Uncertainty quantification and the calibration of numerical models

Peter Challenor
University of Exeter

Numerical models have reached the stage where our simulations are believed to be fairly accurate representations of the real world, and recently the term 'digital twin' has been coined to describe such simulators. However it should be remembered that all simulations are models of the real world not the real world itself. The underlying equations of our simulators are the result of good scientific understanding, which may itself be partial. In addition we solve numerical approximations to these equations ,not the equations in a continuum, and parameterise many processes because of discretisation or lack of knowledge. The difference between the simulator and reality is often known as the model *discrepancy*. In addition there are usually unknown parameters (or other inputs) in the simulators which we need to estimate either from external (expert) knowledge or by fitting the simulators to data using some statistical methodology. We will refer to this problem as *calibration* (or *inverse modelling*). Thus our simulator output is always uncertain, in a number of distinct ways and any form of calibration not only needs to estimate the values of the simulator inputs but also the associated uncertainty. Although the quality and quantity of measurement continues to improve, data are also always uncertain. So the calibration problem involves estimating parameters in uncertain models with uncertain data. The simple way of solving such a problem is maximum likelihood (or least squares) or Bayesian calibration. Unfortunately such methods are flawed as they do not take the discrepancy into account. The nearest point to the data on the model manifold is found, even though this may be a long way from the true solution. Even worse the uncertainty on the estimator reduces as the amount of data increases, going to zero as the number of data points goes to infinity, giving a completely false impression of the true accuracy. It is possible to create a better methodology that includes the model discrepancy, for example see Kennedy and O'Hagan (2001), who model the real world as the sum of the simulator and the discrepancy both of which are modelled as Gaussian processes; one representing the simulator, and one representing the discrepancy. The Kennedy and O'Hagan formulation has proved very popular, but suffers from a huge drawback - the two Gaussian processes are not separately identifiable. This isn't a problem for prediction, where we are only interested in the sum of the two processes, but if we want to gain understanding about the simulator and discrepancy we need to be able to distinguish them. A number of solutions have been proposed, including using strong prior information and restricting the form that the discrepancy can take. We suggest a different approach known as history matching. In history matching rather than trying to find a point estimate for the simulator inputs (or equivalently their joint posterior distribution) we find those sets of inputs that give simulator outputs so far from the data that we can rule them out as implausible. Once we have ruled out all the implausible input values what is left must include the 'best' value, if such a value exists. As we will see, it is possible to rule out all possible input values in which case it is not possible to make the simulator and the data agree.

History matching is based on an implausibility statistic which we can calculate for all values of the simulator. Since most simulators are too computationally expensive to allow us to do this the first part of the process is to build a Gaussian process emulator. We run the simulator in a carefully designed experiment to fill the input space. It is important to fill space as much as possible so that we can predict what the simulator would produce across the whole of space. The Gaussian process emulator is a stochastic surrogate model that allows us to predict the output of the simulator at any point with its expectation but also gives the uncertainty at any point arising from the stochastic interpolation. With the emulator we can now predict the simulator for any set of input values and so we can also calculate the implausibly. Note at this point in the analysis we should find out which of the input values actually change the simulator outputs and only include those that are

important in the analysis. Once we have built the Gaussian process emulator on the reduced input space we can calculate the implausibility. This is defined by

$$Imp(x) = \sqrt{\frac{(y_{data} - E[y_{emul}(x)])^2}{\sigma^2_{emul}(x) + \sigma^2_{data} + \sigma^2_{disc}}}$$

where y_{data} is the data value, $y_{emul}(x)$ is the emulator for y evaluated at the inputs x, $E[.]$ is the expectation operator, $\sigma^2_{emul}(x)$ is the variance of the emulator at the inputs x, σ^2_{data} is the variance of the data and σ^2_{disc} is the simulator discrepancy expressed as a variance.

The implausibility is simply a scaled distance between the estimated simulator value from the emulator and the data. Let us consider each of the terms of the denominator in turn. The first term is simple, it is the uncertainty in our emulator. This will vary with the values of the inputs but is easily calculated from our emulator. Large values of this term imply a poor emulator in those regions of space and the implausibility will be small, thus we cannot rule them out, even if the numerator is large. The other two terms in the denominator do not vary with the inputs but are important. The uncertainty on the data (σ^2_{data}) is clearly important. If we have poor data we will be less confident in rejecting simulator inputs than if our data is of very high quality. We will discuss some issues with biological data and how these this term below. The third term is the most controversial. This gives the additional uncertainty that comes from the discrepancy between the simulator and the real world. We usually think of discrepancy in terms of a bias whereas here we are describing it as a variance. It may be better to think of this as a mean square error rather than a variance so in effect it is the bias squared. History matching doesn't estimate the discrepancy as the Kennedy and O'Hagan method does (and thus avoids the identifiably problem), we need to elicit it from experts.

Given the emulator and values for $\sigma^2_{data}$and σ^2_{disc}we can now calculate the implausibility across the domain of the inputs. We deem implausible any set of inputs for which the implausibility is greater than 3. This leaves a region known as NROY (Not Ruled Out Yet). Within this NROY space we now do another space filling design and rebuild the emulator. Because we have reduced the space over which we are building the emulator and removed extreme differences between the model and data often the corners of the input domain this new emulator has reduced uncertainty compared to the original one. We now recalculate the implausibility and reduce the current NROY to a new one. This process is carried out recursively; each step is referred to as a *wave*.

The process of history matching is terminated when a stopping criterion is reached. The stopping criteria are (1) that a new wave does not reduce NROY in a meaningful way; (2) that NROY has been reduced to a region so small that we do care to reduce it further; or (3) the NROY space goes to zero. If we consider each of these in turn. The first implies that we cannot improve the emulator any further, the dominant terms in the implausibility are now $\sigma^2_{data}$and $\sigma^2_{disc}$and further runs of the simulator would not improve the calibration. At this point we need to either acce[pt that this is as good a calibration as we are going to get or collect more data to reduce σ^2_{data}. The second option occurs when we preset a limit on how good a calibration needs to be for the task or decision in hand. Once this limit is reached we stop the calibration. The third option is in many ways the most interesting. If the NROY space goes to zero it means there are no values of inputs that can make the simulator agree with the data within the specified tolerance. The simulator does not fit. This often happens when the discrepancy variance has been set to zero, a perfect nodal assumption. One way around this problem is to increase the discrepancy term until a non-empty NROY is produced. This discrepancy value is the amount of discrepancy in the simulator you are prepared to accept to produce a

simulator that 'fits' the data. For this reason the discrepancy term is sometimes described as the *tolerance to error*.

It should be noted that history matching does not give us a point estimate for inputs to the simulator. We get a region of not implausible values. There is no guarantee that this region is simply connected and we have no information of whether one value is less or more plausible than another. The method is purely geometric and makes no assumptions about probability distributions or likelihoods. If a more Bayesian interpretation is required then it is possible to cast history matching as a version of Approximate Bayesian Computation (ABC), but this isn't necessary and standard history matching is a purely geometric process.

Having considered what might be described as classical history matching let us consider the special characteristics of biological data that might make us vary the methodology. We will initially concentrate on the data error term. History matching has been developed for applications mainly in oil reservoir modelling and climate. In both cases there is no concept of a population of oil reservoirs or climates. There is only one oil reservoir in a particular geographical location and we only have one planet with a climate, even in engineering applications all aero-engines are designed and built to be as identical as possible. With biological systems this is not the case. For example consider a cardiac model. We might want to calibrate this model so that it represents the whole adult population of Britain. Alternatively we might want a simulator that is representative of the female population. Or we may wish to concentrate our attention on a single individual and personalise the model. The entire population of Britain will have more variability than the female population which in turn will have more variability than a single individual. But unlike a jet engine even the same measurements on the same patient will exhibit some variability in time. Therefore the variance σ^2_{data} represents not only the measurement error but also the variability within the population we are considering. This variance can be decomposed into within groups variances and between group variances. For example we might decompose the variance into

$$\sigma^2_{data} = \sigma^2_{between\ gender} + \sigma^2_{within\ gender} + \sigma^2_{within\ individual} + \sigma^2_{measurement}$$

we could then choose how far up the hierarchy of variability we wish to. If we are trying to fit our simulator to an individual (personalised medicine) we would only need to consider the last two terms - the measurement error $\sigma^2_{measurement}$ and variability within the single patient $\sigma^2_{within\ individual}$. This will be smaller than the variance for whole population and hence the the NROY for individuals will be smaller than the corresponding NROY for populations. In fact it would be possible for some individuals to have empty NROY spaces while the population NROY is non-empty, i.e. it is not possible for the simulator to fit those individuals while it is possible to fit the population.

One criticism of history matching is that it does not give a point estimate, or posterior distribution, of the inputs. It is possible to do history matching followed by a Kennedy and O'Hagan analysis on the reduced NROY space. Such an analysis would be much more efficient than doing the Kennedy and O'Hagan calibration on the whole space, identifiability problems would be reduced by limiting the domain to the relatively well behaved region defined by the NROY space. I would argue that, at least for simulators of biological and medical processes, a range of input values may be more useful than a single point estimate. We have discussed above the variability in the data collected from a single patient, not simply the measurement error that would also be present in engineering, but also natural variation between measurements, possibly due to unmeasured covariates. We know for environmental models that inputs calibrated for precipitation data and not the same as those calibrated for temperature and in biological systems we expect this effect to be even greater. Rather than expect there to be a single best simulator input, sometimes referred to as x^* in the

literature, we should think more of clouds of acceptable sets of inputs all of which produce reasonable fits to the data. These clouds should not be confused with a posterior uncertainty distribution around a single best value. The members of an NROY produced by history matching do not have a probabilistic interpretation. Although we can show that one point has a smaller or larger implausibility value this should not be interpreted as one set of inputs being more likely to fit the data than another.

In this paper I have presented an alternative to traditional statistical calibration where we try to find the 'best' set of inputs. This is known as 'history matching' and consists of rejecting those simulator inputs that are implausible given the data. In addition to explaining how history matching works we have described how the method may need to be extended to work better with biological systems.

Kennedy, M., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, **63**(3), 425–464.