New and Notable

New and Notable: Uncertainty Quantification

James B. Bassingthwaighte[1,*]
[1]Department of Bioengineering, University of Washington, Seattle, Washington

Johnston et al. (1), in this issue of the *Biophysical Journal*, show us that simple calculations are not so simple when there is uncertainty in the underlying input data. They illustrate this using an on-line CALADIS calculator where the uncertainty in a variable or parameter is represented using a probability density function (pdf). CALADIS calculations are done using Monte Carlo, drawing 20,000 (or other) successive samples from the pdfs for the components of the equations, and using them to add, subtract, multiply, or divide. (CALADIS provides a variety of pdfs, so one can avoid those that, like Gaussian random, spread into negativity.) The answers are provided graphically as a pdf of the 20,000 results, plus some statistics (mean, standard deviation (SD), quartiles, etc.). These results are based on assuming that each pdf describes an independent, identically distributed (i.i.d.) set of numbers; they demonstrate that the spread of the answers tends to be greater than that of the input data.

For example, summing numbers drawn from two Gaussian pdfs, 2.0 ± 0.4 and 2.0 ± 0.4 on 20,000 trials, gives 4.0 ± 0.565 (mean $\pm$ SD)—a coefficient of variation, *CV*, of $0.565/4 = 0.141$. Subtracting numbers from these same two pdfs gives 0.0 ± 0.566. Multiplication gives 4.0 ± 1.14, doubling the SD compared to addition, $CV = 0.253$. The estimated means are very close to the point calculations, i.e., operations on the mean values alone. As expected, results on addition, subtraction, and

multiplication with Gaussian pdfs agree with analytic predictions.

However, uncertainty in the denominator results in bias and skew. When numbers drawn from dividing a Gaussian normal (0.4) (2) with numbers drawn from the same pdf, 2.0 ± 0.4, the results from several trials gave means of 1.04–1.05 and SDs of 0.324–0.329 ($CV = 0.31$) and notable skewness. (It would be useful if CALADIS provided the estimates of skewness, although one can download the result pdf and do such calculations outside of CALADIS.) While the result of adding or subtracting Gaussian pdfs is a symmetric distribution, multiplying and dividing always produces right-skewed distributions with larger CV values.

CALADIS helps us to understand that Gaussian processes are not the norm. Most measures of populations (people's heights, concentration, mass, reaction rates, channel opening intervals, volumes) are not really Gaussian: they cannot have negative values and therefore cannot have symmetric tails. Useful pdfs for nonnegative distributions where the SD can exceed the mean are right-skewed (e.g., Poisson, γ-variate, log normal).

Modeling in biology is usually an inverse problem: from observing inputs and outputs to a system, one attempts to characterize the nature of the system and its transfer function. It is not adequate to provide its numerical descriptor, as from a deconvolution: one needs to define a mechanism. Consequently one uses a forward technique, using the observed inputs to drive the model, then adjusting the mechanistic parameters to fit the observed output data. Modeling is to seek out the mechanisms, relating cause and effect. Quantifying uncertainty of predictions is key to a model's utility.

We can generalize from CALADIS: it is a model whose outputs depend on the inputs, their uncertainty, and the chosen operations. The Monte Carlo sampling is one method of determining effects of the uncertainty inherent in all

modeling efforts. The same approach can be used in more complicated models, ones with spatial and temporal dependencies. Models of biological systems need to account for uncertainties in defining characteristics and predicting future behavior. Fig. 1 suggests three ways to incorporate uncertainty:

1. input functions and initial and boundary conditions (*left* in figure),
2. parameters values (*bottom*), and
3. model configuration, or its internal stochastic nature, or in the numerical solutions (inside the *box*).
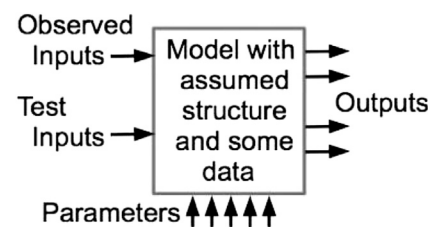

FIGURE 1 The modeling processes should account for variation and noise in inputs, parameter values, and insecurity in model structure and calculations.

Uncertainties influence the responses to perturbations, transients, and the steady-state profiles, e.g., in biochemical concentrations, distributions of flow, and channel currents.

In pharmacokinetic pharmacodynamic studies for FDA approval, uncertainty quantification is becoming an expected last element of the template for model-centered research, Verification, Validation, and Uncertainty Quantification (VVUQ).

Verification is testing to determine that the models are coded and solved correctly. Validation is testing against real-life observations—fitting data or predicting outcomes showing that the model is not obviously wrong. Models are never proven correct, but if the model is not invalid, it has value as a working hypothesis. For uncertainty quantification, of the three types of

uncertainty diagrammed, parameter uncertainty is the easiest to handle:

1. use Monte Carlo;
2. run 1000 solutions of the model with the parameter values set by random selection with the a priori pdf of values for every parameter simultaneously;
3. observe and evaluate the model outputs; and
4. search the outputs for correlations among parameters, especially those with such high correlation that the model should be simplified to improve identifiability (3).

Input uncertainty is harder to define: current pulses to drive a neuron may have little variation, but dietary input or other time-dependent input uncertainty is more difficult, requiring more personal choices of how to characterize the variability.

Model structural uncertainty is at the scientific heart of the matter. Comparisons among variously configured models, in the style of Platt's (4) strong inference tends to work well by encouraging design strategies that produce data distinguishing between a pair of hypotheses, so that a well-executed experiment eliminates at least one hypothesis. The Akaike information criterion (5) and alternatives are limited to measures of the goodness of fit of model to data, and do not evaluate validity, i.e., adherence to reality. Its virtue, echoing Occam's razor or Albert Einstein's admonition, "Make the model as simple as possible, but not too simple," is to remind us that overparameterization may give a better fit but masks the identification of key components. Its vice is that it requires parameters be independent, a near-impossibility in model systems.

Uncertainty quantification is central to predicting a hurricane trajectory, planning financing, assessing environmental impacts, handling epidemics, and making accurate prognoses. Continuity, in the form of a priori correlation, momentum, accumulations, periodicity, or feedback regulation,

is the basis for prediction. Fractal processes (Nile floods (6), sunspots (7), long memory processes (8), and regional myocardial blood flow (9) demonstrate that time series and spatial profiles are often not i.i.d. processes, but exhibit scale-independent autocorrelation, and accordingly allow prediction from prior or local behavior. (These are called "long memory processes", a bit of a euphemism inasmuch as they are best used for short-range prediction: near-neighbors tend to be alike, or, in other words, tomorrow's weather is most likely to be like today's). Long memory processes provide a statistical description of long-term likelihood, but are almost useless to predict infrequent events like earthquakes.

A final caveat on the CALADIS tool is that its calculations rely on independence, such that if $C = A + B$, and the process is i.i.d., then the means and the variances sum. This is no longer true if parameters are correlated. For example, if parameters x_A and x_B are variable but their sum is constrained so that the corresponding ith elements $x_{Ai} + x_{Bi} = 1.0 \pm 0.2$, they are necessarily correlated negatively. Then the sum of their variances is narrower than the Gaussian expectation and depends on the correlation

$$\mathrm{Var}_C = \mathrm{Var}_A + \mathrm{Var}_B \\ + 2\,\rho\,\sqrt{(\mathrm{Var}_A \cdot \mathrm{Var}_B)},$$

where ρ, the correlation coefficient for ordered elements in A and B, in this example is negative. Then one cannot sample randomly from pdfs but must draw simultaneously from ordered pdfs providing the correct degree of correlation. One can create ordered sets with correlated parameter values though a different Monte Carlo approach: add noise to observed data sets (e.g., a few percent proportional Gaussian), optimize to find the best-fitting parameter set, and repeat 1000 times. Regression analysis shows the correlations among parameters. The multiparameter ordered arrays can then be sampled, linearly adjusted to exemplify the desired conditions, and

used to create the 1000 new solutions around the model best-fit solution; the correlation structure is not changed by linear scaling, and the uncertainty quantitation is provided through the variance in the solutions. The remaining problem is that the result is relevant only for the local region in state space, like parameter sensitivity functions at the point of best fit in state space.

Smith's book (10) provides insight into the mathematics of new developments in this accelerating field. There are many strategies. Ferson and Hajagos (11) demonstrate a probability box, one that defines lower and upper exceedance probabilities, which are the complementary cumulative distribution functions bounding the expected results. The probability box region, $0 < p < 1$ and between the lower and upper exceedance complementary cumulative distribution functions, confines the expected result of a computation. The approach allows interdependence among parameters, but does not define exact probabilities for a parameter.

CONCLUSIONS

Uncertainty quantification is an underdeveloped science, emerging from real-life problems. Johnston et al. (1) illustrate how important it is to account for uncertainty in making estimates from simple arithmetic operations, and thereby provoke us to consider their ideas in the larger context of the biological sciences that commonly deviate from i.i.d. processes. Modeling analysis needs a concerted effort in this direction. In the nether regions beyond i.i.d. processes: here be dragons!

REFERENCES

1. Johnston, I. G., B. C. Rickett, and N. S. Jones. 2014. Explicit tracking of uncertainty increases the power of quantitative rule-of-thumb reasoning in cell biology. *Biophys. J.* 107:2612–2617.

2. Bassingthwaighte, J. B., L. S. Liebovitch, and B. J. West. 1994. Fractal Physiology. Oxford University Press, New York.

3. Carson, E., and C. Cobelli. 2014. Modeling Methodology for Physiology and Medicine, 2nd Ed. Elsevier, London, UK.

4. Platt, J. R. 1964. Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science.* 146:347–353.

5. Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19:716–723.

6. Hurst, H. E. 1951. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* 116:770–808.

7. Watari, S. 1995. Fractal dimensions in solar activity. *Sol. Phys.* 158:365–377.

8. Beran, J. 1994. Statistics for Long-Memory Processes. Chapman & Hall, New York.

9. Bassingthwaighte, J. B., R. B. King, and S. A. Roger. 1989. Fractal nature of regional myocardial blood flow heterogeneity. *Circ. Res.* 65:578–590.

10. Smith, R. 2014. Uncertainty Quantification: Theory, Implementation, and Applications. SIAM Press, New York.

11. Ferson, S., and G. J. Hajagos. 2004. Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliab. Eng. Syst. Saf.* 85:135–152.