Assessing Model Adequacy

Michael Goldstein, Allan Seheult & Ian Vernon

Department of Mathematical Sciences, Durham University,
Science Laboratories, Stockton Road, Durham DH1 3LE, UK

February 12, 2010

1 Introduction

Environmental models are simplified representations of complex physical systems. The implementation of any such model, as a computer simulator, involves further simplifications and approximations. The value of the resulting simulator, in giving scientific and practical insights into the functioning of the corresponding physical system, depends both on the nature and degree of these simplifications and also on the objectives for which the model is to be used.

This chapter provides an introduction to some basic general techniques for assessing the adequacy of a computer model for its intended purpose. There are many ways to approach this question. We will take the view that the aim of the model is to provide some, necessarily partial, information about the behaviour of the system, and we will consider the model adequate for an intended task if the information that is provided by the simulator is sufficient to allow us to carry out this task. We would usually prefer precise forecasts of system behaviour but we may often be able to tolerate probabilistic forecasts, provided that we are able to quantify the level of uncertainty with which these forecasts should be interpreted, and to confirm that this uncertainty is not so large as to prevent us from achieving our objectives. This will, inevitably, be a pragmatic judgement.

Therefore, in our account, we will outline some basic methods for assessing the degree of uncertainty that it would be reasonable to associate with model outcomes. It is beyond the scope of this account to produce precise quantifications of predictive uncertainty, as such analysis requires rather more technical machinery than we have space to describe. Instead, we offer some basic tools for making order of magnitude quantifications for such uncertainties, which should indicate whether the limitations of the model are likely to render it unfit for the task at hand.

This is by no means a complete account, even for our stated goal, as such analysis is strongly dependent both on the scientific context and also on the size and complexity of the model. In the next section, we outline the general methods that we suggest and, in the following sections, we illustrate how the methods can be used in practice, by applying them to a rainfall runoff model.

2 General issues in assessing model adequacy

For the purposes of this chapter, we consider that a model is a description of how system properties affect system behaviour. We may represent the model in the general form

$$y = f(x) \tag{1}$$

where model inputs x corresponds to a vector of system properties; for example, in a rainfall runoff model, x might be a description of the physical characteristics of a particular catchment area. Some of the elements of x might be control or tuning parameters. To simplify our account, we will not make such distinctions. The model output vector y is a description of corresponding system behaviour; for example, y might be a time series description of water flows in the catchment area. The function f is a description of the way in which system properties determine system behaviour, based partly on the mathematical equations which determine y from x and partly from relevant initial and boundary conditions, forcing functions and so forth. Usually, f is implemented in the form of a computer simulator. We suppose also that we have some system field data z comprising observations made on the system corresponding to some sub-vector of y.

There are two main reasons for interest in such a model. Firstly, we may want to gain insights into the general behaviour of the model; for example, to assess which features of the system properties are most important for determining the system behaviour, how sensitive are such relationships to mis-specification and other factors. If f in (1) is a new version of a pre-existing model, then we will want to assess the form and magnitude of the changes between versions. Similarly, we may want to compare the model to other pre-existing models for the same phenomenon. There are many ways to gain such insights. One of the simplest, if the model is fast to evaluate, is to make many evaluations of the model at widely differing choices of input parameters and to carry out a careful data analysis of the resulting joint collections of input and output values. In such an analysis, we will also look for anomalous and counter-intuitive behaviour in the model which may enable us to detect errors in the computer simulator, namely features which are wrong in ways that we are able to fix. These may be simple coding errors, data transcription errors, mistakes in our implementation of numerical solvers or problems with the science used in our problem formulation for which we can see ways to formulate effective alternatives within the limitations of time and resource which are available.

Secondly, when we have completed this analysis, then we often pass to a further stage of using the model to make inferences about specific physical systems; for example, to help to understand actual water flow for specified catchment areas. We will have much greater confidence in our use of model predictions for an actual system if we have a good intuitive feel for the general behaviour of the model, and we have carried out a careful error analysis for the code. In this chapter, we will focus attention on this second stage, as it is natural to consider model adequacy in the context of practical purposes for which the model is to be used.

We therefore consider whether a model is adequate to represent a given physical system for some specified purpose. In all but the most elementary problems, the behaviour of the model will not be precisely the same as the behaviour of the system.

Partly, this is because we must simplify our description of the system properties, partly because we cannot fully describe the science determining the effect of system properties on system behaviour, partly because, even with the simplified science that we choose to implement, we will typically need to approximate the solution of the equations required to determine the relationships between system properties and behaviour and partly because the forcing functions, initial conditions, boundary conditions and so forth are rarely known with certainty. This irresolvable difference between the output of the model and the performance of the physical system is often termed *model discrepancy.*

A crucial part of the assessment of model adequacy comes from assessing the magnitude of model discrepancy and then deciding whether it is so large that this renders the model unfit for the intended uses. It is rare that we can place a precise value on this discrepancy, as, otherwise, we would have incorporated this assessment directly into the model itself. Therefore, we must usually carry out an uncertainty analysis. Rather than considering that the model makes deterministic predictions about system behaviour, we consider that the model offers probabilistic predictions for such behaviour. The level of uncertainty associated with these predictions will determine whether the model is adequate for the intended purposes.

The sources of uncertainty that we must usually deal with are: (i) input uncertainty, as we are unsure as to which is the appropriate value of the inputs at which to evaluate the model, or even whether there is any meaningful choice of input parameters; (ii) functional uncertainty, as, for complex, slow-to-run models, there will be large areas of the input space which will be explored only very lightly; (iii) observational error, complicating our ability to assess the quality of model fit to historical field data; (iv) forcing function, initial condition and boundary condition uncertainty; (v) general aspects of model uncertainty, for example problems arising when we train a model on data in one context but we intend to use the model in a very different context. We may view a model as adequate in principle if model discrepancy is small. However, all sources of uncertainty should be included in a composite uncertainty analysis, as the model will only be adequate in practice if we can control all of the relevant sources of uncertainty to a level where predictions are sufficiently accurate for the purpose in hand.

There are different views as to what constitute appropriate formulations for an uncertainty analysis. We shall describe our analysis from a Bayesian viewpoint. In this view, all uncertainties may be expressed as best current judgements in probabilistic form and then combined with observational data by the usual probabilistic rules. The advantage of this approach is that it places all of the uncertainties in relating model to system behaviour within a common framework and produces a probabilistic assessment which represents the best current judgements of the expert in a form which is appropriate for use in subsequent decision analysis.

As with any other aspect of the modelling process, we can make such a probabilistic assessment with different degrees of detail and care. It may be enough to make a rough order of magnitude assessment of the most important aspects of model discrepancy or we may need to carry out a more careful analysis. As a simple rule of thumb, the more that we intend to rely on the model to make decisions with important consequences, under substantially different conditions to those for which we have available historical data, for example, to extrapolate over large time scales, then the more careful we will need to be in our assessments of model discrepancy. We will also be limited in our ability to make a full uncertainty analysis by factors such as the dimension and complexity of the model, the time that it takes to carry out a single model evaluation, whether there are any other models against which we may compare our analysis and the nature and extent of any historical data which we may use to assess the performance of the model. In our account, we will introduce some basic analyses that we may wish to carry out. The uncertainties that we shall refer to may be assessed as variances, as full probability distributions or as an uncertainty description at some intermediate level of complexity. In our example analyses, we will illustrate some particular forms that such calculations might take.

There are two basic aspects to model discrepancy. First, we may assess intrinsic limitations to the model whose order of magnitude we may quantify by direct computer experimentation. We refer to these as *internal model discrepancies*, and quantify them by analysis of the computer output itself. There are two general types of internal discrepancy. The first type is due to lack of precise knowledge of the values of certain quantities which are required in order to evaluate the model, but which it is inappropriate to treat as part of the model input specification x. For example, if we judge that the elements of the forcing function for the system are only determined within, say, 10%, then we may assess the effect on the output of the model of making a series of model evaluations with varying values of the forcing function within the specified limits. The second type of internal discrepancy is due to acknowledged limitations in the ways in which the model equations transform system properties into system behaviour. For example, a common practical modelling structure is to determine a spatio-temporal series of system responses by propagating a state equation across time and space. Each propagation step involves a level of approximation. Provided that we have access to the governing equations of the model, we can directly assess the cumulative effect of such approximations by introducing an element of uncertainty directly into the propagation step in the equations for the system state, reimposing system constraints as necessary after propagation, and making a series of evaluations of the model based on simulating the variation in overall system behaviour with differing levels of propagation uncertainty.

The second aspect of model discrepancy concerns all of those aspects of the difference between the model and the physical system which arise from features which we cannot directly quantify by operations on the computer model. We refer to such aspects as *external model discrepancies*. Some external discrepancies may correspond to features which we acknowledge to be missing from the model and whose order of magnitude we may consider directly, at least by thought experiments. However, our basic means of learning about the magnitude of many aspects of external discrepancy is by comparing model outputs to historical field data. The difference between the historical field observations z on the system and the corresponding model outputs $f(x)$, when evaluated at the appropriate choice of inputs to represent the system properties, is the sum of the observational error and the internal and external model discrepancy errors. Provided that we have already quantified uncertainty for observational and internal model error, any further lack of fit is due to external model error, and the magnitude of such mismatch between model output and field data is therefore a guide to external model uncertainty, for historical outcomes. The extent to which this may be considered informative for such uncertainties when using the model to forecast future outcomes is a matter of scientific judgement dependent on the context of the problem in question.

In practice, we usually do not know the appropriate choices of inputs at which to evaluate the model, as achieving a good fit to historical observations is itself a common method for estimating appropriate values of the input parameters. Model calibration or tuning is a subject with an extensive literature; see, for example, Rougier (2009) and Kennedy and O'Hagan (2001). All that we are looking for at this stage is to be reasonably confident that the model is sufficiently reliable to merit such a tuning effort. A simple approach for making such an assessment is to make many evaluations of the model using a space filling design in the input parameters and to determine which choices of input parameter lead to the best fits to the field data. For high dimensional input spaces, it may not be directly feasible to make evaluations over all areas of the input space to an acceptable level of concentration. In such cases, we often use an iterative design, eliminating all input choices within the first stage design which give very poor fits and placing second stage designs centred on those evaluations which have given more reasonable fits and continuing in this manner until a collection of relatively good fits have been found.

This process is sometimes referred to as *history matching*; see, for example, Craig et al. (1997). We are not trying to determine the best choice of input parameters but simply to determine if there is some sub-collection which gives an acceptable match to historical data. It might be that every evaluation that we make of the model provides such a poor fit to the historical data that we reach the conclusion that external discrepancy is so large as to render the model unacceptable for practical use. Otherwise, assessment of the order of magnitude discrepancy between model and data in regions of good fit gives us a guide to the magnitude of external discrepancy. This method of tuning is only likely to give meaningful results if we have access to a large quantity of field data relative to the number of parameters that we may vary; otherwise, it is highly likely that we will over-fit the model to the data. If our assessment of external variance appears to be negative for many components of z, because the differences between $f(x)$ and z are small compared to observational plus internal discrepancy errors, then this suggests we have possibly over-fitted the model, and further investigation may be required.

In order to carry out the above analysis, we must make many evaluations of the model within a reasonable length of time. For many problems, this is not a realistic possibility. In such cases, we may employ the method of *model emulation.* Emulation refers to the expression of our beliefs about the function $f(x)$ by means of a fast stochastic representation, which we can use both to approximate the value of the function over the input space and also to assess the uncertainty that we have introduced from using this approximation. For example, we might represent our beliefs about the i-th component of $f(x)$ in the form

$$f_i(x) = \sum_j g_j(x)\,\beta_{ij} + u_i(x) \tag{2}$$

where each $g_j(x)$ is a known deterministic function of x, for example a polynomial term in some sub-collection of the elements of x, the β_{ij} are unknown constants to estimate and $u_i(x)$, the residual function, is specified as having zero mean and constant variance σ_i^2 for each x, with a correlation function $c_i(x, x') = \text{corr}(u_i(x), u_i(x'))$ which only depends on the distance between x and x'. There are many possible choices for the form of the $c_i(x, x')$. If we want to carry out a full probabilistic analysis, then we may suppose, for example, that $u_i(x)$ is a Gaussian process, so that the joint distribution of any sub-collection of values of $u_i(x)$ for different choices of x is multivariate normal.

There is an extensive literature on the construction of emulators for computer models, based on a collection of model evaluations.; see, for example, O'Hagan (2006) and MUCM (2009). Given these evaluations, we may choose our functional forms $g_j(x)$ and estimate the coefficients β_{ij} using standard model building techniques from multiple regression, and then assess the parameters of the residual process $u(x)$ using, for example, variogram methods on the estimated residuals from the fitted model. Given the emulator, we can then carry out the history matching procedures described above, but, instead of evaluating the function at each input choice, we evaluate the emulator expectation $\text{E}[f_i(x)]$ at each chosen x. We therefore need to add the emulator variance $\text{Var}[f_i(x)]$ to the observational variance and model error variance terms when making the comparison between z_i and $\text{E}[f_i(x)]$, but otherwise the analysis is the same as for fast-to-run models.

3 Assessing model adequacy for a fast rainfall runoff model

We consider a rainfall runoff model described in Iorgulescu et al. (2005) (henceforth IBM), that simulates fluctuations in water discharge and Calcium and Sodium concentrations over time. We illustrate our methods with its application to a particular sub-catchment of the Haute-Mentue research catchment (Switzerland); see IBM who refer to other studies and runoff models. Each model run simulates three time series: discharge (D) and the tracers Calcium (Ca) and Silicon (Si) over 839 consecutive hours. Any such simulation may be compared to the corresponding 839 hours of field data collected at the sub-catchment between August and September 1993. The field data also includes hourly rainfall which is used as a forcing function (RAIN) to the model. There is a second forcing function, actual evapotranspiration (AET), an evaporation rate, which is modelled as a deterministic sinusoidal function of time.

3.1 Mathematical Model

The model, depicted in Fig. 1, comprises three compartments with parallel transfer, whereby water, input as rain, may enter three compartments representing three different soil types, "Direct Precipitation" (DP), "Acid Soil" (AS) and "Ground Water" (GW). The water is stored in each compartment for a fast or slow amount of time before being discharged into the streams. The water can instead enter the "Ineffective Storage" compartment, in which case it will not be discharged and can only leave the system via actual evapotranspiration (AET). Six parameters $a_{soil}, b_{soil}, k_{soil}, p_{soil}, c^f_{soil}$ and c^s_{soil} characterise the fluid dynamics of water flow through each *soil* (DP, AS, GW), subject to the constraint $k_{DP} + k_{AS} + k_{GW} = 1$, leaving 17 functionally independent input parameters. Details of parameter descriptions, ranges and units are given in IBM.

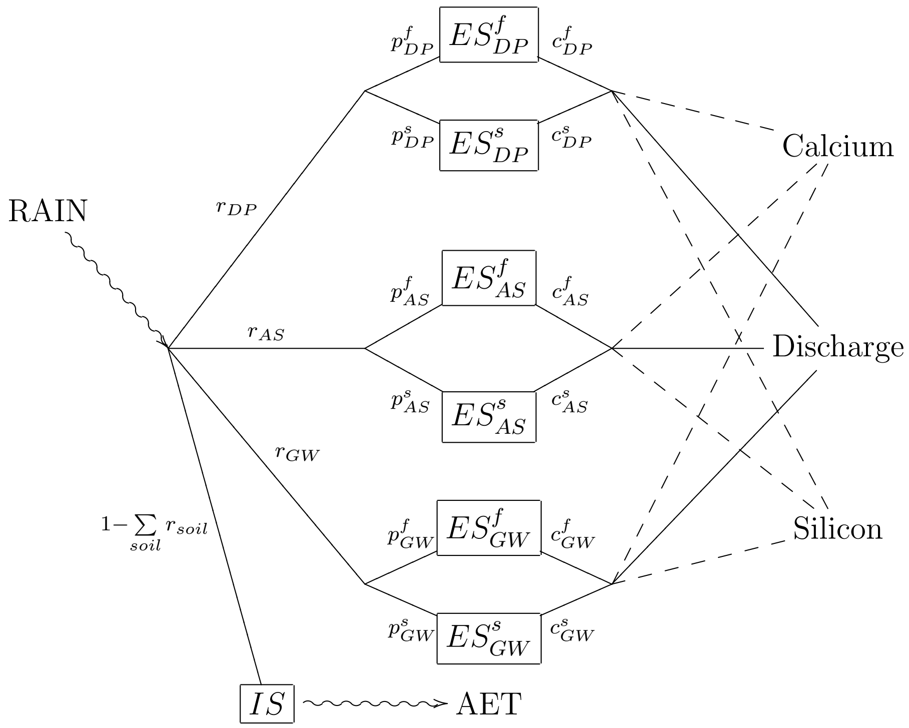

Figure 1: Three compartment rainfall runoff model.

Thus, in terms of the general description given in Section 2, the input vector x has 17 components, y represents the three time series for discharge, Calcium and Sodium, and z represents the corresponding field data. The function $f(\cdot)$ relating y to x develops as follows:

There is a fast f and a slow s sub-compartment for each of the three soil-type compartments DP, AS and GW. Updating the effective water stored from hour t to $t+1$ for each sub-compartment is governed by the equations

$$ES^f_{soil}(t+1) = ES^f_{soil}(t) + r_{soil}(t)p^f_{soil}\,\mathrm{RAIN}(t) - c^f_{soil}ES^f_{soil}(t)$$
$$ES^s_{soil}(t+1) = ES^s_{soil}(t) + r_{soil}(t)p^s_{soil}\,\mathrm{RAIN}(t) - c^s_{soil}ES^s_{soil}(t)$$

where *soil* is one of DP, AS and GW, $p^f_{soil} + p^s_{soil} = 1$

$$r_{soil}(t) = \frac{k_{soil}}{1 + \exp\left[a_{soil} - b_{soil}S(t)\right]}$$

with $k_{DP} + k_{AS} + k_{GW} = 1$ and $S(t)$, the total water stored in the system at time t is given by

$$S(t) = \sum_{soil}[ES^f_{soil}(t) + ES^s_{soil}(t)] + IS(t)$$

That is, the *total water storage* S in the system at any time is the sum of the *effective storages* for each soil type, both fast and slow, plus the overall residual *ineffective storage* IS. Physical interpretations of the six parameters for each compartment will emerge in the next subsection. Updating the total storage from t to $t+1$ is governed by the equation

$$S(t+1) = S(t) + \mathrm{RAIN}(t) - \mathrm{AET}(t) - \sum_{soil} F_{soil}(t)$$

where the $F_{soil}(t) = c^f_{soil} ES^f_{soil}(t) + c^s_{soil} ES^s_{soil}(t)$ are the *flows* out of each soil-type compartment. Similarly, updating the ineffective storage from t to $t+1$ is governed by the equation

$$IS(t+1) = IS(t) + \mathrm{RAIN}(t)[1 - \sum_{soil} r_{soil}(t)] - \mathrm{AET}(t)$$

Hourly model outputs, discharge $D(t)$, Calcium $Ca(t)$ and Silicon $Si(t)$ are given by

$$\begin{aligned} D(t) &= \sum_{soil} F_{soil}(t) \\ Ca(t) &= \sum_{soil} T^{Ca}_{soil} F_{soil}(t)/D(t) \\ Si(t) &= \sum_{soil} T^{Si}_{soil} F_{soil}(t)/D(t) \end{aligned}$$

where the T^{tracer}_{soil} govern the tracer concentrations of Ca and Si emanating from each soil-type compartment.

Thus, to run the model $y = f(x)$ we need (i) a computer code implementation of $f(\cdot)$; (ii) valid values for the 17 components of x; (iii) the forcing functions RAIN and AET; (iv) the initial conditions ES^f_{soil}, ES^s_{soil} and IS at $t = 0$; and (v) the values of the six tracer concentrations T^{tracer}_{soil}.

3.2 Informal model exploration

As an illustration of the many types of data analysis that we may carry out to explore the qualitative behaviour of the model, we focus on the water discharged at hour 620, and investigate its sensitivity to changes in a selection of some of the 17 model input parameters.

We illustrate the process by observing in Fig. 2 (left panel) how the logarithm of discharge at hour 620 varies over the range of c^f_{DP} for a selection of four values of p^f_{DP} and in Fig. 2 (right panel) how it varies with b_{AS} for four values of c^s_{AS}, where in both illustrations the other inputs were held fixed at their mid-range values.

As hour 620 is shortly after a large rainfall between 610 and 619 hours, peaking at hour 615, increasing c^f_{DP} from its minimum value of 0.1 initially increases discharge, as more water will flow out of the fast DP compartment; see Fig. 2. However, increasing c^f_{DP} past 0.2 leads to a decrease in discharge, because lots of the water will have drained away before 620, resulting in less flow. Increasing p^f_{DP} increases the amount of water entering the fast DP compartment (as opposed to the slow DP compartment), which leads to a corresponding increase in discharge.

Fig. 2 (right panel) shows that as b_{AS} is increased, the system approaches saturation and more water is directed into the fast and slow AS sub-compartments, with less going into the ineffective storage (IS) compartment. This leads to larger flows out of the AS sub-compartments, resulting in an increased discharge, which tends to an asymptotic value. Increasing c^s_{AS} increases the flow out of the slow AS compartment, which results in a small increase in discharge. Several additional plots were considered and they all demonstrated sensible model behaviour.

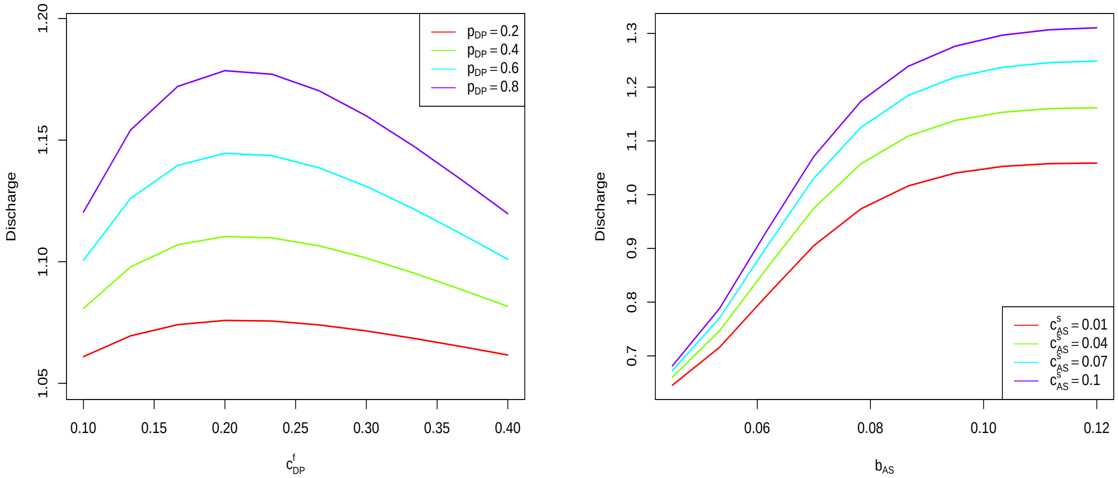

Figure 2: Left panel: logarithm of discharge at hour 620 versus c^f_{DP} for four different values of p^f_{DP}. Right panel: logarithm of discharge at hour 620 versus b_{AS} for four different values of c^s_{AS}.

3.3 Internal model discrepancy

We consider assessment of the internal model discrepancy contribution to overall model discrepancy for the runoff model. To do this, we perturb different features of the model and focus on how they perturb the discharge output $D(t)$. There are several distinct model features that we consider perturbing, including the six input parameters for each soil-type compartment, the initial flow conditions and the output tracer concentrations, the transfer functions $r_{soil}(t)$ and the two forcing functions RAIN and AET.

To illustrate our approach, we focus on perturbing (a) the initial conditions; (b) the forcing function RAIN; (c) the parameters a_{soil} in the transfer functions $r_{soil}(t)$ which influence the amount of water entering each compartment; and (iv) the input parameters c^f_{soil} and c^s_{soil} governing the flow rates out of the three compartments. Note that (c) is a simple example of perturbing the propagation step in the equations for the system state, while retaining the water conservation constraint.

We adopt a similar formulation for each of the four perturbations.

3.3.1 Initial condition contribution

First consider the condition specified by IBM that the initial flow out of the slow groundwater sub-compartment equals the observed initial discharge and the other initial flows are all zero. This implies that the initial storage of water in each of the 7 sub-compartments is zero except for the slow, ground water sub-compartment ($ES^s_{GW}(t=0)$), which is chosen to ensure that initial flow matches the observed flow. This is not an unreasonable specification, as there was an extensive dry period prior to the study. We will perturb the initial slow groundwater content $ES^s_{GW}(t=0)$, which we write as w. We do this by replacing w by ηw, where η is a positive random quantity with expectation $\mathrm{E}[\eta] = 1$ and standard deviation $\mathrm{SD}[\eta] = p$ corresponding to a small percentage, such as $100p = 5\%$. Thus, $\mathrm{E}[\eta w] = w$ and $\mathrm{SD}[\eta w] = pw$. We further assume, for convenience, that η has a log-normal distribution; that is, $\log \eta$ has a normal distribution with some mean μ and variance σ^2. It is reasonably straightforward to show that our expectation and standard deviation conditions on η imply that $\mu = -0.5 \log(1+p^2)$ and $\sigma^2 = \log(1+p^2)$. Thus, a convenient way to sample a value of η, is to sample $\log \eta$ from the normal distribution with this mean and variance, and then exponentiate the result.

Now suppose we (i) fix values for the 17 input parameters x; (ii) sample a value η_1 of η; and (iii) run the model with initial condition $\eta_1 w$ and inputs x. Let $D^1_1(x), \ldots, D^1_{839}(x)$ denote the resulting discharge output time series: actually, we take the logarithm of discharge to be the model output y. Now repeat the above with each of another $K-1$ independent η values, so that for η_k with initial condition $\eta_k w$ we have discharge model outputs $D^k_1(x), \ldots, D^k_{839}(x)$ for $k = 1, \ldots, K$. In our implementation, we set the components of x to be equal to the middle of the ranges specified by IBM, $p = 0.1$ and $K = 400$

Next, we calculate for each hour t the sample variance $V_t(x)$ of $D^1_t(x), \ldots, D^K_t(x)$. The 839×839 diagonal matrix V^{INIT}_x with diagonal elements $V_1(x), \ldots, V_{839}(x)$ is an estimate of the initial condition contribution to the overall internal model discrepancy variance. To simplify the discussion, we have chosen not to estimate the off-diagonal covariance terms, setting them to be zero instead. Fig. 3 plots the standard deviations $SD_t(x) = \sqrt{V_t(x)}$ against t. Notice that the effect of perturbing the initial condition eventually decreases to a constant value.

We repeated the above perturbation exercise for a few other fixed values of the inputs and discovered that the pattern and magnitude of the initial condition contribution was essentially the same, the biggest differences occurring at essentially infeasible input combinations.

3.3.2 RAIN contribution

We treat the forcing function RAIN similarly, except we perturb RAIN(t) for each hour $t = 1, \ldots, 839$ and also introduce a dependency between the perturbations as follows. Write $\xi(t) = \log \eta(t)$, where the perturbation is $\eta(t)\mathrm{RAIN}(t)$ and, as before, we assume $\mathrm{E}[\eta(t)] = 1$, $\mathrm{SD}[\eta(t)] = p$ and $\xi(t)$ has a normal distribution with mean $\mu = -0.5 \log(1+p^2)$ and variance $\sigma^2 = \log(1+p^2)$, the same values for each hour t. We now need to model the distribution of the collection $\eta(1), \ldots, \eta(839)$ or equivalently the collection $\xi(1), \ldots, \xi(839)$.

The simplest assumption would be to treat the ξ-collection as independent normal random quantities and proceed as for the initial condition perturbation. However, it makes sense to introduce a time dependency which we do here by assuming the ξ-collection to have a multivariate normal distribution with a correlation between $\xi(s)$ and $\xi(t)$ for any two hours s and t of the form

$$\exp\left[-\left(\frac{s-t}{\theta}\right)^2\right]$$

where the number of hours θ is to be chosen. Notice that for any given choice of θ, the correlation decreases as the time difference $|s-t|$ increases. On the other hand, the correlation decreases as θ decreases when the time difference is held fixed. In our implementation, we set $p = 0.1$ and $\theta = 5$ hours, reflecting the belief that the correlation in rainfall measurement error will not persist over the duration of an average storm. The 839 values of $\xi(t)$, hence those of $\eta(t)$, can be simulated, for example, using the function `mvrnorm` in the `R` library `MASS`; see, Venables and Ripley (2002). We now run the model at some input x, using the original initial condition and perturbed forcing function values $\eta(1)\mathrm{RAIN}(1), \ldots, \eta(839)\mathrm{RAIN}(839)$ and record the perturbed discharge series. We repeat this procedure K times and, exactly as we did with the perturbation of the initial condition above, estimate a diagonal variance matrix V_x^{RAIN}. Fig. 3 plots the standard deviations (the square roots of the diagonal elements of V_x^{RAIN}) against t when the components of x are chosen to be the mid-range values specified by IBM.

3.3.3 Structural inflow contribution

The amount of water flowing into each soil sub-compartment at each hour t is governed by its transfer function $r_{soil}(t)$ and p_{soil}. There are many possible perturbations: for illustrative purposes we chose to perturb the three a_{soil} parameters in a similar way as we did for RAIN. Specifically, we used the *same* perturbation process for η_t for a_{DP}, a_{AS} and a_{GW} with $p = 0.1$ and $\theta = 100$, reflecting slowly varying changes in the physical system. As previously noted, this is a simple example of perturbing the propagation step in the equations for the system state, while retaining the water conservation constraint. Fig. 3 shows the standard deviation of the logarithm of discharge for each hour for this internal error contribution.

3.3.4 Parameter outflow contribution

The flow out of each soil-compartment is governed by c_{soil}^f and c_{soil}^s. We perturb these six parameters as we did for the a_{soil} parameters in Section 3.3.3 using the *same* η_t process for each of them. Fig. 3 shows the standard deviation in the logarithm of discharge for each hour for this internal error contribution.

Overall, the patterns of the RAIN, structural and flow contributions to internal model discrepancy shown in Fig. 3, are similar with flow lagging a few hours behind the other two: they all increase significantly during periods of heavy rainfall.

Fig. 4 shows three traces: (i) the logarithm of observed discharge; (ii) three standard deviation intervals of observed error in the logarithm of discharge (which was chosen to be 5%); and (iii) three standard deviation intervals of internal model discrepancy, where the standard deviations are the root mean square of the variances of the four contributions: the initial condition, the RAIN forcing function, structural inflow and parameter outflow. The calculation in (iii) assumes that these four contributions are uncorrelated, which was confirmed with further runs of the model. Note that in Fig. 4, the internal model discrepancy is significantly smaller than the range of discharges explored by the model output and the observed discharge. This suggests that the model would not be deemed inadequate due to this level of internal discrepancy. Since there are many possible further internal error contributions, the internal model discrepancy based on the four contributions is likely to underestimate that based on a comprehensive overall assessment.

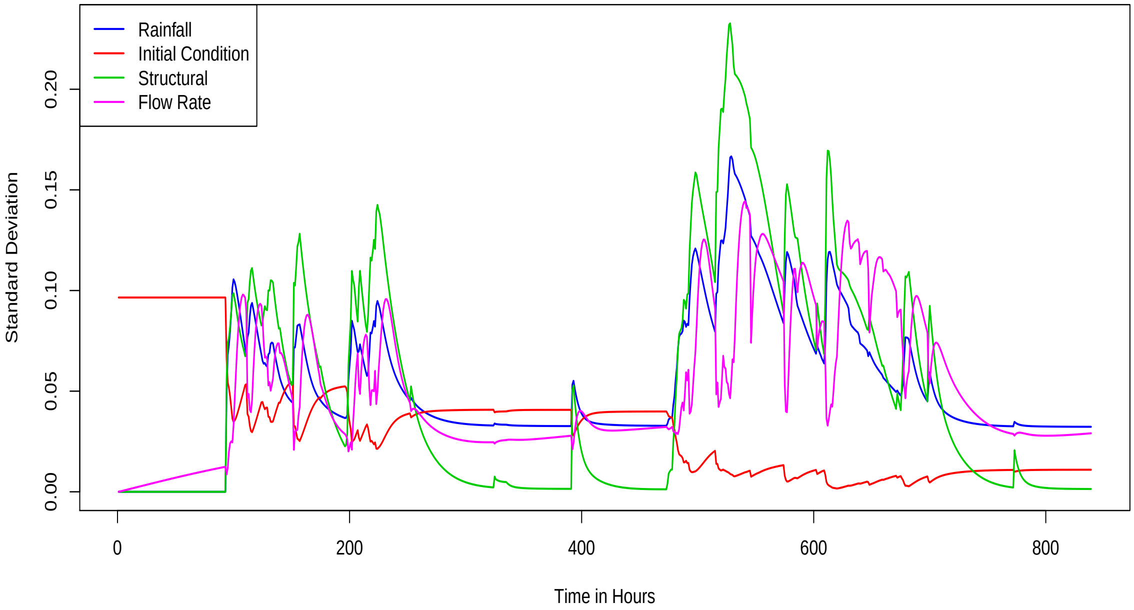

Figure 3: Standard deviations of the logarithm of discharge for four contributions to internal model discrepancy: initial flow condition (red), the RAIN forcing function (blue), structural inflow (green) and parameter outflow (magenta).

3.4 External model discrepancy

We introduce the notion of implausibility as a basis for assessing the external contribution to overall model discrepancy.

Suppose we observe a system at N equally-spaced time points $t = 1, 2, \ldots, N$. In the runoff model, there are $N = 839$ consecutive hourly discharge measurements for the 35 days between 19 August and 22 September 1993. Denote by z_t a field observation at time t. In the runoff model, $z_1, z_2, \ldots, z_N$ are the logarithms of water discharge at each of the 839 hours. Denote by $f_t(x)$ the model output at time t when the model is evaluated at input x. In the runoff model, $f_1(x), \ldots, f_N(x)$ are the 839 logarithms of water discharge simulated by the runoff model at input x, where x comprises 18 parameters subject to their range constraints and a sum-to-one restriction for three of them, leaving 17 inputs that can be varied independently.

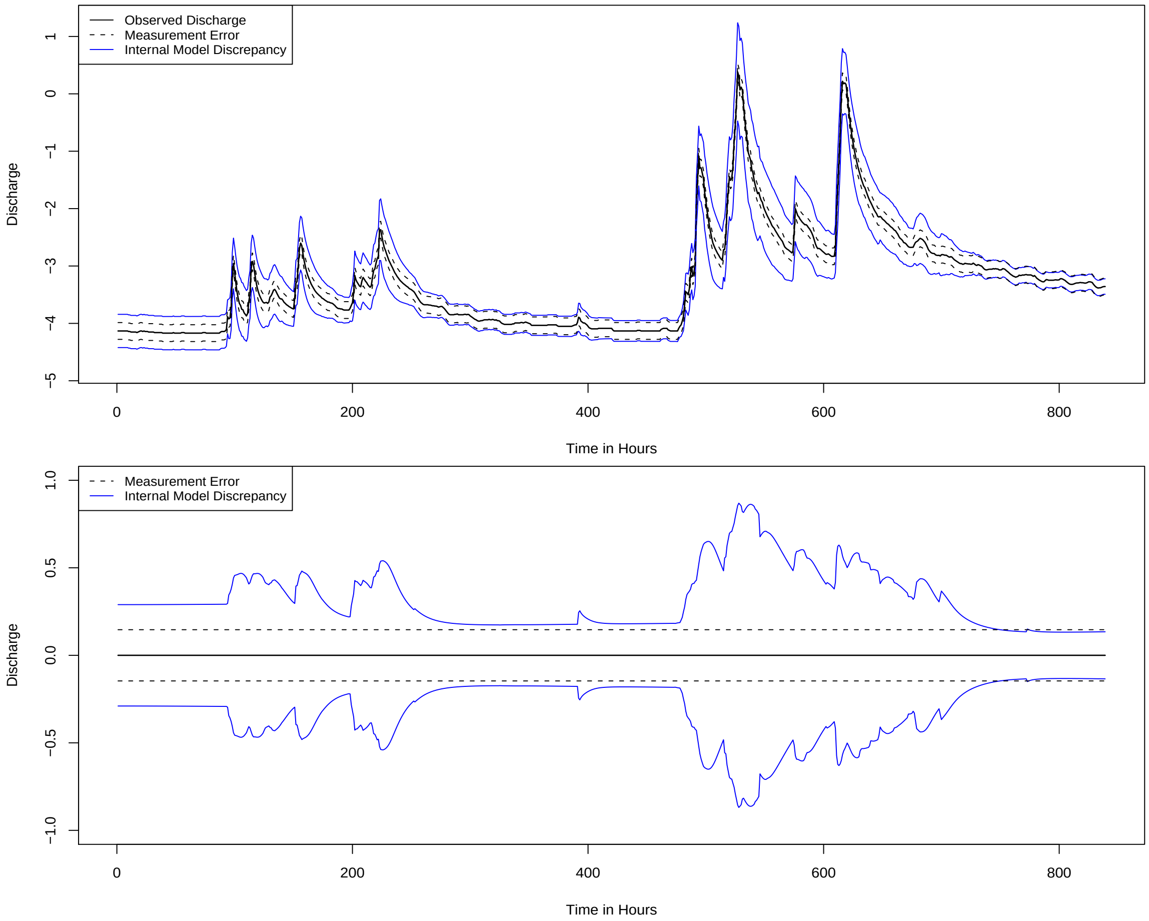

Figure 4: The *upper panel* shows the logarithm of observed discharge, three standard deviation traces of observation error and of corresponding internal model discrepancy, assessed by perturbing an initial condition, the RAIN forcing function, structural inflow and parameter outflow. The *lower panel* shows three-sigma limits for both the internal model discrepancy and the measurement error.

3.4.1 Implausibility

We have structured our uncertainty specification in general form, in which

$$\varepsilon_t = y_t - f_t(x^*) \tag{3}$$

is the model discrepancy at time t, where x^* is taken to be the appropriate model representation of the actual system properties, and which correspond to the actual unobserved system output y_t. We regard the values of x^* and the y_t as random quantities, as their values are unknown. Next, we write

$$z_t = y_t + e_t \tag{4}$$

where z_t is the measurement of y_t and e_t is the associated measurement error. Furthermore, we have decomposed the overall model discrepancy into the sum of internal and external components, which we write as

$$\varepsilon_t = \varepsilon_{I_t} + \varepsilon_{E_t} \tag{5}$$

Putting these relationships together, we obtain

$$z_t = f_t(x^*) + \varepsilon_{I_t} + \varepsilon_{E_t} + e_t \tag{6}$$

We regard the discrepancy and error terms $\varepsilon_{I_t}, \varepsilon_{E_t}$and e_t to be uncorrelated random (uncertain) quantities each with expectation zero and respective variances $\sigma^2_{I_t}$, $\sigma^2_{E_t}$ and σ^2_e. We will assume that the value of the measurement error variance σ^2_e is known, whereas $\sigma^2_{I_t}$ and $\sigma^2_{E_t}$ need to be carefully assessed, preferably in conjunction with a system expert, taking into account the limitations of the model in describing the actual system. We define the implausibility $I(x)$ of a model input x to be

$$I(x) = \max_{1 \le t \le N} \left| \frac{z_t - f_t(x)}{\sigma_t} \right| \tag{7}$$

where

$$\sigma^2_t = \text{Var}[(z_t - f_t(x^*))] = \sigma^2_{I_t} + \sigma^2_{E_t} + \sigma^2_e \tag{8}$$

Note that $I(x)$ is scale-free and the σ^2_t do not depend on x. Other definitions of implausibility are possible; for example, the average of the deviations in (7) or the average of their squares. The definition in (7) is more stringent than these two: imposing a constraint upon $I(x)$ would demand that the maximum deviation between model output and observed data was small.

Our aim is to "rule out" any input x for which $I(x)$ is "too large" when compared to a threshold based on a reasonable calibration for $I(x)$. One such calibration is based on assuming independent standard normal distributions for the signed standardised deviations in (7), deeming an input x to be implausible if say $I(x)$ exceeds the upper 5% point of its distribution in the null case when $x = x^*$. Then, the distribution of $I(x)$ is such that

$$p \;=\; \text{P}\left[I(x) \ge m \,|\, x = x^*\right] = 1 - \left[2\Phi(m) - 1\right]^N \tag{9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Hence, we want to choose m so that the probability p in (10) is "small"; that is, choose m such that

$$\Phi(m) = \frac{1 + (1-p)^{1/N}}{2} \tag{10}$$

When $p = 0.01$ and $N = 839$ we find that $m = 4.38$. At the other extreme, when the signed standardised deviations in (7) are completely dependent, corresponding to $N = 1$ in (10), we find that $m = 2.58$ when $p = 0.01$. The actual result will be somewhere between these two extremes. The corresponding values of m for $p = 0.05$ are 4.01 and 1.96. We adopt the conservative, stringent independence assumption with $p = 0.01$. Thus, we deem an input x implausible if $I(x) > 4.38$.

We applied this implausibility criterion to the logarithm of discharges from $100,000$ runs of the runoff model, where the inputs were from a subset of a Latin hypercube design chosen to accommodate the sum-to-one restriction. The σ^2_t in (7) were modified to be the sum of the measurement error variance and the internal model discrepancy variance contribution to the overall component-wise model discrepancy variance; that is, $\sigma^2_t = \sigma^2_{I_t} + \sigma^2_e$. The intention was to see if we could find some non-implausible inputs (without introducing any external model discrepancy) to help assess the external discrepancy variance contribution to the overall model discrepancy. However, we found that without the external discrepancy, every one of the $100,000$ inputs were implausible (given zero external model discrepancy): the lowest implausibility is about 4.7 with only two runs less than 5.0. In fact, we observed that for all $100,000$ runs the model consistently over-reacted to short periods of rain and reacted too quickly (or too slowly) to major peaks in rain, demonstrating that its predictive adequacy may be regarded as questionable for such rainfall patterns.

To obtain an order of magnitude assessment of the external model discrepancy $\sigma_{E_t}^2$, we might choose a small number n of the least implausible inputs $x_1, \ldots, x_n$ of the $100,000$ runs, and consider the corresponding model outputs $f_t(x_i)$. We then choose $\sigma_{E_t}^2$ in (8) so that

$$\max_{1\leq i\leq n} \left| \frac{z_t - f_t(x_i)}{\sigma_t} \right| \leq 3 \tag{11}$$

Note that this choice of $\sigma_{E_t}^2$ can be zero. Fig. 5 shows the results when we choose $n = 8$. The upper panel shows plots for the logarithm of observed discharge z_t, the mean $\bar{f}_t$ of the corresponding eight model outputs and the $3\sigma_t$ limits about that mean. The lower panel shows the residual plot $z_t - \bar{f}_t$, the same $3\sigma_t$ limits as in the upper panel and three-sigma limits for both the internal and the external discrepancy. Note how frequently the external discrepancy is zero.

A large external model discrepancy standard deviation σ_{E_t} indicates that the model fails to predict well for reasons not explained by measurement error or internal error. These occur here mainly when either the model reacts too quickly or too slowly during heavy rainfalls, for example around 490 hours, or when the model overreacts to smaller rainfall events, such as at 395 and 690 hours. We might expect such deficiencies in a simple compartment model of a complex physical runoff system.

The forms of the external and internal model discrepancy traces are very different, as they are measuring different things. The green trace, which shows the sum of the total model discrepancy and the measurement error, is of primary interest in assessing model adequacy. Even though it is large in places, it is much smaller than the overall range of both the observed and model discharge, suggesting that that the model is mostly adequate for describing discharge, except during periods of heavy rainfall, where the green trace spikes, which is particularly evident in the residual plot, displayed in the lower panel.

4 Slow computer models

The above analysis was based on an ability to modify the computer code and to carry out very many evaluations of the model. We now describe how to modify our analysis when neither of these conditions applies. For purposes of comparison, we re-analyse the runoff model of Section 3, but we now suppose that we have no access to the computer code and that the runoff model has a long run time. Therefore, we used only 250 carefully chosen training runs with which to build an emulator of the computer code implementation of the model. As discussed in Section 2, an emulator is a fast stochastic approximation of the model. We can evaluate the expectation

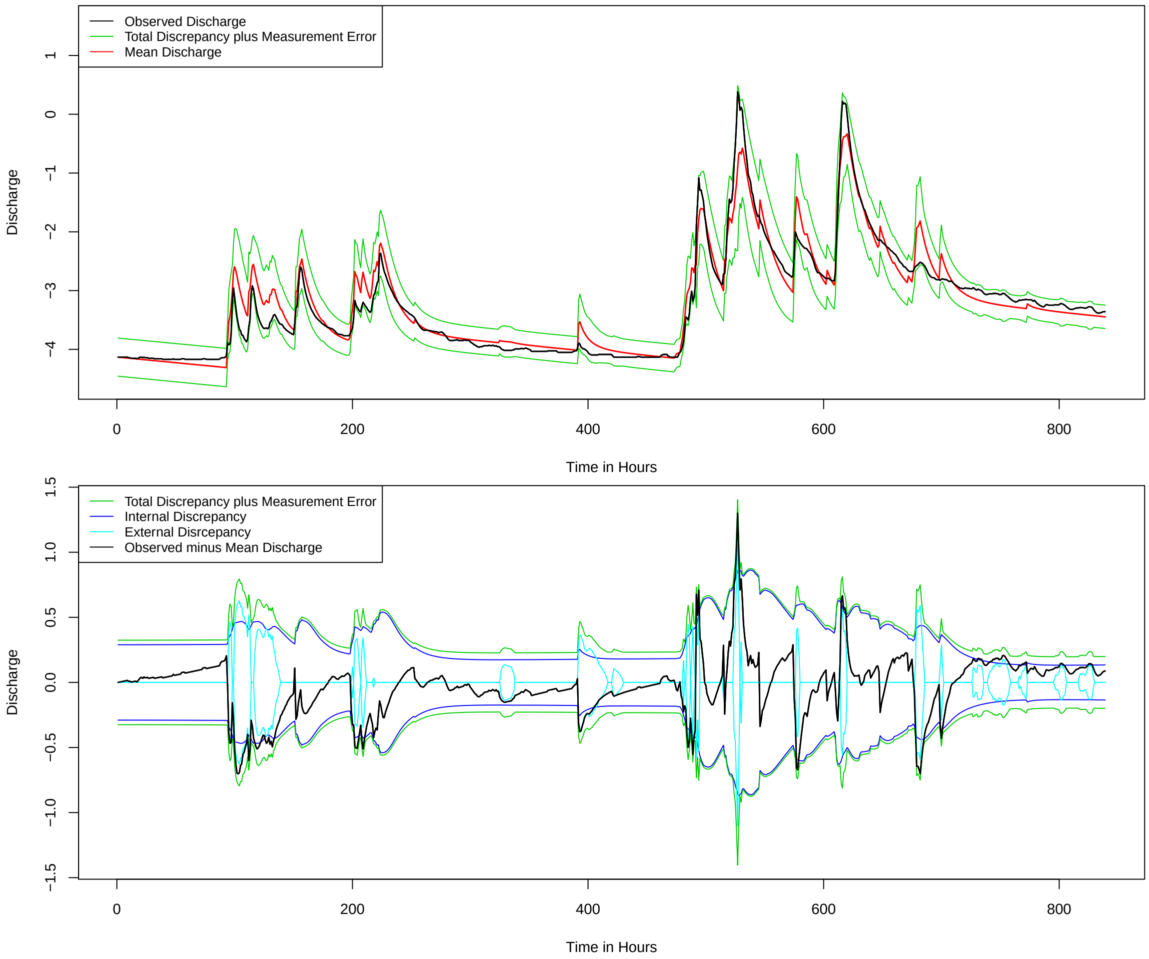

Figure 5: The *upper panel* shows the logarithm of observed discharge, the mean of the corresponding eight 'best' model outputs and overall three-sigma limits ($3\sigma_t$) about that mean; while the *lower panel* shows the residual about about that mean, the same three-sigma limits as in the upper panel and three-sigma limits for both the internal and the external discrepancy. See the text for the definition of each sigma.

and variance of the emulator: the former mimics the model's behaviour while the later represents our uncertainty in the approximation; see, for example, Craig et al. (1997), Craig et al. (2001), Kennedy and O'Hagan (2001), O'Hagan (2006) and MUCM (2009).

To illustrate emulation, we consider the logarithm of the discharge at each of the 13 equally-spaced hours $100, 160, \ldots, 760, 820$. The following procedure is used to construct an emulator for the logarithm of discharge at each of these 13 hours.

(i) We select a Latin hypercube of 250 points over the 17 functionally independent inputs and run the model at each of them. To construct a Latin hypercube of n points, the range of each of the inputs is divided into n equal intervals, and the points are then chosen randomly so that no two points occupy the same interval for any of the inputs.; see, for example, MUCM (2009).

(ii) Next we fit a linear model of the form (12) to the logarithm of the 250 model discharges, using the `lm` function in `R`; see the R Development Core Team (2008). Each model input choice x has 17 components $x^{(1)}, \ldots, x^{(17)}$ and, in the first instance, we choose $g_j(x) = x^{(j)}$ for $j = 1, \ldots, 17$ and $g_0(x) = 1$.

(iii) We then use the `step` function in `R` to carry out a backward step-wise selection procedure to identify a subset of *active inputs* x_a of the inputs x that account for a high percentage of the total variation in the logarithm of model discharge in relation to the fitted model. A further reduction of the subset can be achieved by removing statistically significant inputs which otherwise have little practical impact on model output. For simplicity, we kept the same number of active inputs for each output, and found that 12 inputs were sufficient, although a different 12 were chosen for each of the 13 outputs.

(iv) We then fit a quadratic in the active inputs determined in (iii); that is, with the $g(\cdot)$ in (12) of the form $g_{ij}(x_a) = x_a^{(i)} x_a^{(j)}$ for $0 \leq i \leq j$, where $g_{00}(x_a) = 1$ and $g_{0j}(x_a) = x_a^{(j)}$

(v) If the multiple correlation R^2 for the fitted quadratic model is substantial, in excess of 90% say, then it should be a useful predictor of model output at untried inputs. However, as an emulator of the model, the quadratic regression fit will not agree with the model outputs at the 250 inputs. As explained after (12), current emulator research treats the residuals as a "smooth" random process instead of the "rough" residuals from the quadratic regression fit, acknowledging that the model is likely to be a continuous, differentiable function of the inputs x. Thus, the emulator for a single output $f(x)$ of the runoff evaluated at x, has the form

$$f(x) = \sum_{0 \leq i \leq j \leq 12} x_a^{(i)} x_a^{(j)} \beta_{ij} + u(x) \tag{12}$$

(vi) The actual emulator for the computer model at any input x is obtained by assessing (a) σ^2 to equal the residual mean square from the least squares fit to (12); (b) the β_{ij} to equal to their least squares estimates; and (c) the variances and covariances of the β_{ij} to equal their estimated values resulting from the least squares fit.

Furthermore, we usually decompose $u(x)$ to be of the form $u(x) = \epsilon(x_a) + \nu(x)$, where $\nu(x)$, called a"nugget" residual, accounts for the absence of variation due to the inactive inputs: two *different* inputs x and $x^{'}$ may have the *same* values for their active input components. We assume that $\nu(x)$ has zero expectation and variance $\delta\sigma^2$ for all x, and $\nu(x)$ and $\nu(x^{'})$ are uncorrelated, unless $x = x^{'}$ when they are perfectly correlated: we take $\delta = 0.05$. The other residual component $\epsilon(x_a)$ has zero expectation and variance $(1 - \delta)\sigma^2$ for all x_a, and the correlation between any two residuals $\epsilon(x_a)$ and $\epsilon(x_a^{'})$ is taken to be of the form

$$\exp\left[-\sum_k \left(\frac{x_a^{(k)} - x_a^{'(k)}}{\theta_k} \right)^2 \right]$$

for any two inputs x and x' with active input components $x_a^{(k)}$ and $x_a^{'(k)}$, where θ_k, which is either chosen or estimated, controls the contribution to the overall correlation between the corresponding two outputs in the direction of the kth active input component $x_a^{(k)}$. We chose each $\theta_k = 0.33$, one-third of the length an input interval, a choice based on previous experience of fitting quadratics to computer model output

(vii) We check emulator accuracy by evaluating it at the inputs of an additional set of *evaluation* or *diagnostic* model runs to see whether the emulator evaluations at these inputs are "close" to the corresponding model outputs, where for each evaluation, closeness is assessed with respect to the standard deviation of the emulator at the evaluation input. We would normally choose a small number of diagnostic runs (about 100) with inputs in a Latin hypercube, modified to accommodate the sum-to-one restriction. However, for demonstration purposes we use the 13 emulators to obtain emulator expectation and variances at the same 100,000 points used in Section 3 to obtain a more detailed assessment of the emulators in comparison with the model outputs. Fig. 6 illustrates results (using just $1,000$ randomly selected points from the $100,000$) for the emulators at 460 hours (left panel) and 820 hours (right panel). Each panel shows the emulated logarithm of discharge with three emulator standard deviation limits *versus* the corresponding 100,000 model values, the 45 degrees line and the field observation value with three measurement error standard deviation limits. Clearly the emulator at 460 hours is more accurate than that for 820 hours. It can also be seen that both emulators are satisfactory in that a large number of prediction intervals (red) do indeed cover the correct model discharge values represented by the green line. Notice that for both hours there are model runs which match the field data within the measurement error limits, suggesting good fits. However, while we find that this is also true for the other 11 hours, we cannot be sure there is a common set of inputs at which the model runs for all 13 hours fit well, or indeed for all 839 hours, a point we address in the next section.

4.1 Implausibility

The definition of implausibility for slow computer models is similar to that for fast models given in (7); see, for example, Craig et al. (2001). We define the implausibility $I(x)$ of an input x to be

$$I(x) = \max_{1 \le t \le N} \left| \frac{z_t - \mathrm{E}[f_t(x)]}{\sigma_t(x)} \right| \tag{13}$$

where $\mathrm{E}[f_t(x)]$ denotes the mean of the emulator at time t for input x and $\sigma_t^2(x)$ is the sum of three variances, those of measurement error, model discrepancy and the emulator at x: in our example $N = 13$. Cut-off considerations for $I(x)$ are similar to those for fast models. Note that, as internal model error assessment is not possible for slow computer models, external model error implicitly includes the internal contribution.

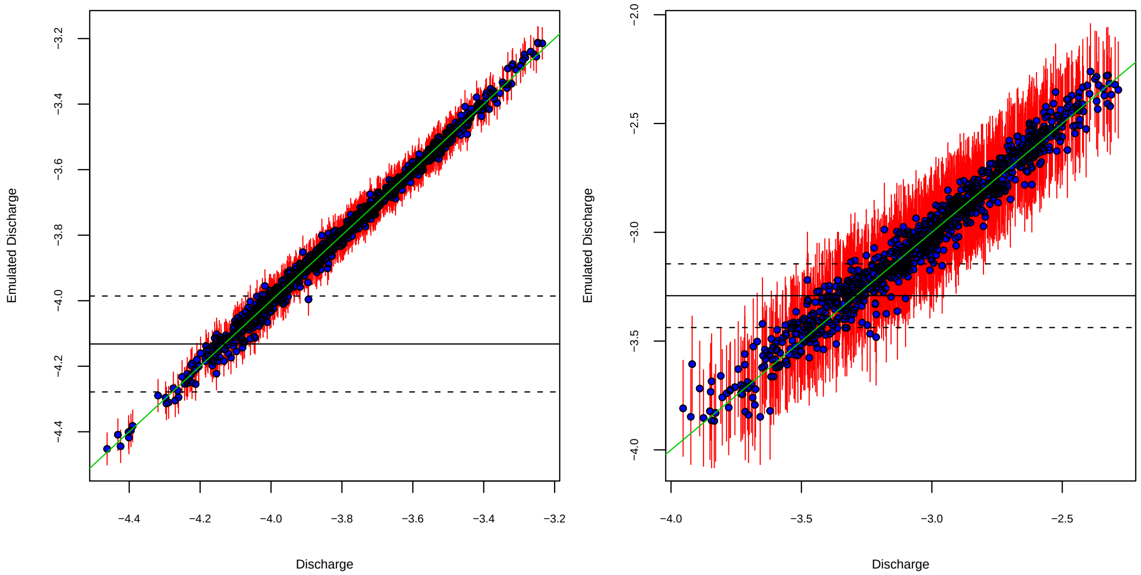

Figure 6: Emulated logarithm of discharge (blue dots) with three emulator standard deviation limits (red line segments) *versus* the corresponding randomly chosen 1000 runoff model values from $100,000$ runs; the 45 degrees line (green); and field observation value (black line) with three measurement error standard deviation limits (black dotted lines) for the emulators at 460 hours (left panel) and 820 hours.

Since an emulator run time will be fast compared to that for the computer model it emulates, we can evaluate it at many inputs (as we did for fast models) to help determine implausible inputs. As for fast models, we set the model discrepancy variance component term of $\sigma_t^2(x)$ in (13) equal to *zero* to help identify some "non-implausible" inputs with which to help assess model discrepancy error, which in turn can be used to assess whether the model is worth calibrating and adequate for prediction.

In the runoff example, we selected 203 *candidate* runs from the emulated values at the $100,000$ point design used in Section 3.4.1, using an implausibility cut-off value of 6.5. We then evaluated these 203 runs on the actual runoff model and computed the implausibility in (7) with σ_t^2 equal to measurement error variance, and found that their implausibility values were all greater than 8.5. We then chose the eight best of these 203 runs having implausibility less than 10 . These eight runs were used to assess the external model discrepancy, exactly as in Section 3.4.1. Fig. 7 shows the results.

The upper panel in Fig. 7 shows the 250 runs used to build the 13 emulators, the 203 candidate runs and the observed discharges: the error bars are based on the sum of the external model discrepancy and measurement error.

The lower panel in Fig. 7, summarised in Table 4.1, compares the fast and slow model results and shows standard deviations for fast internal model discrepancy, fast external model discrepancy, fast total model discrepancy and slow external model discrepancy, the later equaling the slow total model discrepancy, as there is no internal model discrepancy. Observe that the fast total model discrepancy and slow external model discrepancy are of a similar order of magnitude, with the fast total model discrepancy being mostly larger due to the fast internal discrepancy contribution, which could not be assessed in the slow model situation. Other deviations are due to the best run selection process being slightly different in the fast and slow cases: in particular, the presence of the internal discrepancy alters the definition of an acceptable run.

As a further check on the quality of the emulators, we found that the 203 candidate runs suggested by the emulator did in fact include the best 8 runs that would have been found had we evaluated all $100,000$ runs used in Section 3.4.1 using the runoff model directly. Thus, using the 13 emulators we have only had to evaluate the runoff model $453 = 250 + 203$ times to achieve the same results as running the runoff model $100,000$ times! As the model discrepancy is similar in magnitude to that for the fast simulator, our conclusions regarding model adequacy are consistent with those given in Section 3.4.1.

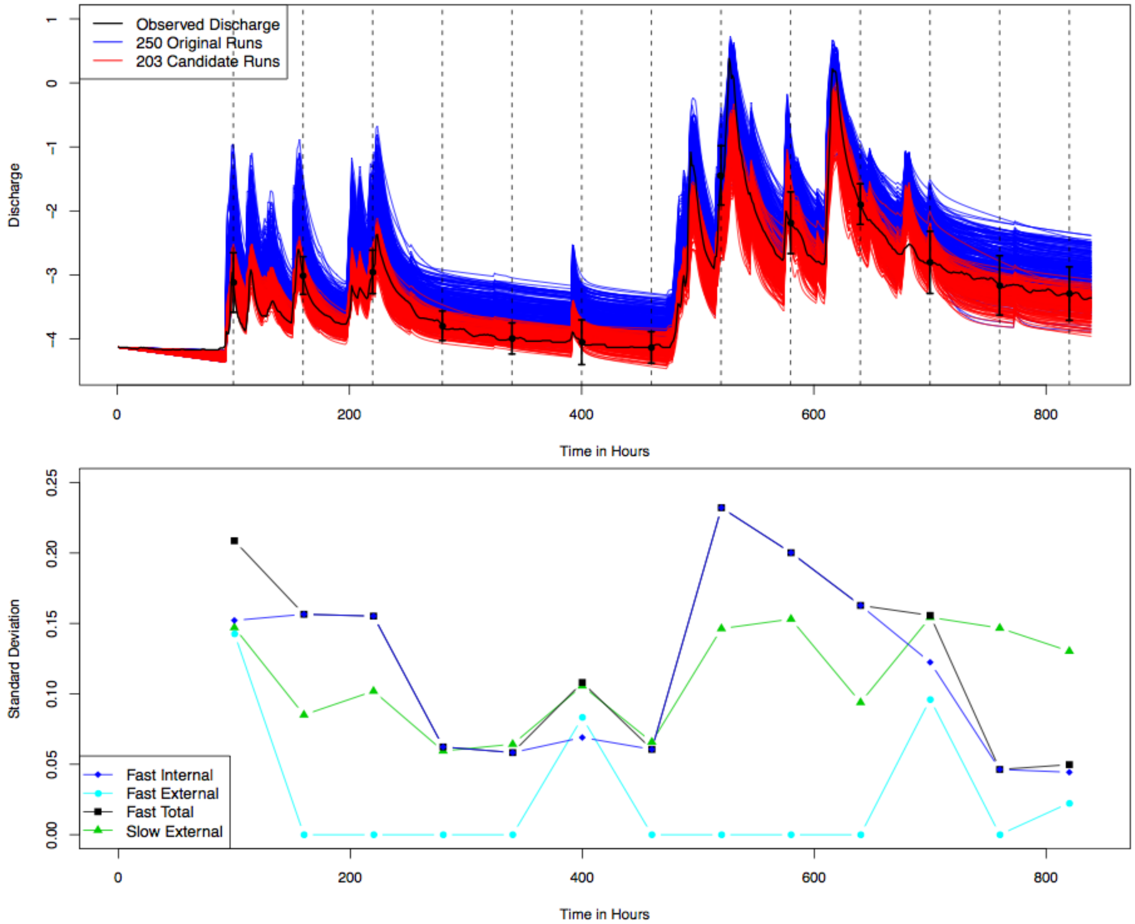

Figure 7: The upper panel shows the 250 runs used to build the 13 emulators (blue), the 203 candidate runs (red), the observed discharges (black) and the error bars are based on the sum of the external model discrepancy and measurement error. The vertical dotted lines are the 13 emulator hours. The lower panel compares the fast and slow model results, showing standard deviations for fast internal model discrepancy (blue) , fast external model discrepancy (turquoise), fast total model discrepancy (black) and slow external model discrepancy (green).

	100	160	220	280	340	400	460	520	580	640	700	760	820
Fast Internal	.15	.16	.16	.06	.06	.07	.06	.23	.20	.16	.12	.05	.04
Fast External	.14	.00	.00	.00	.00	.08	.00	.00	.00	.00	.10	.00	.02
Fast Total	.21	.16	.16	.06	.06	.11	.06	.23	.20	.16	.16	.05	.05
Slow External	.15	.09	.10	.06	.06	.11	.07	.15	.15	.09	.15	.15	.13

Table 1: Standard deviations of internal, external and total model discrepancy at 13 different hours for the fast model and of external model discrepancy for the slow model, as depicted in Fig. 7.

Acknowledgements

We are extremely grateful to Keith Beven for many helpful discussions on the runoff model and its application to the Haute Mentue catchment and to Leanna House for explaining the model and providing her `R` code.

This chapter was produced with the support of the Basic Technology Initiative as part of the "Managing Uncertainty for Complex Models" project, and with the support of an EPSRC Mobility Fellowship (IV).

References

Craig, P. S., Goldstein, M., Rougier, J. C., and Seheult, A. H. (2001), "Bayesian forecasting for complex systems using computer simulators," *Journal of the American Statistical Association*, 96, 717–729.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997), "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments," in *Case Studies in Bayesian Statistics*, eds. Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D., New York: Springer-Verlag, vol. 3, pp. 36–93.

Iorgulescu, I., Beven, K. J., and Musy, A. (2005), "Data-based modelling of runoff and chemical tracer concentrations in the Haute-Mentue research catchment (Switzerland)," *Hydrological Processes*, 19, 2557–2573.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian calibration of computer models," *Journal of the Royal Statistical Society, Series B*, 63, 425–464.

MUCM (2009), *MUCM Toolkit Release 3*, Aston, England.

O'Hagan, A. (2006), "Bayesian analysis of computer code outputs: A tutorial," *Reliability Engineering and System Safety*, 91, 1290–1300.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rougier, J. (2009), "Formal Bayes methods for model calibration with uncertainty," in *Applied Uncertainty Analysis for Flood Risk Management*, eds. Beven, K. and Hall, J., Imperial College Press / World Scientific.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer, 4th ed., iSBN 0-387-95457-0.