Blending mathematical models with observational data

School of Mathematics, University of Edinburgh

Dr Aretha Teckentrup

PiWORKS Seminar, June 7 2022

THE UNIVERSITY of EDINBURGH
School of Mathematics

A little bit about me

My career path:

- MMath at University of Bath, 2005-2009. Focussing on applied mathematics/statistics
- PhD at University of Bath. 2009-2013. Focussing on multilevel Monte Carlo methods for random partial differential equations (PDEs)
- Postdoc at Florida State University, 2013-2014. Focussing on multilevel interpolation for random PDEs
- Postdoc at University of Warwick, 2014-2016. Focussing on inverse problems in PDEs
- Lecturer at University of Edinburgh, 2016-2022
- Reader at University of Edinburgh, 2022-

My research is in computational mathematics, at the interface of statistics, numerical analysis and data science.

Outline of talk

The topic of my talk today is **mathematical data science**, with a particular focus on combining mathematical models with observational data.

I will:

- Give a general overview of topics in the mathematics of data science.
- Discuss an example in numerical weather prediction.
- Discuss an example in geophysics.

Mathematical data science

Data science

The recent explosion in data, driven by the increase in large-scale scientific experiments and the development of sensor technology, is bringing new challenges to the forefront.

Enabling evidence-based decision making requires tools to inform decisions, assess risk and formulate policies based on available evidence.

Data science

Engineering
Natural Sciences
...

Social Sciences
Technology
...

sparse data — vast data

strong mathematical models — lack of mathematical models

Data science

Engineering
Natural Sciences
...

Social Sciences
Technology
...

sparse data — vast data

strong mathematical models — lack of mathematical models

- *Model calibration*: using real data to tune parameters in the model (e.g. reservoir modelling and history matching)
- *Filtering*: blending a dynamical system with sequentially observed data (e.g. weather prediction)
- *Constrained optimisation/Optimal control*: steering the model to a desired state (e.g. aerodynamic design)
- Bayesian inference, high-dimensional sampling, Markov chain Monte Carlo, high-dimensional optimisation

Data science

Engineering
Natural Sciences
. . .

Social Sciences
Technology
. . .

sparse data
strong mathematical models

vast data
lack of mathematical models

- *Data Analysis* and *Feature Extraction*: what patterns are there in the data? (e.g. image processing in medicine and astrophysics)
- *Machine Learning* and *Statistical modelling*: learning a model of the underlying process from data (e.g. healthcare and environmental applications)
- *Software at scale*: modern applications require scalable software

- Neural networks, deep learning, kernel machines, clustering, graphical models, TensorFlow, PyTorch

Data science

Enabling evidence-based decision making presents a demand for accurate, reliable predictions and inferences obtained from the evidence.

Data science

Enabling evidence-based decision making presents a demand for accurate, reliable predictions and inferences obtained from the evidence.

In areas as diverse as climate modelling, manufacturing, energy, life sciences, finance, geosciences and medicine, the available evidence typically comes in the form of observed data together with mathematical models that govern the underlying processes.

Data science

Enabling evidence-based decision making presents a demand for accurate, reliable predictions and inferences obtained from the evidence.

In areas as diverse as climate modelling, manufacturing, energy, life sciences, finance, geosciences and medicine, the available evidence typically comes in the form of observed data together with mathematical models that govern the underlying processes.

My research develops tools and methods that allow for efficient prediction and risk assessment through the integration of real-world data with complex mathematical models, taking into account any source of uncertainty that may influence the accuracy of our outcomes (such as noise in the observed data or incomplete knowledge of the physical system).

Example I:
Numerical weather prediction

Numerical weather prediction

A typical example is in numerical weather prediction, where we have:

- Observational data from weather balloons, satellites, buoys, …
- A system of partial differential equations governing the evolution of the state of the atmosphere (and ocean) in time and space.

Numerical weather prediction

A typical example is in numerical weather prediction, where we have:

- Observational data from weather balloons, satellites, buoys, …
- A system of partial differential equations governing the evolution of the state of the atmosphere (and ocean) in time and space.

To predict the weather in the future, both are crucial:

- Predicting tomorrow's weather knowing only today's weather is impossible.
- Predicting tomorrow's weather without knowing today's weather is impossible.

Numerical weather prediction

A typical example is in numerical weather prediction, where we have:

- Observational data from weather balloons, satellites, buoys, …
- A system of partial differential equations governing the evolution of the state of the atmosphere (and ocean) in time and space.

To predict the weather in the future, both are crucial:

- Predicting tomorrow's weather knowing only today's weather is impossible.
- Predicting tomorrow's weather without knowing today's weather is impossible.

How are we combining data and mathematical model in this case?

- We are using the data to sequentially update the initial conditions for our model.

Numerical weather prediction

Our data is partial and noisy, meaning that there are many possible initial conditions that are consistent with the observed data. Which one should we pick?

Numerical weather prediction

Our data is partial and noisy, meaning that there are many possible initial conditions that are consistent with the observed data. Which one should we pick?

To mitigate risk and quantify uncertainty, we use ensemble methods:

1. Choose an ensemble of N initial conditions consistent with the observed data. (Often chosen randomly).

2. Run the simulation with your PDE model for each of the N initial conditions.

3. Take the ensemble of N predictions made, and use summary statistics such as mean, variance, quantiles, …

Numerical weather prediction

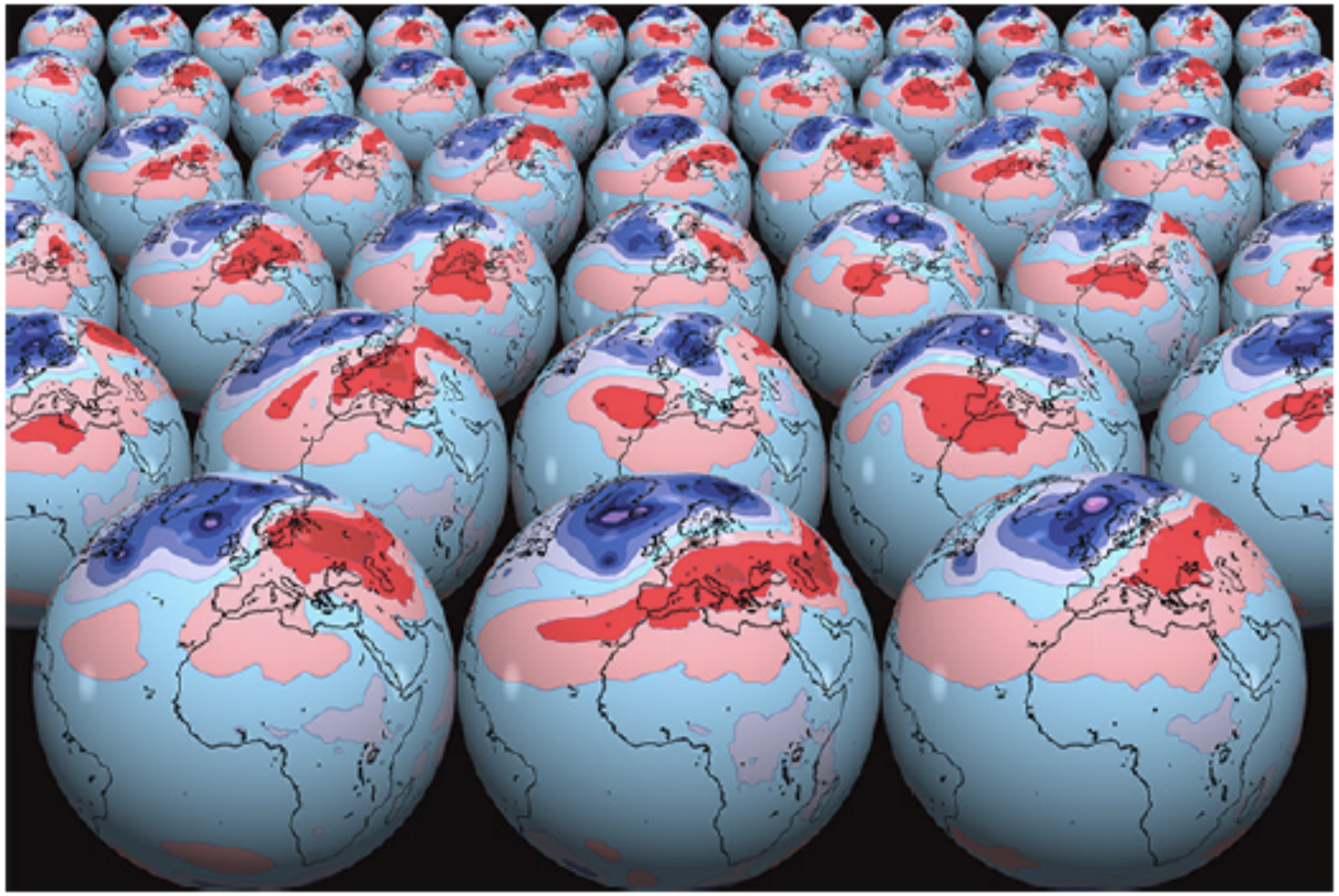
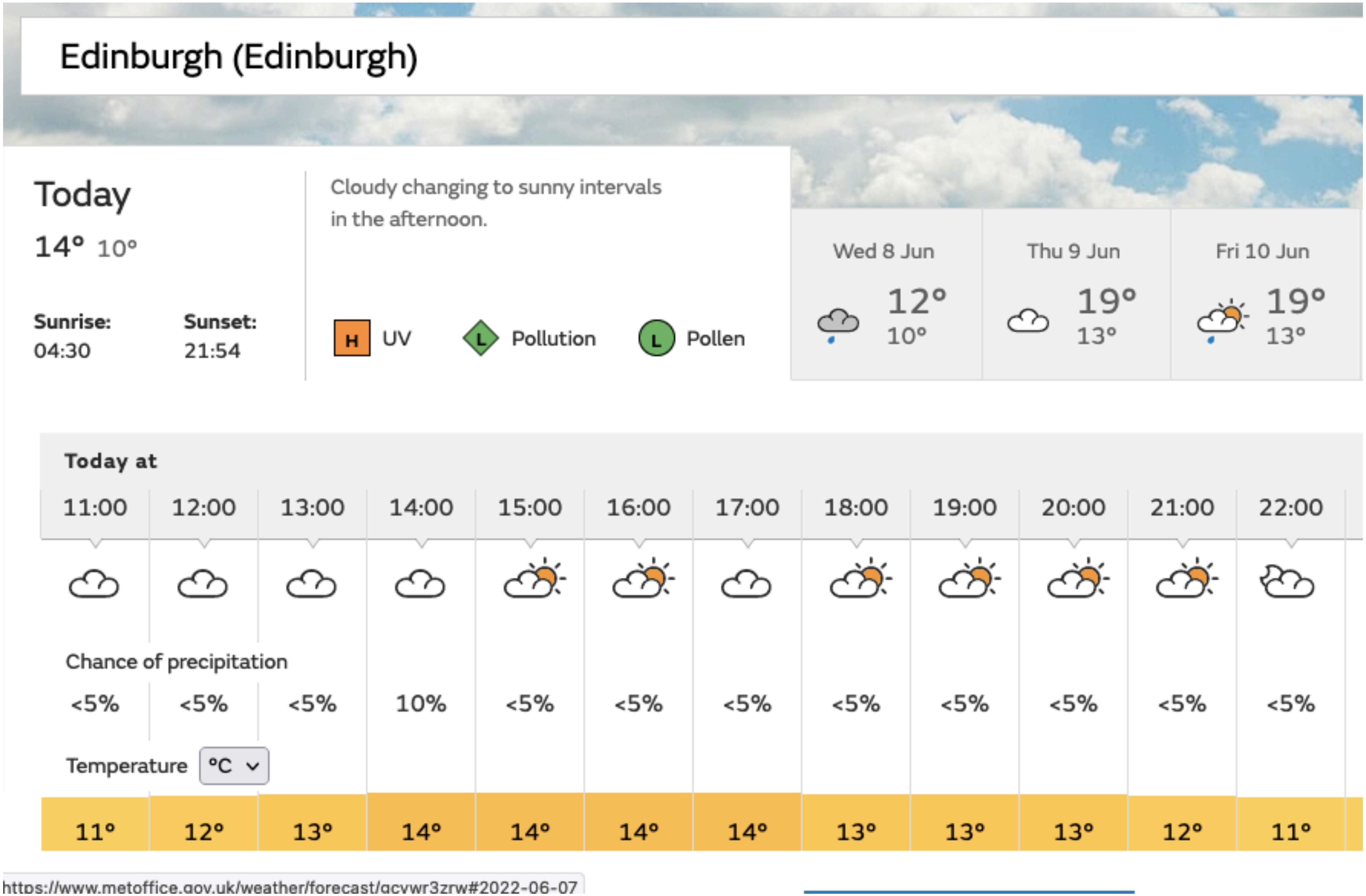


Figure: Source https://www.ecmwf.int

Ensemble of possible future weather predictions

Numerical weather prediction



Numerical weather prediction

- To get accurate predictions, you need to run a large number of expensive simulations.
- The ensemble method to compute $Q = \mathbb{E}[\phi(p)]$ results in the Monte Carlo estimator

$$\widehat{Q}^{\mathrm{MC}}_{h,N} := \frac{1}{N}\sum_{i=1}^{N} \phi(p_h^{(i)})$$

- The mean square error of this estimator is

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{Q}^{\mathrm{MC}}_{h,N} - \mathbb{E}[\phi(p)]\right)^2\right] &= \mathbb{V}[\widehat{Q}^{\mathrm{MC}}_{h,N}] + \left(\mathbb{E}[\widehat{Q}^{\mathrm{MC}}_{h,N}] - \mathbb{E}[\phi(p)]\right)^2 \\ &= \underbrace{\mathbb{V}[\phi(p_h)]\, N^{-1}}_{\text{sampling error}} + \underbrace{\left(\mathbb{E}[\phi(p_h) - \phi(p)]\right)^2}_{\text{FE error (“bias”)}} \end{aligned}$$

Numerical weather prediction

- To get accurate predictions, you need to run a large number of expensive simulations.
- The ensemble method to compute $Q = \mathbb{E}[\phi(p)]$ results in the Monte Carlo estimator

$$\widehat{Q}_{h,N}^{\mathrm{MC}} := \frac{1}{N} \sum_{i=1}^{N} \phi(p_h^{(i)})$$

- The mean square error of this estimator is

$$\begin{aligned} \mathbb{E}\Big[\big(\widehat{Q}_{h,N}^{\mathrm{MC}} - \mathbb{E}[\phi(p)]\big)^2\Big] &= \mathbb{V}[\widehat{Q}_{h,N}^{\mathrm{MC}}] + \big(\mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] - \mathbb{E}[\phi(p)]\big)^2 \\ &= \underbrace{\mathbb{V}[\phi(p_h)]\, N^{-1}}_{\text{sampling error}} + \underbrace{\Big(\mathbb{E}[\phi(p_h) - \phi(p)]\Big)^2}_{\text{FE error (“bias”)}} \end{aligned}$$

- We are currently working on reducing the computational effort required to achieve a given mean square error using multilevel methods.

Numerical weather prediction

- Multilevel methods use simulations of varying numerical accuracy, e.g. using coarse and fine meshes, and do the bulk of simulations with coarse meshes. A few simulations with fine meshes are then used to "correct" the predictions made with coarse meshes:

$$\mathbb{E}\left[\phi(p_{h_L})\right] = \mathbb{E}\left[\phi(p_{h_0})\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[\phi(p_{h_\ell}) - \phi(p_{h_{\ell-1}})\right]$$

$$\longrightarrow \widehat{Q}^{\mathrm{ML}}_{\{h_\ell, N_\ell\}} = \frac{1}{N_0} \sum_{i=1}^{N_0} \phi(p_{h_0}^{(i,0)}) + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \phi(p_{h_\ell}^{(i,\ell)}) - \phi(p_{h_{\ell-1}}^{(i,\ell)})$$

- Terms are estimated independently.

- Individually, the L+1 estimators are cheap to compute since:

  - $\phi(p_{h_0}^{(i,0)})$ is cheap to compute (on a coarse grid),

  - N_ℓ $(\ell > 0)$ can be chosen small (estimating a correction).

Example II:
Modelling subsurface flow

Modelling subsurface flow

Another example is in subsurface flow, where we have:

- Observational data from drilling wells and taking measurements
- A system of partial differential equations governing the flow of water (and/or oil and gas) underground. Darcy's law plus conservation of mass.

$$-\nabla \cdot (k(x)\nabla p(x)) = g(x),$$

Permeability k, pressure head p, sources/sinks g

Modelling subsurface flow

Another example is in subsurface flow, where we have:

- Observational data from drilling wells and taking measurements
- A system of partial differential equations governing the flow of water (and/or oil and gas) underground. Darcy's law plus conservation of mass.

$$-\nabla \cdot (k(x)\nabla p(x)) = g(x),$$

Permeability k, pressure head p, sources/sinks g

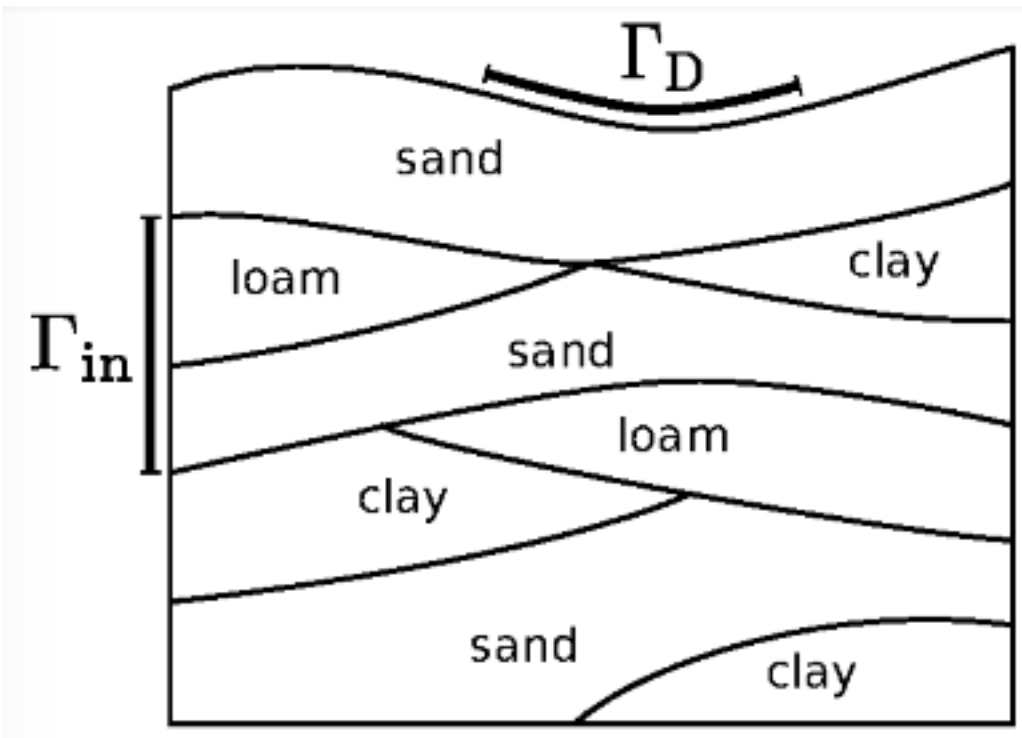
- This has applications in nuclear waste disposal and carbon storage underground, where we need to quantify the risk of leaks back into the human environment.
- We typically want to compute a quantity related to p, e.g outflow through right boundary:

$$-\int_0^1 k \frac{\partial p}{\partial x_1}\Big|_{x_1=1} \mathrm{d}x_2$$

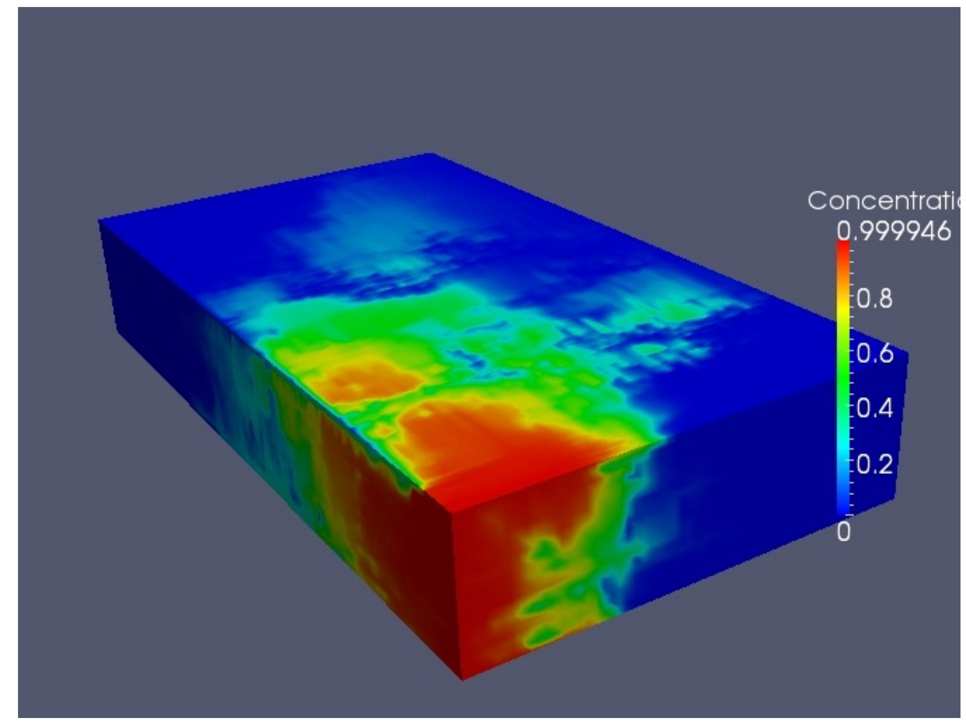
How are we combining data and mathematical model in this case?

- We are using the data to infer the diffusion coefficient k in our model.

Modelling subsurface flow



Possible configuration of subsurface geology



Spread of pollutant in a porous medium

Modelling subsurface flow

Consider the problem of predicting the outflow of water through parts of the boundary of an aquifer:

$$Q := -\int_0^1 k \frac{\partial p}{\partial x}\bigg|_{x_1=1} \, dx_2.$$

Modelling subsurface flow

Consider the problem of predicting the outflow of water through parts of the boundary of an aquifer:

$$Q := -\int_0^1 k \frac{\partial p}{\partial x}\Big|_{x_1=1} \, dx_2.$$

This can be computed in the following way:

1. Assign a probability distribution to k that is consistent with expert knowledge and observed data on k and p. This is usually done using Bayesian statistics.

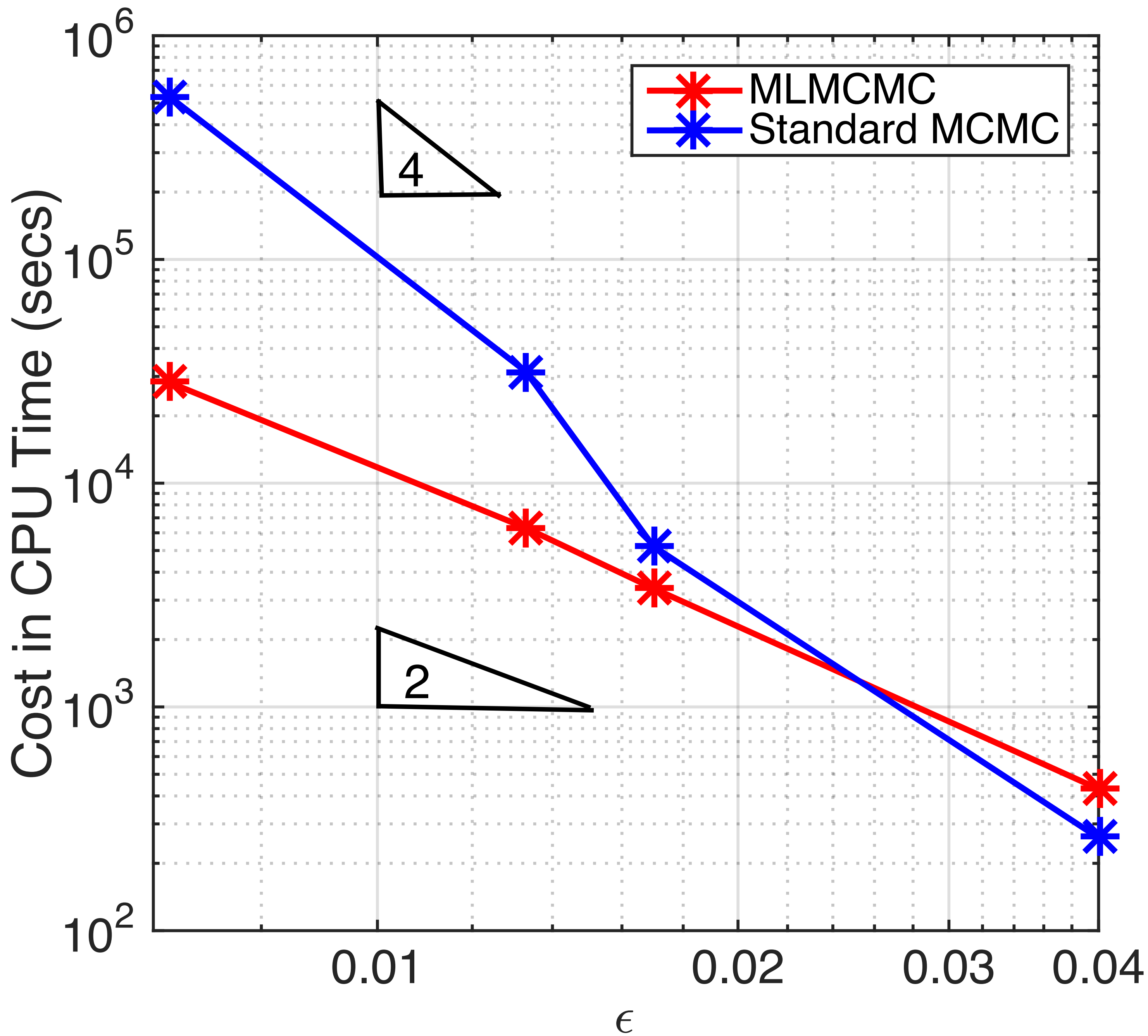
2. Sample from the distribution on k using Markov chain Monte Carlo methods, such as Metropolis Hastings.

3. Compute the outflow Q for each value of k .

4. Compute the ensemble average to get an estimate for Q.

Modelling subsurface flow

Standard methods quickly become infeasible for large scale applications, and a new state-of-the-art is required. 140 hours vs 5 hours



Summary

In the era of big data, there are many new challenges for mathematicians to tackle.

Blending mathematical models with observational data requires efficient algorithms that can cope with complex, real-world systems and heterogeneous data.