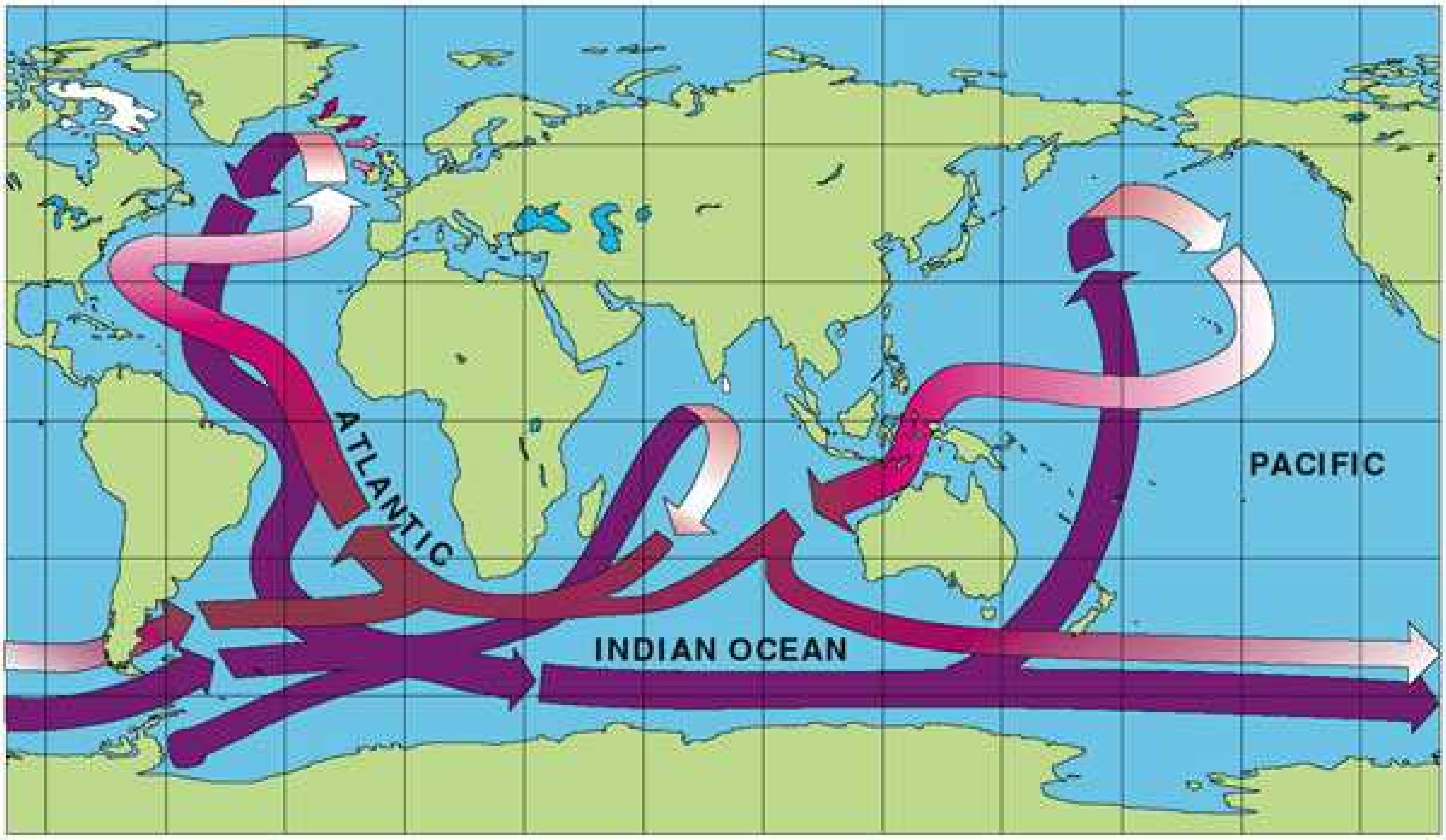
General Issues in Uncertainty Analysis for Complex Models

Michael Goldstein
Department of Mathematical Sciences,
Durham University

Global circulation



Thermohaline shutdown

Many atmosphere-ocean models show a slowdown of thermohaline circulation in simulations of the 21st century with the expected rise in greenhouse gases. [This is due to a combination of effects which reduce the density of surface waters, which makes it harder for them to sink.]

Thermohaline shutdown

Many atmosphere-ocean models show a slowdown of thermohaline circulation in simulations of the 21st century with the expected rise in greenhouse gases. [This is due to a combination of effects which reduce the density of surface waters, which makes it harder for them to sink.]

“ The weight of evidence makes it clear that climate change is a real and present danger. The Exeter conference was told that whatever policies are adopted from this point on, the Earth’s temperature will rise by 0.6F within the next 30 years. Yet those who think climate change just means Indian summers in Manchester should be told that the chances of the Gulf stream - the Atlantic thermohaline circulation that keeps Britain warm - shutting down are now thought to be greater than 50%.”

[**Burying carbon** Leader Column Thursday February 3, 2005 The Guardian]

Climate uncertainty

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Climate uncertainty

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Questions

- What does this statement mean?

Climate uncertainty

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Questions

- What does this statement mean?
- What analysis was actually done to reach this conclusion?

Climate uncertainty

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Questions

- What does this statement mean?
- What analysis was actually done to reach this conclusion?
- What analysis could possibly be done to justify (or contradict) this conclusion?

Climate uncertainty

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Questions

- What does this statement mean?
- What analysis was actually done to reach this conclusion?
- What analysis could possibly be done to justify (or contradict) this conclusion?
- What do we learn about physical systems from the analysis of (necessarily imperfect) models?

Models and Best Current Judgements

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Models and Best Current Judgements

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Exploring the behaviour of our favourite physical model gives us qualitative and semi-quantitative insights about the behaviour of the corresponding physical system.

Models and Best Current Judgements

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Exploring the behaviour of our favourite physical model gives us qualitative and semi-quantitative insights about the behaviour of the corresponding physical system.
However, behaviour of our model is not the same as the behaviour of actual system. Analysing models helps us to make judgements, but model analyses are not the same as judgements about physical systems.

Models and Best Current Judgements

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Exploring the behaviour of our favourite physical model gives us qualitative and semi-quantitative insights about the behaviour of the corresponding physical system.
However, behaviour of our model is not the same as the behaviour of actual system. Analysing models helps us to make judgements, but model analyses are not the same as judgements about physical systems.

[2] When we consider what actions we should take, we are concerned with actual physical systems. The fundamental question is:
what does the collection of models, scientific theories, observations and analysis of the likely implications arising from our imperfect knowledge, [model deficiency, observation error, uncertainty about physical constants, etc.] tell us about actual system behaviour?

Models and Best Current Judgements

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Exploring the behaviour of our favourite physical model gives us qualitative and semi-quantitative insights about the behaviour of the corresponding physical system.
However, behaviour of our model is not the same as the behaviour of actual system. Analysing models helps us to make judgements, but model analyses are not the same as judgements about physical systems.

[2] When we consider what actions we should take, we are concerned with actual physical systems. The fundamental question is:
what does the collection of models, scientific theories, observations and analysis of the likely implications arising from our imperfect knowledge,
[model deficiency, observation error, uncertainty about physical constants, etc.]
tell us about actual system behaviour?
Such analysis results in our Best Current Judgements as to future system behaviour, expressed as uncertainties.

Making good judgements

To form our best current judgements, we must first consider many specific detailed modelling questions - each of which is hard. There will always be more work to be done on such questions. But we can’t wait forever.

Making good judgements

To form our best current judgements, we must first consider many specific detailed modelling questions - each of which is hard. There will always be more work to be done on such questions. But we can’t wait forever.

In principle, we may envisage knowledge evolving through three distinct stages.

Making good judgements

To form our best current judgements, we must first consider many specific detailed modelling questions - each of which is hard. There will always be more work to be done on such questions. But we can’t wait forever.

In principle, we may envisage knowledge evolving through three distinct stages.

Stage 1 The best current judgements of an individual expert (expressed as probabilities).
(This is a subjective Bayes analysis)

Making good judgements

To form our best current judgements, we must first consider many specific detailed modelling questions - each of which is hard. There will always be more work to be done on such questions. But we can’t wait forever.

In principle, we may envisage knowledge evolving through three distinct stages.

Stage 1 The best current judgements of an individual expert (expressed as probabilities).
(This is a subjective Bayes analysis)

Stage 2 A careful analysis of the range of uncertainty judgements that it would be reasonable to hold given the differing views of experts.
(This is a scientific Bayes analysis)

Making good judgements

To form our best current judgements, we must first consider many specific detailed modelling questions - each of which is hard. There will always be more work to be done on such questions. But we can't wait forever.

In principle, we may envisage knowledge evolving through three distinct stages.

Stage 1 The best current judgements of an individual expert (expressed as probabilities).
(This is a subjective Bayes analysis)

Stage 2 A careful analysis of the range of uncertainty judgements that it would be reasonable to hold given the differing views of experts.
(This is a scientific Bayes analysis)

Stage 3 An analysis so clear and compelling that it would command agreement from all knowledgeable experts.
(This is an objective Bayes analysis (note non-standard use of "objective"!), and the only case where we can talk about, eg THE probability of THC collapse.)

Model analyses in practice

Despite all of the enormous amounts of very hard science that is being done, and the detailed knowledge that we are acquiring, it is very unusual to see analysis of these types. This is partly because of unfamiliarity and partly because the required tools may be hard to use (and some are still under development).

Model analyses in practice

Despite all of the enormous amounts of very hard science that is being done, and the detailed knowledge that we are acquiring, it is very unusual to see analysis of these types. This is partly because of unfamiliarity and partly because the required tools may be hard to use (and some are still under development).

'Best current judgements' is a high standard to set (though why aim for less?). What we require is care and clarity. This is challenging, but no more challenging, in principle, than the process of collecting system data and building and analysing models themselves. However, this does require a different tool-set and proper resources to carry through.

Model analyses in practice

Despite all of the enormous amounts of very hard science that is being done, and the detailed knowledge that we are acquiring, it is very unusual to see analysis of these types. This is partly because of unfamiliarity and partly because the required tools may be hard to use (and some are still under development).

'Best current judgements' is a high standard to set (though why aim for less?). What we require is care and clarity. This is challenging, but no more challenging, in principle, than the process of collecting system data and building and analysing models themselves. However, this does require a different tool-set and proper resources to carry through.

To understand the tool-set, we first need to look at the tool-set for individual models, and then consider how this tool-set extends to collections of models.

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Issues If the model is slow to evaluate, this is computationally expensive.

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Issues If the model is slow to evaluate, this is computationally expensive.

[2] Same as [1], but expert has chosen input parameters to match various historical climate features. As history is measured with error, this implies a probability distribution on the inputs. Drawing from this distribution, more than half of the input choices show THC shutdown.

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Issues If the model is slow to evaluate, this is computationally expensive.

[2] Same as [1], but expert has chosen input parameters to match various historical climate features. As history is measured with error, this implies a probability distribution on the inputs. Drawing from this distribution, more than half of the input choices show THC shutdown.

Issues The model calibration step is statistically and computationally challenging (it is a very high-dimensional ill-posed inverse problem).

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Issues If the model is slow to evaluate, this is computationally expensive.

[2] Same as [1], but expert has chosen input parameters to match various historical climate features. As history is measured with error, this implies a probability distribution on the inputs. Drawing from this distribution, more than half of the input choices show THC shutdown.

Issues The model calibration step is statistically and computationally challenging (it is a very high-dimensional ill-posed inverse problem).

[3] Same as [2], but expert has carefully distinguished between the model and the true physical system and therefore assessed a probability of actual THC shutdown given model THC shutdown.

Judgements based on a model

The chances of the Gulf stream shutting down are now thought to be greater than 50%

[1] Expert has model for THC. To evaluate the model, we need to specify a vector of inputs. More than half of the input choices show THC shutdown.

Issues If the model is slow to evaluate, this is computationally expensive.

[2] Same as [1], but expert has chosen input parameters to match various historical climate features. As history is measured with error, this implies a probability distribution on the inputs. Drawing from this distribution, more than half of the input choices show THC shutdown.

Issues The model calibration step is statistically and computationally challenging (it is a very high-dimensional ill-posed inverse problem).

[3] Same as [2], but expert has carefully distinguished between the model and the true physical system and therefore assessed a probability of actual THC shutdown given model THC shutdown.

Issues Serious assessment of model discrepancy is hard.

The state of the art in climate modelling

Large climate models take months to run on supercomputers. The biggest computer in the world is the Earth Simulator in Japan, which is often used for running climate models.

Leading climate models

One leading climate model at the moment is HadCM3, based at the UK Met Office. One component of this model is HadAM3, the atmospheric module. In a simple experiment to study the effect of CO2-doubling (Murphy et al, 2004, Nature), this is coupled with simple mixed-layer ocean sea-ice models.

Leading climate models

One leading climate model at the moment is HadCM3, based at the UK Met Office. One component of this model is HadAM3, the atmospheric module. In a simple experiment to study the effect of CO2-doubling (Murphy et al, 2004, Nature), this is coupled with simple mixed-layer ocean sea-ice models.

The climate model (HadSM3) has about 100 uncertain parameters, including:

1. *Large scale cloud.* Six parameters
2. *Convection.* Six parameters
3. *Sea ice.* Two parameters
4. *Radiation.* Four parameters
5. *Dynamics.* Four parameters
6. *Land surface.* Four parameters
7. *Boundary layer.* Four parameters

Leading climate models

One leading climate model at the moment is HadCM3, based at the UK Met Office. One component of this model is HadAM3, the atmospheric module. In a simple experiment to study the effect of CO2-doubling (Murphy et al, 2004, Nature), this is coupled with simple mixed-layer ocean sea-ice models.
The climate model (HadSM3) has about 100 uncertain parameters, including:

1. *Large scale cloud.* Six parameters
2. *Convection.* Six parameters
3. *Sea ice.* Two parameters
4. *Radiation.* Four parameters
5. *Dynamics.* Four parameters
6. *Land surface.* Four parameters
7. *Boundary layer.* Four parameters

We have a few hundred evaluations of HadSM3, made over a period of about three years. These evaluations will be one of the main resources for the UK Climate Impacts Programme 2008 (UKCIP08), which is intended as a fairly definitive statement about how climate change will impact the UK.

How to relate a model and a physical systems?

- Basic ingredients:

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)

How to relate a model and a physical systems?

- Basic ingredients:

  x^*: system properties (unknown)

  y: system behaviour (influenced by x^*)

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)
  y: system behaviour (influenced by x^*)
  z: partial observation of y (with error)

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)
  y: system behaviour (influenced by x^*)
  z: partial observation of y (with error)
- Ideally, we would like to construct a deterministic model f, embodying the **laws of nature**, which satisfies

$$y = f(x^*)$$

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)
  y: system behaviour (influenced by x^*)
  z: partial observation of y (with error)
- Ideally, we would like to construct a deterministic model f, embodying the **laws of nature**, which satisfies

$$y = f(x^*)$$

- In practice, however, the our actual model f is inadequate:
  - f simplifies the physics;
  - f approximates the solution of the physical equations

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)
  y: system behaviour (influenced by x^*)
  z: partial observation of y (with error)
- Ideally, we would like to construct a deterministic model f, embodying the **laws of nature**, which satisfies

$$y = f(x^*)$$

- In practice, however, the our actual model f is inadequate:
  - f simplifies the physics;
  - f approximates the solution of the physical equations
- This raises the basic question as to what does the imperfect f tell us about the system values (x^*, y)?

How to relate a model and a physical systems?

- Basic ingredients:
  x^*: system properties (unknown)
  y: system behaviour (influenced by x^*)
  z: partial observation of y (with error)
- Ideally, we would like to construct a deterministic model f, embodying the **laws of nature**, which satisfies

$$y = f(x^*)$$

- In practice, however, the our actual model f is inadequate:
  - f simplifies the physics;
  - f approximates the solution of the physical equations
- This raises the basic question as to what does the imperfect f tell us about the system values (x^*, y)?
- In particular, input and output very high dimensional and evaluating $f(x)$ for any x may be VERY expensive.

Relating the model and the system

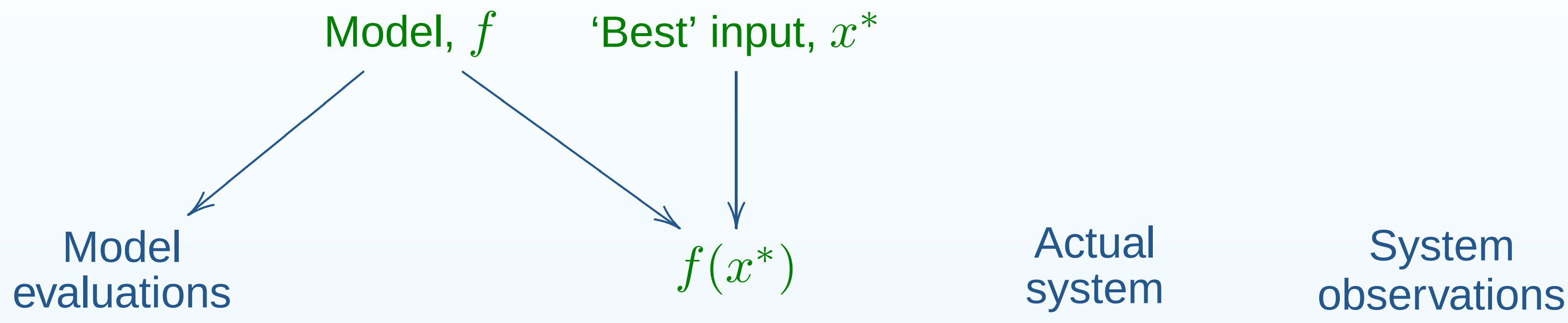
Model evaluations

Actual system

System observations

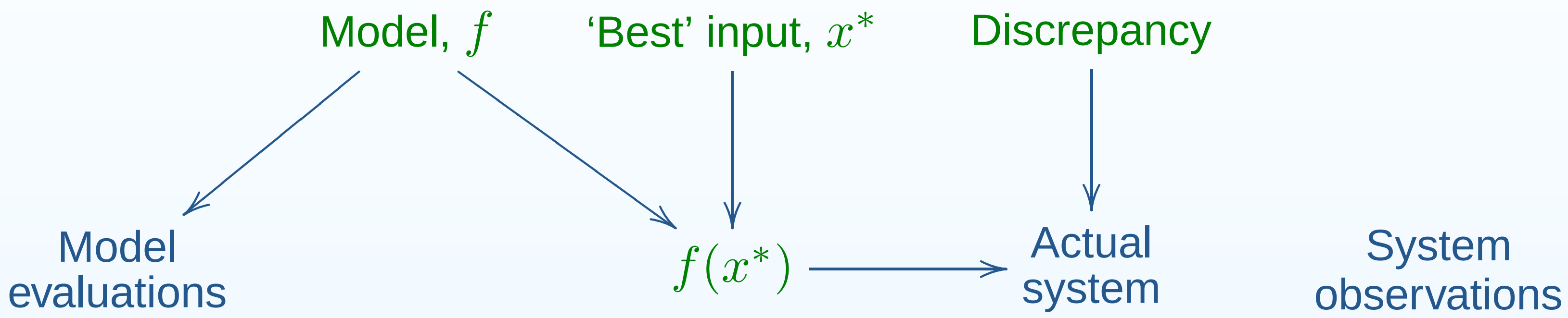
1. We start with a collection of model evaluations, and some observations on actual system
2. We link the evaluations to the notion of a 'best' evaluation
3. We link the 'best' evaluation to the actual system
4. We incorporate measurement error into the observations
5. There are many useful statistical approaches to assess each individual aspect of uncertainty. The key strength of the Bayesian approach is in the unified treatment of all of the sources of uncertainty.

Relating the model and the system



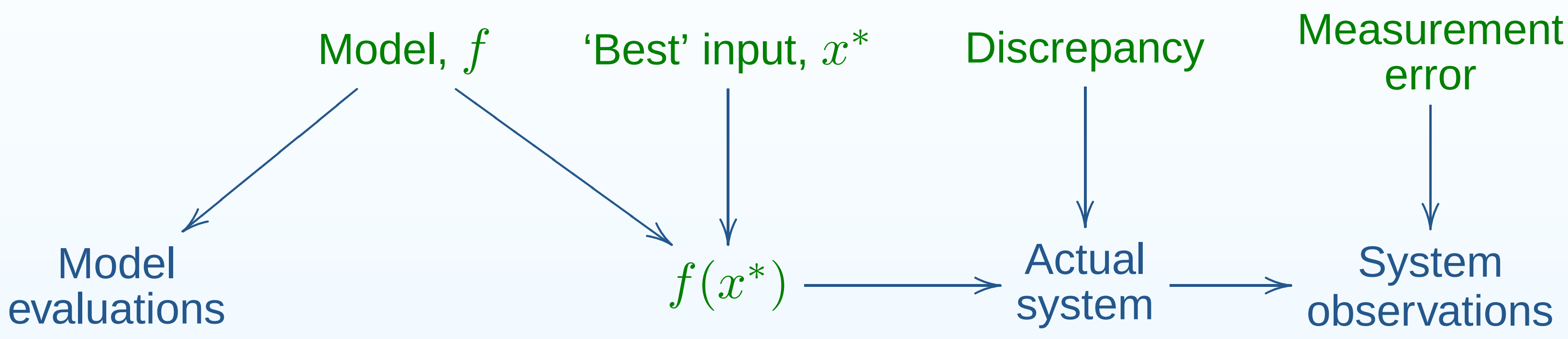
1. We start with a collection of model evaluations, and some observations on actual system
2. We link the evaluations to the notion of a 'best' evaluation
3. We link the 'best' evaluation to the actual system
4. We incorporate measurement error into the observations
5. There are many useful statistical approaches to assess each individual aspect of uncertainty. The key strength of the Bayesian approach is in the unified treatment of all of the sources of uncertainty.

Relating the model and the system



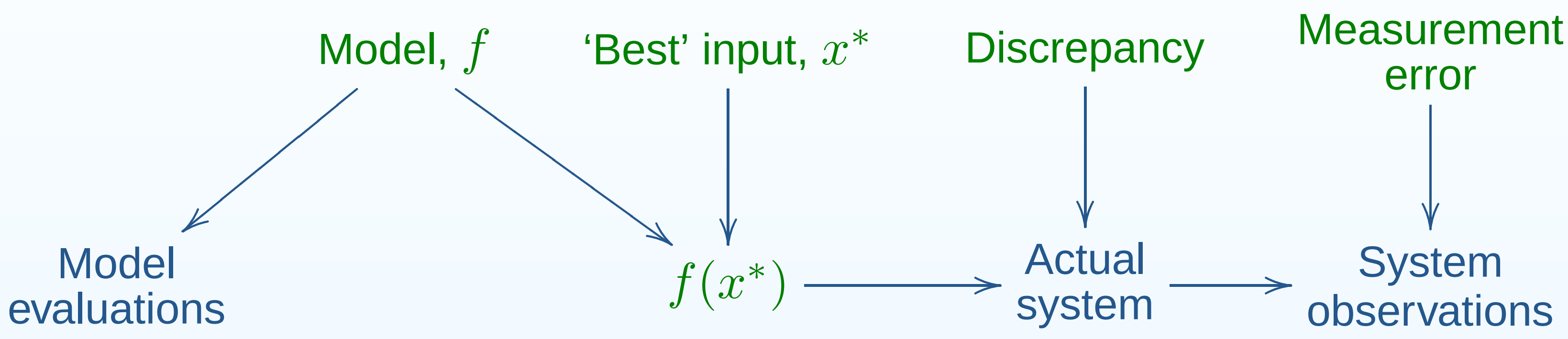
1. We start with a collection of model evaluations, and some observations on actual system
2. We link the evaluations to the notion of a 'best' evaluation
3. We link the 'best' evaluation to the actual system
4. We incorporate measurement error into the observations
5. There are many useful statistical approaches to assess each individual aspect of uncertainty. The key strength of the Bayesian approach is in the unified treatment of all of the sources of uncertainty.

Relating the model and the system



1. We start with a collection of model evaluations, and some observations on actual system
2. We link the evaluations to the notion of a 'best' evaluation
3. We link the 'best' evaluation to the actual system
4. We incorporate measurement error into the observations
5. There are many useful statistical approaches to assess each individual aspect of uncertainty. The key strength of the Bayesian approach is in the unified treatment of all of the sources of uncertainty.

Relating the model and the system



1. We start with a collection of model evaluations, and some observations on actual system
2. We link the evaluations to the notion of a 'best' evaluation
3. We link the 'best' evaluation to the actual system
4. We incorporate measurement error into the observations
5. There are many useful statistical approaches to assess each individual aspect of uncertainty. The key strength of the Bayesian approach is in the unified treatment of all of the sources of uncertainty.

Representing beliefs about f using emulators

An *emulator* is a probabilistic belief specification for a deterministic function. Our emulator for component i of f might be

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x) + u_i(x)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u(x)$ is a weakly stationary stochastic process.
[A simple case is to suppose, for each x, that $u_(x)$ is normal with constant variance and $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$.]

Representing beliefs about f using emulators

An *emulator* is a probabilistic belief specification for a deterministic function. Our emulator for component i of f might be

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x) + u_i(x)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u(x)$ is a weakly stationary stochastic process.
[A simple case is to suppose, for each x, that $u_(x)$ is normal with constant variance and $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$.]

Interpretation: in this formulation, $Bg(x)$ expresses global variation in f. $u(x)$ expresses local variation in f

Representing beliefs about f using emulators

An *emulator* is a probabilistic belief specification for a deterministic function. Our emulator for component i of f might be

$$f_i(x) = \sum_j \beta_{ij}\, g_{ij}(x) + u_i(x)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x, and $u(x)$ is a weakly stationary stochastic process.
[A simple case is to suppose, for each x, that $u_(x)$ is normal with constant variance and $\mathrm{Corr}(u_i(x), u_i(x'))$ is a function of $\|x - x'\|$.]

Interpretation: in this formulation, $Bg(x)$ expresses global variation in f. $u(x)$ expresses local variation in f

When the input dimension is high, relative to the number of function evaluations we can make, then most of what we may learn about the function comes through the global component. For simplicity, we therefore often suppose that the simulator behaviour can be summarised by the global behaviour (as we dont learn much about local behaviour).

Emulator details

We fit the emulator, $f = Bg(x) + u(x)$, given a collection of model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes - with a generous helping of expert judgement.

Emulator details

We fit the emulator, $f = Bg(x) + u(x)$, given a collection of model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes - with a generous helping of expert judgement.

So, we need careful experimental design to choose which evaluations of the model to make, and detailed diagnostics, to check emulator validity.

Emulator details

We fit the emulator, $f = Bg(x) + u(x)$, given a collection of model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes - with a generous helping of expert judgement.

So, we need careful experimental design to choose which evaluations of the model to make, and detailed diagnostics, to check emulator validity.

We have some useful backup tricks - for example, if we can only make a few evaluations of our model, we may be able to make many evaluations of a simpler approximate version of the model to get us started.

Emulator details

We fit the emulator, $f = Bg(x) + u(x)$, given a collection of model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes - with a generous helping of expert judgement.

So, we need careful experimental design to choose which evaluations of the model to make, and detailed diagnostics, to check emulator validity.

We have some useful backup tricks - for example, if we can only make a few evaluations of our model, we may be able to make many evaluations of a simpler approximate version of the model to get us started.
The emulator expresses prior uncertainty judgements about the function. These are modified by function evaluations. From the emulator, we may extract uncertainty statements for the function, at each input value x, e.g.

$$\mu_i(x) = \mathrm{E}(f_i(x))$$
$$\kappa_i(x, x') = \mathrm{Cov}(f_i(x), f_i(x')),$$

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x), \; y = f(x^*) + \epsilon, \; z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs.

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x), \; y = f(x^*) + \epsilon, \; z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs. We may then use our collection of computer evaluations and historical observations to analyse the physical process

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs. We may then use our collection of computer evaluations and historical observations to analyse the physical process

- to determine “correct” settings for simulator inputs (calibration);

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x), \; y = f(x^*) + \epsilon, \; z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs. We may then use our collection of computer evaluations and historical observations to analyse the physical process

- to determine “correct” settings for simulator inputs (calibration);
- to assess the future behaviour of the system (forecasting);

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x), \; y = f(x^*) + \epsilon, \; z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs.
We may then use our collection of computer evaluations and historical observations to analyse the physical process

- to determine “correct” settings for simulator inputs (calibration);
- to assess the future behaviour of the system (forecasting);
- to “optimise” the performance of the system (control).

Uncertainty analysis for complex models

$f(x) = Bg(x) + u(x),\ y = f(x^*) + \epsilon,\ z = y + e$

The Bayesian treatment of this structure involves:

- a prior distribution for best input x^*
- a probabilistic emulator for the computer function f
- a probabilistic discrepancy measure relating $f(x^*)$ to the system y
- a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesise expert prior judgement, historical data and a careful choice of simulator runs. We may then use our collection of computer evaluations and historical observations to analyse the physical process

- to determine "correct" settings for simulator inputs (calibration);
- to assess the future behaviour of the system (forecasting);
- to "optimise" the performance of the system (control).

For problems of medium size, this approach is very powerful and effective.

Bayes linear approach

For very large scale problems a full Bayes analysis is very hard because
(i) it is difficult to make meaningful probability specifications over high dimensional spaces;
(ii) the computations, for learning from data (observations and computer runs), particularly when choosing informative runs, may be technically difficult;
(iii) the likelihood surface is extremely complicated, and any full Bayes calculation (based on emulation) may be extremely non-robust.

Bayes linear approach

For very large scale problems a full Bayes analysis is very hard because
(i) it is difficult to make meaningful probability specifications over high dimensional spaces;
(ii) the computations, for learning from data (observations and computer runs), particularly when choosing informative runs, may be technically difficult;
(iii) the likelihood surface is extremely complicated, and any full Bayes calculation (based on emulation) may be extremely non-robust.
However, the idea of the Bayesian approach, namely capturing our expert prior judgements in stochastic form and modifying them by appropriate rules given observations, is conceptually appropriate (and there is no obvious alternative).

Bayes linear approach

For very large scale problems a full Bayes analysis is very hard because
(i) it is difficult to make meaningful probability specifications over high dimensional spaces;
(ii) the computations, for learning from data (observations and computer runs), particularly when choosing informative runs, may be technically difficult;
(iii) the likelihood surface is extremely complicated, and any full Bayes calculation (based on emulation) may be extremely non-robust.
However, the idea of the Bayesian approach, namely capturing our expert prior judgements in stochastic form and modifying them by appropriate rules given observations, is conceptually appropriate (and there is no obvious alternative).
The Bayes Linear approach is (relatively) simple in terms of belief specification and analysis, as it is based only on the mean, variance and covariance specification which, following de Finetti, we take as primitive.
For a full account, see
Michael Goldstein and David Wooff (2007) Bayes Linear Statistics: Theory and Methods, Wiley.

Calibration via history matching

History Matching is concerned with learning about best inputs, x^*, using simulator evaluations and data, z. Using the emulator we obtain, for each input choice x, the adjusted values of $\mathrm{E}(f(x))$ and $\mathrm{Var}(f(x))$. We rule out regions of x space for which $f(x)$ is likely to be a very poor match to observed z.

Calibration via history matching

History Matching is concerned with learning about best inputs, x^*, using simulator evaluations and data, z. Using the emulator we obtain, for each input choice x, the adjusted values of $\mathrm{E}(f(x))$ and $\mathrm{Var}(f(x))$. We rule out regions of x space for which $f(x)$ is likely to be a very poor match to observed z.
To achieve this, we calculate, for each output $f_i(x)$, the implausibility:

$$I_{(i)}(x) = |\mathrm{E}(f_i(x)) - z_i|^2 / \mathrm{Var}(f_i(x) - z_i)$$

Calibration via history matching

History Matching is concerned with learning about best inputs, x^*, using simulator evaluations and data, z. Using the emulator we obtain, for each input choice x, the adjusted values of $\mathrm{E}(f(x))$ and $\mathrm{Var}(f(x))$. We rule out regions of x space for which $f(x)$ is likely to be a very poor match to observed z.
To achieve this, we calculate, for each output $f_i(x)$, the implausibility:

$$I_{(i)}(x) = |\mathrm{E}(f_i(x)) - z_i|^2 / \mathrm{Var}(f_i(x) - z_i)$$

This calculation can be performed univariately, or over sub-vectors. The implausibilities are then combined, such as by using $I_M(x) = \max_i I_{(i)}(x)$, and can then be used to identify regions of x with large $I_M(x)$ as unlikely to be good choices for x^*.

Calibration via history matching

History Matching is concerned with learning about best inputs, x^*, using simulator evaluations and data, z. Using the emulator we obtain, for each input choice x, the adjusted values of $\mathrm{E}(f(x))$ and $\mathrm{Var}(f(x))$. We rule out regions of x space for which $f(x)$ is likely to be a very poor match to observed z.

To achieve this, we calculate, for each output $f_i(x)$, the implausibility:

$$I_{(i)}(x) = |\mathrm{E}(f_i(x)) - z_i|^2 / \mathrm{Var}(f_i(x) - z_i)$$

This calculation can be performed univariately, or over sub-vectors. The implausibilities are then combined, such as by using $I_M(x) = \max_i I_{(i)}(x)$, and can then be used to identify regions of x with large $I_M(x)$ as unlikely to be good choices for x^*.

We iteratively refocus on the 'non-implausible' regions of the input space, by further model runs and refitting our emulator over the sub-region and repeating the analysis. This process is a form of iterative global search aimed at finding all choices for x^* which would give acceptable fits to historical data.

Calibration via history matching

History Matching is concerned with learning about best inputs, x^*, using simulator evaluations and data, z. Using the emulator we obtain, for each input choice x, the adjusted values of $\mathrm{E}(f(x))$ and $\mathrm{Var}(f(x))$. We rule out regions of x space for which $f(x)$ is likely to be a very poor match to observed z.

To achieve this, we calculate, for each output $f_i(x)$, the implausibility:

$$I_{(i)}(x) = |\mathrm{E}(f_i(x)) - z_i|^2 / \mathrm{Var}(f_i(x) - z_i)$$

This calculation can be performed univariately, or over sub-vectors. The implausibilities are then combined, such as by using $I_M(x) = \max_i I_{(i)}(x)$, and can then be used to identify regions of x with large $I_M(x)$ as unlikely to be good choices for x^*.

We iteratively refocus on the 'non-implausible' regions of the input space, by further model runs and refitting our emulator over the sub-region and repeating the analysis. This process is a form of iterative global search aimed at finding all choices for x^* which would give acceptable fits to historical data.

If all values of x are implausible, this is important diagnostic information!

Forecasting without calibration

For the computer model emulator, the mean and variance of $f^* = f(x^*) = Bg(x^*) + u(x^*)$ are obtained from the mean function and variance function of the emulator for $f(x)$. Therefore, we can compute the mean and variance of f^* by conditioning on x^* and then integrating with respect to a prior distribution on x^*.

Forecasting without calibration

For the computer model emulator, the mean and variance of $f^* = f(x^*) = Bg(x^*) + u(x^*)$ are obtained from the mean function and variance function of the emulator for $f(x)$. Therefore, we can compute the mean and variance of f^* by conditioning on x^* and then integrating with respect to a prior distribution on x^*.

Given $\mathrm{E}(f^*)$, $\mathrm{Var}(f^*)$, $\mathrm{Var}(\epsilon)$, $\mathrm{Var}(e)$, it is now straightforward to compute the joint mean and variance of the collection $(y = f^* + \epsilon, \;\; z = y_h + e)$.

Forecasting without calibration

For the computer model emulator, the mean and variance of $f^* = f(x^*) = Bg(x^*) + u(x^*)$ are obtained from the mean function and variance function of the emulator for $f(x)$. Therefore, we can compute the mean and variance of f^* by conditioning on x^* and then integrating with respect to a prior distribution on x^*.
Given $\mathrm{E}(f^*)$, $\mathrm{Var}(f^*)$, $\mathrm{Var}(\epsilon)$, $\mathrm{Var}(e)$, it is now straightforward to compute the joint mean and variance of the collection $(y = f^* + \epsilon,\ \ z = y_h + e)$.
We now evaluate the adjusted mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae. This analysis gives system forecasts without model calibration, and therefore is tractable even for large systems.

Forecasting without calibration

For the computer model emulator, the mean and variance of $f^* = f(x^*) = Bg(x^*) + u(x^*)$ are obtained from the mean function and variance function of the emulator for $f(x)$. Therefore, we can compute the mean and variance of f^* by conditioning on x^* and then integrating with respect to a prior distribution on x^*.

Given $\mathrm{E}(f^*)$, $\mathrm{Var}(f^*)$, $\mathrm{Var}(\epsilon)$, $\mathrm{Var}(e)$, it is now straightforward to compute the joint mean and variance of the collection $(y = f^* + \epsilon, \;\; z = y_h + e)$.

We now evaluate the adjusted mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae. This analysis gives system forecasts without model calibration, and therefore is tractable even for large systems.

In particular, if we are cautious about assigning physical meaning to the inputs, then we may view the model as a device for deriving relationships between past and future observables which we directly exploit, as above.

Forecasting without calibration

For the computer model emulator, the mean and variance of $f^* = f(x^*) = Bg(x^*) + u(x^*)$ are obtained from the mean function and variance function of the emulator for $f(x)$. Therefore, we can compute the mean and variance of f^* by conditioning on x^* and then integrating with respect to a prior distribution on x^*.

Given $\mathrm{E}(f^*)$, $\mathrm{Var}(f^*)$, $\mathrm{Var}(\epsilon)$, $\mathrm{Var}(e)$, it is now straightforward to compute the joint mean and variance of the collection $(y = f^* + \epsilon, \;\; z = y_h + e)$.

We now evaluate the adjusted mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae. This analysis gives system forecasts without model calibration, and therefore is tractable even for large systems.

In particular, if we are cautious about assigning physical meaning to the inputs, then we may view the model as a device for deriving relationships between past and future observables which we directly exploit, as above.

Where appropriate, we can improve accuracy by adding a Bayes linear calibration stage to the forecasting (while retaining tractability).

The Best input

How does learning about f inform us about y?

The Best input

How does learning about f inform us about y?

The simplest (and therefore most popular) way to relate uncertainty about the simulator and the system is the so-called “Best Input Approach”.
We proceed as though there exists a value x^* independent of the function f such that the value of $f^* = f(x^*)$ summarises all of the information that the simulator conveys about the system. This means that we consider the model discrepancy, $\epsilon = y - f^*$, to be independent of f, x^*.

The Best input

How does learning about f inform us about y?

The simplest (and therefore most popular) way to relate uncertainty about the simulator and the system is the so-called “Best Input Approach”.
We proceed as though there exists a value x^* independent of the function f such that the value of $f^* = f(x^*)$ summarises all of the information that the simulator conveys about the system. This means that we consider the model discrepancy, $\epsilon = y - f^*$, to be independent of f, x^*.

This formulation raises serious questions.
In particular, does x^* correspond to “true” system properties?

The Best input

How does learning about f inform us about y?

The simplest (and therefore most popular) way to relate uncertainty about the simulator and the system is the so-called “Best Input Approach”.
We proceed as though there exists a value x^* independent of the function f such that the value of $f^* = f(x^*)$ summarises all of the information that the simulator conveys about the system. This means that we consider the model discrepancy, $\epsilon = y - f^*$, to be independent of f, x^*.

This formulation raises serious questions.
In particular, does x^* correspond to “true” system properties?

If so, why should they give best fit to our imperfect model?

The Best input

How does learning about f inform us about y?

The simplest (and therefore most popular) way to relate uncertainty about the simulator and the system is the so-called "Best Input Approach".
We proceed as though there exists a value x^* independent of the function f such that the value of $f^* = f(x^*)$ summarises all of the information that the simulator conveys about the system. This means that we consider the model discrepancy, $\epsilon = y - f^*$, to be independent of f, x^*.

This formulation raises serious questions.
In particular, does x^* correspond to "true" system properties?

If so, why should they give best fit to our imperfect model?

If not, why should there even be any such "best" inputs, and what does it mean to have expert judgements about x^*?

The Best input

How does learning about f inform us about y?

The simplest (and therefore most popular) way to relate uncertainty about the simulator and the system is the so-called “Best Input Approach”.
We proceed as though there exists a value x^* independent of the function f such that the value of $f^* = f(x^*)$ summarises all of the information that the simulator conveys about the system. This means that we consider the model discrepancy, $\epsilon = y - f^*$, to be independent of f, x^*.

This formulation raises serious questions.
In particular, does x^* correspond to “true” system properties?

If so, why should they give best fit to our imperfect model?

If not, why should there even be any such “best” inputs, and what does it mean to have expert judgements about x^*?

Further, surprising contradictions arise when we try to construct joint specifications linking collections of models to the physical system in this way.

Reification

Why should we consider that there is a ‘best input’ x^*?

Reification

Why should we consider that there is a 'best input' x^*?

What does a simulator f really tell us about a physical system y?

Reification

Why should we consider that there is a 'best input' x^*?

What does a simulator f really tell us about a physical system y?

How do we combine the information about y from a collection of simulators?

Reification

Why should we consider that there is a ‘best input’ x^*?
What does a simulator f really tell us about a physical system y?
How do we combine the information about y from a collection of simulators?

Consider both our inputs x and the simulator f as abstractions/simplifications of real physical quantities and processes (through approximations in physics, solution methods, level of detail, limitations of current understanding) to a much more realistic simulator f^*, for which real, physical x^* would be the best input, in the sense that $(y - f^*(x^*))$ would be judged independent of (x^*, f^*).

Reification

Why should we consider that there is a 'best input' x^*?
What does a simulator f really tell us about a physical system y?
How do we combine the information about y from a collection of simulators?

Consider both our inputs x and the simulator f as abstractions/simplifications of real physical quantities and processes (through approximations in physics, solution methods, level of detail, limitations of current understanding) to a much more realistic simulator f^*, for which real, physical x^* would be the best input, in the sense that $(y - f^*(x^*))$ would be judged independent of (x^*, f^*).

We call f^* the reified simulator (from reify: to treat an abstract concept as if it was real).

Reification

Why should we consider that there is a 'best input' x^*?
What does a simulator f really tell us about a physical system y?
How do we combine the information about y from a collection of simulators?

Consider both our inputs x and the simulator f as abstractions/simplifications of real physical quantities and processes (through approximations in physics, solution methods, level of detail, limitations of current understanding) to a much more realistic simulator f^*, for which real, physical x^* would be the best input, in the sense that $(y - f^*(x^*))$ would be judged independent of (x^*, f^*).

We call f^* the reified simulator (from reify: to treat an abstract concept as if it was real).

Reifying principle
[1] Simulator f is informative for y, because f is informative for f^* and $f^*(x^*)$ is informative for y.

Reification

Why should we consider that there is a 'best input' x^*?
What does a simulator f really tell us about a physical system y?
How do we combine the information about y from a collection of simulators?

Consider both our inputs x and the simulator f as abstractions/simplifications of real physical quantities and processes (through approximations in physics, solution methods, level of detail, limitations of current understanding) to a much more realistic simulator f^*, for which real, physical x^* would be the best input, in the sense that $(y - f^*(x^*))$ would be judged independent of (x^*, f^*).

We call f^* the reified simulator (from reify: to treat an abstract concept as if it was real).

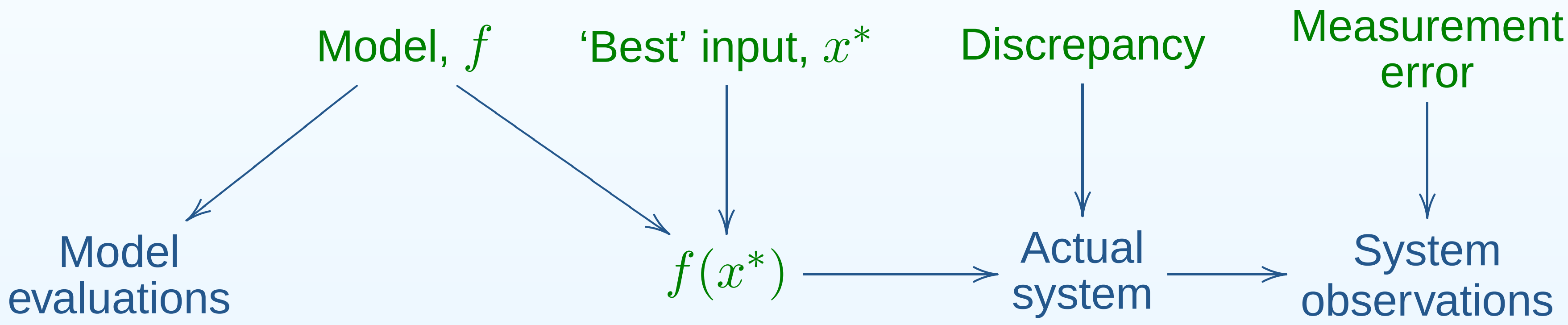
Reifying principle
[1] Simulator f is informative for y, because f is informative for f^* and $f^*(x^*)$ is informative for y.
[2] A collection of simulators $f_1, f_2, \ldots$ is jointly informative for y, as the simulators are jointly informative for f^*.

Relating the model and the system (2)

Our model f is informative for y because f is informative for a more elaborate model f^*

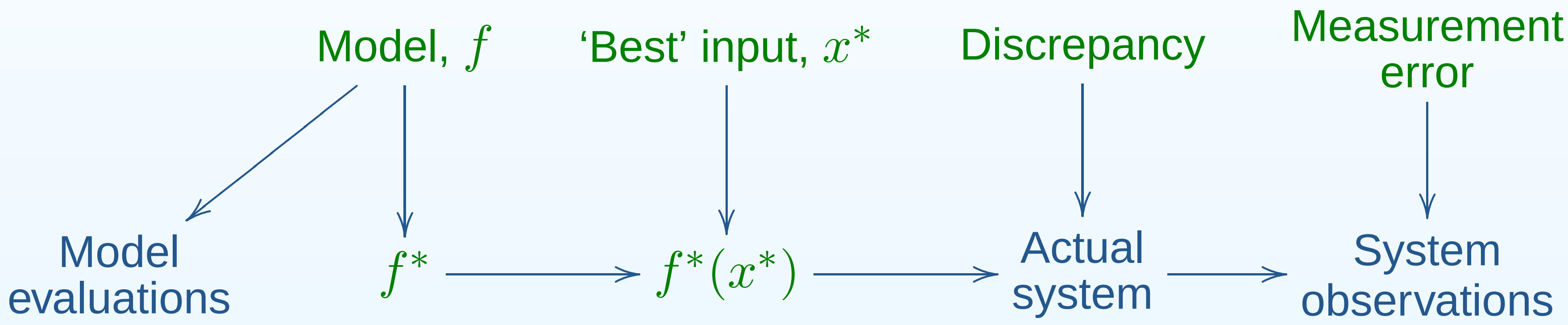
Relating the model and the system (2)

Our model f is informative for y because f is informative for a more elaborate model f^*



Relating the model and the system (2)

Our model f is informative for y because f is informative for a more elaborate model f^*



Linking f and f^* using emulators

Suppose that our emulator for f is

$$f(x) = Bg(x) + u(x)$$

Linking f and f^* using emulators

Suppose that our emulator for f is

$$f(x) = Bg(x) + u(x)$$

Our simplest emulator for f^* might be

$$f(x, w) = B^* g(x) + u^*(x) + u^*(x, w)$$

where we might model our judgements as $B^* = CB + \Gamma$ for known C and uncertain Γ, correlate $u(x)$ and $u^*(x)$, but leave $u^*(x, w)$, involving any additional parameters, w, uncorrelated.

Linking f and f^* using emulators

Suppose that our emulator for f is

$$f(x) = Bg(x) + u(x)$$

Our simplest emulator for f^* might be

$$f(x, w) = B^* g(x) + u^*(x) + u^*(x, w)$$

where we might model our judgements as $B^* = CB + \Gamma$ for known C and uncertain Γ, correlate $u(x)$ and $u^*(x)$, but leave $u^*(x, w)$, involving any additional parameters, w, uncorrelated.

Structured reification: systematic probabilistic modelling for all those aspects of model deficiency whose effects we are prepared to consider explicitly.

Linking f and f^* using emulators

Suppose that our emulator for f is

$$f(x) = Bg(x) + u(x)$$

Our simplest emulator for f^* might be

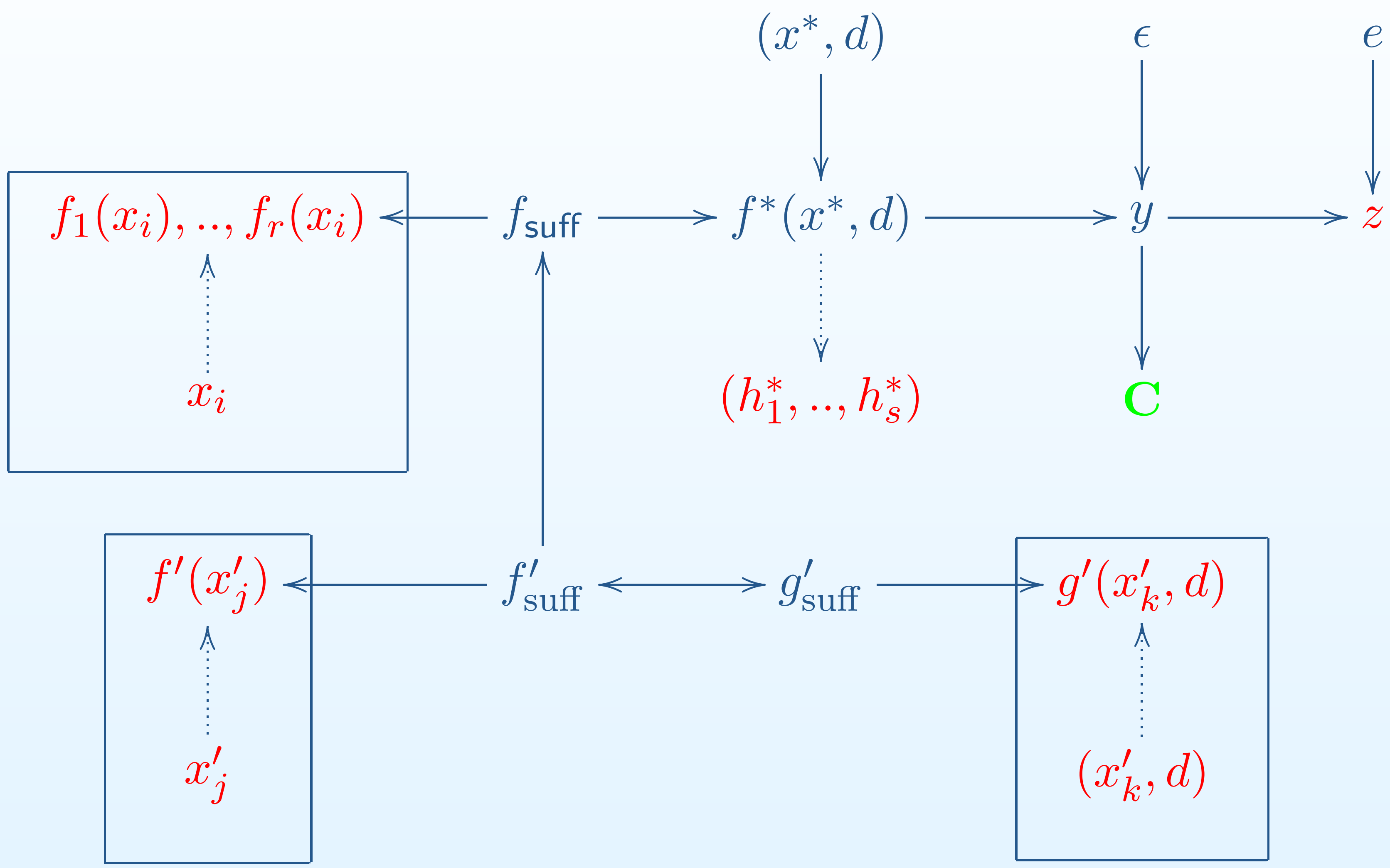
$$f(x, w) = B^* g(x) + u^*(x) + u^*(x, w)$$

where we might model our judgements as $B^* = CB + \Gamma$ for known C and uncertain Γ, correlate $u(x)$ and $u^*(x)$, but leave $u^*(x, w)$, involving any additional parameters, w, uncorrelated.

Structured reification: systematic probabilistic modelling for all those aspects of model deficiency whose effects we are prepared to consider explicitly.

Comment: All our previous methods are unchanged - all that has changed is our description of the joint covariance structure.

A reified influence diagram



Reified simulator f^*; $f_1, \ldots f_k$ are exchangeable refinements of f'; g' is a simulator at level of f' adding decision variables d; (h_i^*) is ensemble of tuned model runs. C is cost to the planet.

Best current judgements for complex systems

To assess best current judgements about complex systems, it is enormously helpful to have an overall framework to unify all the uncertainties arising from

Uncertain model parameters, outputs and discrepancies

Uncertain observations/initial conditions/forcing functions

Uncertain relationships between different modelling approaches

Uncertain effects of our attempts to influence the system

Best current judgements for complex systems

To assess best current judgements about complex systems, it is enormously helpful to have an overall framework to unify all the uncertainties arising from

Uncertain model parameters, outputs and discrepancies

Uncertain observations/initial conditions/forcing functions

Uncertain relationships between different modelling approaches

Uncertain effects of our attempts to influence the system

We have suggested a conceptual/graphical framework for unifying our qualitative and quantitative knowledge about all such uncertainties within a structure which is both logical and tractable, so that we can focus on science rather than technical/computational issues.

We need new methodology to construct the general language and tool kit required for making this synthesis in principle, and a close joint effort between statisticians and scientists to achieve the synthesis in practice.

Best current judgements for complex systems

To assess best current judgements about complex systems, it is enormously helpful to have an overall framework to unify all the uncertainties arising from

Uncertain model parameters, outputs and discrepancies

Uncertain observations/initial conditions/forcing functions

Uncertain relationships between different modelling approaches

Uncertain effects of our attempts to influence the system

We have suggested a conceptual/graphical framework for unifying our qualitative and quantitative knowledge about all such uncertainties within a structure which is both logical and tractable, so that we can focus on science rather than technical/computational issues.

We need new methodology to construct the general language and tool kit required for making this synthesis in principle, and a close joint effort between statisticians and scientists to achieve the synthesis in practice.

Such analysis poses serious challenges, but they are no harder than all of the other modelling, computational and observational challenges involved with studying large scale physical systems.

References

P.S. Craig, M. Goldstein, J.C.Rougier, A.H. Seheult, (2001) Bayesian Forecasting Using Large Computer Models, JASA, 96, 717-729

M. Goldstein and J.C.Rougier (2008). Reified Bayesian modelling and inference for physical systems (with discussion), JSPI, to appear, .

Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). Journal of the Royal Statistical Society, B,63, 425-464

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91, 12901300.

J.C. Rougier (2007), Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, Climatic Change, 81, 247-264.

Santner, T., Williams, B. and Notz, W. (2003). The Design and Analysis of Computer Experiments. Springer Verlag: New York.

References

P.S. Craig, M. Goldstein, J.C.Rougier, A.H. Seheult, (2001) Bayesian Forecasting Using Large Computer Models, JASA, 96, 717-729

M. Goldstein and J.C.Rougier (2008). Reified Bayesian modelling and inference for physical systems (with discussion), JSPI, to appear, .

Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). Journal of the Royal Statistical Society, B,63, 425-464

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91, 12901300.

J.C. Rougier (2007), Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, Climatic Change, 81, 247-264.

Santner, T., Williams, B. and Notz, W. (2003). The Design and Analysis of Computer Experiments. Springer Verlag: New York.

And don't foget to check out the website for the
Managing Uncertainty in Complex Models (MUCM) project
[A consortium of Aston, Durham, LSE, Sheffield and Southampton all hard at work on developing technology for computer model uncertainty problems.]