An Introduction to Physical-Statistical Modelling Using Bayesian Methods

Technical Report 2004/49

Edward P. Campbell
CSIRO Mathematical & Information Sciences
Centre for Environment and Life Sciences
Private Bag 5, PO Wembley, WA 6913.
Tel: (08) 9333 6203
eddy.campbell@csiro.au

March 2004

Abstract

In this report we discuss the use of Bayesian hierarchical modelling as a tool to integrate physical and statistical models in geophysics. We demonstrate Bayes' theorem as a means to learn about physical processes, developing a simple example concerned with friction on slopes to illustrate the ideas. Some more complex examples drawn from the literature are summarised; we then use these ideas to develop a candidate framework for climate forecasting. We also consider methods for fitting physical-statistical models.

Keywords: Bayes' theorem; hierarchical methods; physical-statistical models; boundary values; uncertainty.

1 Introduction

Uncertainty is fundamental to the physical world, from grand physics such as Heisenberg's Uncertainty principle to the simple ubiquity of measurement error in observational work. In many cases deterministic rules are sufficient to capture the essence of a physical phenomenon. We advocate a framework known as *Bayesian Hierarchical Modelling* (BHM) within which to develop physical-statistical models that can incorporate both stochastic and deterministic elements. As we will see, such models require both physical and statistical skills if they are to be built. Bayesian methods have a long-established placed in geophysical science. For an on-line exploration see http://astrosun.tn.cornell.edu/staff/loredo/bayes/ .

Bayesian methods are used in statistical science to update information on parameters of interest in the light of new sources of information. The approach is sometimes termed 'Bayesian Learning' as a result. Fundamental to the modelling philosophy to be followed is that models are developed in a *hierarchical* fashion. That is, whilst we may be interested in modelling complex multi-dimensional phenomena, they may be broken down using a clear rule into products of simpler conditional phenomena.

BHM is now finding use in novel geophysical applications. An introduction to the field is provided by Berliner (2003). Berliner et al. (2000) used BHM to develop a forecasting scheme for Pacific Ocean sea-surface temperatures, incorporate nonlinear behaviour due to ocean regimes. An application with more explicit physical modelling is provided by Berliner et al. (2003), who handle uncertain boundary values using the approach described by Wikle et al. (2003). An excellent review of hierarchical methods in environmental science is provided by Wikle (2003)

All of the examples cited above are characterised by problems that have components amenable to physical, or physically inspired, modelling, and some components that are distinctly stochastic or subject to uncertainty of some kind. In the next section we will explore Bayes' theorem for learning about physical processes. In section 3 we will develop a simple example, and then explore some more complex applications drawn from the literature. In section 4 we develop a framework for climate forecasting. Section 5 considers issues of model-fitting, and identifies a potential way forward (with technical

statistical details largely relegated to an appendix). We finish with a discussion and some conclusions.

2 Learning About Physical Processes via Bayes' Theorem

2.1 Bayes' Theorem

There are two major objectives of interest to us:

i. Calculation of a predictive probability distribution for a physical phenomenon of interest, in light of the observed data and process modelling;

ii. Updated information on a physical process in light of the observed data.

Bayes' theorem will be of use to us in this endeavour. Suppose that a physical process is denoted by P and we collect data D on this process. Before observing D we can specify available knowledge and uncertainty about P via a probability model, denoted $\left[P|E\right]$[1]. The only source of information is expert knowledge E. To learn about P we seek to calculate $\left[P|D,E\right]$. That is, the probability model for P updated with new data D, additional to the expert knowledge E. Bayes' theorem from mathematical statistics states:

$$\left[P|D\right]=\frac{\left[D|P\right]\left[P\right]}{\left[D\right]},$$

where we have suppressed dependence on expert knowledge E for ease of exposition.

For most purposes the normalising factor $\left[D\right]$ is not required, and Bayes' theorem is written as:

$$\left[P|D\right]\propto\left[D|P\right]\left[P\right]. \qquad (1)$$

In words, the posterior model for P given D is proportional to the product of the *data model* (conditional on the process) and the prior *process model.* This requires us to model the data given our process knowledge, and to quantify our prior information.

An alternative expression of Bayes' theorem will offer a clue to the hierarchical approach to modelling:

$$\left[D|P\right]=\frac{\left[P,D\right]}{\left[P\right]},$$

where $\left[P,D\right]$ denotes the joint probability model for the data and the process. The joint probability model for the process and the data may therefore be factored as:

[1] The [ ] notation used here is a contemporary way of writing "probability of" or "distribution of", as appropriate.

$$[P, D] = [D|P][P].$$

So to model the complex joint relationship of the process and the data we calculate the product of the data model and the prior process model. We will employ this factorisation technique below.

2.2 Hierarchical Models

We now take a step closer to practical application. We assume that there are statistical parameters Θ (typically error variance, but possibly parameters accounting for factors other than variability) and physical parameters η. We may now apply Bayes' theorem to model the joint uncertainty of all the quantities involved, making some sensible modelling assumptions along the way, which should be tested in any particular application.

Thus,

$$\begin{aligned}[P, D, \Theta, \eta] &= [D|P, \Theta, \eta][P, \Theta, \eta] \\ &= [D|P, \Theta, \eta][P|\Theta, \eta][\Theta, \eta].\end{aligned}$$

The first term is the *data model*, whilst the second term is a *prior physical process* model incorporating uncertainty. The last term is a prior model for the statistical and physical parameters, called the *prior parameters model*. We now consider each of these model components in turn.

Data Model:

Conditional on the process *P* and the statistical parameters Θ it is reasonable to assume that there is no further information in the physical parameters. We may therefore simplify this term to $[D|P, \Theta]$. The most basic data model is a measurement error model, and would take the form

$$D = P + N.$$

Here *N* represents error, and would be a random variable with mean 0 and variance Θ.

Prior Process Model:

Conditional on the physical parameters, η, it is reasonable to assume that there is no further information on the process *P* contained in the statistical parameters. We may therefore simplify this term to $[P|\eta]$.

Prior Parameters Model:

Applying Bayes' theorem we find $[\Theta, \eta] = [\Theta|\eta][\eta]$. It seems reasonable to assume that the physical and statistical parameters are *a priori* independent, so that $[\Theta|\eta] = [\Theta]$. Note that this does not mean that Θ and η are independent conditional on the observed data (*a posteriori*, that is). With this assumption the prior parameters model simplifies to $[\Theta, \eta] = [\Theta][\eta]$.

□

Thus the joint probability model becomes

$$[P,\ D,\ \Theta,\ \eta] = [D|P,\Theta][P|\eta][\Theta][\eta]. \qquad (2)$$

This is a general hierarchical model for a physical process subject to uncertainty. Whilst a model for the collection P, D, Θ, η is necessarily very complex, equation (2) demonstrates that we can break it into simpler, conditional models that are simply multiplied together. This is the essence of hierarchical thinking

3 An Example

We examine here a simple example to expose the details of how Bayesian methods work in terms of developing models. The focus in this example will be on the data model and prior parameters model. Some moderately sophisticated examples from the recent literature are then described, with a greater focus on the prior process model.

Consider the motion of a body of mass m down a slope of angle θ, with a friction coefficient of μ. Our objective is to infer information about the friction coefficient using observations and physical reasoning. We may represent our physical knowledge via the following force diagram, where R represents resistive forces and g acceleration due to gravity:

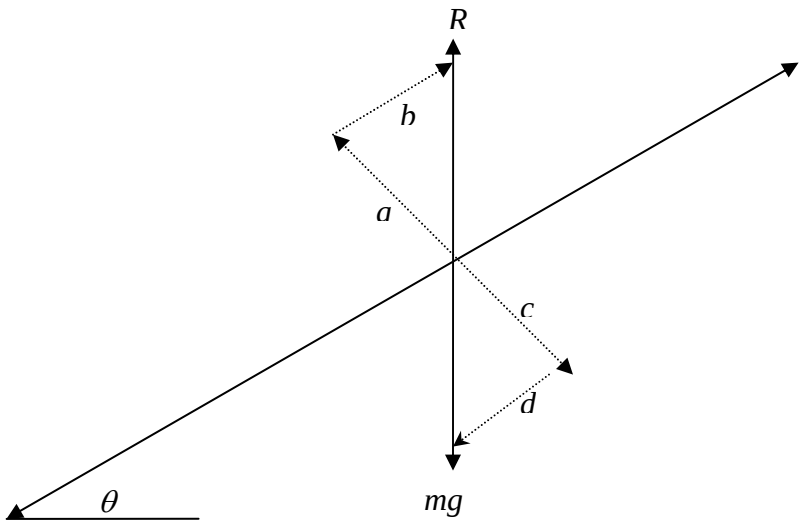

Figure 1 Force diagram for a body moving down a slope, starting at rest.

From the force diagram the resultant force down the slope is given by c-a, which follows readily as $mg\sin\theta - R\cos\theta$. If we define $X(t)$ to be the body's distance at time t down

the slope from the starting point, and define the boundary conditions $X(t)=0$ and $\dot{X}(t)=0$:

$$\ddot{X}(t) = mg\sin\theta - R\cos\theta$$
$$\Rightarrow X(t) = c_1 + c_2 t + (mg\sin\theta - R\cos\theta)t^2/2.$$

Applying the boundary conditions we find

$$X(t) = (mg\sin\theta - R\cos\theta)t^2/2\,.$$

The objective of our study is to model the resistive forces *R*. We assume here that $R \propto mg$, with constant of proportionality given by the friction coefficient μ applicable for the surfaces in contact. Thus,

$$X(t) = mg(\sin\theta - \mu\cos\theta)t^2/2\,.$$

At fixed times $t_1,\ldots,t_n$ we collect measurements $x(t_1),\ldots,x(t_n)$ on the distance down the track by some means. These are however subject to error- we assume that this error is not biased, and so is equally likely to be positive or negative. We may model the observations as a normal distribution having mean $mg(\sin\theta - \mu\cos\theta)t^2/2$ and some error variance σ^2, which we denote as $X(t) \sim N\left(mg(\sin\theta - \mu\cos\theta)t^2/2,\, \sigma^2\right)$. The better the measurement method, the smaller the value of σ^2. A more realistic error model is for a constant coefficient of variation, so that larger absolute errors are observed for larger observations. However, this simple case allows straightforward explicit calculations.

We may now formulate a hierarchical model.

Data Model

This is the joint probability of the observed data, which under reasonable assumptions is just the product of the individual observations' probability distributions:

$$\left[X(t_1),\ldots,X(t_n)\middle|\sigma^2,\mu\right] = \prod_{i=1}^{n}\left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[X(t_i) - \tfrac{1}{2}mg(\sin\theta - \mu\cos\theta)t_i^2\right]^2\right\}$$
$$= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[X(t_i) - \tfrac{1}{2}mg(\sin\theta - \mu\cos\theta)t_i^2\right]^2\right\}.$$

Prior Process Model

We are free now to describe any information we have on the process, and we choose to quantify our knowledge of the friction coefficient μ. In this case we assume that we can specify a mean value (η) and a variance (τ^2), and it seems once again not unreasonable to assume a normal distribution. However, given that we expect movement down the slope (else why do the experiment), we know that the resultant force is positive:

$$mg\left(\sin\theta - \mu\cos\theta\right) > 0$$
$$\Rightarrow \mu < \tan\theta.$$

The variance parameter in cases like this will normally be a consequence of expert assessment of a reasonable range for the parameter of interest. Thus,

$$\left[\mu|\eta\right] \sim N\left(\eta, \tau^2\right), \qquad \mu < \tan\theta.$$

Clearly we could be more sophisticated, although there is a strong physical inspiration in the data model.

Prior Parameters Model

We now need to specify our prior uncertainty in all the parameters we have used- η, σ^2 and τ^2, denoted by $\left[\eta, \sigma^2, \tau^2\right]$. It seems reasonable to assume that these quantities are *a priori* independent, so that $\left[\eta, \sigma^2, \tau^2\right] = \left[\eta\right]\left[\sigma^2\right]\left[\tau^2\right]$. It is common to use previous data to define prior distributions for variance parameters, but it is often the case that no previous data are available. In such cases it is common to assign so-called *vague* priors, which in this case (for technical reasons) would be

$$\left[\eta < \tan\theta\right] \propto 1, \left[\sigma^2\right] \propto 1/\sigma^2, \left[\tau^2\right] \propto 1/\tau^2.$$

Where more information is available it is normally found that positively skewed distributions, such as the gamma or inverse chi-squared distributions are appropriate prior models.

□

Appealing to equations (1) and (2), the posterior probability distribution for all the parameters is proportional to the product of the data model, the prior process model and the prior parameters model:

$$\left[\mu, \eta, \tau^2, \sigma^2 \middle| X\left(t_1\right), \ldots, X\left(t_n\right)\right] \propto \left[X\left(t_1\right), \ldots, X\left(t_n\right) \middle| \sigma^2, \mu\right]\left[\mu|\eta, \tau^2\right]\left[\eta, \tau^2, \sigma^2\right]. \quad (3)$$

3.1 Numerical Example

To find the distribution for μ from (3) we must in general integrate out the unwanted parameters η, τ^2 and σ^2. In practice we rarely do this as the multi-dimensional integral involved is only analytically tractable in rare cases. We would also have to calculate the constant of proportionality implied by (3). There are in fact methods for generating a sample from the probability distribution for μ, and the most popular of these is known as Markov chain Monte Carlo (MCMC). Such methods are beyond the scope of the present report; suffice it to say that this is possible, and the algorithm was in fact first developed in the physics literature (Metropolis et al. (1953)). It was subsequently generalised in the statistics literature by Hastings (1970).

We will consider a simpler case here for illustration, assuming that the measurement error is known to be $\sigma^2 = \left(10\text{mm}\right)^2$ from past experiments. We assume further that good prior

information on μ is available, which is captured as $\mu \sim N\left(0.1,\, 0.1^2\right)\ \mu < \tan\theta$, eliminating the parameter τ^2 in effect. The posterior distribution then is

$$\left[\mu \middle| X\left(t_1\right), \ldots,\ X\left(t_n\right)\right] \propto \exp\left\{-\frac{1}{2.\sigma^2}\sum_{i=1}^{n}\left[X\left(t_i\right) - mg\left(\sin\theta - \mu\cos\theta\right)t_i^2\right]^2\right\} \times \exp\left\{-\frac{1}{2.\tau^2}\left(\mu - 0.1\right)^2\right\}, \tag{4}$$

retaining only terms in μ.

It can be shown that[2]

$$\mu \middle| X\left(t_1\right), \ldots,\ X\left(t_n\right) \sim N\left(\frac{\sigma^2\eta - 0.5\tau^2\sum b_i}{\sigma^2 + \tau^2\sum a_i},\ \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2\sum a_i}\right),$$

where $a_i = \frac{1}{4}m^2g^2t_i^4\cos^2\theta$ and $b_i = mgX_it_i^2\cos\theta - \frac{1}{4}m^2g^2t_i^4\sin 2\theta$.

We have simulated a case for a 45° slope with friction coefficient 0.09, and the results are shown in Figure 2 below. Because of the size of the slope there is insignificant prior probability greater than $\tan 45^\circ = 1$, so we can neglect the upper bound on μ. The posterior mean in this case turns out to be 0.076 with standard deviation 0.0178, so a substantial decrease compared to the prior standard deviation. The Bayesian learning process is depicted in Figure 3 where we show the prior and posterior distributions for the friction coefficient. It is evident that the posterior is more concentrated than the prior, and shifted a little to the left, in keeping with the observations. As more data are collected the posterior will tend to be less influenced by the prior.

As implied above, many improvements could be made to this example, which is simplified to illustrate the ideas involved. We could employ a positive-valued or otherwise truncated prior distribution for μ to avoid negative values, and more realistic measurement error models are possible, as noted above. These cases are no more complex from a statistical perspective, but we require different approaches to quantifying the posterior distribution. Rather than dwell on technical issues, it is perhaps more beneficial at this stage to consider some more substantive, and necessarily complex, applications studied in the literature.

3.2 An Extension

It could be argued that the composition of the resistive forces *R* is, in practice, subject to substantial uncertainty (Frank Dehoog, P*ers. Comm.*). In such cases we might be able to suggest alternative compositions for *R*, to be tested against available data. Alternatively, we might simply develop a statistical model for the resistive forces that is inspired by our physical understanding.

---

[2] This is done by forming a quadratic in μ and then 'completing the square' to form the kernel of a normal probability density. By the uniqueness property of density functions, we can identify the distribution.

First, the alternative compositions approach. There are a number of ways that this might be done. Suppose that

$$R = \sum_{i=1}^{r} \delta_i R_i \left(\psi_i \right),$$

where we assume that there are r different compositions, the presence or absence of the i^{th} indicated by the indicator variable

$$\delta_i = \begin{cases} 1, & \text{present} \\ 0, & \text{absent.} \end{cases}$$

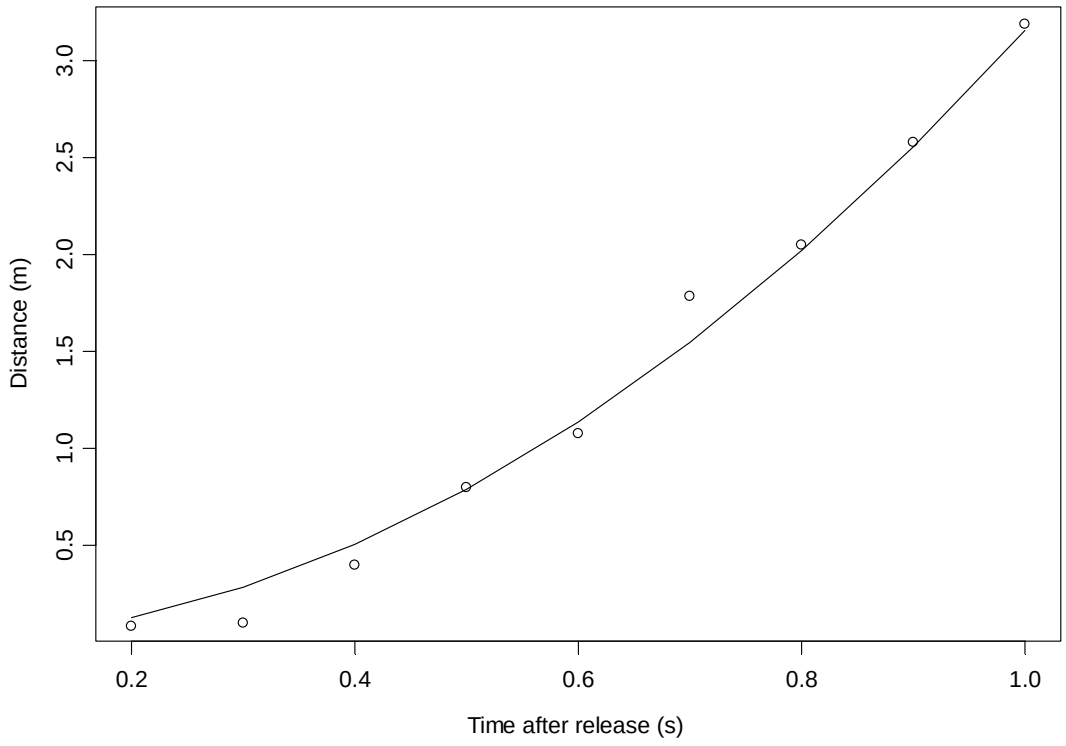

Figure 2 Simulated points and expected distance (solid curve).

The parameters for the i^{th} composition are denoted by ψ_i. In the hierarchical approach we would define a prior $[\Delta, \Psi]$ for these indicator variables and parameters, and the prior could incorporate sophisticated information such as not all compositions are mutually exclusive. The posterior for Δ would give an indication of support for the various resistive force compositions.

Now for the statistical approach; the possibilities really are endless, so let's examine an approach to modelling nonlinear effects. If we write $u = mg$ then in the example we used $R(u) = u$. An obvious extension is $R(u) = \beta_0 + \beta_1 u + \beta_2 u^2$, or perhaps some kind of spline. Bayesian methods could then be used to select an optimal model, in light of the available data.

It seems entirely feasible to mix these approaches to see if there are any interesting patterns over and above our physical formulation of the physical problem. That is,

$$R = R_i(u; \psi) + \beta u^2,$$

say, to look for an additional nonlinear effect. Clearly such a modelling exercise would have to be guided by physical insight to avoid modelling noise, in effect.

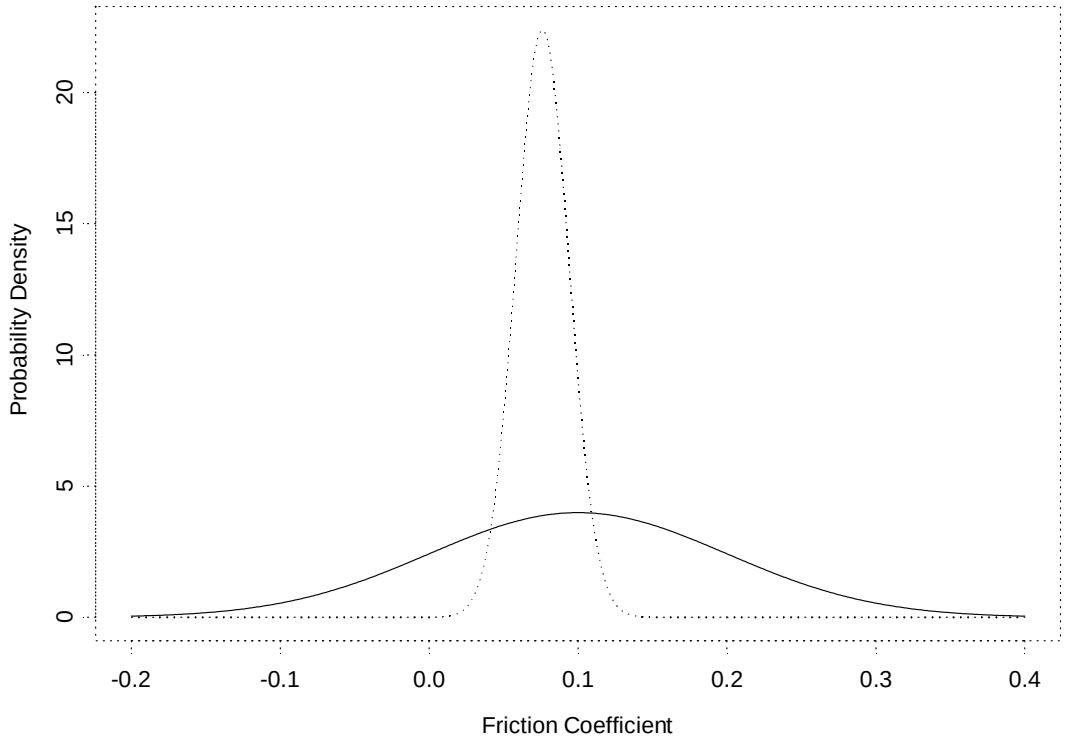

Figure 3 Prior (solid line) and posterior (dashed line) distributions for the friction coefficient.

3.3 Examples From the Literature

Having looked in some detail at the simple example above, we now return in a little more detail to the geophysical applications noted in the introduction. The first application to

receive attention in the climate literature was the prediction of Pacific Ocean sea-surface temperatures (SST) using hierarchical methods (Berliner et al. (2000)). This work used a qualitative description of SST dynamics to build a process model that was essentially statistical. The model incorporated regime-dependent behaviour of the El Niño-Southern Oscillation phenomenon using a threshold model for the southern oscillation index (SOI). Whilst the modelling was physically-inspired, there was no explicit physical modelling. The results obtained are impressive, nonetheless[3].

A more explicitly physical process model of air-sea interaction was explored by Berliner et al. (2003). The model was developed as the product of an atmosphere-given-ocean model and an atmospheric model. The ocean component used a 'basic quasi-geostrophic' model of the space-time evolution of the gridded ocean streamfunction values. To implement the model the spatial derivatives required were approximated empirically via finite differences. A wind stress term is introduced into the evolution equation for the streamfunction, linking the ocean and atmosphere components.

A feature of the modelling is that the evolution of interior points in space is conditional on the boundary of the modelled space, which is very easy to handle in a Bayesian framework. The method for handling boundary values and uncertainty in boundary conditions is based on Wikle et al. (2003). Boundary condition uncertainty is a common feature of geophysical applications, and this paper shows how to integrate this with other sources of uncertainty via a hierarchical model. A statistically technical feature of this paper is the development of a model calibration algorithm suitable for a coupled physical model. This is complicated by conventional statistical standards, but is not computationally intensive by the standards of a physical coupled ocean-atmosphere model.

The scope for hierarchical approaches in environmental modelling is clearly very large. A very readable introduction to the field is provided by Wikle (2003). The focus of this paper is on general methods of modelling processes that evolve in space and time, using hierarchical methods to capture the complexity. An introductory paper targeted explicitly for a geophysics audience is Berliner (2003). In this latter paper the ideas are introduced through a study of the kinematics of fluids. The paper then considers applications in climate modelling, fingerprinting of climate change and near-surface ocean winds.

4 A Suggested Framework for Climate Forecasting

We are typically interested in learning about and forecasting climate phenomena using observed data and physical knowledge. Suppose we denote by Y the climate output of interest, with observations on Y denoted D_Y. Possibilities for Y include rainfall, temperature etc. We assume that the climate output is influenced by a collection of physical processes or climate inputs, say, $Z = \{Z_i : i = 1, \ldots, p\}$, with data on some or all of these processes denoted by D_Z. We denote by Θ_Y and Θ_Z parameters associated with

[3] See http://www.stat.ohio-state.edu/~sses/collab_enso.php for the latest predictions.

the measurement processes- typically measurement error. We further denote by η_Z and η_Y parameters associated with the physical process models for *Z* and *Y* respectively.

In summary, we have defined the following quantities:

Table 1 Nomenclature for a climate forecasting framework.

Processes	**Data**	**Parameters**
Y- Climate output	D_Y	Θ_Y - Parameters of measurement model η_Y - Parameters of physical process model for climate output.
Z- Physical processes influencing *Y*	D_Z	Θ_Z - Parameters of measurement model η_Z - Parameters of physical process model for climate inputs.

We seek a joint probability model for all the quantities shown in Table 1, and applying hierarchical thinking:

$$\begin{aligned}\left[D_Y, D_Z, Y, Z, \Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right] &= \left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right]\left[Y, Z, \Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right] \\ &= \left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right]\left[Y, Z \middle| \Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right]\times \\ &\qquad \left[\Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right].\end{aligned}$$

Conditional on *Z* there is no further information in the physical process parameters, so the first term on the right hand side simplifies to $\left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z\right]$. Conditional on the physical parameters, the statistical parameters yield no further information on the *Y* and *Z* processes, so the second term simplifies to $\left[Y, Z \middle| \eta_Y, \eta_Z\right]$. This yields a factorisation of the joint probability model as

$$\left[D_Y, D_Z, Y, Z, \Theta_Y, \Theta_Z, \eta_Z\right] = \left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z\right]\left[Y, Z \middle| \eta_Y, \eta_Z\right]\left[\Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right],$$

the product of a data model, a process model and a prior parameter model. We consider these components in more detail below.

4.1 The Data Model

The data model is

$$\left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z\right].$$

A reasonable assumption at this stage is that each data set is conditionally independent, given their respective statistical and process parameters. Applying Bayes' theorem with this assumption we find:

$$\left[D_Y, D_Z \middle| Y, Z, \Theta_Y, \Theta_Z\right] = \left[D_Y \middle| Y, \Theta_Y\right]\left[D_Z \middle| Z, \Theta_Z\right].$$

This does not mean that the data sets are independent in an absolute sense- they are inter-connected through the process models.

4.2 The Prior Process Model

The process model is

$$\left[Y, Z \middle| \eta_Y, \eta_Z\right].$$

Given the climate input-output formulation introduced earlier, it makes sense to factorise this as

$$\left[Y, Z \middle| \eta_Y, \eta_Z\right] = \left[Y \middle| Z, \eta_Y\right]\left[Z \middle| \eta_Z\right].$$

The first component is a process model for how the climate inputs *Z* influence the climate output *Y*. We might, for example, capture some physical description of the rainfall process here. Even if this only crude, it will serve to constrain the data model to be physically reasonable. The second component is a pure climate-input process model, perhaps capturing broad scale ocean and atmosphere processes.

It may be that the climate-inputs process model partitions naturally into ocean and atmosphere components Z_o and Z_a say, then

$$\begin{aligned}\left[Z \middle| \eta_Z\right] &= \left[Z_o, Z_a \middle| \eta_o, \eta_a\right] \\ &= \left[Z_a \middle| Z_o, \eta_a\right]\left[Z_0 \middle| \eta_o\right].\end{aligned}$$

The first component is an atmospheric model conditional on ocean state, whilst the second component is an (unconditional) ocean model. This factorisation was employed by Berliner et al. (2003) in their application. It is evident that significant statistical-geophysical collaboration is required to develop a model of this sort.

4.3 The Prior Parameters Model

The prior parameters model is

$$\left[\Theta_Y, \Theta_Z, \eta_Y, \eta_Z\right].$$

Generally speaking we will assume *a priori* independence of the parameters, and just factor this into a product of individual prior distributions. Bayes' theorem can be used to develop more complex prior parameter models if needed.

5 Fitting Physical-Statistical Models

If we have just one physical process *P*, then in keeping with our earlier notation the posterior distribution of the physical process and all parameters given the data is

$$\left[P, \eta, \Theta \middle| D\right] \propto \left[D \middle| P, \Theta\right]\left[P \middle| \Theta\right]\left[\eta, \Theta\right]. \qquad (5)$$

In order to summarise properties, such as posterior means and standard errors, we must be able to carry-out what are likely to be complex integrations. It is rarely possible to do this analytically, so it has become common practice to use simulation techniques. The most common method for developing suitable algorithms is known as Markov chain Monte Carlo (MCMC- see section 3.1).

In general MCMC will require us to generate many random realisations from the physical model. This will work well if this model is relatively simple, but it is likely to be inherently nonlinear and complex in the applications of interest to us. This means that MCMC is likely to be computationally inefficient and time-consuming.

An alternative approach is based on the idea of importance sampling-resampling (IS-R). In essence, IS-R is a method of generating a sample from a desired (typically complex) distribution given a sample from some other (simpler) distribution. The algorithm can be written as:

1. Generate a sample $\{x_i\}$ of M observations from an easy to simulate distribution g.
2. Calculate $w_i = f(x_i)/g(x_i)$ $i = 1, \ldots, M$, where f is the target distribution, known up to a constant of proportionality.
3. Calculate $q_i = w_i / \sum w_j$.
4. Repeat from 1 to N
    a. Sample with replacement one of the $\{x_i\}$ at random.
    b. Accept as a sample from the target distribution g with probability q_i.
5. End

In the physical-statistical model (5) we can apply this algorithm as follows:

1. Generate a sample from the prior parameters model.
2. Using this sample, generate an ensemble of physical process realisations.
3. Use the data model in IS-R to generate a sample approximately from the target posterior distribution using the ensemble simulated from the prior process model.

This is essentially the model-fitting approach used by Berliner et al. (2003), with the complication of two interacting processes (ocean and atmosphere). An ensemble from the (unconditional) atmospheric model was generated first using MCMC, then realisations of ocean streamfunction were generated for each member of this ensemble. This gives a combined ensemble for the physical and prior components. Importance sampling using the ocean data model was then used to summarise the posterior quantities of interest. The algorithm is termed importance sampling Monte Carlo- Markov chain Monte Carlo (ISMC-MCMC) by the authors. Further technical details are given in the appendix.

A weakness of importance sampling in general is that if the information in the data is inconsistent with the prior specifications then the algorithm can be very inefficient The ISMC-MCMC algorithm appears to be very promising however because the MCMC-derived sample that provides the input stream for importance sampling is guided in part by the data. We would therefore expect a corresponding gain in efficiency compared to a sample drawn from a prior distribution alone.

6 Discussion and Conclusions

We have seen that Bayesian hierarchical methods can be used to integrate physical and statistical models in one framework. This enables us to directly produce probability forecasts of system evolution, bringing together all sources of uncertainty. Because the statistical and physical components are inter-connected they 'borrow strength' from each other. This also means that total uncertainty is not simply the sum of all uncertainties in the system.

We explored a very simple example of friction on a slope to illustrate the ideas. In this case physical modelling came mostly through the data model, conditional on the process. The prior process model was very simple. We then examined some examples from the literature where the prior process model is much more complex. Having explored these ideas we have suggested an outline framework for climate forecasting.

Fitting physical-statistical models using available data is a difficult problem. At this point it seems that a hybrid approach using importance sampling is the most promising. It seems unlikely that conventional approaches based on Markov chain Monte Carlo (MCMC) alone will be more efficient, unless some special insights can be used in constructing the proposal distribution that drives this approach. Determining appropriate sample sizes for simulation procedures is an open question at this stage.

There is a growing literature on adaptive MCMC that is still in its infancy, and could well provide a way forward. One possibility is the use of the Hastings Coupling method. If efficient MCMC approaches can be developed then the question of convergence arises immediately, which is an open question of some complexity.

A major challenge of modelling using hierarchical methods is that a truly collaborative effort between statisticians and geophysicists is required to develop models in this way. It's a new way of building models.

Acknowledgements The ideas described here are being developed as part of the Indian Ocean Climate Initiative (IOCI) to be applied to climate forecasting problems. IOCI is supported financially and in-kind by the WA State Government The author is grateful to helpful comments and suggestions provided by Brent Henderson and Noel Cressie in particular.

References

Berliner, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research- Atmospheres,* **To appear**.

Berliner, L. M., Milliff, R. F. and Wikle, C. K. (2003). Bayesian hierarchical modeling of air-sea interaction. *Journal of Geophysical Research,* **108,** 1-1:1-18.

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

Hastings, W. K. (1970). Monte Carlo sampling using Markov chains and their applications. *Biometrika,* **186**.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* **21,** 1087-1092.

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

Wikle, C. K., Berliner, L. M. and Milliff, R. F. (2003). Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions. *Monthly Weather Review,* **131,** 1051-1062.

Appendix: Importance Sampling for Fitting Physical-Statistical Models

Using our general framework the posterior distribution typically of interest is

$$\left[P,\Theta,\eta|D\right]\propto\left[D|P,\Theta\right]\left[P|\eta\right]\left[\eta,\Theta\right].$$

In geophysical applications the prior process model $\left[P|\eta\right]$ is likely to be nonlinear and complicated, at least in terms of the physical parameters η. This means that conventional MCMC algorithms are likely to be inefficient[4]. The importance sampling algorithm of Berliner et al. (2003) is as follows:

1. Generate a sample $\left\{\Theta_i,\eta_i\right\}$ from the prior parameters distribution.
2. Generate an ensemble $\left\{P_i\right\}$ from the prior physical model using the $\left\{\eta_i\right\}$.
3. Resample the $\left\{P_i,\Theta_i,\eta_i\right\}$ with the acceptance probability

[4] Gibbs sampling would appear to be ruled out in all but trivial cases on the grounds that full conditionals are not accessible. A Metropolis-Hastings algorithm would require very careful tuning to achieve a reasonable acceptance rate. It seems likely that at least some of the physical parameters will be correlated, and reparameterisation is unlikely to be an option.

$$q_i = \frac{\left[D\middle|P_i,\Theta_i\right]}{\sum_j \left[D\middle|P_j,\Theta_j\right]}.$$

The resulting sample tends in distribution to the posterior.

[]

An immediate issue that arises is that this algorithm is heavily dependent on how 'close' the prior is to the posterior. If we were to replace the sample from the prior distribution with a more general distribution $g(\eta,\Theta)$, say, then the acceptance probability becomes

$$q_i = \frac{\left[D\middle|P_i,\Theta_i\right]\left[P_i\middle|\eta_i\right]\left[\eta_i,\Theta_i\right]\Big/ g(\eta_i,\Theta_i)}{\sum_j \left[D\middle|P_j,\Theta_j\right]\left[P_j\middle|\eta_j\right]\left[\eta_j,\Theta_j\right]\Big/ g(\eta_j,\Theta_j)}.$$

We can make this algorithm more efficient by matching *g* to the posterior as best we are able. Berliner et al. (2003) allude to 'sequential' application of MCMC in appendix A3 to their paper. One way is to tune the density function *g* using an initial run of the algorithm, using the basic acceptance probability and sampling from the prior parameters model. For example, we could use a multivariate normal form for *g*, with transformed parameters to take account of finite ranges or skewness for example. This seems a promising approach to make the importance sampling approach adaptive, and could be made essentially automatic. An alternative path to explore is the use of density estimation techniques.

Berliner et al.'s Model:

This model is for air-sea interaction, with components *U, V* for wind direction and ocean streamfunction Ψ. The posterior of interest in their notation is

$$\left[U,V,\Psi,\Theta_w,\Theta_\psi,\eta_w,\eta_\psi\middle|D_\psi,D_w\right] \propto \left[D_\psi\middle|\Psi,\Theta_\psi\right]\left[D_w\middle|U,V,\Theta_w\right]\left[\Psi\middle|U,V,\eta_\psi\right]\times \left[U,V\middle|\eta_w\right]\left[\Theta_w,\Theta_\psi,\eta_w,\eta_\psi\right],$$

albeit expressed a little differently. Applying Bayes' theorem to the second line yields their equation (B2):

$$\left[U,V,\Psi,\Theta_w,\Theta_\psi,\eta_w,\eta_\psi\middle|D_\psi,D_w\right] \propto \left[D_\psi\middle|\Psi,\Theta_\psi\right]\left[D_w\middle|U,V,\Theta_w\right]\left[\Psi\middle|U,V,\eta_\psi\right]\times \left[\eta_\psi,\Theta_\psi\right]\left[U,V,\Theta_w,\eta_w\middle|D_w\right].$$

Their algorithm then proceeds as follows:

1. use MCMC to generate a sample from $\left[U,V,\Theta_w,\eta_w\middle|D_w\right]$.
2. Simulate a matching sample from $\left[\eta_\psi,\Theta_\psi\right]$.
3. Generate an ensemble $\{\Psi_i\}$ using the sampled values.
4. Resample these simulated samples using the acceptance probability

$$q_i = \frac{\left[D_{\psi} \middle| \Psi_i, \Theta_{\Psi_i} \right]}{\sum_j \left[D_{\psi} \middle| \Psi_j, \Theta_{\Psi_j} \right]}.$$

[]

As before, we could generalise to a case where we sample from some distribution other than the prior parameters model at step 2.

First Draft:	16/12/03
Second Draft:	9/1/04
Incorporating 2nd draft comments:	22/1/04
Incorporating some thoughts on partial knowledge:	10/2/04
More details on model-fitting algorithms:	27/2/04
Final edits:	5/3/04