Modeling Uncertainty in the Earth Sciences

Or, Why I Fear the Priors and Nonlinearities and Model Coupling and Numerical Error Propagation

Based on Jef Caers book with same title
http://uncertaintyES.stanford.edu
This tutorial is more about questions than answers

How can we possibly know what we don't know?

Why would one model of uncertainty be any better than another?

What types of uncertainty are there?

What are ways in which people try to get anywhere?

And many other unanswered questions

Uncertainty quantification: a philosophy and a thorny challenging issue particularly for the nonlinear problems we encounter in practice.
Why I am interested in uncertainty

Performance prediction and optimization of reservoir processes:
Highly nonlinear processes
  Multi-scale, with iffy subgrid scale models that are calibrated
  Multi-objectives
  sensitivity of objectives to data
  Data and model uncertainties
  Various classes of observations
  Numerical models prone to error
  Error behavior not always known
  Unknown error propagation in model

Highly desirable to do more than just guess what decisions are best
Even deterministic models are challenging

- Multi-scale
- Strongly nonlinear
- Complex domains
- Unstructured data
- Ill-conditioned matrices
- \( O(10^7 - 10^8) \) unknowns
  - not uncommon

So models are:
- Large
- Nonlinear
- Ill-behaved
- highly sensitive to perturbations
A SOURCE OF UNCERTAINTY: GEOLOGY

Have some prior knowledge of structure
Create realizations that fit the observations

Make decisions under these and other uncertainties
Uncertainty: incomplete understanding of what we like to quantify

How do we quantify what we do not know?
Quantifying what we do not know is subjective and can not be tested

There is no true uncertainty
There are only models of uncertainty
And ultimately they are only as good as the intuition of the designer
Sources of uncertainty

Typically, two sources of uncertainty are considered:

1. As a result of process randomness (inherent randomness in nature)

2. As a result of limited understanding
   • Measurement errors
   • Insufficient data to understand behavior (miss out on features)
   • Subjective and different interpretations
   • Unknown processes, both physical and computational
   • Unmeasurable data
Information and knowledge (experts) informs all of the above
Interaction/symbiosis, always, between data and models:
You process data in a way that reflects your knowledge
You create a model in a way that reflects your knowledge
Bayes can provide a framework

Given a model A of reality, what is the probability of data B (data)?

In other words, what is the likelihood $P(B | A)$ of B?

Given data B, what is the probability $P(A | B)$ of a model A (inverse modeling)?

Using that $P(A \text{ and } B) = P(A | B)P(B) = P(B | A)P(A)$

we get $P(A | B) = \frac{P(B | A)P(A)}{P(B)}$
**Example**

A = profitable gas extraction from a shale reservoir

Suppose we know (from global data) that \( P(A) = 0.1 \), so \( P(\sim A) = 0.9 \)

Let \( B = \text{permeability} > 0.5 \text{ Darcy} \)

We know also that \( P(B|A) = 0.8 \), and \( P(B|\sim A) = 0.4 \)

What is \( P(A|B) \)?

Here, \( P(A)=0.1 \), \( P(B|A)=0.8 \) and \( P(B) = P(B|A)P(A)+P(B|\sim A)P(\sim A) = 0.8*0.1 + 0.4*0.9 \)

So, \( P(A|B) = 0.18 \)
Bayes for complex models

A = earth model we are after (permeability, production numbers, ....)

B = all available and used data (in sets Bi)

P(A|B) is the desired uncertainty. It depends on some prior uncertainty P(A), as well as the likelihood probability

Prior selection critical and completely up to the imagination of the modeler

Likelihood can often be determined from models

Choice of Bi typically done through sensitivity analyses
The big picture

Information and knowledge (experts) informs all of the above
Interaction/symbiosis, always, between data and models:

You process data in a way that reflects your knowledge
You create a model in a way that reflects your knowledge
Falsification, not validation

Deterministic model matches data? No guarantee of truth
Maybe we could say that a model is "valid" if
• we cannot detect apparent flaws
• we have (some) consistency between model and observed data

Karl Popper: "We can never prove a model is correct, we can only ever prove that it is false"

The prior consists of all possibilities imagined by human modeler + computer
The posterior includes only those possibilities that can not be falsified with data as modeled in the likelihood
Many more questions than answers

How to select priors?
Based on prior expertise modeler. How to quantify intuition/expertise?
Easy when systems are linear and data Gaussian

What data to be included?
Based on sensitivity analysis using prior models. Linearization dangerous

How do we ensure consistency between models and data?
Particularly if different experts generate different parts.
What does it mean to mismatch data?

Inherently, the overall design process is iterative, or should be
Musings on climate models

- Large, but still relatively coarse. Subgrid modeling key
- Coupling of O(10) models
- Raw observations require intense filtering and processing
- Subgrid models by definition local and imprecise (calibration)
- Numerical modeling very challenging (Skamarock)

How complex should we make it? As computer power grows, tendency is to include more and more models (physics)

Simple models are more easily falsifiable

Complex models include more parameters that all introduce uncertainty and stronger dependence on the priors
Let's have a look at spatial models

Critical for earth models
Modeling spatial continuity

• Variograms – crude, takes samples values 2 at a time only

• Object models – introduce realistic shapes (objects) given by a set of parameters

• 3D training images – conceptual model rendering of major variations
Variograms

Measures geological distance between 2 points in n-dimensions (crude)
An Earth sciences approach: 3D training images

Explicit representation of what you expect. No mathematical formulae.
Training images derived from stratigraphy, for example.
Uncertainty manifests itself in multiple training images and multiple realizations that match data.
Let's have a look at spatial models

Critical for earth models.
Basic thought: there can be multiple models that satisfy the data
We look for a probability distribution of such models
Constraining to training images to construct priors

A case with a 2x2 grid

A training image

Another possible model
For another choice for first cell

Step 1. Pick a cell
Step 2. Assign probability
Step 3. Assign color
Using random draw based on given prob.

Step 4. Pick a cell
Step 5. Assign color
Based on previously assigned color and training model.

Step 6. Final result
For this choice of first cell

April 11, 2011  IMA - uncertainty
Constraining priors to cell data

Probability for central value to be “Cat 1” = 1/3
Constraining spatial uncertainty models with *multiple data*
Different data at different scales

In earth sciences, we typically have

• hard data or samples
detailed info on small scale

• soft information, e.g., remote sensing
data that give indication only

How can soft info be used to reduce uncertainty?
Both types of data are "partial". How do we combine these?

Here, we look at a possible approach
Calibration

First we need to assess what is the information content of a source.

Example: how much does seismic impedance (B) tell us about porosity (A)?

Fit a calibration function $F$ to pairs $(a_i, b_i)$, that is, find

$$F(s,t) = P(\text{porosity} < t \mid s < \text{impedance} < s + ds)$$

Assuming we have several $P(A \mid B_i)$ for $i=1,\ldots, n$. What is $P(A \mid B_1, B_2, \ldots)$?

Calibration approach unreasonable for multiple data (higher dimensions) unless behavior linear and interactions are non-existing.
**Integration Example**

Will it rain tomorrow (event A)?

Assume we have B1 and B2, with P(A|B1) = 0.7 and P(A|B2) = 0.6

We must model redundancy: here assume that B2 contributes regardless of B1.

\[
\begin{align*}
b_1 &= \frac{1 - P(A|B_1)}{P(A|B_1)}, \quad b_2 = \frac{1 - P(A|B_2)}{P(A|B_2)}, \quad a = \frac{1 - P(A)}{P(A)}
\end{align*}
\]

We'd like to know

\[
x = \frac{1 - P(A|B_1, B_2)}{P(A|B_1, B_2)} \Rightarrow P(A|B_1, B_2) = \frac{1}{1 + x}
\]

\[
\frac{b_2 - a}{a} = \text{the relative contribution of information source } B_2 \text{ when not having } B_1
\]

\[
\frac{x - b_1}{b_1} = \text{the relative contribution of information source } B_2 \text{ when } B_1 \text{ is available}
\]

\[
\frac{b_2 - a}{a} = \frac{x - b_1}{b_1} \Rightarrow x = \frac{b_1 b_2}{a} \quad \text{or} \quad P(A|B_1, B_2) = \frac{a}{a + b_1 b_2}
\]

\[
P(A|B_1, B_2) = 0.91
\]
Now, a more complicated example

Presence or absence of sand at each grid block. We are given probability derived from seismic. We'd like more sand channels where sand prob is higher.
A possible process

1. Assign hard data to the earth model grid
2. Define a random path that loops over all grid cells
3. At each grid cell:
   • Determine $P(A|B_1)$ by scanning
   • Find $P(A|B_2)$ (sand probability)
   • Use previous equation to find $P(A|B_1, B_2)$
   • Sample a value from this distribution function and assign to grid cell
Probability is often not the only info

In previous approach, we need conditional probabilities modeled from calibration data sets. They are not always available, and not always good. What if we knew that the data, or part of the data, were physical responses? How to take that into account?
**Inverse Modeling Takes into Account Physical Data Responses**

If unacceptable mismatch between response and observed data, then adjust input parameters for earth or forward model, or physical laws.

What to adjust?
Adjust parameters/models that impact data fit
and/or
Adjust parameters/models that impact decision questions
**Example**

M has 10 possible outcomes. A prior distribution is known

<table>
<thead>
<tr>
<th>$m$</th>
<th>$P(M = m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>60</td>
<td>10</td>
</tr>
<tr>
<td>&gt;60</td>
<td>0 in %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m$</th>
<th>$g(m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>60</td>
<td>3</td>
</tr>
<tr>
<td>&gt;60</td>
<td>4</td>
</tr>
</tbody>
</table>

A physical law $g(M)$ gives forward response $g(m)$

If the observed data for this forward model is $D=1$, then $m=0, 10$ or $20$. What are the probabilities of these outcomes?
We have prior for $M$

Do we have prior for $D$? Nope, but probs for all $M$ must add to one:

\[
P(M = 10 | D = 1) = 20\% \\
P(M = 20 | D = 1) = 80\% \\
P(D = d) = \sum_{m=0}^{100} P(D = d | M = m) P(M = m) \\
P(D = 1) = 100\% \times 0\% + 100\% \times 5\% + 100\% \times 20\% = 25\%
\]
What if there are measurement errors?

We model it somehow. Say

\[
P(D = 1| M = 10) = 75\% \text{ and } P(D = 1| M = 20) = 25\% \]

Now, Bayes gives

\[
P(M = 10| D = 1) = \frac{75 \times 5}{75 \times 5 + 25 \times 20} = 0.428 \\
P(M = 20| D = 1) = \frac{25 \times 20}{75 \times 5 + 25 \times 20} = 0.572
\]
Example

Want to find a model for hydraulic conductivity: either in a channel or not
Style of channeling is available in training model

Model: vector of 100x100 binary variables
We have pressure data at 9 locations
Pressure not a direct parameter of spatial model: it results from forward model
What model matches those values?
Find prior earth models that fit training images
Nine possibilities (but there are many more):
Amongst the prior models, search for those that match the data 
If no models can be found, it may be that there is inconsistency between priors and the forward model (developed by different experts)
Amongst the prior models, search for those that match the data. \textbf{HOW?}

A simple approach: rejection sampling (Popper would like this)

1. Generating a (prior) Earth model $m$ consistent with any prior information.
2. Evaluating $g(m)$.
3. If $g(m) = d$, then keeping that model; if not, reject it.
4. Go to step 1 until a desired amount of models has been found.

We could also accept the model with some defined probability, rather than requiring an exact match

Or, we could use Metropolis sampling instead
But, we can also use optimization methods such as gradient-based optimization, or Ensemble Kalman filtering.

Here, primarily looking to find an optimal fit to the data.

In sampling, $m$ is changed consistent with prior information.

In optimization methods, it may happen that models are created that are not part of the initial prior set.

Be very careful when choosing such methods!
OTHER COMPONENTS: RESPONSE AND DECISIONS MODEL
How about the response?

- Surrogate model (proxy) can be used if simulation model too expensive
- Is OK typically for ranking exercises
- Experimental design -> create response surfaces
- Sampling from data (Monte Carlo) too expensive: be smarter
- Response surface: typically smoothed out. How well does it capture nonlinearities? Scenarios difficult to test with experimental design
- Need to come up with quantitative measures for scenarios (eg realizations)
Decision making under uncertainty

• Should really happen over extended time period

• Identification of realistic alternative decisions dependent again on expertise

• Multiple objectives in different "currencies"

• Once you have alternatives and objectives you estimate payoff for objective \( j \) is we use alternative \( i \). In payoff matrix include those objectives that distinguish amongst alternatives

• Then we evaluate preference for objectives relative to each other. May be nonlinear and are subjective

• Combine in some structured way (eg decision tree). Conflicting objectives? Then evaluate tradeoffs: typically trade-off in risk and return ("are you willing to trade-off more risk for more return?")