The Bayes linear approach to uncertainty analysis for complex computer models

Michael Goldstein
Durham University *

*Thanks for support from NERC under the PURE project, and Nathan Huntley for the model analyses

Probability, Uncertainty & Risk in the Environment (PURE)

Durham University

PURE is a new action that has been prioritised by NERC in order to increase the impact of NERC natural hazards research, and to take a national leadership role in changing the way in which uncertainty and risk are assessed and quantified across the natural hazards community.

NERC is funding two consortia. Each brings together experts in many different areas of environmental science, alongside statisticians and others who specialise in uncertainty and risk. We are members of

RACER (Richard Chandler, UCL, project leader)

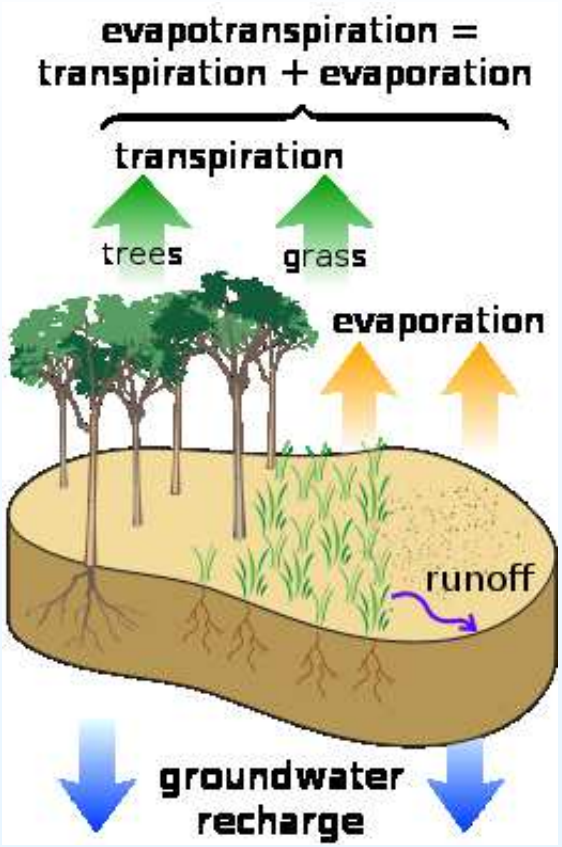
[Robust Assessment and Communication of Environmental Risk]

The other consortium is CREDIBLE (Thorsten Wagener, Bristol, Project Leader) [Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation]

We are concerned initially with Flood uncertainty and risk assessment.

In this talk, we'll address our investigations for flood models as a way of introducing the general Bayes linear approach for uncertainty analysis for computer simulators.

Water cycle



Water cycle of the Earth's surface, showing the individual components of transpiration and evaporation that make up evapotranspiration.
Other closely related processes shown are runoff and groundwater recharge.
This cycle is driven by rainfall (among other things).

Flood modelling

Durham University

The initial objective of the flooding strand is to investigate uncertainties using the FUSE rainfall runoff model, particularly those arising from non-stationary features such as applications to different catchments and time-related changes such as changing land use.

The first stage of the project is to adapt and apply computer model techniques to FUSE, and to compare with current methods used in hydrology.

The sites to be examined are the Pontbren catchment in Wales and the Eden catchment in Cumbria. This talk focuses on the Pontbren site.

Pontbren Catchment

The Pontbren catchment, in mid-Wales near Welshpool, is a fairly small catchment (around 32 square kilometres; 12 square miles).
A diagram of the catchment can be seen below (from the documentation of the Flood Risk Management Research Consortium, who provided the data). There are six rain gauges, marked by white triangles.

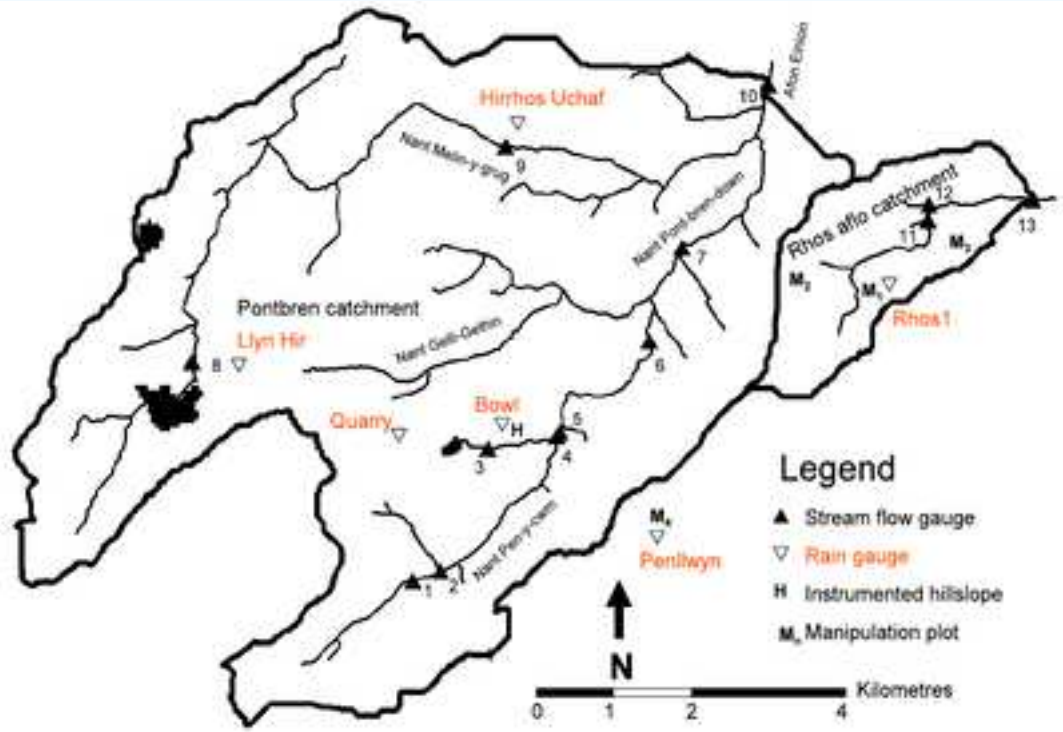


Figure 1. Showing the instrument locations across the Pontbren study site.

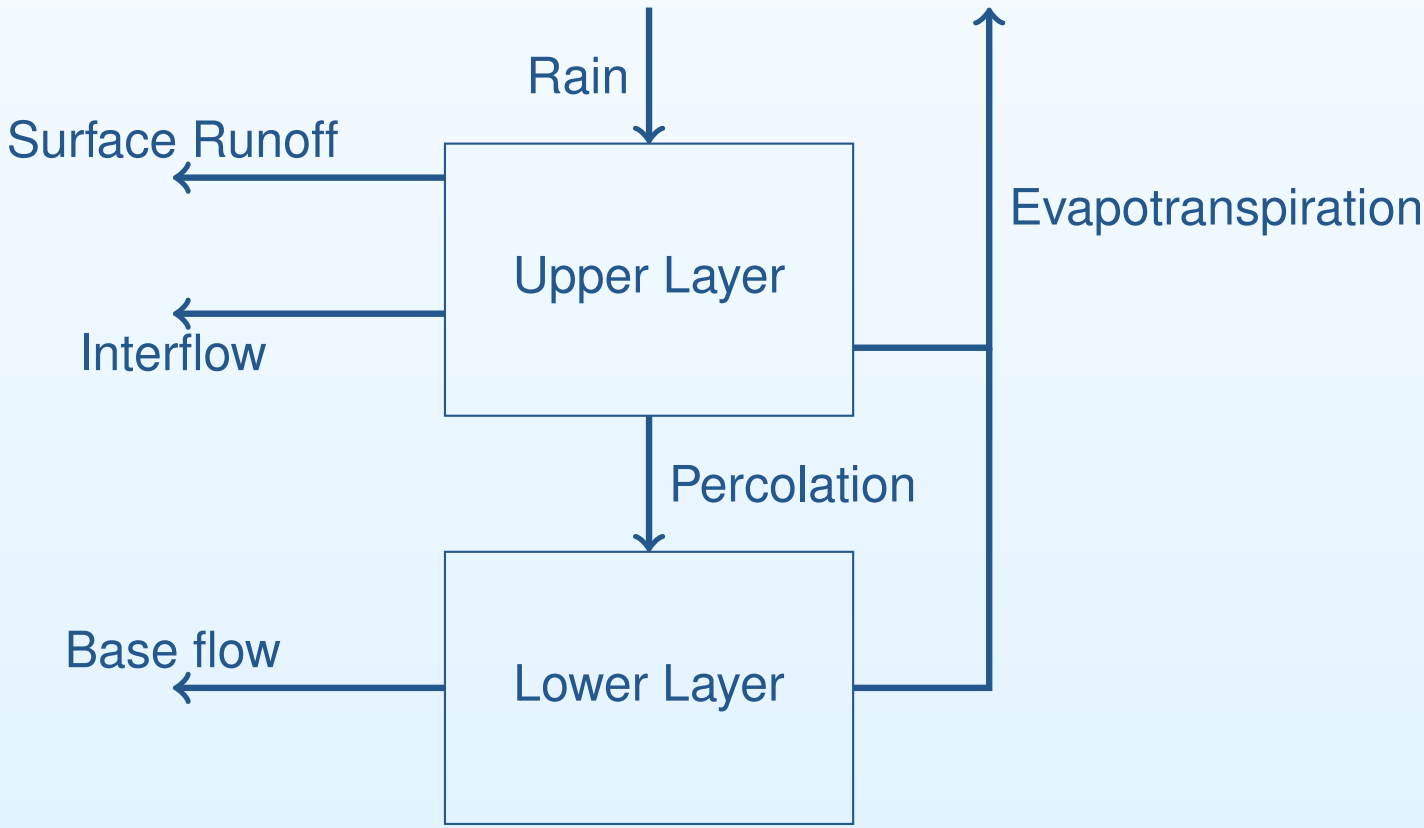
The FUSE model

Durham University

FUSE (Framework for Understanding Structural Errors) is a collection of simple models to predict rainfall runoff, designed as a toolbox to allow users to explore the suitable sub-models to use for each process in the model.
It allows the creation of very many (about 1200) models that are fast to run (about 7 seconds for a year of output on our desktop computers).
In most catchments it will not be clear which of the many possible models for the processes will describe the catchment best. For this presentation, this aspect is not relevant so we consider only one of the models.
It was presented in the paper M. P. Clark, A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02
FUSE has been implemented in FORTRAN by Martyn Clark, and then into R by the RHydro team, which includes RACER members Wouter Buytaert, Nataliya Bulygina, and Claudia Vitolo from Imperial College (our collaborators for this strand of the RACER project).

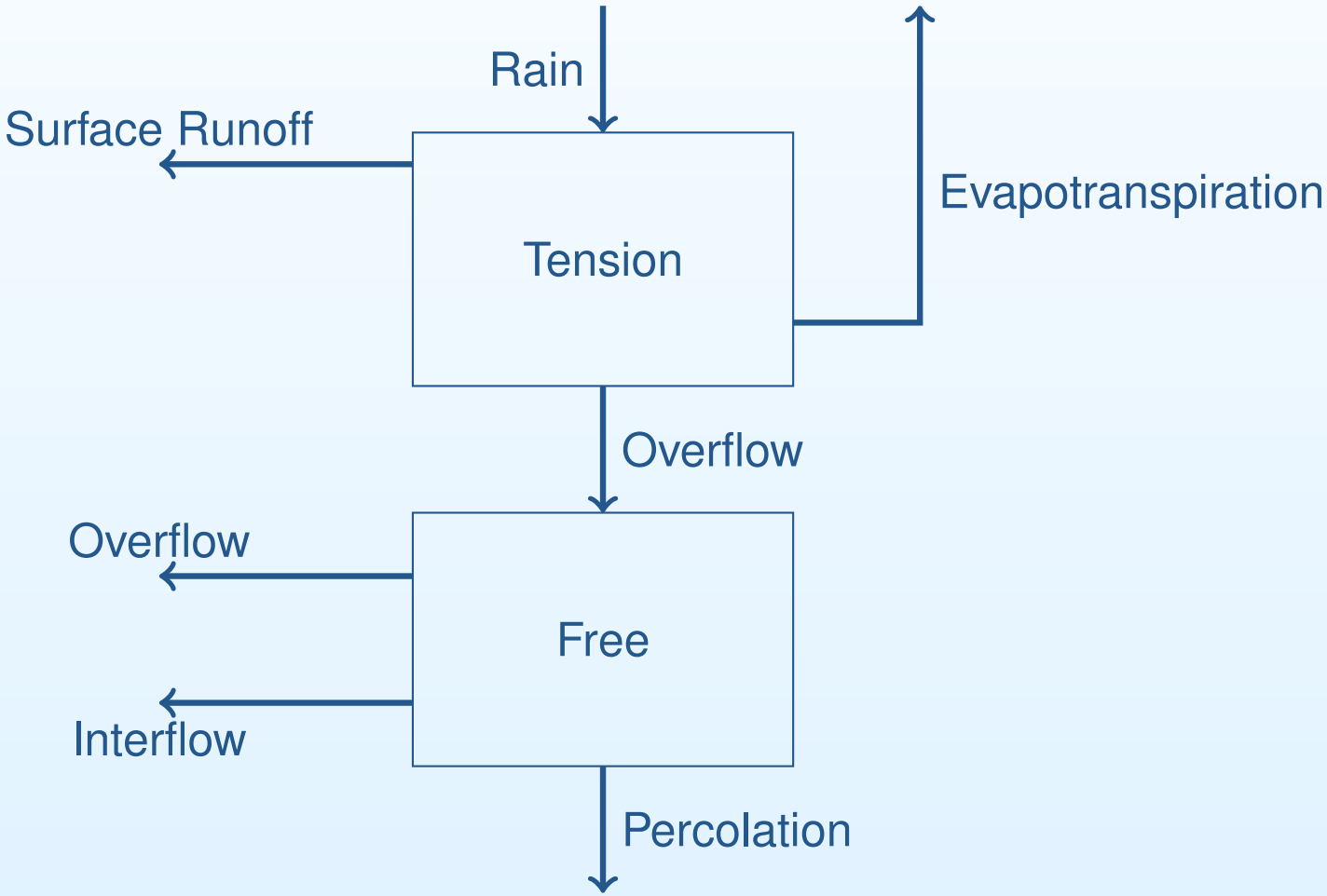
Model Outline

The FUSE models all have the same basic form: the ground is divided into an upper layer and a lower layer, as shown below. Note that the minor interflow process can be disabled; all other processes are always active but their parametrisation depends on the model structure used.



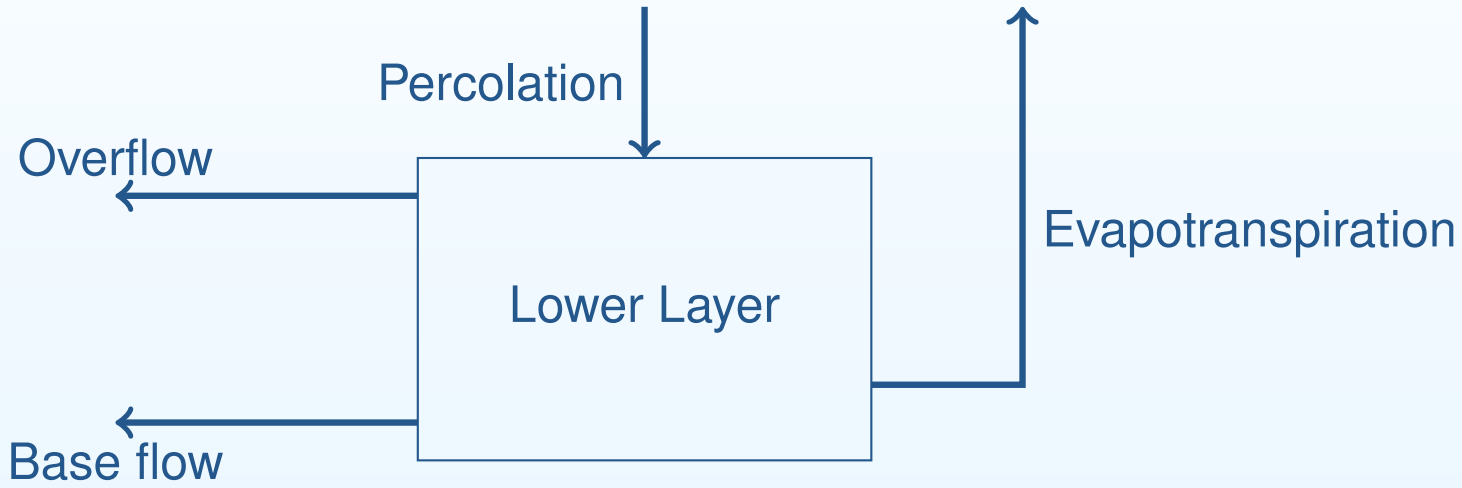
Upper Layer Structure

The upper layer can be represented in several different ways. In the model we consider, it is split into a tension compartment and a free compartment. Most of the fluxes apply only to the free compartment.



Lower Layer Structure

In our model, the lower layer is represented by a single compartment of fixed size. If this overflows, the excess water is added to base flow.



It is also possible for the lower layer to be modelled as an infinitely-large compartment. In this case, there can be no overflow, and also the evapotranspiration flux does not occur at this layer.

The FUSE model

Durham University

To run FUSE we need to provide:

- A time series of average rainfall across the catchment;
- A time series of potential evapotranspiration;
- A model structure choice: FUSE contains 8 switches each with 2–4 values that determine the structure of the model and the equations governing the fluxes. This leads to around 1200 possible models. We consider only model structure 17;
- Up to 16 parameters (described further later). For model structure 17, 9 parameters are used.

Parameters

Durham University

Parameters in FUSE govern the sizes of the various compartments, and the constants in the flux equations.
There are 22 parameters in total, but most only have influence for certain choices of switches.
The actual number of parameters needed to run the model ranges from 7 to 16.
The roles of the parameters can be broken down into:

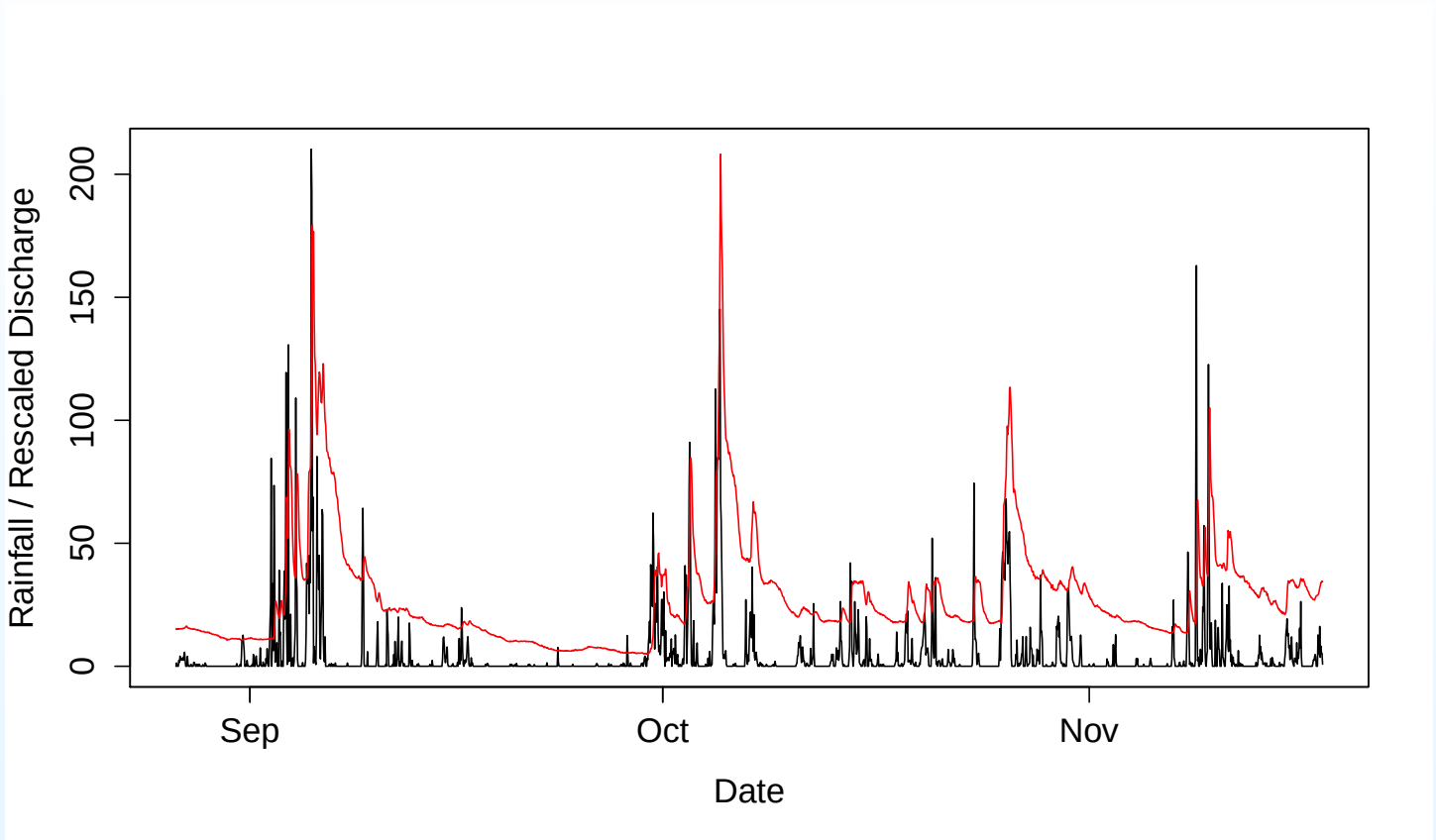
- Five governing the sizes of the compartments;
- One governing evapotranspiration;
- Five governing percolation from the upper layer to the lower layer;
- Six governing base flow (the movement of water in the lower layer);
- Four governing surface runoff (the movement of water that does not enter the upper layer at all);
- One governing the time delay of runoff.

Available Data

Durham University

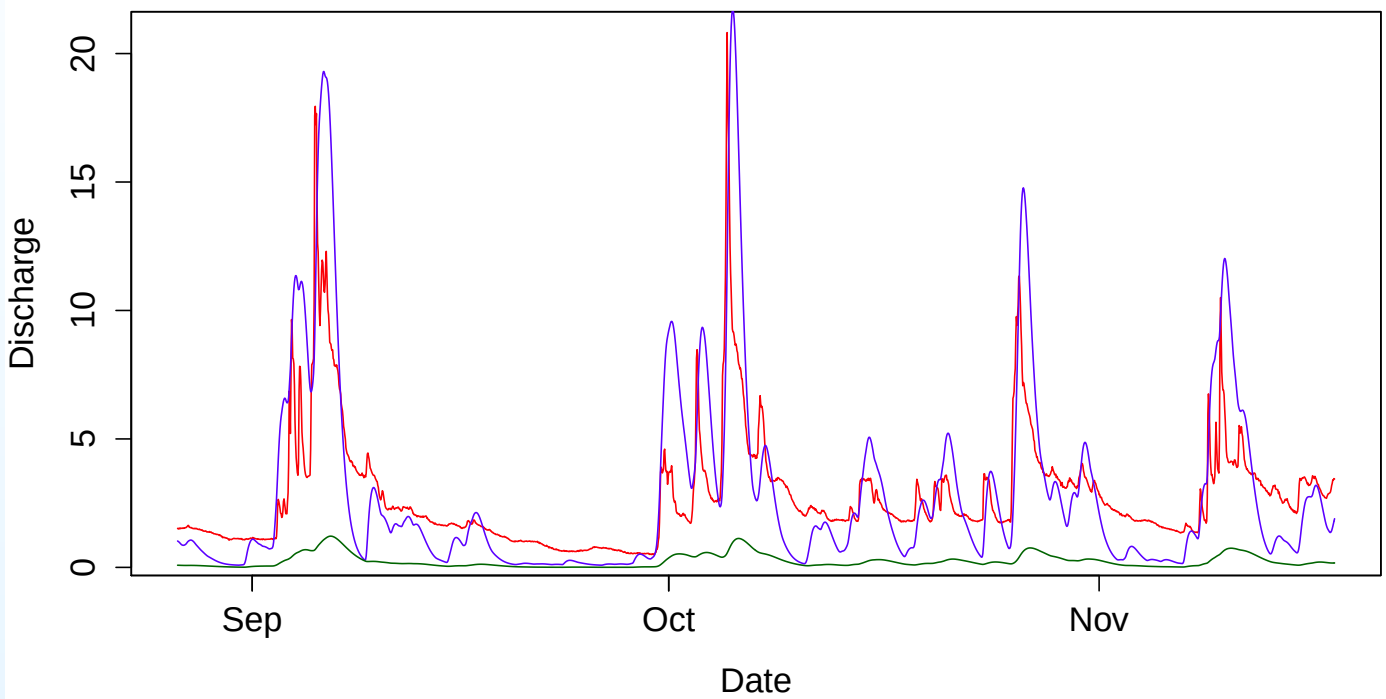
The potential evapotranspiration time series is estimated from hourly weather data using the Penman-Monteith equation.
This equation calculates evapotranspiration from temperature, wind speed, humidity, solar radiation, and conductivity of air and plants.
The areal rainfall time series is calculated from the readings of six rain gauges, which provide rainfall data every 10 minutes (or so), not always synchronised
These measurements were used to produce hourly rainfall estimates for each gauge (to synchronise with evapotranspiration); when a gauge's measurements overlapped two hours, the measurement was distributed proportionally between the hours.
We then estimated the average over the whole catchment using kriging.
The discharge is measured by a stream flow gauge every fifteen minutes (or so). This was switched to hourly values in the same way as the rain gauge data.
We concentrate on about 1 1/3 years of data where all the equipment appeared to be both present and functioning correctly for the most part.

Rainfall and Observed Discharge



This plot shows how rainfall (black line; units are mm per day) drives observed stream flow discharge (red line; litres per second rescaled in this plot to match the scale of the rainfall), over the time period of interest in this talk.

Model Runs and Observations



This plot shows observed discharge (red line), a model run at a fairly good parameter choice (blue line), and a model run at a very poor parameter choice (green line) over the time period of interest.

Complex physical models

Durham University

Most large and complex physical systems are studied by mathematical models, implemented as high dimensional computer simulators.

Examples

Systems biology simulators

Oil reservoir simulators,

Climate simulators

Galaxy formation simulators

Natural hazards simulators

Energy systems simulators

To use complex simulators to make statements about physical systems (like climate), we need to quantify the uncertainty involved in moving from the model to the system.

Sources of Uncertainty

Durham University

(i) parametric uncertainty (each model requires a, typically high dimensional, parametric specification)
(ii) condition uncertainty (uncertainty as to boundary conditions, initial conditions, and forcing functions),
(iii) functional uncertainty (model evaluations take a long time, so the function is unknown almost everywhere)
(iv) stochastic uncertainty (either the model is stochastic, or it should be),
(v) solution uncertainty (as the system equations can only be solved to some necessary level of approximation).
(vi) structural uncertainty (the model only approximates the physical system),
(vii) measurement uncertainty (as the model is calibrated against system data all of which is measured with error),
(viii) multi-model uncertainty (usually we have not one but many models related to the physical system)
(ix) decision uncertainty (to use the model to influence real world outcomes, we need to relate things in the world that we can influence to inputs to the simulator and through outputs to actual impacts. These links are uncertain.)

Current state of the art

Many people work on different aspects of these uncertainty analyses

Great resource: the Managing Uncertainty in Complex Models web-site

http://www.mucm.ac.uk/ (for references, papers, toolkit, etc.)

[MUCM is a consortium of U. of Aston, Durham, LSE, Sheffield, Southampton - with Basic Technology funding. Now mutating into MUCM community.]

However, in practice, it is extremely rare to find a serious quantification of the total uncertainty about a complex system arising from the all of the uncertainties in the model analysis.

Without an informed judgement for the reliability of a model based analysis, there is no basis for identifying appropriate real world decisions based on such an analyses.

This is

(i) partly a problem of attitude/awareness/resource

(ii) partly a problem of lack of supporting technology

(iii) partly a problem of fundamental difficulty

General form of the problem

Different physical models vary in many aspects, but the formal structures for analysing the physical system through computer simulators are very similar (which is why there is a common underlying methodology).

Each simulator can be conceived as a function $f(x)$, where

x: input vector, representing unknown properties of the physical system;

$f(x)$: output vector representing system behaviour.

Interest in general qualitative insights plus some of the following.

the "appropriate" (in some sense) choice, x^*, for the system properties x,

how informative $f(x^*)$ is for actual system behaviour, y.

the use that we can make of historical observations z, observed with error on a subset y_h of y, both to test and to constrain the model,

the optimal assignment of any decision inputs, d, in the model.

[In a rainfall runoff model, y_h might correspond to historical measurements of stream flow discharge, y to current and future discharge, and the "decisions" might correspond to different policy relevant choices such as changes in land use.]

"Solving" these problems

Durham University

If observations, z, are made without error and the model is a perfect reproduction of the system, we can write $z = f_h(x^*)$, invert f_h to find x^*, learn about all future components of $y = f(x^*)$ and choose decision elements of x^* to optimise properties of y.

COMMENT: This would be very hard.

In practice, the observations z are made with error. The simplest way is to represent the data as

$$z = y_h + e$$

where e (observation error) has some appropriate probabilistic specification independent of y_h.

We must now carry out statistical inversion/optimisation

COMMENT: This is much harder.

COMMENT And we still haven't accounted for the difference between the simulator and the physical system, condition uncertainty, multi-model uncertainty, etc.

The meaning of an uncertainty analysis

Durham University

A Bayesian analysis results in a collection of uncertainty statements about quantities such as future system behaviour.

What do these uncertainty statements mean?

This quote from the BBC web-site is typical:
'Fortunately, rapid climate change is one area that the UK has taken the lead in researching, by funding the Rapid Climate Change programme (RAPID), the aim of which is to determine the probability of rapid climate change occurring.'

This means what exactly?

RAPID-WATCH

Durham University

What are the implications of RAPID-WATCH observing system data and other recent observations for estimates of the risk due to rapid change in the MOC? In this context risk is taken to mean the probability of rapid change in the MOC and the consequent impact on climate (affecting temperatures, precipitation, sea level, for example). This project must:

* contribute to the MOC observing system assessment in 2011;
* investigate how observations of the MOC can be used to constrain estimates of the probability of rapid MOC change, including magnitude and rate of change;
* make sound statistical inferences about the real climate system from model simulations and observations;
* investigate the dependence of model uncertainty on such factors as changes of resolution;
* assess model uncertainty in climate impacts and characterise impacts that have received less attention (eg frequency of extremes).

The project must also demonstrate close partnership with the Hadley Centre.

RAPID-WATCH

Durham University

What are the implications of RAPID-WATCH observing system data and other recent observations for estimates of the risk due to rapid change in the MOC? In this context risk is taken to mean the probability of rapid change in the MOC and the consequent impact on climate (affecting temperatures, precipitation, sea level, for example). This project must:

* contribute to the MOC observing system assessment in 2011;
* investigate how observations of the MOC can be used to constrain estimates of the probability of rapid MOC change, including magnitude and rate of change;
* make sound statistical inferences about the real climate system from model simulations and observations;
* investigate the dependence of model uncertainty on such factors as changes of resolution;
* assess model uncertainty in climate impacts and characterise impacts that have received less attention (eg frequency of extremes).

The project must also demonstrate close partnership with the Hadley Centre.

RAPID-WATCH

Durham University

What are the implications of RAPID-WATCH observing system data and other recent observations for estimates of the risk due to rapid change in the MOC? In this context risk is taken to mean the probability of rapid change in the MOC and the consequent impact on climate (affecting temperatures, precipitation, sea level, for example). This project must:

* contribute to the MOC observing system assessment in 2011;
* investigate how observations of the MOC can be used to constrain estimates of the probability of rapid MOC change, including magnitude and rate of change;
* make sound statistical inferences about the real climate system from model simulations and observations;
* investigate the dependence of model uncertainty on such factors as changes of resolution;
* assess model uncertainty in climate impacts and characterise impacts that have received less attention (eg frequency of extremes).

The project must also demonstrate close partnership with the Hadley Centre.

Subjectivist Bayes

Durham University

In the subjectivist Bayes view, a probability statement is the uncertainty judgement of a specified individual, expressed on the scale of probability. This interpretation has a well-defined and operational meaning.
In this interpretation, any probability statement is the judgement of a named individual, so we should speak not of “the probability of rapid climate change”, but instead of this expert’s probability or that expert’s probability (or this group of experts shared probability) of rapid climate change.
So, a natural objective of a scientifically rigorous uncertainty analysis should be probabilities which are
asserted by at least one person, expert in the area
for reasons that are open to outside scrutiny
including consideration of the range of alternative probability judgements which could plausibly be reached by other experts in the area

Whenever you discuss an uncertainty, please be clear whether
(i) you mean the uncertainty of a (named) individual, or
(ii) you mean something else (in which case please explain what this is).

Bayesian uncertainty analysis for complex models

Durham University

Aim: to tackle previously intractable problems arising from the uncertainties inherent in imperfect computer models of highly complex physical systems, using a Bayesian formulation. This involves

- prior probability distribution for best inputs x^*
- a probabilistic uncertainty description 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 elicitation, 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 values for simulator inputs (calibration; history matching);
- assess the future behaviour of the system (forecasting).
- "optimise" the performance of the system

Approaches for Bayesian analysis

Durham University

For very large scale problems a full Bayes analysis is very hard because
(i) it is difficult to give a meaningful full prior probability specification 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 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).
Within the Bayesian approach, we have two choices.
(i) Full Bayes analysis, with complete joint probabilistic specification of all of the uncertain quantities in the problem, or
(ii) Bayes linear analysis, based on a prior specification of the means, variances and covariances of all quantities of interest, where we make expectation, rather than probability, the primitive for the theory, following de Finetti “Theory of Probability”(1974,1975).

Bayes linear approach

Durham University

de Finetti chooses expectation over probability as, if expectation is primitive, then we can choose to make as many or as few expectation statements as we choose, whereas, if probability is primitive, then we must make all of the probability statements before we can make any of the expectation statements, so that we have the option of restricting our attention to whatever subcollection of specifications we are interested in analysing carefully.

The approach to Bayesian analysis based on expectation as primitive is termed (by me) Bayes linear analysis (because of the linearity properties of expectation).

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 we take as primitive.

Full Bayes analysis can be very informative if done extremely carefully, both in terms of the prior specification and the analysis. Bayes linear analysis is partial but easier, faster, more robust particularly for history matching and forecasting.

Bayes linear adjustment

Durham University

Bayes Linear adjustment of the mean and the variance of y given z is

$$\mathsf{E}_z[y] = \mathrm{E}(y) + \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}(z - \mathrm{E}(z)),$$
$$\mathsf{Var}_z[y] = \mathrm{Var}(y) - \mathrm{Cov}(y, z)\mathrm{Var}(z)^{-1}\mathrm{Cov}(z, y)$$

$\mathsf{E}_z[y]$, $\mathsf{Var}_z[y]$ are the expectation and variance for y adjusted by z.
Bayes linear adjustment may be viewed as:
an approximation to a full Bayes analysis;
or
the "appropriate" analysis given a partial specification based on expectation as primitive (with methodology for modelling, interpretation and diagnostics).
The foundation for the approach is an explicit treatment of temporal uncertainty, and the underpinning mathematical structure is the inner product space (not probability space, which is just a special case).

Function emulation

Uncertainty analysis, for high dimensional problems, is particularly challenging if the function $f(x)$ is expensive, in time and computational resources, to evaluate for any choice of x. [For example, large climate models.]

In such cases, f must be treated as uncertain for all input choices except the small subset for which an actual evaluation has been made.
Therefore, we must construct a description of the uncertainty about the value of $f(x)$ for each x.
Such a representation is often termed an emulator of the function - the emulator both suggests an approximation to the function and also contains an assessment of the likely magnitude of the error of the approximation.
We use the emulator either to provide a full joint probabilistic description of all of the function values (full Bayes) or to assess expectations variances and covariances for pairs of function values (Bayes linear).

Form of the emulator

Durham University

We may represent beliefs about component f_i of f, using an emulator:

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

$Bg(x)$ expresses global variation in f: $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x (for example, polynomials)
$u(x)$ expresses local variation in f: $u_i(x)$ is a weakly second order stationary stochastic process, with (for example) correlation function

$$\mathrm{Corr}(u_i(x), u_i(x')) = \exp(-(\frac{\|x-x'\|}{\theta_i})^2)$$

We fit the emulators, given a collection of carefully chosen model evaluations, using our favourite statistical tools - generalised least squares, maximum likelihood, Bayes - with a generous helping of expert judgement.
We need careful (multi-output and level) experimental design for informative model evaluations, and detailed diagnostics to check emulator validity.

Linked emulators

Durham University

If the simulator is really slow to evaluate, then we emulate by jointly modelling the simulator with a fast approximate version, f^c, plus older generations of the simulator which we've already emulated and so forth.
So, for example, based on many fast simulator evaluations, we build emulator

$$f_i^c(x) = \sum_j \beta_{ij}^c g_{ij}(x) \oplus u_i^c(x)$$

We use this form as the prior for the emulator for $f_i(x)$.
Then a relatively small number of evaluations of $f_i(x)$, using relations such as

$$\beta_{ij} = \alpha_i \beta_{ij}^c + \gamma_{ij}$$

lets us adjust the prior emulator to an appropriate posterior emulator for $f_i(x)$.
[This approach exploits the heuristic that we need many more function evaluations to identify the qualitative form of the model (i.e. choose appropriate forms $g_{ij}(x)$, etc) than to assess the quantitative form of all of the terms in the model - particularly if we fit meaningful regression components.]

Emulating FUSE

Durham University

(All analyses carried out by Nathan Huntley)

The FUSE output is initially quite difficult to emulate well, so our first choice of summary measure was governed by what gave an acceptable and useful fit.

We found that the log of the maximum discharge over one of the largest rainfall events (August and September 2008) worked well for this purpose.

We used model structure 17 for this example. In this structure, the switches that are expected to have minor impact are set at the simplest choices: interflow disabled (no parameters), evapotranspiration governed by the simplest model (no parameters), percolation governed by two parameters, and surface runoff governed by one parameter.

We want the switches thought to have major impact to be set so as to give the most general model possible: the upper layer is composed of three compartments (3 parameters governing the sizes); time delay is enabled (1 parameter governing the average delay). However, the three-compartment upper layer in practice doesn't differ much from the two-compartment version, so the latter was chosen.

Parameters for Model Structure 17

Durham University

The parameters that influence the chosen model structure

Parameter	Description
x_1	Maximum storage in the upper layer
x_2	Maximum storage in the lower layer
x_3	Fraction of total storage that is tension storage
x_7	Percolation rate
x_8	Percolation exponent
x_{13}	Baseflow rate
x_{14}	Baseflow exponent
x_{19}	Exponent for surface runoff
x_{22}	Mean time delay

Emulation

Durham University

We ran the model at 1000 points using a Latin hypercube design with ranges for the parameters taken from Clark et al.

After transforming a parameter, we found that a quadratic fit (with interactions) proved to give a satisfactory fit, giving an adjusted R^2 of about 0.97.

A Gaussian correlation was chosen, with the distance parameter selected by first choosing a plausible value, checking how many of the model runs could be predicted adequately from all the other runs, and adjusting the distance parameter accordingly, settling on a value of 1/3.

The parameters x_2, x_7, x_{13}, x_{14} (involving lower-layer procedures) made little difference to the fit, so were removed.

When variables are removed from the emulator, we must assess the increase in variance, for example by fixing all active parameters and running the model at a variety of settings of the inactive parameters, calculating the maximum each time. This gives an assessment of the variance due to this “nugget” effect.

Emulator Performance

Durham University

We assessed the performance of the emulator by predicting 100 of the model runs from the remaining 900.

The emulator predicted about 95% of these within two standard deviations and about 96% within three standard deviations.

Examining the cases where the emulator failed to get close to the true run showed that it was overestimating the maximum for a region of the parameter space that yielded extremely low model output.
Even in this space, the emulator's predictions were much lower than the observed maximum, even though they didn't reach as low as the actual model runs.

In the regions where model output was similar to the observations or much higher, the emulator performed better, approached 99% accuracy.

Uncertainty analysis for perfect models

Durham University

Minimal formulation for a full Bayesan treatment. Suppose

$$z = y_h + e,\ y_h = f(x^*)$$

Specify
a prior distribution for x^*,
a likelihood for the observational error e,
a probabilistic emulator for f,

Carry out Bayesian updates for emulator given evaluations of the function and for x^* by observation of the data z.

This is an analysis of all of the uncertainties arising directly in the minimal Bayesian description of the problem. It is a common 'state of the art' method of carrying out an uncertainty analysis based on a computer simulator.

The analysis is conceptually straightforward, though it may be technically challenging, requiring particular care when constructing the emulators for all the functional outputs and dealing with the computational difficulties arising from high dimensional and often highly multimodal likelihood functions

Limitations of ‘perfect model’ analysis

Durham University

A physical model is a description of the way in which
system properties (the inputs to the model)
affect system behaviour (the output of the model).
This description involves two basic types of simplification.

(i) we approximate the properties of the system (as these properties are too complicated to describe fully and anyway we don’t know them)

(ii) we approximate the rules for finding system behaviour given system properties (because of necessary mathematical simplifications, simplifications for numerical tractability, and because we do not fully understand the physical laws which govern the process).

Neither of these approximations invalidates the modelling process. Problems only arise when we forget these simplifications and confuse the internal uncertainty analysis of the model with the corresponding uncertainty analysis for the physical system itself.
Basic principle: it is always better to recognise than to ignore uncertainty, even if modelling and analysis of uncertainty is difficult and partial.

Internal and external uncertainty

Durham University

Models do not produce statements about reality, however carefully they are analysed. Such statements require **structural uncertainty** assessment, taking account the mismatch between the simulator and the physical system.
We may distinguish two types of model discrepancy.

(i) **Internal discrepancy** This relates to any aspect of structural discrepancy whose magnitude we may assess by experiments on the computer simulator. For example, if the simulator requires the values of a forcing function (like rainfall) only known with error, then assessing variation on simulator outomes for a collection of forcing functions simulated within this magnitude of error, gives order of magnitude error for this aspect of structural discrepancy. Internal discrepancy analysis gives a lower bound on the structural uncertainty that we must introduce into our model analyses.

(ii) **External discrepancy** This relates to inherent limitations of the modelling process embodied in the simulator. There are no experiments on the simulator which may reveal this magnitude. It is determined by a combination of expert judgements and statistical estimation.

Structural uncertainty analysis

Durham University

One of the simplest, and most popular, approaches is to suppose that there is an appropriate choice of system properties x^* (currently unknown), so that $f(x^*)$ contains all the information about the system:

$$y = f(x^*) + \epsilon$$

where ϵ, the model or structural discrepancy, has some appropriate probabilistic specification, possibly involving parameters which require estimation, and is independent of f, x_0, e.

We modify this form as follows:

$$y = f(x^*) + \epsilon^*$$

where $\epsilon^* = (\sigma(x^*) + c)\epsilon$

$\sigma(x^*)$ is a vector of scale parameters, dependent on x^*, which we will learn about using internal discrepancy experiments,

c is a vector of constants reflecting a direct external disrepancy specification,

ϵ is a vector with unit variance, independent of everything else.

Internal Discrepancy for FUSE

Durham University

We considered four sources of internal discrepancy:
initial conditions, potential evapotranspiration, rainfall, and the effects of varying the parameters with time.
[There are other sources we have yet to consider. For example, the internal discrepancy from modifying the propagation of the state vector.]
We perturb each source, observing the effects on the model output.
[1] Generate (independently) 300 perturbations of evapotranspiration, rainfall, and parameter vectors. Combine these to form 300 settings at which to run the model.
[2] Choose 50 input parameter sets at which to run the model.
[3] Run the model at each input parameter for each of the 300 perturbations.
For each parameter choice, this leads to a standard deviation for the influence of the perturbations on the maximum runoff in the region of interest.
Of these 50 standard deviations, we take the maximum to be our initial working value for internal discrepancy. This is sufficient for our first history match. We will make a more careful assessment later.

Review of Variances

Durham University

We now have several variances calculated: the emulator variance, the nugget effect, and the estimate of the internal discrepancy. We also need to consider the streamflow measurement error (assessed by experts) and the "external" variance that picks up the model discrepancies we haven't modelled (in this case chosen based on experience with the model).

Type of variance	Approximate value
Emulator variance	0.04^2
Nugget effect	0.006^2
Internal discrepancy	0.12^2
Stream flow gauge measurement error	0.03^2
External discrepancy	0.05^2

We can now consider using these variances to check whether the emulator's prediction at a given parameter choice is "close enough" to the observation.

History matching

Durham University

Model calibration aims to identify the best choices of input parameters x^*, based on matching data z to the corresponding simulator outputs $f_h(x)$.
However
(i) we may not believe in a unique true input value for the model;
(ii) we may be unsure whether there are any good choices of input parameters
(iii) full probabilistic calibration analysis may be very difficult/non-robust.
A conceptually simple procedure is "history matching", i.e. finding the collection, $C(z)$, of all input choices x for which you judge the match of the model outputs $f_h(x)$ to observed data, z, to be acceptably small, taking into account all of the uncertainties in the problem.
If $C(z)$ is non-empty, then an analysis of its elements reveals the constraints on the parameter space imposed by the data.
Further the model projections $f(x) : x \in C(z)$ over future outcomes, reveal the futures consistent with the model physics and the historical data.
If the data is informative for the parameter space, then $C(z)$ will typically form a tiny percentage of the original parameter space, so that even if we do wish to calibrate the model, history matching is a useful prior step.

History matching by implausibility

Durham University

We use an 'implausibility measure' $I(x)$ based on a probabilistic metric (eg. number of sd between z and $f_h(x)$) where $z = y_h + e,\ y_h = f_h(x^*) + \epsilon^*$ for observational error e, and model disrepancy ϵ^*.
For example, if we are matching a single output, then we might choose

$$I(x) = \frac{(z - \mathrm{E}(f_h(x)))^2}{\mathrm{Var}(z - \mathrm{E}(f_h(x)))}$$

Here $\mathrm{Var}(z - \mathrm{E}(f_h(x)))$ is the sum of measurement variance, $\mathrm{Var}(e)$, structural discrepancy variance, $\mathrm{Var}(\epsilon^*)$, and emulator variance $\mathrm{Var}(f_h(x))$.
The implausibility calculation can be performed univariately, or by multivariate calculation 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 implausible, i.e. unlikely to be good choices for x^*.
With this information, we can then refocus our analysis on the 'non-implausible' regions of the input space, by (i) making more simulator runs & (ii) refitting our emulator over such sub-regions and repeating the analysis. This process is a form of iterative global search.

History Matching for FUSE

Durham University

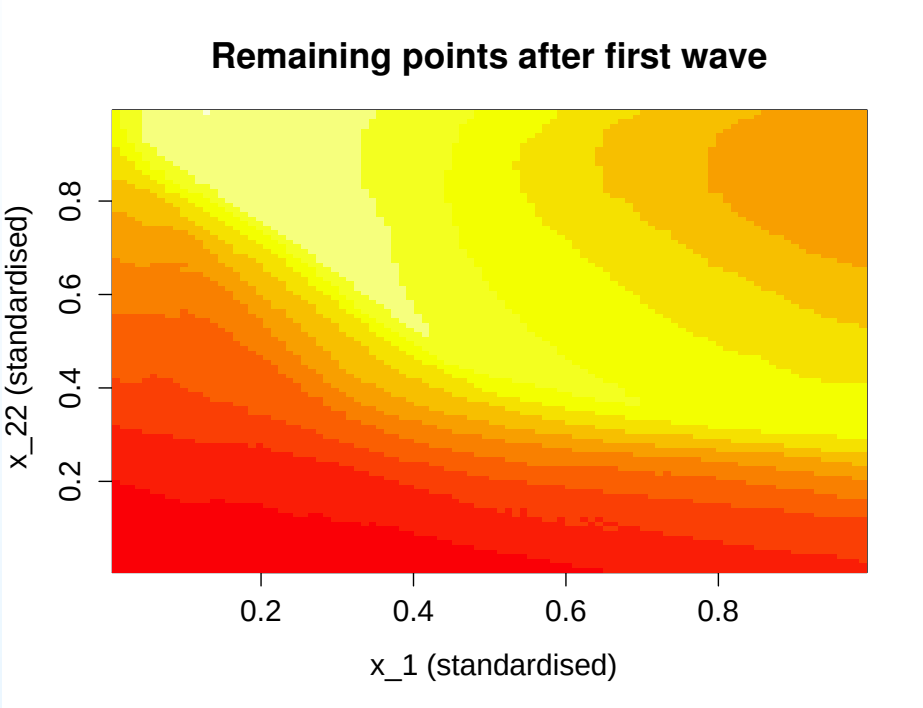
We may calculate $I(x)$ for any x. We apply the 3-σ rule to set our cutoff for $I(x)$ (at least 95% of any continuous unimodal distribution is within 3-σ of the mean.)

So, for any point x we can quickly judge whether it is implausible or non-implausible.

Thus, we can divide the parameter space into two regions: implausible and non-implausible. It is considered that the best input x^* is unlikely to be in the implausible region. This allows us to focus our analysis on a reduced region.

Visualising the non-implausible region in all 9 dimensions is difficult, but we can consider two-dimensional plots. If we fix two parameters and vary the other seven over a range of values, we can calculate what proportion are implausible. We can calculate such a proportion across a grid of the two parameters, thus producing a heat map of implausibility.

Visualising History Matching: Wave 1



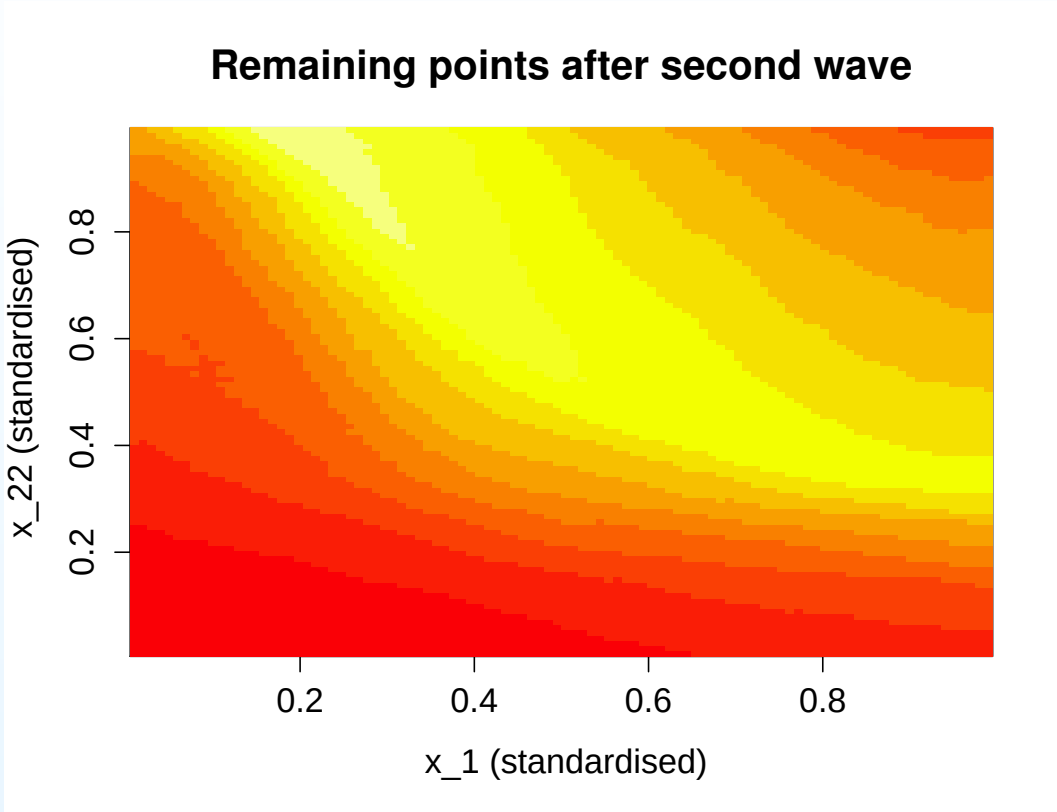
Plots such as those above illustrate the regions that history matching removes. Each point on the heatmap was generated by fixing x_1 (maximum upper layer storage) and x_{22} (time delay), and varying the remaining parameters over a grid, and calculating the proportions that would be rejected by history matching. In redder areas, most points were rejected. In whiter areas, most points were accepted.

History Matching: Second Wave

Durham University

We generate a new collection (around 1000) of model runs, all of which are within the non-implausible region, and build a second emulator.
The second stage emulator provides a better fit, allowing some points that are not rejected by the first emulator to be rejected.
We also generate a new collection (around 40) of non-implausible parameter choices to run further internal discrepancy experiments.
At this stage, instead of using the maximum, we emulate the internal discrepancy variance.
For every new parameter choice x, this will give an expectation and a variance for the output of the model at x, and an expectation and a variance for the outcome of the internal discrepancy experiments at x.
Internal discrepancy is, typically, a smoother function that the model output itself. This allows us to build a good emulator (R^2 around 0.9) for the internal discrepancy standard deviation.
For a candidate parameter choice x, we again calculate the implausibility measure $I(x)$, but this time instead of using the maximum internal discrepancy observed, we use a value derived from the internal discrepancy emulator.

Visualising History Matching



In this plot, a point is judged implausible if it is judged implausible by either the first emulator or the second emulator. We can see that, for x_1 and x_{22} at least, the second wave of history matching does not dramatically alter the shape of the regions.

History Matching: Third Wave

Durham University

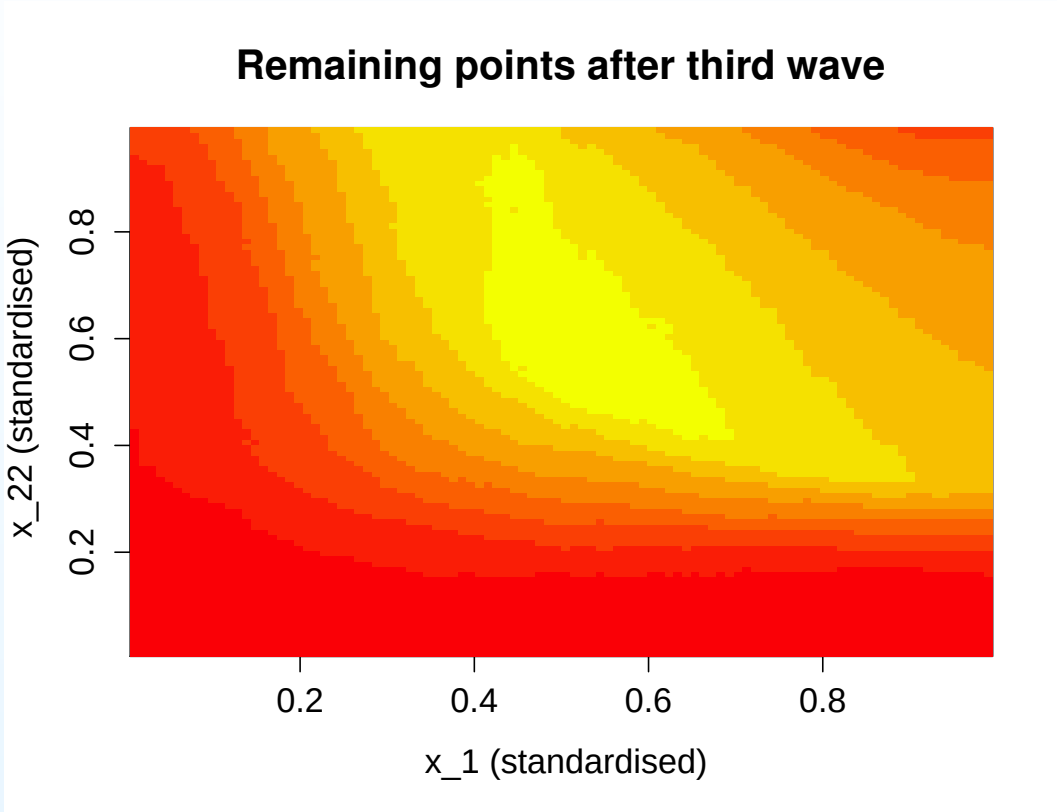
Once again we generate around 1000 non-implausible parameters for new model runs, and around 40 non-implausible parameters for internal discrepancy runs.

At this stage we make two further changes.

After a number of waves of history matching, continuing to match on the maximum will no longer remove points. This is the case by the third wave. Instead we emulate and match on the total rainfall over the period of interest. Initially this was impossible to emulate well (R^2 around 0.6) but in the reduced space the total rainfall is better behaved (R^2 around 0.9). This demonstrates another feature of history matching: new quantities may be emulated that couldn't be initially.

The second change is that at this stage the magnitude parameter perturbations for the internal discrepancy runs were allowed to vary, so this becomes a tenth parameter. This allows us to examine how the non-implausible region changes when we change the perturbation magnitude.

Visualising History Matching



Again, the new wave refines the regions rather than greatly changes their shape.

We have now eliminated about 2/3 of the original parameter space.

Forecasting

Durham University

Suppose that we want to predict some future outcome y_p, given observed historical data on earlier values z_h.
The Bayes linear update of y_p given z_h requires the joint variance structure of y_p, z_h.
This can be derived directly from the decomposition

$$y = f(x^*) + \epsilon^*, \;\; z_h = y_h + e$$

(i) computing the joint mean, variance and covariance of $f_h(x^*), f_p(x^*)$, by first conditioning on x^* and then integrating out x^* using the emulator for f.
(ii) assessing the mean, variance and covariance of ϵ^* directly (largely from our internal discrepancy experiments).
This analysis is tractable even for large systems. It is fast and flexible enough to function as a real time control system, if we need to make decisions to optimise forecast system behaviour.
When the forecast variance is large, then we have methods to improve forecast accuracy.

Discrepancy correlation and forecasting in FUSE

Durham University

Our internal discrepancy experiments, at the final stage, allow us to assess the correlation matrix for internal discrepancies across historical and future outputs. Here is the correlation matrix relating the past log max and sum to the log max and sum for the next major event (assessed for one parameter choice; most other parameters gave essentially the same differing by no more than 0.05).

	historical max	historical sum	future max	future sum
historical max	1.00	0.91	0.73	0.79
historical sum		1.00	0.79	0.89
future max			1.00	0.91
future sum				1.00

Forecasting the log max

Without considering covariance from internal discrepancies:

Expected value: 2.933 Standard deviation: 0.070

Including the internal discrepancy correlation:

Expected value: 2.943 Standard deviation: 0.066

Observed value: 3.0357 with measurment sd: 0.033

Concluding comments

Durham University

Problems in science and technology arising from computer simulators for physical systems are extremely common. Each such problem involves a substantial uncertainty quantification task. Statisticians have a central role to play in this task.

To assess our uncertainty about complex systems, it is enormously helpful to have an overall (Bayesian) framework to unify all of the sources of uncertainty.

Within this framework, all of the scientific, technical, computational, statistical and foundational issues can be addressed in principle.

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

In particular,
Bayes (linear) multivariate, multi-level, multi-model emulation,
careful structural discrepancy modelling
and iterative history matching and forecasting
gives a great first pass treatment for most large modelling problems.

References:applications

P.S. Craig, M. Goldstein, A.H. Seheult, J.A. Smith (1997). Pressure matching for hydocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in Case Studies in Bayesian Statistics, vol. III, eds. C. Gastonis et al. 37-93. Springer-Verlag.

J. Cumming, and M. Goldstein (2009). Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments. In The Oxford Handbook of Applied Bayesian Analysis. O'Hagan, and West (eds), Oxford University Press. 241-270.

I. Vernon, M.Goldstein, and R. Bower (2010) Galaxy Formation: a Bayesian Uncertainty Analysis (with discussion), Bayesian Analysis, 5, 619-670

D. Williamson, M. Goldstein, et al (2013) History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. Climate Dynamics, 41 (7-8). 1703-1729

References:general

Goldstein, M., Seheult, A. & Vernon, I. (2013). Assessing Model Adequacy. In Environmental Modelling: Finding Simplicity in Complexity. Mulligan, Mark. & Wainwright, John. Wiley-Blackwell. 435-449.

M. Goldstein and J.C.Rougier (2008). Reified Bayesian modelling and inference for physical systems (with discussion), JSPI, 139, , 1221-1239

M. Goldstein and J.C.Rougier (2006) Bayes linear calibrated prediction for complex systems, JASA, 101, 1132-1143

D. Williamson, M. Goldstein, A. Blaker (2012) Fast linked analyses for scenario-based hierarchies, Journal of the Royal Statistical Society. Series C: Applied Statistics, volume 61, no. 5, pages 665-691.

M. Goldstein (2006) Subjective Bayesian analysis: principles and practice (with discussion) Bayesian Analysis 1 403-420

M. Goldstein and D. Wooff (2007) Bayes linear Statistics, Wiley