From *SIAM News*, Volume 45, Number 7, September 2012

Uncertainty Quantification 2012

Statistical Approaches to Combining Models and Observations

By L. Mark Berliner

Continual improvements in both computational assets and observational data are revolutionizing science and engineering. However, models, computations, and observations are subject to a variety of sources of uncertainty, mandating the need for quantification and management of uncertainty. Bayesian hierarchical modeling is a framework for combining diverse datasets, mechanistic and statistical models, and computation in a fashion that manages uncertainty (see, for example, [1,5]).

Hierarchical probability models are sequences of conditional distributions that correspond to a joint distribution. Let X, Y, and Z be three random quantities (scalars, vectors, or space–time fields), and let $p(x,y,z)$ denote their joint probability density. This density admits the factorizations

$$p(x,y,z) = p(x \mid y,z)\, p(y,z) = p(x \mid y,z)\, p(y \mid z)\, p(z),$$

where $p(x \mid y,z)$ is the density of X given $Y = y$ and $Z = z$. This is elementary mathematics but suggests a powerful applied modeling strategy, in which we form models in three primary steps: (1) *Data Model*: a probability distribution of observations Y conditional on the processes or state variables X of interest and on model parameters θ_Y; (2) *Process Model*: a prior distribution for X conditional on parameters θ_X; and (3) *Parameter Model*: a prior distribution for θ_Y and θ_X. Bayes' theorem provides the posterior distribution of X and the parameters conditional on the observed data $Y = y$. The posterior distribution is the Bayesian answer. From it, we derive probabilities of hypotheses and events of interest, estimates, confidence intervals, predictions and associated intervals, etc.

The data model $p(y \mid x, \theta_Y)$ is typically a "measurement error model." For example, we might consider a model based on $Y = x + \varepsilon$, where ε is a random, unobservable error. The parameter θ_Y might include unknown measurement error variances, measurement biases, and so forth. The power of the strategy is the ability to treat diverse datasets. Suppose, for example, that $Y = (Y_w, Y_\psi)$, where Y_w are wind measurements and Y_ψ are pressure measurements over some region. $X = (W, \psi)$ are true winds and pressures. We expect that (Y_w, Y_ψ) would display a complicated, difficult-to-model dependence structure. If the lion's share of that structure arises from the underlying relationship between W and ψ, however, we may be able to defend the data model

$$p(y_w, y_\psi \mid w, \psi, \theta_Y) = p(y_w \mid w, \theta_Y)\, p(y_\psi) \mid \psi, \theta_Y);$$

that is, Y_w, Y_ψ are conditionally independent. Notice that $p(y_w \mid w, \theta_Y)$ does not include ψ in the conditioning; this does not mean that Y_w and ψ are independent, but rather that they are conditionally independent given $W = w$.

The process model offers the opportunity to incorporate scientific modeling of the quantities of interest. Often, we formulate models from underlying differential equations or discretized versions of them [4,10]. For our wind–pressure example, the geostrophic approximation suggests that winds are proportional to the gradient of the pressure field. We can incorporate this notion in a stochastic geostrophic approximation,

$$p(w, \psi \mid \theta_X) = p_g(w \mid \psi, \theta_X)\, p(\psi \mid \theta_X),$$

where p_g is based on the actual geostrophic relation [8]. This example indicates how we can incorporate mechanistic models among the quantities in X. In some examples we model the process of interest conditional on boundary and/or initial conditions and then model those conditions. Finally, with the parameter model we can incorporate further information (calibration studies, for instance, lead to priors for θ_Y) in a fashion that allows for uncertainty. For example, physical theory may suggest the values, or at least interpretations, of some quantities in θ_X. This information is used to construct the prior, but allows for uncertainty. Moreover, that uncertainty responds to the data through the posterior distribution.

Analysis of Bayesian hierarchical modeling is often compute-intensive. Such advances as Markov chain Monte Carlo, sequential Bayes, and particle filtering have made serious BHM applications possible (e.g., [7]). However, use of process models requiring runs of large-scale, supercomputer models for single iterations of a Monte Carlo Bayesian calculation are typically feasible. This suggests the need for approaches that can incorporate ensembles from large models. Let $\boldsymbol{O} = (O_1, \ldots, O_n)$ denote an ensemble of size n. (We can account for ensembles from different models and/or generated from various model parameterizations, but I do not do so here.) The following potential strategies are organized around the BHM skeleton presented earlier.

First, consider modeling $\boldsymbol{O}$ *as if* the data were observational [2]. That is, we form a data model $p(Y, O \mid x, \theta_Y, \theta_O)$. The parameter θ_O includes variation from the ensembling, model-to-model differences, and model biases (or offsets), thereby allowing us to learn about these features based on Y. This framework also lends itself to the design of hybrid experiments involving both observational data and computer models. (See [6].)

Next, we can use model output to formulate a process model prior in a variety of ways. Much of the literature in the design and analysis of computer experiments (e.g., [9]) begins with a Gaussian process model for model output: $p(o \mid \theta)$ for some collection of parameters θ. In many cases, θ are unknown parameters in a covariance function characterizing the dependence structure of output as a function of model inputs. This

model is then updated to produce $p(o, \theta \mid O)$. Related ideas are known as "model emulators." In any case, transferring such results to form priors on true processes (X) remains a challenge. In yet another possibility, data analysis on model output can lead to parameter prior models (e.g., [3]). Finally, various combinations of these modeling strategies are feasible.

References

[1] L.M. Berliner, *Physical–statistical modeling in geophysics*, J. Geophy. Res., 108:D24 (2003), 1–10, doi: 10.1029/2002JD002865.

[2] L.M. Berliner and Y. Kim, *Bayesian design and analysis for superensemble based climate forecasting*, J. Climate, 21 (2008), 1891–1910.

[3] L.M. Berliner, R.A. Levine, and D.J. Shea, *Bayesian climate change assessment*, J. Climate, 13 (2000), 3805–3820.

[4] L.M. Berliner, R.F. Milliff, and C.K. Wikle, *Bayesian hierarchical modeling of air–sea interaction*, J. Geophy. Res., 108:C4 (2003), 1–18, doi:10.1029/ 2002JC001413.

[5] N. Cressie and C.K. Wikle, *Statistics for Spatio-Temporal Data*, John Wiley & Sons, Hoboken, New Jersey, 2011.

[6] D. Higdon, M.C. Kennedy, J. Cavendish, J. Cafeo, and R.D. Ryne, *Combining data and computer simulations for calibration and prediction*, SIAM J. Sci. Comput., 26:2 (2004), 448–466.

[7] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer, New York, 2004.

[8] J.A. Royle, L.M. Berliner, C.K. Wikle, and R.F. Milliff, "A hierarchical spatial model for constructing surface winds from scatterometer data in the Labrador Sea," in *Case Studies in Bayesian Statistics IV*, C. Gatsonis, R.E. Kass, A. Cariquiry, and B. Carlin, eds., Springer, New York, 1999.

[9] T.J. Santner, B.J. Williams, and W.I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.

[10] C.K. Wikle, R.F. Milliff, D. Nychka, and L.M. Berliner, *Spatiotemporal hierarchical Bayesian blending of tropical ocean surface wind data*, J. Amer. Statist. Assoc., 96 (2001), 382–397.

L. Mark Berliner is a professor and chair of the Department of Statistics at Ohio State University.