

POSITION PAPER

WORKSHOP ON HIERARCHICAL MODELING IN ENVIRONMENTAL STATISTICS (COLUMBUS, OH, MAY 14-16, 2000)

Submitted to U.S. Environmental Protection Agency

Principal Investigator

Noel Cressie

Department of Statistics
The Ohio State University
1958 Neil Avenue, #404
Columbus OH 43210-1247

tel: (614) 292-5194

fax: (614) 292-2096

e-mail: ncressie@stat.ohio-state.edu

July 21, 2000

Preamble

The Workshop on Hierarchical Modeling in Environmental Statistics (WHIES) was held on May 14-16, 2000 at The Ohio State University in Columbus, OH. During WHIES, break-out discussion groups met on several occasions to consider questions posed, in advance, on hierarchical modeling in environmental statistics. Each discussion-group leader prepared a draft of their discussion and these have been combined into the position paper given below.

Topics were chosen in the following manner. The conference organizer (Noel Cressie) asked for suggestions from the discussion-group leaders (George Casella, Leo Knorr-Held, Louise Ryan, Tom Santner, and Mark Schervish) in advance of the workshop. He also asked the rest of the Scientific Program Committee (Mark Berliner and Tim Gregoire) and the U.S. Environmental Protection Agency (Larry Cox) for suggestions. After some filtering and discussion by e-mail, Cressie and the discussion-group leaders prepared a list of more than 20 topics to be considered by workshop participants. At a plenary meeting on the first day, participants added about 5 more topics and then voted on the topics they thought were the most interesting. That evening, Cressie and the discussion-group leaders reduced the number of topics down to 11. During subsequent break-out discussion sessions, each group was asked to discuss 5 topics, chosen so that each topic was discussed by 2 or 3 groups.

List of Groups and Topics

- **Group 1** (led by Leo Knorr-Held) discussed topics A, D, E, H, and I.
- **Group 2** (led by Mark Schervish) discussed topics E, F, I, J, and K.
- **Group 3** (led by Louise Ryan) discussed topics, C, D, G, J, and K.
- **Group 4** (led by George Casella) discussed topics, A, B, F, G, and K.
- **Group 5** (led by Tom Santner) discussed topics, A, B, C, H, and J.

Text of Discussion Topics (Topics A-K)

- A. Hierarchical models of environmental phenomena are typically highly parameterized and require (Markov Chain) Monte Carlo methods in their fitting. They tend not to lend themselves to traditional measures of lack-of-fit

such as residuals analysis, and model-comparison measures such as BIC and AIC can be impractical or meaningless. What diagnostics can be used for checking assumptions in the hierarchical model? (Cross-validation? Posterior-predictive checks?)

- B. How should one model multiple, related response functions? How should one design the study to collect the training data in such a setting (keeping in mind the need for computational efficiency and modeling flexibility)?
- C. In the context of environmental risk assessment, one is often faced with ‘less than perfect’ data sets from different sources. Invariably, one will encounter problems with missing data and measurement error. Sometimes, important covariates might be missing all together. What kinds of statistical models are useful in such settings? Hierarchical models seem like a natural candidate. What pitfalls should we be aware of in attempting to fit such models?
- D. How should one approach model selection in the context of environmental risk assessment? In the classic toxicologically based risk-assessment setting, current wisdom suggests basing regulatory decisions on a so-called “benchmark dose” analysis, which involves finding the dose level that corresponds to a moderate (5% or 10%) increase above background rates. Because the benchmark dose is usually within the experiment dose range, it tends to be fairly robust to model-choice issues. However, in the context of epidemiologically based risk assessment, issues such as exposure measurement error and the general uncertainty inherent in human data mean that model choice can be a critical issue, even for selecting the 5% or 10% excess risk. A particular problem for the epidemiological setting is that unexposed event rates are often poorly defined. How should one approach model selection in such settings? Does Bayesian model averaging provide a good approach?
- E. The construction of realistically complex models for space-time data seems to remain a mixture of art and science. There are a number of useful principles in classical statistics (e.g., invariance, identifiability, parsimony), but in Bayesian hierarchical modelling there seems to be less attention paid to them. Can we nevertheless agree on a list of basic requirements and principles that any such Bayesian hierarchical model should fulfill?
- F. Many of the models that have been developed for environmental data ben-

enefit greatly from incorporation of geographical and/or ecological background information. For example, the USGS has developed models of pesticide concentration in drinking water that incorporate water-flow patterns. Much is known about volatility and solubility of chemicals and patterns of wind flow. What are appropriate ways to incorporate such scientific and engineering knowledge into statistical models (or vice versa) in order to produce models that provide the best of both worlds?

- G. Almost any reasonably complex model is a mixture of empirical and subjective components. When evaluating such models, we might consider both frequency and Bayesian measures. However, it may be the case that the empirical and subjective pieces are so intertwined as not to allow separate evaluations (for example, in empirical Bayes models the prior parameters are estimated from the data). Can we agree on an evaluation strategy for complex models that would result in an inference (no doubt a mixture of Bayes and frequentist) that would be acceptable? Such a scheme should include
- an assessment of long-run properties,
 - an assessment of Bayesian robustness,
 - a sensitivity analysis.
- H. Hierarchical models often involve modeling “local” structure in order to analyze spatially extensive data. The level of localness is often chosen for practical reasons having to do with the granularity of the data, rather than for reasons having to do with the natural scale of the phenomenon. This leads to two questions:
- Does the hierarchical statistical model aggregate consistently, at least approximately?
 - Is the hierarchical statistical model flexible enough to allow inference at a disaggregated level?
- I. When making a spatial map of some summary of the full posterior (e.g., pointwise spatial medians), what meaningful measures of variability can we attach to the map? Alternatively, how can we summarize 1,000 maps drawn from the full posterior distribution?
- J. Expert opinion can be crucial to hierarchical statistical modeling, particularly as one goes deeper into the hierarchy where process-behavior is less

well known. Different experts will likely differ in their opinions. Can we formalize a way to combine expert opinion?

- K. Good hierarchical statistical modeling to solve environmental problems requires teamwork between statistical scientists and substantive scientists. However, statisticians are often put in a consulting role where quick answers are sought. Important questions may require a longer, more collaborative effort, and hierarchical statistical models demand that effort. How can we convince scientists that a hierarchical statistical analysis is “worth waiting for”?

RESPONSES

Topic A

Hierarchical models of environmental phenomena are typically highly parameterized and require (Markov Chain) Monte Carlo methods in their fitting. They tend not to lend themselves to traditional measures of lack-of-fit such as residuals analysis, and model-comparison measures such as BIC and AIC can be impractical or meaningless. What diagnostics can be used for checking assumptions in the hierarchical model? (Cross-validation? Posterior-predictive checks?)

Group 1 Response to Topic A:

- Cross validation was proposed as a diagnostic tool which is both easily interpretable and widely applicable when the goal is prediction. Typically, this might proceed by splitting the data into two portions, X_1 and X_2 , comparing the posterior predictive distribution based on X_1 to the empirical distribution of X_2 , and vice versa. X_1 and X_2 are often selected via a random sample which is stratified with respect to covariates. However, problems with more complicated data structure, for example spatial or spatio-temporal dependencies, may require more sophisticated techniques for splitting the sample. Furthermore, the usual practice of splitting the data by half is not necessarily optimal, and alternative schemes should be investigated.
- Models with k parameters typically have an “effective number of parameters” (less than k) which is not well understood. This results from constraints imposed by the prior/model. This frustrates some “*IC” (e.g. AIC, BIC) approaches.

- The DIC criteria provides a way of quantifying the complexity of the model. However, it is not transformation invariant. A suggestion was to replace means with medians.
- It may be possible to estimate the effective number of parameters via cross-validation.
- A suggestion was made to perform model diagnostics at intermediate levels of the hierarchy. It may be possible to generate pseudo data (from deterministic models, for example) for this purpose.

Group 4 Response to Topic A:

The main question here is “What diagnostics can be used for checking assumptions in the hierarchical model?”

Some initial concerns and comments:

- We are not convinced that BIC and AIC are meaningless.
- A possible problem with hierarchical models are that they are too highly parameterized. Can diagnostics, or other methods, tell us when to stop adding levels?
- “Why should we be penalized for fitting more complicated models?” One possible answer to this is that if the resulting model is not robust, or overly sensitive, then a less complex model may be preferred. Remember that the AIC/BIC criterion, which penalizes complexity, is derived directly from the criterion of predictive loss.
- Not all models can always be diagnosed. Many disciplinary models have to rely on expert judgment.

Some suggested solutions:

- We agree that an important first step is to define the goals of the diagnostic. For example, is the interest in checking predictive efficacy, assessing sensitivity to assumptions, etc? This is part of a general overall recommendation that the statistician needs to focus the experimenter on the ultimate goals and policy decisions that are to be obtained from the analysis. Often, consideration of such goals will help focus the appropriate analysis.
- There are two different kinds of diagnostics. AIC/BIC, etc. are “summary measures” of overall fit, in contrast to specific, prediction measures like “ $O - E$ ”.

- We must also differentiate between “comparison of models” versus “absolute fit.” The former can be judged with a statistical criterion, but the latter needs expert assessment.
- Models based on physical principles can still be tuned by the statistician, possibly to achieve a goal such as robustness. It may also happen that a simpler model is better for prediction. This suggests that AIC/BIC can be important even if the model is based on physical principles, and again underscores that it is important to know the goals of the experimenter. But for geophysical (and other) models based on scientific principles, we might need a variation of, or weighting in, an AIC-type criterion. The experimenter might need to break down the hierarchy to get at the principles, regardless of what AIC or anything else indicates.
- The model diagnostics we have discussed have been of the form of summary measures (AIC/BIC) or predictive measures. Are there other ideas beyond cross validation and prediction? For example, we could use simulation to check estimates of parameters or latent variables. Are there guidelines for such procedures?
- Criteria such as AIC/BIC are based on the likelihood function, while predictive checks do not necessarily need the likelihood. If the likelihood is “hard to get”, how do we assess comparative fit?
- Hierarchical models are necessarily complex, and often have a large number of parameters. This does not translate directly into degrees of freedom, however. So, how do we count degrees of freedom in complex models, and how do we assess the interaction between the complexity of a model (depth of a hierarchy) and information criteria/fit? (A step in this direction is DIC - Bayes predictive goodness of fit with automatic df adjustment.) It should be noted that Hodges and others have done work on this issue.

Group 5 Response to Topic A:

- *This is a fundamental problem.* There are at least three points of interest: model fit, prediction and forecasting. The diagnostic used will depend on the outcome of interest.
- To assess model fit, several diagnostics were suggested. One technique is to use the fitted model to generate new data and perform informal comparisons with the “real” data. Another technique is to compute generalized residuals (Cox and Snell); the difficulty with this approach is to determine

the comparison distribution since different stochastic spatial models can give rise to different theoretical distributions for the residuals, the determination of which is an area of open research. Another aspect of model fit is to explicitly assess fit in the lower levels of the hierarchy. Model interpretability is a critical component in creating a hierarchy for which one can elicit expert opinion. In many subject areas, it is feasible to ask experts to think conditionally about A given B and B given C in a hierarchical fashion, thereby improving the accuracy of the modeling process. We also discussed AIC, BIC and likelihood type measures that may be useful for comparing two models, but may not be so good for telling whether or not the model is appropriate. For example, one might be able to compare a model based on one prior with a model based on another prior. Another technique for verifying fit is to collect a validation dataset and compare those values with model predictions.

- To assess prediction (i.e., guesses of new data within the support of the observations), cross-validation and bootstrapping are the obvious “automatic” diagnostics. Of course one can split the training data in other ways (90%/10%, for example) and it is not clear what is the most powerful method. An example is Jim Zidek’s PM-10 airport prediction based on city data.
- To assess forecasting (i.e., guesses of new data outside the support of the observations), use validation.
- Sensitivity analysis: cannot be done comprehensively in high-dimensional problems. Process must be guided by expert opinion. For example, their input would be used to determine the critical priors whose effect on the posterior we wish to study. Importance sampling can be used to permit approximate sampling from the posterior based on a single set of MCMC draws, say. In low dimensional problems, the use of a fractional factorial experimental design with the prior distributions as the factors might be useful to examine sensitivity to the priors.
- Useful reference: Hodges, James S., (1998). Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics, *Journal of the Royal Statistical Society, Series B*, **60**, 497–521.

Topic B

How should one model multiple, related response functions? How should one design the study to collect the training data in such a setting (keeping in mind the need for computational efficiency and modeling flexibility)?

Group 4 Response to Topic B:

By mutual group decision, this topic was not discussed during the break-out session.

Group 5 Response to Topic B:

- One method, used by econometricians and psychometricians, creates multivariate models by using latent variables. From a hierarchical model viewpoint, this is equivalent to adding another layer to the hierarchy.
- Markov-random-field models were proposed as a promising method to modeling multivariate data. The difficulty in taking this approach is to make sure that a joint distribution exists, which can be nontrivial in most circumstances. But, in principle, joint distributions that have marginals with vastly different structures are feasible.
- Principal components can be applied to reduce dimensionality of a multivariate response. Perhaps univariate modeling can then be applied to each component, or at least the multivariate analysis can be simplified. This technique can be especially useful in exploratory analysis where one is looking for “signal” and interpretability is less important.
- Design is important. For example, estimating correlations at both small and large distances requires training data at both types of locations. What are good designs for estimating conditional distributions? What are good models for improving assessment of model fit? Where do we sample today in order to best forecast tomorrow’s weather? These appear to be open questions to the discussion group.
- One additional design consideration is to try to have a design that allows sampling of the multiple processes at the same locations *and* at different locations.

Topic C

In the context of environmental risk assessment, one is often faced with ‘less than perfect’ data sets from different sources. Invariably, one will encounter problems with missing data and measurement error. Sometimes, important covariates

might be missing all together. What kinds of statistical models are useful in such settings? Hierarchical models seem like a natural candidate. What pitfalls should we be aware of in attempting to fit such models?

Group 3 Response to Topic C:

Statisticians working on environmental applications invariably encounter *less than perfect data*, for example, missing values, measurement error, poorly defined outcomes. A good way to approach the analysis of such data is to sit down with the subject-matter scientists and talk about all the problems with the data, as well as what ideal analysis the subject-matter scientist would like to do if she actually had perfect data. Such discussions will enhance communication between the subject-matter and statistical scientists and help the statistician to learn more about the science behind the problem. Often, a conversation about the “ideal data setting” will lead to specification of a model that links data and process in a hierarchical model, similar to the settings talked about in Sunday’s Short Course on Bayesian Hierarchical Statistics. In a time series analysis, for example, the ideal might be daily observations, while the reality might be only weekday measurements or even sporadic measurements. A useful construction might involve a hierarchical model that treats observed data as independent observations around an unobserved true daily process that in turn follows an autoregressive or even more complicated model.

As discussed further in our response to Topic K, classical statistical training does not prepare us well to handle the practical challenges of *less than perfect data*. Many statisticians are lost in messy real world settings, because we are trained in the solution of well defined problems that have a clear right or wrong answers. Our profession is at an important crossroad, since if we cannot adapt ourselves to cope with large, complex real world problems, other disciplines, for example Computer Science, will step in with more practical solutions.

Having stressed the importance of considering data imperfections as part of the model-building strategy, it is important to keep in mind that some studies will simply be too poorly designed or conducted to warrant any sophisticated statistical analysis. Caution is needed when building complex models to make sure that the answers are driven appropriately by the data and expert opinion rather than unjustified modeling assumptions.

Perhaps one of the most common and specific examples of *less than perfect data* is the problem of *missing covariates*. Sometimes, an important covariate may

be completely unavailable. An appropriate collaboration between statistical and subject-matter scientists might lead to a theoretical model (e.g. path or structural equations model) which takes account of this missing factor, though the results are likely to be sensitive to modeling assumptions in such settings. A more common, and perhaps easier problem, corresponds to the setting where some covariates are missing for a subset of subjects. Such missingness is almost inevitable, no matter how well run a study might be. Discussion within our group revealed that few of us have ever applied in practical settings any of the myriad of available methods to handle missing data covariates. Hierarchical methods are in fact very well suited to this problem, though of course such methods require careful application to ensure that inappropriate assumptions are not made regarding the mechanism of missingness. That being said, even simplistic “ad hoc” missing covariate methods (e.g. complete case analysis, imputation of mean values, inclusion of missingness indicators) make implicit strong and sometimes inappropriate assumptions! When we talked about why more appropriate methods are not used in practice, several group members pointed to the lack of reliable, easy to use software in widely used packages as SAS. Another group member commented that the formulation and application of such models is time consuming and usually such time is not budgeted properly into projects that support environmental statistics work. Our profession needs to do a better job of convincing subject-matter scientists of the potential importance of such issues. A good approach might be a two-stage approach of giving a quick answer, followed by a more thoughtful one at a later time. There is an important connection between missing data and measurement error problems. This is another area where we need to do a better job in convincing our subject matter colleagues of the value of waiting for a more sophisticated, albeit time-consuming analysis.

Action Items:

- Hierarchical models provide an excellent framework for building models that relate observable to “ideal” data. The profession needs to find ways to support the development and subsequent support of (interactive) software.

Group 5 Response to Topic C:

- It is important to find out as much as possible about reasons *why* data is missing. More generally, understanding the mechanisms used to pre-process data is important not only for assessing missingness mechanisms but censoring, truncation, and other data issues required for modeling as well. Standard methods of imputation, such as the missing at random as-

- sumption, may or may not be appropriate, and it may be possible to determine this by discussions with appropriate subject matter experts.
- Use multiple imputation to produce multiple posteriors to show the amount of “information” in the observed data about parameters of interest.
 - Examine hyperparameters marginally and jointly to assess the effect of the missing data. Highly correlated joint posteriors can be indicative of a lack of information in the observed data.

Topic D

How should one approach model selection in the context of environmental risk assessment? In the classic toxicologically based risk-assessment setting, current wisdom suggests basing regulatory decisions on a so-called “benchmark dose” analysis, which involves finding the dose level that corresponds to a moderate (5% or 10%) increase above background rates. Because the benchmark dose is usually within the experiment dose range, it tends to be fairly robust to model-choice issues. However, in the context of epidemiologically based risk assessment, issues such as exposure measurement error and the general uncertainty inherent in human data mean that model choice can be a critical issue, even for selecting the 5% or 10% excess risk. A particular problem for the epidemiological setting is that unexposed event rates are often poorly defined. How should one approach model selection in such settings? Does Bayesian model averaging provide a good approach?

Group 1 Response to Topic D:

The fundamental issue in this case is that estimation in data poor regions is always risky, and model averaging can not be expected to remedy this entirely. However, a model averaging approach does allow quantification of model uncertainty, which is essential for this problem. One of the primary advantages may be a more appropriate variance estimate.

One of the primary concerns is appropriate choice of the model space and the prior probabilities for each model. It was felt that there is need for subjective elicitation of prior model probabilities based on expert knowledge.

Two suggestions were made as to how this averaging might be carried out:

- Fit models separately and assign model probabilities based on some fit criteria such as BIC. This approach allows for inclusion of both Bayesian

and non-Bayesian models in the averaging process.

- The reversible jump MCMC technique allows simultaneous fitting and averaging of all models.

The question was also raised whether this problem might be partially circumvented by improvements in study design.

Group 3 Response to Topic D:

Classical training teaches us to specify “the model” and do our inference from there. We need to enlarge our modeling paradigm to include model uncertainty. This is not simply a matter of building larger, nested models, but perhaps expanding the model space to allow for completely different approaches. While there have been good theoretical developments in this area, practical guidelines are badly needed. Reversible jump methods are difficult to program and understand. More straightforward approaches (e.g., Carlin and Chib) can have problematic convergence.

Action Items:

- Further applied and theoretical research on model averaging needed.

Topic E

The construction of realistically complex models for space-time data seems to remain a mixture of art and science. There are a number of useful principles in classical statistics (e.g., invariance, identifiability, parsimony), but in Bayesian hierarchical modelling there seems to be less attention paid to them. Can we nevertheless agree on a list of basic requirements and principles that any such Bayesian hierarchical model should fulfill?

Group 1 Response to Topic E:

- “If enough data were available, could one assign a reasonable interpretation to the hyperpriors?” It was proposed that in many problems the answer to this question should be “yes”. That is, a vague hyperprior should represent lack of knowledge regarding a parameter as opposed to a lack of interpretability of that parameter.
- Hierarchical Bayes models are often not invariant with respect to certain types of re-parameterizations. For example, identifiability may require the constraint $\sum \theta_i = 0$, in which case it may be necessary to select a reference

group. In this case, the posterior will not be invariant with respect to different choices of the reference group. This and similar problems can often be avoided by more careful specification of the prior.

- With respect to identifiability, an obvious basic requirement is that the final results be interpretable. However, identifiability at intermediate stages of the model may be less relevant if the primary goal is prediction.
- A general concern is that optimality criteria reduce flexibility in the modeling process, flexibility being one of the primary reasons for adopting a hierarchical approach. “Constrained flexibility” or “pragmatic flexibility” was proposed as a general ideal. That is, models must be constrained enough that there will be reasonable consensus on the results, but otherwise the modeling process should be as flexible as possible. Such consensus might be achieved through model averaging (or by just looking at many models), however the computational expense may be considerable.

Group 2 Response to Topic E:

It is difficult to prescribe *general* principles for Bayesian hierarchical models. Indeed some of the principles used in non-Bayesian models may not be compelling. For example, identifiability becomes an issue only when we want to interpret the parameters, but not for prediction, as it may get integrated out in other situations.

It will be very useful to have formal tools for the comparison of hierarchical models that keep the bias-variance trade-off in mind, similar to what we have in the GLM setup. It was suggested that perhaps we could use *predictive densities* or *Bayes factors* or the *parametric bootstrap*. But there was not a lot of experience using these methods. Indeed, one participant found that *predictive densities* can be very difficult to compute for complicated models that require MCMC.

Topic F

Many of the models that have been developed for environmental data benefit greatly from incorporation of geographical and/or ecological background information. For example, the USGS has developed models of pesticide concentration in drinking water that incorporate water-flow patterns. Much is known about volatility and solubility of chemicals and patterns of wind flow. What are appropriate ways to incorporate such scientific and engineering knowledge into statis-

tical models (or vice versa) in order to produce models that provide the best of both worlds?

Group 2 Response to Topic F:

For building science into the model Mark Berliner uses what he calls a *Bayesian spectrum*. On one end of this Bayesian spectrum is a purely mechanistic model (i.e., pure science) and the other end of it is a purely statistical (Bayesian) model that does not utilize any knowledge of the underlying physical process. At both of these extremes, uncertainty exists due to parameters being unknown. For example, a differential equation may have coefficients that are unknown, and hence can become parameters of a statistical model. In addition, data might be the inputs and/or outputs of mechanistic models perturbed by random noise. Interior to the spectrum, there are examples like the model used by Chris Wikle in his presentation. In this model, simplified equations for wind motion are modified by replacing sinusoidal functions by unknown functions that then have a prior distribution with prior means equal to the sinusoids. Of course, non-Bayesian methods could also be applied at points along this same spectrum.

Perhaps there is one more co-ordinate to this Bayesian spectrum - “how hard the problem is”. We need to be able to balance the desire for parsimony with the need for accurate modeling. The degree to which we make our models more complicated (realistic?) should depend on how much effort is involved and how important it is to improve the model.

Group 4 Response to Topic F:

Some initial concerns and comments:

- We are agreed that to incorporate expert opinion, we should use Bayes’ Theorem in some way.
- The problem of aggregating data from various scales is an extremely important case where expert opinion is needed. A related problem concerns change of support.
- With data on different scales, different levels of aggregation can change the correlation matrix (among other things). Moreover, the inferences from different aggregation levels may not be consistent (Simpson’s paradox).
- Can problems due to aggregation be diagnosed similarly to those arising in the usual hierarchical model (using the ideas presented in Topic A). Can we think of aggregation as another level of hierarchy?

- Not only is aggregation an issue, but one must also be aware of the statistical issues relating to the support of the data and the support of the inference desired by the subject-matter scientist. The subject-matter scientist may desire knowledge of continuous processes while the data are supported at different, possibly discrete, scales. The subject-matter scientist must be made aware of the influence of the area of support.
- It should be noted that the same scale issues apply to temporal/time series data.
- How does one do enough runs of large models to accurately assess error in physical models? (Some physical models can take days or more to run on a computer, for instance atmospheric models.) Perhaps a Latin hypercube sampling of parameter techniques can partially address this is, but it is unclear what the complete solution may be.

Some suggested solutions:

- Clearly it is important that part of the goal be to appropriately embed the physical models into the stochastic systems.
- It is of utmost importance to specify the goals of your analysis. For example, in the Scotland influenza data, aggregation will be at different levels if the goal is a public policy decision as opposed to a decision about physician preferences. But also, the data are very sparse - so aggregation is important.
- Expert opinion is important in designing the scale of aggregation. However, it is important to ask the expert specific questions: “What goal do you have?” “What level of aggregation is important to you?” Experts can have difficulty expressing the scale of interest. (Ecologists are on the forefront of understanding scale.)
- It may be the case that the data are aggregated at one level, and the results are aggregated at a different level. The issues bearing on the level of aggregation may be different.
- Aggregation problems result in models that are not nested (data are often also misaligned). This causes additional problems.
- Issues beyond aggregation?
 - o An expert hands you a scientific equation and says, “fit this model”.

- Work with the expert to add error in an appropriate way. Don't just tack an epsilon to a deterministic model.
- Train students to ask the right questions to be able to do this themselves.
- Get the expert to tell you the story and build the variance model and the hierarchical model appropriately.

Topic G

Almost any reasonably complex model is a mixture of empirical and subjective components. When evaluating such models, we might consider both frequency and Bayesian measures. However, it may be the case that the empirical and subjective pieces are so intertwined as not to allow separate evaluations (for example, in empirical Bayes models the prior parameters are estimated from the data). Can we agree on an evaluation strategy for complex models that would result in an inference (no doubt a mixture of Bayes and frequentist) that would be acceptable? Such a scheme should include

- *an assessment of long-run properties,*
- *an assessment of Bayesian robustness,*
- *a sensitivity analysis.*

Group 3 Response to Topic G:

Evaluating a complex model with empirical and subjective components is very challenging. Although properties like Bayesian robustness and good long-run frequentist properties may sometimes make sense, our group felt that the usefulness of such approaches is highly context dependent. For example, what do good frequentist properties tell us about the usefulness of a model when we can essentially “cook” our priors so that our models fit the data very well. Having good frequentist properties can make sense, however, in contexts (e.g. climate forecasting) where this natural replication built into the system. Such properties will make less sense, however, in the context of a study such as one focused on environmental epidemiology, which is unlikely to be ever repeated again. Similarly, it might be important, in some contexts, to have Bayesian robustness (i.e. a lack of sensitivity to prior specification). Sometimes, however, it will be inevitable for the prior to have a strong influence on the results. Hence, the important issue there is not to simply ask the question of whether the prior matters, but to characterize

and quantify the role of the prior in affecting scientific conclusions based on the model.

The group spent a considerable amount of time talking about ways to accomplish effective sensitivity analyses in the context of a hierarchical model. Often, this is a challenge since informative prior informative can be imbedded deeply in the hierarchy and also take on multiple dimensions. It would be useful to develop strategies for effectively identifying the sensitive nodes in a hierarchical model (that is, identify the critical points in the hierarchy that have the potential to affect results). It would be really useful to come up with effective ways to graphically represent how model conclusions are affected by prior specifications. Some work along these lines has been done, for example, in the frequentist setting where non- or barely identifiable models have been suggested to account for informative dropout in a longitudinal study. If we are willing to specify a parameter that characterizes the relationship between the dropout probability and an unobserved future observations, then it is possible to then estimate an exposure effect, adjusting for informative dropout. Because the true dropout mechanism can never be known exactly, then it might be appropriate to construct a plot showing the estimated exposure effect (and perhaps associated confidence limits) as a function of the unknown parameter characterizing the degree of informative dropout. Hierarchical modeling would seem to be a natural framework within which to generalize such ideas. It would be really helpful to develop techniques that help to visualize the effect of prior specification on model conclusions. There is apparently some software that allows one to interactively assess sensitivity to prior distributions by using a slide bar to alter the priors and then re-weight the posteriors to reflect changes in the posteriors.

If we are not careful, our profession is at risk of being left behind in the current internet and technology revolution that that is fundamentally changing the way that people store, access and handle data. To keep up, we need to find better ways to synthesize and represent high-dimensional data and complex models. As technology becomes more “visual”, it is important that we develop innovative new ways to effectively communicate complicated models and results, and also to quantify and present uncertainty.

A caveat to all this discussion is that in many settings, a simple model will do as well as a more complex one. The challenge is knowing when this is true! Often, statistician cannot be fully confident that the simple answer is adequate until a more complex analysis has been pursued as well.

Action Items:

- Develop visualization and communication tools to allow statistical and subject-matter scientists to better collaborate on model construction.
- Develop theory and applied tools for sensitivity analysis, for example, ways to “quickly” recompute posteriors under changed model assumptions.
- Develop tools to visualize model fit.
- Work needed on “flexible” or even non-parametric hierarchical modeling techniques. Some good ideas from the generalized additive model framework (e.g., splines) would be useful to develop in the hierarchical modeling context.

Group 4 Response to Topic G:

Some initial concerns and comments:

- The key question here seems to be: “What type of inference do we ask for and when are we satisfied with our assessment?”
- There are many sources of variation in hierarchical models. In addition, we are mixing Bayesian and frequentist approaches. When evaluating hierarchical techniques under these conditions, it is important to define criteria that “adjust” to the complexity and mixture of the inference approaches.

Some suggested solutions:

- The answer to this set of questions again seems to be dependent on the setting. Certain stakeholders decisions or policy making applications require all of the criterion, while other applications may be less stringent.
- It may be true that traditional inference does not work well. The traditional P -value may be inappropriate. As a result, we may need a new paradigm.
- Is the “model averaging” approach an appropriate one? While model averaging is useful for designing models, what we are fundamentally looking for here is a way to evaluate models. Again, focusing on the goals of the study may help to answer these questions. Model averaging may be good for prediction, for instance, but less useful for understanding particular applications.
- We tend to use a lot of pictures to judge our models. It is important to be clear in our statements of where these pictures and graphs come from.

It is further important to quantify how to interpret these images and to determine how useful or how rigorous these techniques may be.

- As statisticians, we all condition, at some point, in our modeling. We need to be aware of where that conditioning is being done in the model and how to communicate this conditioning.
- In attempting to understand the complexities of evaluating hierarchical models, it may be time to be statisticians again, and no longer to debate whether we are Bayesians or frequentists.
- In the spatial context, different types of asymptotics are available, infill and increasing-domain asymptotics. These are not equally appropriate. The study context and goals of the analysis determine the pertinent framework for asymptotic inference. Increasing domain asymptotics may not save hierarchical models when the ways hyperparameters were incorporated are considered. However there are some possibilities using infill asymptotics for inference of hierarchical models.

Topic H

Hierarchical models often involve modeling “local” structure in order to analyze spatially extensive data. The level of localness is often chosen for practical reasons having to do with the granularity of the data, rather than for reasons having to do with the natural scale of the phenomenon. This leads to two questions:

- *Does the hierarchical statistical model aggregate consistently, at least approximately?*
- *Is the hierarchical statistical model flexible enough to allow inference at a disaggregated level?*

Group 1 Response to Topic H:

The issue of aggregation consistency is of particular importance when data arise at different spatial resolutions. In contrast, there are problems for which inference which is conditional on a given scale of resolution is readily interpretable and seems sufficient. This appears to be the case in time-series models, where there is surprisingly little literature on this subject. This area may benefit from future development of aggregation consistent models in spatial statistics. However, the price paid for aggregation consistent formulations may currently be too high. More work needs to be done to properly link continuous spatial point process

models with formulations for aggregated data. This seems particularly important for problems which require a global view of local time scale behaviour, for example in remote sensing.

Group 5 Response to Topic H:

- Model consistency is desirable in an ideal analysis. In some situations, where physical conservation laws apply, the scientific impact of the model may *depend* on having this property. This would be the case, in small-area estimation or geophysical applications, for example. In some ecological studies having this property is less critical.
- Some widely used models do not aggregate/disaggregate consistently (e.g., Poisson-lognormal).
- One explanation for models that do not aggregate properly is the presence of confounders at the fine level of aggregation.
- The inference at disaggregated levels from a model is an open research area.

Topic I

When making a spatial map of some summary of the full posterior (e.g., point-wise spatial medians), what meaningful measures of variability can we attach to the map? Alternatively, how can we summarize 1,000 maps drawn from the full posterior distribution?

Group 1 Response to Topic I:

As a simple solution, one can map marginal posterior probabilities of interest. However, single maps are often inadequate because they fail to represent correlations of the response at different points in space (or more crucially, in space-time). A proposal was to generate “movies” animated with respect to different values of the parameters from the posterior distribution. The process of creating and presenting (for example in a journal context) must be made easier. In particular, possibilities with web-based journals must be explored. Some software already exists for related purposes, for example GIS in combination with Xgobi.

Group 2 Response to Topic I:

Liberal use of color and shadings can be used in drawing maps. Additional symbols can be added to maps to display more information. One suggestion is

to overlay a symbol onto various parts of the map to denote a measure of spread associated with the value of the map at that point. For example, one could place a dot whose area is proportional to the prediction standard deviation or the posterior standard deviation of a parameter for that point on the map. This could become cluttered if there are lots of dots per unit area on the map. Another possibility is to consider a function of the map, such as some critical rate. Then simulate the map many times, each time computing the same function. Then show the variation in the values of that function. The function can be more complicated than a single number, but plotting the variation in more complicated functions can become messy.

Of course, information is always lost when drawing a single map. Covariation between locations is particularly difficult to display on a single map. An alternative is to provide an animation in which values of the plotted quantities vary together according to the model. If such an animated map were to be made available on the web, one could even allow the user to specify (perhaps by clicking a mouse) which values should change and then he/she could watch the effect such a change has on the entire map. Of course it would take some training and practice to be able to learn effectively from such a process, but we think that it has the potential of being highly informative. Even if journals are unwilling to accept animations as part of a publication (is Java JASA in our future?), an agency like the EPA might find such animations useful.

The question also arose as to how to compare maps. Various one-dimensional summaries have been used, but were generally considered to be inadequate because they do not take proximity in the map into account (e.g. grid box-by-grid box sum of squared deviations, measures of overlap of the submap sections). Procrustes analysis might be extended to a method for forcing one map to look like another, keeping track of the amount of distortion required to make the change. In the final analysis, it might be that we have to rely upon judgment to compare maps.

Topic J

Expert opinion can be crucial to hierarchical statistical modeling, particularly as one goes deeper into the hierarchy where process-behavior is less well known. Different experts will likely differ in their opinions. Can we formalize a way to combine expert opinion?

Group 2 Response to Topic J:

Expert opinion arises both a priori (in the formulation of priors and models) and a posteriori (in the interpretation of results). Prior elicitation can be difficult in hierarchical models because the parameters whose distributions are being elicited are so far removed from the observables, that is, so little is known about them. It is more difficult to do “what if” analysis. After an analysis has been performed, people have been known to vote on which interpretation to give to the results. If experts offer opinions that suggest that different models be fit to the data, an approach using Bayesian model averaging might be called for. On the other hand, if the competing models make drastically different predictions, it might be better to report the results of several models and recommend either additional study or a search for more appropriate models.

Group 3 Response to Topic J:

Our group explored this question in two different ways. Someone suggested the thought experiment of six different statisticians being assigned to work on the same applied problem. It is natural to expect that six different solutions will be presented. While in many cases the variations will be relatively minor, sometimes major differences will result. What is the “right” answer in such settings? Assuming that none of the approaches are blatantly wrong, one could think of synthesizing the analysis in some way. This is one way to think about combining expert opinion.

A broader, and in many ways more difficult, aspect of this question is how to combine fundamentally different types of information. For example, statisticians may be able to suggest a broader range of suitable models, toxicologists may suggest other models based on biological considerations, toxicologists may also be willing to purport prior distributions on the values of specific model parameters that cannot be easily estimated from the data. More generally, the group may be willing to hypothesize a range of unobserved mechanisms that might be explaining the truth behind the setting of interest. Expert panels building consensus are a fascinating example of how disparate sources of information can be synthesized. It would be nice if statistical modeling frameworks could be developed to do a similar sort of synthesis.

The group felt that there is fertile ground for statistical research on strategies for building, fitting and evaluating models that combine data and theory. Hydrology, physiologically based pharmacokinetic models, and air-pollution-dispersion

models were suggested as good examples of modeling settings that attempt to combine a theoretical model and some empirical. Although very different on the surface, these areas have in common a general approach of hypothesizing a theoretical model, then “tuning” it using some real data that should be well fit by the model. Many field suffer from so-called “physics envy” wherein they want to believe that their theoretical models are rich enough to explain all observable variability in the data. In practice, models usually need to be “hybrid” and also bring in a stochastic component. A Bayesian framework is a natural one for such problems. Hierarchical models can be thought of as “bookkeeping for uncertainty”. Hierarchical models provide a great framework for a conversation about uncertainty.

Group 5 Response to Topic J:

- In some areas, for example medicine, expert opinion can be skewed. Physicians can be unduly optimistic or pessimistic about treatments.
- Eliciting expert opinions about joint distributions is typically difficult. Elicitation of prior information is most successful when model parameters have clear physical interpretations.
- Conditional modeling typically leads to easier expert solicitation.
- Can we formalize a way to combine expert opinion? In general, NO!
- When there is a range of expert opinion, it is better to determine the range of posterior solutions then to attempt to combine the expert opinion into a single prior.
- Panel studies may or may not be a good thing???

Topic K

Good hierarchical statistical modeling to solve environmental problems requires teamwork between statistical scientists and substantive scientists. However, statisticians are often put in a consulting role where quick answers are sought. Important questions may require a longer, more collaborative effort, and hierarchical statistical models demand that effort. How can we convince scientists that a hierarchical statistical analysis is “worth waiting for”?

Group 2 Response to Topic K:

It is often the case that subject matter experts are either not familiar with or not comfortable with statistical models. We have, here, another case of the *Bayesian spectrum*. Dialog between statisticians and researchers can help to bring the extremes of the spectrum together. Of even greater potential is the relating of success stories, that is, examples in which the added effort of statistical modeling have provided additional insight that were not available with less sophisticated analyses. In order to be convincing, the success stories should be based on the researcher's own data.

Group 3 Response to Topic K:

This very broad question is of critical importance for the field of environmental statistics.

Communication is critically important. Statisticians tend to separate themselves from subject-matter scientists. We need to take a more aggressive role in explaining our methods, techniques and results. We need to develop better ways of presentation that clearly establish the value of our role. We need to be an integral part of the research team from study conceptualization to analysis and reporting. We need to be able to clearly explain why complicated models are needed (if they are!) and help our colleagues to interpret the results.

What steps can we take to make these things happen?

- Educational paradigms need to evolve to encompass broader training about uncertainty, role of policy and practical complexities in design. We are behind in statistical methods for handling large data sets.
- Priority should be to develop high-tech methodologies for portraying results of complex modeling.
- Become involved in the subject matter journals. Perhaps work with jour-

nals from other societies to get representation on the editorial boards. Should get the societies involved.

- Involve subject-matter scientists in our statistical publications.
- Good user-friendly software can bridge the gap between theoretician and applied statistician.
- Balancing theory and application - changing the reward systems.

Challenges:

- User population is changing quite rapidly and becoming quite sophisticated.
- Tukey - “the role of a good statistician is essentially ephemeral”. After a really good statistician has done their work, the results are often so simple and obvious that it is not clear that you needed anything sophisticated in the first place!

Group 4 Response to Topic K:

Reformulated Question:

How can we determine whether development of a complex hierarchical model will provide enough additional information to be worth the effort? If answers from simple and complex analyses differ, how do we assess which is correct?

Some initial concerns and comments:

- How do we know how far you have to go? What is the cost/benefit of pursuing a more complex analysis? How do we assess this?
- How good are your data? Our contribution comes in the gray area between great data and useless data. At either extreme, $\bar{X} \pm 2S$ is all you can do.
- Is hierarchical modeling completely new/essentially different than what we as statisticians have done historically?
- There is a feeling that additional computational power, and availability of information, may be driving more complex models. Are we doing more complex models just because we can?
- Does this relate back to Topic A? By adding complexity, we are adding hierarchical levels.
- Grant writing is also related - and leads to more complex models. You

have to identify your model type and the problems related to that model type before you do the work.

- Time marches on. Five to ten years ago the more complex models and analyses were not taught in classes, certainly not in first year classes, nor were they available in software packages. Now they are available to statisticians, leading users to more complex analyses as a matter of course.
- We are agreed that there is a need to familiarize subject-area scientists with hierarchical models.

Some suggested solutions/directions:

- Research into model uncertainty is important. We need to define the criterion that help us decide when it is important to use complex models. In particular, what are the signs/tools that tell us that it is important to use hierarchical models or to stay away from hierarchical models?
- Modeling is an interactive art, and should be realized as such. Where does the line between the experimenter and the statistician get drawn? Should statisticians be doing the modeling?
- The decision between simple and complex models is often non-statistical. In a walk-in consulting center that does not do the statistical analyses for their clients, a large percent of problems will be solved with simple methods. Thus, the decision is based on whether the stakeholder can do, or pay for, a more complex analysis, and not if the data demand such.
- Computing packages are key to this issue, for example, a “Naive vs. Robust” setting will lead to a more complex analysis. But this is generally good, and such software needs to be more widely developed.
- It is the “Progress of Science” that questions are getting more complex, but we must remember that “Some problems are simple and should not be done in a hard way.”