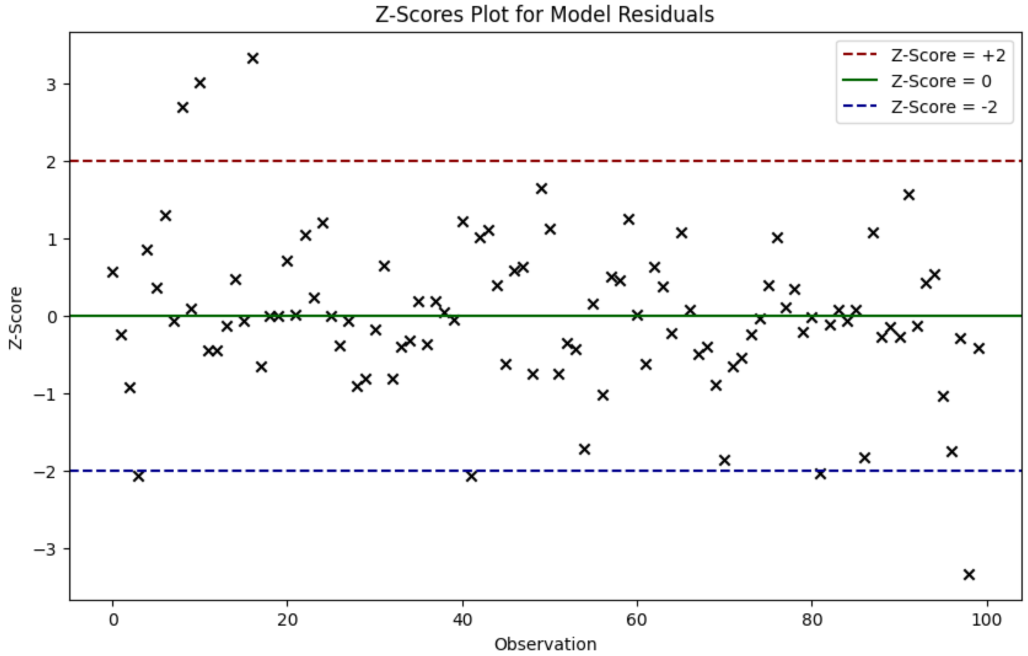
Probability: calibration vs sharpness

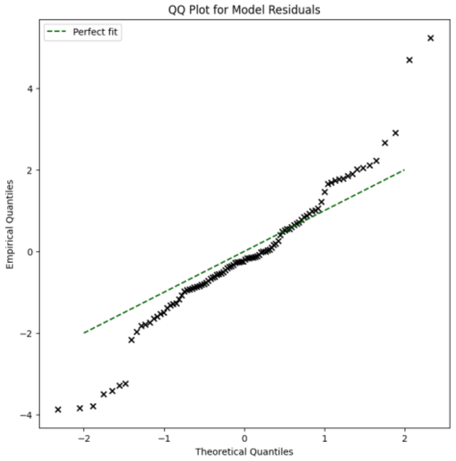
Approximately right vs precisely wrong?

- Probabilistic methods seek to produce the sharpest distribution, subject to calibration
  - when we model, too often we only look at the expectation (mean) and related metrics (RMSE, Std Error)

- Checking for calibration is tricky
  - you need repeated measurements to validate your probabilistic model
  - but reality only happens once so you need to consider ergodicity or exchangeability (c.f. iid)
    - a sufficiently large collection of random samples from a process represents the statistical properties of the entire process
  - basically, can you assume errors from different locations / times can be compared







A list of questions to ask (yourself)

Avoiding bias and making decisions

- We don't see our biases – blind spot effect
- We often put more weight on the first piece of information we see – anchoring bias
  - similarly, some prefer 'trusted' older data – conservatism bias
  - while some put undue weight on more recent data – recency bias
- We are prone to over-weighting data supporting our view – confirmation bias
  - this extends to us seeing what we expect as being more important – choice-supportive bias
  - and tending to put more weight on our successes – survivor bias
- We also tend to be influenced by others – bandwagon effect

- Overall, while we try to be objective, almost all studies show that even if you think you are, you are probably wrong!

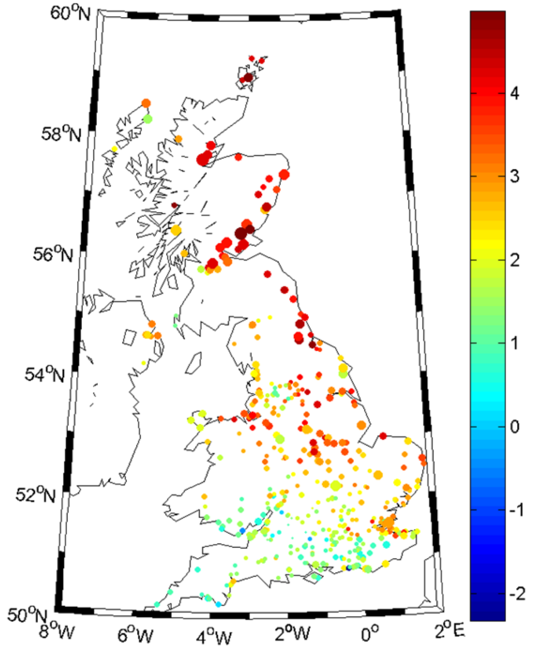
all people are biased, some are honest

- It is important to note that these biases are not malicious
  - the best method I know to minimise them is to justify your evidence publicly
  - and work in teams to challenge each other

Practical implications

Keep it simple …

- Elicitation and bias
  - where possible involve multiple experts
    - try to be aware of your biases
  - once you have elicited your priors treat the models generatively
    - do the outcomes look plausible?
- Selection of models
  - prefer simpler explanations over complex ones (Occam’s razor)
    - case by case consideration, depending on level of knowledge and available data
  - don’t forget about model error!
  - emulation / surrogate models provide a solution to using numerically intensive models
- Design of experiments
  - if you can, don’t forget the power of choosing where to observe
    - careful selection of measurements can be just as important as selection of good models
- Validation – calibration …
  - don’t only focus on expectation
    - try to validate the uncertainty too



Key take homes

- Uncertainty is subjective – it depends on what you know, so think of beliefs
    - probability is a natural, consistent framework to represent uncertainty
        - be aware of potential biases
    - **reality is not uncertain**, but is unknown

- Being Bayesian is a state of mind
    - not just $p(y|X) \propto p(X|y)p(y)$

- The goal is often a decision problem → cost/loss
    - solving your specific problem is often simpler than solving all problems!
        - so you may be better severed building a model to solve each problem
- Prefer models you can understand, or at least have some intuition about
    - you can't elicit what you don't know about
    - linear (in parameter) models where plausible
        - as simple as possible, but no simpler
    - treat your model **generatively** to check your assumptions
- If you have a lot of data, you can worry less about your model
    - but extrapolation will always depend on your model

Thanks for your attention