22 Physical-Statistical Models for Predictions in Ungauged Basins

EDWARD P. CAMPBELL

INTRODUCTION

Uncertainty is ever present in hydrological modelling, and hydrologists have always been keen to make use of new tools for better modelling. Bayesian methods have traditionally been seen as providing a framework to incorporate expert knowledge explicitly in modelling uncertainty, and so these methods have found widespread use in the hydrological literature (Vicens *et al.*, 1975; Valdés *et al.*, 1979; Qian & Richardson, 1997; Campbell *et al*., 1999). Recent advances in the statistical and geophysics literature are demonstrating that there is a new way of modelling that is neither statistical nor physical, but a fusion of both. This approach has been termed physical-statistical modelling by Berliner (2003).

Many of the applications to date have been in climate and other environmental fields. Berliner *et al*. (2000) developed a model for the evolution of Pacific Ocean sea surface temperatures with a view to forecasting the El Niño-Southern Oscillation phenomenon. The physical model in their application is in fact a physically-inspired statistical model. A much more clearly physical model was developed by Berliner *et al*. (2003), driven by a partial differential equation describing quasi-geostrophic flow for ocean streamfunction.

A feature of the work of Berliner *et al*. (2003) is the ease with which complex boundary conditions are handled, using an approach developed by Wikle *et al.* (2003). This follows from the hierarchical nature of the modelling, where the uncertainty of collections of variables is broken down into a product of simpler conditional models using Bayes' theorem (Bernardo & Smith, 1994, p.2). Thus interior locations to be modelled can be expressed conditional on the boundary values in a straightforward way. A very readable introduction to hierarchical methods applied to environmental modelling is provided by Wikle (2003).

This chapter suggests a candidate framework for building physical-statistical models to apply to prediction problems in ungauged basins. By integrating multiple data sources, physical models and expert knowledge, we seek to provide optimal predictions. The next section provides a broad framework. Then we develop some theoretical details for the broad framework as an illustrative example. We do not consider this definitive, but seek to demonstrate the approach and inspire others to apply this sort of thinking to the problem and so develop better solutions. Some thoughts on model-fitting are also included. The fourth section provides outlines of some physical-statistical models in the literature, focusing particularly on the physical component. A discussion and some conclusions complete the chapter.

A SUGGESTED MODELLING FRAMEWORK

We conceive of a framework for prediction in ungauged basins that links gauged and ungauged streamflow processes with available data and parameters, both physical and statistical in nature. The core idea is to develop probability models for all uncertain quantities using a Bayesian hierarchical approach (Berliner, 2003). We illustrate our thinking by assuming a physical process model linking streamflow to physical watershed characteristics that are measurable. This modelling implies the existence of physical parameters, and measurement implies the existence of statistical parameters, if only to represent measurement error.

Let D represent the available data, P the physical streamflow processes, Θ the statistical parameters and η the physical parameters. We use the contemporary notation $[D,P,\Theta,\eta]$ to denote uncertainty of these quantities, then applying Bayes' theorem repeatedly we find:

$$\begin{aligned}[D, P, \Theta, \eta] &= [D \mid P, \Theta, \eta]\,[P, \Theta, \eta] \\ &= [D \mid P, \Theta, \eta]\,[P \mid \Theta, \eta]\,[\Theta, \eta] \end{aligned} \quad (1)$$

We see that the joint probability model may be written as the product of three simpler models, and this is the essence of hierarchical thinking. Whilst building a model for $[D,P,\Theta,\eta]$ directly may seem a daunting prospect, the conditional models derived via Bayes' theorem are much less so. Joint complexity derived from relatively simple conditional models is an appealing feature of the hierarchical approach.

The first component on the right-hand side (rhs) of equation (1) is the so-called *data model*, and is essentially a likelihood function. The second component is a probability model for the physical streamflow process, and is termed the *prior physical process*. The final component is the *prior parameters model*; typically the statistical and physical parameters are considered *a priori* independent, and this model is factored into a product of individual probability distributions. In essence, the physical model can be used to constrain the data model to be physically reasonable.

A modelling perspective

In the derivation of equation (1) we simply applied Bayes' theorem; we now apply a modelling perspective. Conditional on the process P and the statistical parameters Θ it seems reasonable to assume that the physical parameters η are of no further importance in modelling the data, so we reduce the data model to $[D \mid P,\Theta]$. Typically there will be a number of sources of data, and we explore an example application below to see how this is handled. Following the same logic, the statistical parameters are of no importance in the prior process model, which reduces to $[P \mid \eta]$. The hierarchical model therefore becomes:

$$[D, P, \Theta, \eta] = [D \mid P, \Theta]\,[P \mid \eta]\,[\Theta, \eta] \quad (2)$$

AN ILLUSTRATIVE EXAMPLE

We now explore these ideas more deeply by developing a candidate framework for prediction in ungauged basins. The model developed here is only one possibility, and is designed primarily to illustrate the hierarchical approach. Many other choices are possible. We assume that P is comprised of a gauged and an ungauged streamflow

process, denoted R and $\overline{R}$ respectively. Data are available on gauged streamflow (D_R) and watershed characteristics for gauged and ungauged streamflow (D_W and $\overline{D}_W$). We assume a statistical model is used linking gauged streamflow and watershed characteristics, having parameters Θ_W. Gauged and ungauged streamflow is calibrated to observed data, with parameters Θ_R and $\overline{\Theta}_R$ we seek to find the posterior distribution for $\overline{\Theta}_R$. A physical model linking gauged and ungauged streamflow to watershed characteristics has parameters η_W and $\overline{\eta}_W$, respectively. It is important to note that we assume that this model is derived from physical considerations, and is not based on statistical modelling. These symbols are summarized in Table 10.1.

Table 10.1 *Nomenclature for the hierarchical model.*

PROCESSES	R	Gauged streamflow process
	$\overline{R}$	Ungauged streamflow process
DATA	D_R	Gauged streamflow data
	D_W	Gauged streamflow watershed data
	$\overline{D}_W$	Ungauged streamflow watershed data
PARAMETERS	Θ_W	Statistical parameters linking gauged streamflow to watershed characteristics, assumed common to gauged and ungauged basins.
	Θ_R	Statistical parameters describing streamflow in gauged basin (rainfall–runoff model perhaps)
	$\overline{\Theta}_R$	Statistical parameters describing streamflow in ungauged basin, not observable.
	η_W	Physical parameters linking gauged streamflow and watershed characteristics
	$\overline{\eta}_W$	Physical parameters linking ungauged streamflow and watershed characteristics

The data model

From equation (2) the data model becomes $[D_R, D_W, \overline{D}_W \mid R, \overline{R}, \Theta_R, \overline{\Theta}_R, \Theta_W]$. It seems reasonable to assume conditional independence of the gauged and ungauged basins, so we may factor this as:

$$[D_R, D_W, \overline{D}_W \mid R, \overline{R}, \Theta_R, \overline{\Theta}_R, \Theta_W] = [D_R, D_W \mid R, \Theta_R, \Theta_W][\overline{D}_W \mid \overline{R}, \overline{\Theta}_R, \Theta_W] \quad (3)$$

where we retain Θ_W in each conditional distribution as it is considered to be common across gauged and ungauged basins. We may apply Bayes' theorem once again to the first component, $[D_R, D_W \mid R, \Theta_R, \Theta_W] = [D_W \mid R, D_R, \Theta_R, \Theta_W][D_R \mid R, \Theta_R, \Theta_W]$.

The second component on the rhs is a model for the gauged streamflow data, so it seems reasonable to neglect the parameter Θ_W. In typical regionalization applications the streamflow data would also be neglected in favour of the calibration parameters Θ_R, but we will retain the general expression for the first component. The data model therefore becomes:

$$\begin{aligned}[D_R, D_W, \overline{D}_W \mid R, \overline{R}, \Theta_R, \overline{\Theta}_R, \Theta_W] = \\ [D_W \mid R, D_R, \Theta_R, \Theta_W][D_R \mid R, \Theta_R][\overline{D}_W \mid \overline{R}, \overline{\Theta}_R, \Theta_W]\end{aligned} \quad (4)$$

The first component is essentially a regional model describing watershed characteristics in terms of streamflow. The second component may be thought of as a calibration model for gauged streamflow. The final component can be thought of as a regional prediction model since it includes the unobservable parameter $\overline{\Theta}_R$. Because no streamflow data are available, the prior process model makes a key contribution here. We will return to this point in a later section, providing an example that may be helpful. Note that a feature of the data model is the sharing of parameters between the respective model components.

In conventional applications of regionalization (e.g. Dyer *et al.*, 1994) we would typically calibrate a rainfall–runoff model to the gauged watersheds, the resulting parameters would then be linked to watershed characteristics via a statistical model to quantify a regional relationship. This model would then be used to determine rainfall–runoff parameters at the ungauged sites. Thus the conventional process takes place in stages, and a complete integration of uncertainty can be problematic. The data model defined by equation (4) incorporates all these elements in one step. Thus a form of feedback is possible during model-fitting, so each model component can borrow strength from the other components. We would expect to be able to achieve better predictions at ungauged sites as a result, even without the physical model.

The prior process and parameters models

The prior process model becomes: $[R, \overline{R} \mid \eta_W, \overline{\eta}_W]$, with the prior parameters model being: $[\Theta_W, \Theta_R, \overline{\Theta}_R, \eta_R, \overline{\eta}_R]$

It is beyond the scope of this chapter to explore these models in detail, although we will return to some issues in building process models in the next section and during the discussion. It does seem reasonable in the PUB context to assume conditional independence of the streamflow process components: $[R, \overline{R} \mid \eta_W, \overline{\eta}_W] = [R \mid \eta_W][\overline{R} \mid \overline{\eta}_W]$.

Model-fitting and making predictions

To make a prediction of ungauged streamflow $\overline{R}$ we need the posterior distribution of $\overline{\Theta}_R$. In theory this is available as follows by applying Bayes' theorem: $[P, \Theta, \eta \mid D] \propto [D \mid P, \Theta][P \mid \eta][\Theta, \eta$, and we have already developed a model framework for the rhs of this expression. The posterior for $\overline{\Theta}_R$ follows by integrating over the unwanted quantities to give: $[\overline{\Theta}_R \mid D] = c^{-1} \int_{P,\Theta/\overline{\Theta}_R,\eta} [D \mid P, \Theta][P \mid \eta][\Theta \mid \eta] dP d(\Theta / \overline{\Theta}_R) d\eta$, where $c = \int_{P,\Theta,\eta} [D \mid P, \Theta][P \mid \eta][\Theta \mid \eta] dP d\Theta d\eta$ ensures that the resulting density function integrates to 1.

Using the posterior $[\overline{\Theta}_R \mid D]$ it is possible to generate probabilistic predictions for the ungauged streamflow process, integrating all sources of information available. However, it is rarely possible to calculate the integrals involved analytically, so simulation approaches are typically employed. It is not the purpose of this chapter to delve deeply into the practicalities of the suggested approach, but we note here some strands in the literature on algorithms to calculate posterior distributions of interest.

The conventional simulation approach is known as Markov chain Monte Carlo (MCMC; Smith & Roberts, 1993). MCMC algorithms are potentially inefficient if the physical model is nonlinear with respect to the physical parameters, which is commonly the case. Berliner *et al.* (2003) suggested an approach using importance sampling (Bernardo & Smith, 1994, p.350–352) to avoid this problem. The basic algorithm generates an approximate sample from the posterior distribution as follows:

(a) Generate a sample from the prior distribution $[\eta,\Theta]$.
(b) Using the sampled $\{\eta_i, \Theta_i\}$ generate an ensemble $\{\eta_i, \Theta_I, P_i\}$ from the physical model.
(c) Resample the ensemble into the posterior distribution with acceptance probability: $q_i = [D \mid P_i, \Theta_i, \eta_i] / \sum [D \mid P_j, \Theta_j, \eta_j]$.

Given a posterior sample it is possible to generate rainfall–runoff ensembles, for example, so forming a predictive distribution for streamflow.

PHYSICAL-STATISTICAL MODELS IN THE LITERATURE

We have mentioned a number of applications in the literature. We look here in a little more detail at a couple of examples that have a substantive physical model motivation.

A model for air–sea interaction

Berliner *et al*. (2003) developed a model for air–sea interaction, driven by a partial differential equation describing quasi-geostrophic flow for upper ocean streamfunction ψ, incorporating wind stress τ:

$$\left(\nabla^2 - \frac{1}{r^2}\right)\frac{\partial \psi}{\partial t} = -J(\psi, \nabla^2\psi) - \beta\frac{\partial \psi}{\partial x} + \frac{1}{\rho H} curl_z \tau - \gamma\nabla^2\psi + a_h\nabla^4\psi \tag{5}$$

Here ρ is density, H is depth and effects due to bottom friction ($-\gamma\nabla^2\psi$) and lateral dissipation ($a_h\nabla^4\psi$) are incorporated; J denotes a Jacobian operator and x denotes the zonal direction. Further details are provided in the original paper. This is therefore a model for upper ocean streamfunction conditional on wind stress.

The joint distribution for the air and sea components was defined by building a simple unconditional atmospheric model. The model was implemented via a finite difference approach, and a hierarchical model was used to quantify uncertainty in the parameters. The model development here is clearly strongly motivated by physical considerations. A feature of the approach is its capacity to handle boundary conditions, which works naturally via the hierarchical approach by expressing the evolution of interior points conditional on boundary points.

A model for air pressure

Royle *et al*. (1999) developed a hierarchical model in order to produce gridded air pressure using radar scatterometer measures of wind speed. The physical model uses the fact that wind speed is proportional to the derivative of the corresponding pressure field, so that:

$$v \propto \frac{\partial p(x,y)}{\partial x}, \qquad u \propto -\frac{\partial p(x,y)}{\partial y} \tag{6}$$

Here u and v refer to the east–west and north–south wind speed components respec-

tively at coordinates x and y, with the pressure field is denoted by p. The data model is: $[\mathbf{s} \mid \mathbf{W}, \mathbf{P}] \sim N[\mathbf{KW}, \Sigma_s]$, where $\mathbf{s}$ is the observed scatterometer data, $\mathbf{K}$ is a matrix mapping the scatterometer observations to the wind ($\mathbf{W}$) grid; Σ_s is the observed data variance matrix. The process model is: $[\mathbf{W} \mid \mathbf{P}] \sim N[\mathbf{BP}, \Sigma_{w|p}]$, $[\mathbf{P}] \sim N[\mu, \Sigma_p]$.

The matrix $\mathbf{B}$ calculates empirical spatial derivatives of pressure, and pressure is given a prior multivariate normal distribution. This can be derived from a local climatology for example. Even though no pressure data area available, it is possible to produce a map of gridded pressure using this model via the posterior distribution for $\mathbf{P}$.

DISCUSSION AND CONCLUSIONS

We have explored Bayesian hierarchical methods for developing so-called physical-statistical models, and have described a number of applications. It is evident that complex models capable of producing truly integrated sources of uncertainty are now feasible. In the work of Berliner *et al*. (2003) it is clear that these models incorporate advanced physical and statistical concepts. Note, however, that hierarchical methods can be used simply to integrate different data sources, rather than statistical and physical model components. For example, if a true physical model is not viable then perhaps physical indices representing watershed condition could be incorporated via a conceptual model.

To make the most of these models requires careful physical and statistical modelling, not one or the other. We note that much of the regionalization literature still uses classical linear statistical methods, which is perhaps not making the best use of statistical modelling. Additive models have been suggested by Campbell & Bates (2001) as a means to model nonlinear mechanisms in regionalization problems, and this was taken up independently by Latraverse *et al*. (2002) with interesting results.

Physical-statistical modelling is a new way of building models, and is necessarily highly multi-disciplinary in nature. There is much work to be done to turn this into a mature technology. A fundamental issue is the development of efficient model-fitting algorithms, and we have briefly explored one approach based on importance sampling. Much of statistical science has been driven by applications, and the difficulty of prediction in ungauged basins is such that new innovations in hierarchical modelling will result. There is much to be gained by bringing statistical and physical modelling together in one framework.

Acknowledgements

I am grateful to Professor Murugesu Sivapalan for his invitation to participate in the PUB initiative, although my various work commitments have prevented the level of participation I would have liked. I am indebted to my colleague Brent Henderson for a thoughtful review of an earlier draft.

References

Berliner, L. M. (2003) Physical-statistical modeling in geophysics. *J. Geophys. Res.–Atmospheres* **108** (D24), Article no. 8776.

Berliner, L. M., Wikle, C. K. & Cressie, N. (2000) Long-lead prediction of Pacific SST via Bayesian dynamic modeling. *J. Climate* **13**, 3953–3968.

Berliner, L. M., Milliff, R. F. & Wikle, C. K. (2003) Bayesian hierarchical modeling of air-sea interaction. *J. Geophys. Res.* **108**, 1-1–1-18.

Bernardo, J. M. & Smith, A. F. M. (1994) *Bayesian Theory*. John Wiley & Sons, Chichester, UK.

Campbell, E. P. & Bates, B. C. (2001) Regionalization of rainfall–runoff model parameters using Markov Chain Monte Carlo samples. *Water Resour. Res.* **37**, 731–739.

Campbell, E. P., Fox, D. R. & Bates, B. C. (1999) A Bayesian approach to parameter estimation and pooling in nonlinear flood event models. *Water Resour. Res.* **35**, 211–220.

Dyer, R. G., Nathan, R. J., McMahon, T. A. & O'Neil, I. C. (1994) *CRC For Catchment Hydrology*. Melbourne, Victoria, Australia.

Latraverse, M., Rasmussen, P. F. & Bobée, B. (2002) Regional estimation of flood quantiles: parametric *versus* nonparametric regression models. *Water Resour. Res.* **38**, 1–11.

Qian, S. S. & Richardson, C. J. (1997) Estimating the long-term phosphorus accretion rate in the Everglades: a Bayesian approach with risk assessment. *Water Resour. Res.* **33**, 1681–1688.

Royle, J. A., Berliner, L. M., Wikle, C. K. & Milliff, R. F. (1999) In: *Case Studies in Bayesian Statistics IV* (ed. by C. Gatsonis), 367–382. Springer-Verlag, New York, USA.

Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Royal Statistical Soc. B* **55**, 3–23.

Valdés, J. B., Vicéns, G. J. & Rodríguez-Iturbe, I. (1979) Choosing among alternative hydrologic regression models. *Water Resour. Res.* **15**, 347–358.

Vicens, G. J., Rodriguez-Iturbe, I. & Schaake, J. C. Jr (1975) A Bayesian framework for the use of regional information in hydrology. *Water Resour. Res.* **11**, 405–414.

Wikle, C. K. (2003) Hierarchical models in environmental science. *Int. Statistical Review* **71**, 181–199.

Wikle, C. K., Berliner, L. M. & Milliff, R. F. (2003) Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions. *Mon. Weath. Rev.* **131**, 1051–1062.