

Structuring expert input for a knowledge-based approach to watershed condition assessment for the Northwest Forest Plan, USA

Sean N. Gordon · Kirsten Gallo

Received: 4 May 2009 / Accepted: 11 February 2010 / Published online: 20 March 2010
© US Government 2010

Abstract Assessments of watershed condition for aquatic and riparian species often have to rely on expert opinion because of the complexity of establishing statistical relationships among the many factors involved. Such expert-based assessments can be difficult to document and apply consistently over time and space. We describe and reflect on the process of developing a computer-based decision support application from expert judgments for assessing aquatic and riparian conditions over the 100,000 km² managed by the US federal government under the Northwest Forest Plan. The decision support system helped structure and document the assessment process

and provided consistency and transparency to the evaluation methodology. However, many decisions and trade-offs were required in the expert engagement and model-building processes. Knowledge elicitation in an interactive group had a number of benefits over nominal group or Delphi processes, but efficient knowledge capture required considerable planning and expertise in the subject matter and modeling process. Communicating model results for validation was problematic and only effectively accomplished via in-person workshops. The choice to use different expert groups for each biophysical province provided more opportunities for participation and promoted greater ownership in the assessment, but it also led to increased variation among the resulting model structures. We propose three possible approaches for better managing the consistency of assessment models when multiple expert groups are involved.

S. N. Gordon (✉)
Pacific Northwest Research Station,
USDA Forest Service, 333 SW First Avenue,
Portland, OR 97205, USA
e-mail: sgordon@fs.fed.us

K. Gallo
Resource Planning and Monitoring,
Pacific Northwest Region, USDA Forest Service,
Corvallis, OR, USA

Present Address:

K. Gallo
Chihuahuan Desert Inventory and Monitoring
Network, National Park Service, New Mexico State
University, MSC PSL/Anderson Hall,
Las Cruces, NM 88003, USA
e-mail: kirsten_gallo@nps.gov

Keywords Watershed assessment · Composite indicators · Expert judgment · Fuzzy logic · Modeling · Stream condition assessment · Stream habitat

Introduction

One of the earliest and largest ecosystem management applications in the US is the Northwest

Forest Plan (NWFP), a coordinated set of guidelines for federal lands in the continental northwestern USA (USDA and USDI 1944). Although the NWFP area was largely defined by the territories of existing northern spotted owls (*Strix occidentalis caurina*) and other high-quality late-successional and old forest owl habitats in the surrounding area, it was also an attempt to simultaneously address the needs of many species and ecosystem values. Consequently, one piece of the NWFP is an aquatic conservation strategy.

This aquatic conservation strategy and the associated Aquatic and Riparian Effectiveness Monitoring Plan (hereafter, the *Monitoring Plan*; Reeves et al. 2004) pay special attention to anadromous and cold water fish habitats because numerous fish species and populations have been listed as threatened or endangered in the NWFP area. However, evaluating aquatic conditions using statistically correlated relations between habitat indicators and fish populations has proven to be challenging. A wide variety of factors (e.g., large wood, pools, and riparian and upslope vegetation) at multiple scales contribute to habitat, but the data needed to correlate these indicators with fish populations are generally unavailable over large landscapes. Furthermore, the species of concern in the Northwest are mostly anadromous, introducing ocean habitat conditions and migration barriers as confounding variables. As a result, few studies have attempted to address these challenges at larger scales (see Kaufmann and Hughes 2006; Pess et al. 2002). Nevertheless, managers need ways to make use of the best available science and field experience for making management decisions.

Expert panels have frequently been used to bring science to bear on policy and management questions, especially when data are lacking (Johnson et al. 1999; Rauscher 2000). Through the 1990s, assessments of watershed condition in the Northwest relied on expert workshops without a formalized model for assessment (personal communication with Gordon Reeves, Research Fish Biologist, US Department of Agriculture (USDA) Forest Service, Pacific Northwest Region Research Station, Corvallis, OR). While these workshops enabled the integration of diverse knowledge into explicit assessments, it was

difficult to document, communicate, replicate, or transfer their specific evaluation methodologies because much of the rationale for the watershed ratings remained only in the heads of the experts. To address these weaknesses, an assessment of management options for federal lands in the interior Columbia River basin supplemented an initial expert group approach (Sedell et al. 1997) with computer models which encapsulated expert knowledge and assessed each watershed unit in a consistent manner (Rieman et al. 2001). Similarly, the NWFP Aquatic and Riparian Effectiveness Monitoring Plan (Reeves et al.

2004) proposed the use of a knowledge-based systems approach, that is, the encapsulation of expert knowledge in an explicit computer model, which would facilitate and document the consistent and transparent application of a methodology for evaluation.

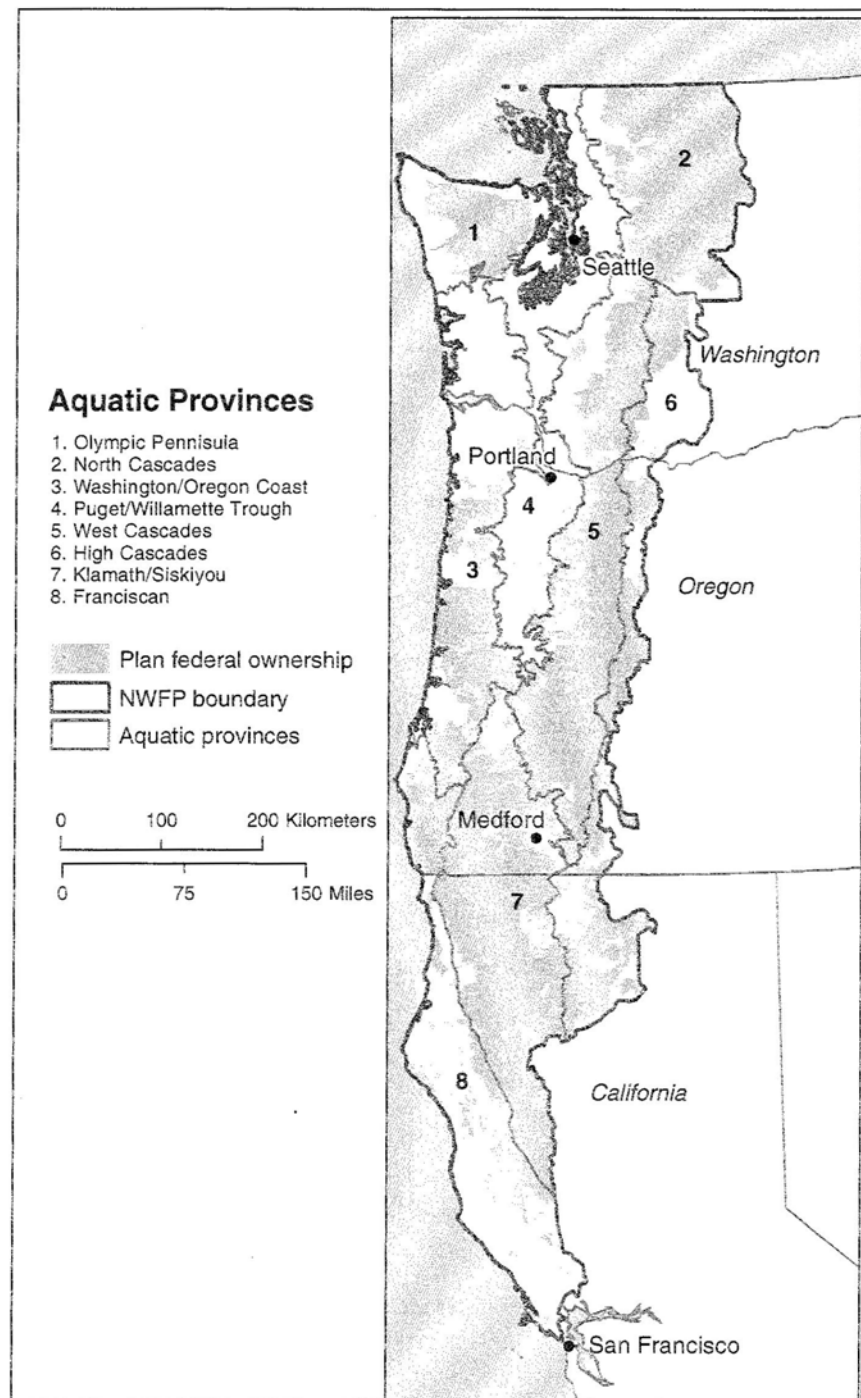
This paper describes the development of a knowledge-based system to evaluate watershed condition in the context of the Northwest Forest Plan. How an assessment is framed and executed can be as important as the data and analytical techniques it uses (Johnson et al. 1999; Mitchell et al. 2006), so here we focus on *the processes* for engaging experts and structuring of our assessment model. Details on our analytical methods and results have been previously published (Gallo et al. 2005; Reeves et al. 2006).

Background

Study area and sampling design

The Northwest Forest Plan applies to approximately 100,000 km² of federal lands in Washington and Oregon, from the eastern slope of the Cascade mountain range to the Pacific Ocean, and an angular slice of northwestern California from the north-central Cascades to the coast just north of San Francisco. Recognizing the inherent ecological variability in an area of this size, eight aquatic provinces were defined by the monitoring team based on the 12 terrestrial physiographic provinces used by the Northwest Forest Plan (Fig. 1)

Fig. 1 Northwest forest plan aquatic province map



Given the extensive area of the plan and the need to measure site-specific habitat attributes (e.g., wood, pools, riparian vegetation), the Mon-

itoring Plan's study design includes random selection of 250 sixth-field hydrologic units (hereafter, subwatersheds), which are typically 42–133 km² in

the mathematics of the fuzzy set theory with a variety of mathematical and logical operators. An assessment objective, or primary topic (often stated as a proposition such as "the watershed is in good condition"), can be modeled by breaking it down into subcomponents until measurable indicators are obtained. Models can be nested, as shown in Fig. 2, where the overall numerical result of the reach-level model (reach condition) feeds in as one input to the watershed-scale model.

A NetWeaver model reads in data for each indicator (also referred to as an elementary topic) and compares the observed or estimated value against evaluation criteria reflected in a fuzzy membership function. Whereas a traditional threshold approach would compare the data to a particular threshold (e.g., water temperature <68°F) and give it either a pass or fail, fuzzy membership functions provide for a multivalent result. A fuzzy membership function would enable a manager to evaluate the resemblance between the measured or estimated water temperature and the 68°F threshold. Thus, a water temperature of 69°F would not simply fail an evaluation but show a considerable resemblance to the desired temperature. Figure 3 shows how two threshold values define a simple evaluation function to convert an observed value for water temperature into a standardized score (NetWeaver typically uses a range between +1 and -1, where +1 indicates full support and -1 indicates no support for a proposition). Once the indicator scores have been standardized in this way, they can be aggregated to successively higher levels of the model using logical operators. The two operators we used most often were MINIMUM (MIN, which treats its arguments as limiting factors) and AVERAGE

(AVE, which treats its arguments as incrementally contributing support, i.e., the arguments are partially compensatory). The different branches of a model are aggregated toward the proposition until an overall condition score is reached. Two other NetWeaver features were important for our models. First, weights can be applied at any point where two or more scores are combined. Second, logic switches enable changing model behavior, depending on contextual data. For example, some models used the presence/absence of bull trout to change the evaluation thresholds for temperature, since bull trout generally require colder water than other fish considered in our models.

The EMDS system produces results in the form of maps and numerical tables within the ArcGIS software. A master data table records the data input values and evaluated scores (on the standardized +1 to -1 scale) at each node, and a color-coded map is also produced for each evaluated node.

Methods

Following guidance in the Monitoring Plan, the project team built two initial models to use as a starting point for the workshops. The first model evaluated in-stream conditions at the various reaches sampled. Results of the reach evaluations were summarized by subwatershed (the overall numerical scores of the reaches were averaged), and a second model then evaluated this summarized reach information together with other subwatershed-level data (Fig. 2). Threshold values for the indicators were derived by the project team from the literature (Gallo et al. 2002). These initial models were useful as a proof of concept and as a practice run for the monitoring team. However, the Monitoring Plan recognized that a variety of experts would need to be engaged to capture the best available knowledge in a suite of models that could address the biophysical variation over the large NWFP area.

Choice of geographic scale

The process for engaging experts in model development began with a decision about geographic

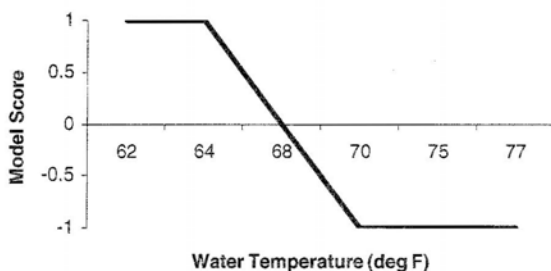


Fig. 3 Fuzzy evaluation function for water temperature

scale. The conceptually simplest scheme would have been to engage a single group of experts to design a single model for the full spatial extent of the NWFP area. However, the NWFP region incorporates areas that are biologically and physically distinct in terms of their geology and geomorphic processes. A group with the breadth of expertise needed to address such differences might be large and difficult to work with, both in logistics and in achieving consensus on model design. Participants in such large group dynamics might also experience considerable *downtime* as areas outside their geographic expertise were analyzed. The resulting model might also prove to be unwieldy, requiring too many logic switches (as described above) to accommodate the physiographically diverse contexts. On the other hand, this choice would be more likely to produce results with the most consistency and comparability across different subregions.

At the opposite extreme, significant physical and geochemical variation can occur between adjacent subwatersheds and even within a single subwatershed. Thus, the argument could be made to organize groups that would have the specific knowledge to evaluate each subwatershed sampled individually. It would be challenging, however, to engage so many small groups (given the sample size of 250 subwatersheds), and the potential resulting diversity of evaluation procedures would have a high potential for inconsistency and a high cost in terms of programming and analysis effort.

Modeling at the level of physiographic provinces was recommended in the Monitoring Plan and ultimately chosen by the project team. Seven expert groups could be engaged in the time available, and province-level models could allow variation for major biological and physiographic differences while not unduly splintering the modeling process. The eighth province, the Willamette/Puget Trough province, is mostly privately owned, and no subwatersheds met the 25% minimum federal land ownership criterion.

Selection of experts

Because the NWFP applies to federal lands, we focused on inviting employees of the federal land

management agencies: USDA Forest Service, Department of the Interior (DOI) Bureau of Land Management (BLM), and DOI National Park Service. The project team identified fish biologists and hydrologists working in the different regions. A few experts from universities and other federal and state agencies were also invited based on team networking and contacts.

Drawing up the list of invitees revealed some overlaps among provinces. Experts are typically employed by National Forests, Parks, BLM districts or research stations, the administrative boundaries of which often straddle physiographic provinces. Given this overlap, we chose to organize four workshops to cover the seven provinces: (1) Oregon/Washington Coast and Olympic Peninsula, (2) Northern Cascades, (3) West and High Cascades, and (4) Klamath, Siskiyou and Franciscan (see Fig. 1).

Knowledge elicitation approach

We reviewed the potential of Delphi and other structured group methods for eliciting expert knowledge (Delbecq et al. 1986). However, we decided that more loosely structured, face-to-face workshops would be the best way to engage experts for the following reasons:

- the new modeling process would require explanation and the ability for interactive question-and-answer sessions,
- discussion and debate among the experts was important for building consensus, and
- physical attendance would capture the time and attention of busy professionals.

Participant preparation

Approximately 1 month before each workshop, we sent participants a prework package to help familiarize them with the modeling objectives and process and to encourage them to gather relevant information. An overview of the process was provided in a word-processing document, graphical representations of the pilot model structures were provided in an electronic presentational file, and a spreadsheet listed tentative indicator threshold values used in the pilot model, as well as a list

of subwatersheds from the province selected for the Aquatic and Riparian Effectiveness Monitoring Program process. We requested that participants suggest alternative threshold values and rate the overall condition of those subwatersheds they were familiar with, for later use as a check on the model results. The experts we engaged were also asked to solicit input from other colleagues in their units. One conference call with a simultaneous computer presentation was held with participants a few weeks before the workshops to review the introductory material and answer any questions.

Model construction

Model specification workshops

The first round of four 3-day workshops was held between April and July of 2003, with approximately one workshop scheduled each month. The workshops were held in the offices of the monitoring team to ensure maximum availability of team members, data, and equipment. The authors facilitated the workshops, with others attending and lending their expertise as needed.

Not all workshop attendees had participated in the pre workshop orientations, so we began each workshop with a presentation of the workshop objectives and modeling basics, followed by questions and discussion. The rest of the workshop was devoted to elaborating the models. The model and evaluation criteria developed by the monitoring team were given to the first workshop participants as a starting point for model refinement, and subsequent workshop participants were provided access to the models developed by previous groups. A summary of the modeling decisions made by each group is presented in Table 1.

Table 1 Model construction decisions available to expert groups

Model structure	Indicators
Indicators to include	Calculation of GIS-based indicators
Hierarchical arrangement	Shape of evaluation functions
Operators/aggregators	Evaluation criteria values
Weights	
Context switches	

The order in which model elements were addressed differed according to the preference of each group, but in general, we tried to settle the overall model structure (which indicators to include) and then addressed how to evaluate each individual indicator. The modeling software was not used in the workshops to design the model structures; instead, models were hand-drawn on a dry erase board to keep the focus more on concepts rather than technology. As project facilitators, we alternated between leading the discussion using the white board and operating a computer linked to a projector to display information such as past model structures or relevant maps and data tables. An additional staff member took notes on the process. The work proceeded as an informal group process during which participants came to a consensus on the model structure and evaluation criteria. To help orient them in the task, participants were asked to imagine themselves standing on a ridge overlooking a watershed and then walking along the stream channel: what information

Table 2 Indicators available for assessing watershed condition

GIS-derived attributes	Field sampling attributes
Roads	Physical habitat
Overall density	Bankfull width: depth
Riparian density	Gradient
Upslope density	Sinuosity
Density by slope class	Entrenchment ratio
Road/stream crossings	Substrate D50
Vegetation/land cover	Percent fines
Vegetation cover (percent)	Wood frequency
Conifer cover	Pool frequency
Broadleaf cover	Pool residual depth
Conifer size class (quadratic mean diameter)	Water chemistry
Urban/agricultural lands	Total Kjeldahl nitrogen
Streams	Total phosphorus
Riparian buffer width (for riparian vegetation or roads analysis)	Dissolved oxygen
	Conductivity
	pH
	Temperature
	Biological sampling
	Periphyton
	Macroinvertebrates
	Amphibians
	Fish

would they use to assess the condition of the watershed?

The selection of indicators was constrained by the available data summarized in Table 2. The only in-stream indicators available were those collected by the monitoring program. Subwatershed-wide indicators were mainly constrained by the availability of data sources extensive enough to cover the whole NWFP area consistently. Workshop participants determined which indicators to include and how they were used in the model. Participants also had flexibility in how some of the indicators were calculated. For example, when looking at riparian vegetation, participants could vary stream buffer width or tree size class. Indicators were organized in a hierarchical structure (as in Fig. 2). Conceptually, this structure facilitated analysis of causal paths affecting watershed and stream condition in the province; for example, grouping road measures on one branch and vegetation measures on another. Functionally, model structure influences the order in which data are read and how scores are calculated. Data at different scales, here, the reach and subwatershed indicators, had to be evaluated in different models. The scores from multiple reaches within one subwatershed were averaged and this summary used as an indicator in the subwatershed-level model.

Model structure also influences how data are evaluated by logic operators and how ultimately the scores that are rendered under primary topics. When two or more indicators contribute to a topic in the model, for example, total nitrogen and phosphorus in Fig. 2, one must decide how to combine them. The NefWeaver software provides a number of logic operators for this purpose. To keep the process simple, we generally limited these to logical analogs for minimum, maximum, and average. That is, given the scores for nitrogen and phosphorus, the model could return the minimum of the two (a limiting factor approach), the maximum (one fully compensates for the other), or the average of the two scores (partially compensatory). Furthermore, for situations in which an average was used, each indicator could be weighted relative to the other, so, for example, the contribution of nitrogen could be weighted twice as much as that of phosphorus. Finally, the

structure could include logical switches to modify model behavior according to additional contextual information. Such switches use an indicator not for direct evaluation but to change how the model operates. For example, two provinces used stream gradient to modify the pool frequency (bankfull widths per pool) evaluation criteria, since the experts expected a naturally higher frequency in lower gradient stream reaches.

Workshop follow-up

After each workshop, the monitoring team consolidated notes and prepared a summary document. Each document began with lists of follow-up tasks (e.g., evaluation curves still to be worked out, data to obtain), and a summary of concerns raised by participants. Most of the subsequent documentation then addressed each major indicator group in detail. For example, the Roads section displayed a graphic of the model structure (including operators and weights), followed by general comments, then conclusions on the evaluation criteria for each indicator and sources for the criteria (whether from expert judgment or a previously published study). This summary was circulated to all participants by e-mail. Comments and corrections from the participants were incorporated into a final version and recirculated. The modeling team then built the provincial models for subwatersheds and reaches in the EMDS software using the templates laid out in the summary document.

Model verification and validation

Oreskes et al. differentiated between model verification and model validation. Verification tests that the model performs as specified computationally (i.e., no errors in the formulas or computer code), while validation has to do with correctly representing the environmental phenomena being modeled. We performed verification by creating three test datasets. One dataset contained values, which should produce a poor rating (-1) for each individual indicator evaluation; one set was designed to produce good ratings (+1) and the third set to produce

intermediate values (0). NetWeaver includes an interactive, color-coded graphical display of model components, which we used to visually check that each evaluation node was producing the expected results.

Model validation as understood in the natural sciences means testing to see if a model produces empirically accurate results with respect to independent, real-world observations. However, knowledge-based systems are often built for situations in which such empirical tests are neither possible nor affordable. For this reason, validation in the expert systems sense is often done by comparing model processes and results back to the judgments of experts (Turban and Aronson 2001).

We addressed validation in two ways. First, we asked the experts to bring a list of subwatersheds with which they were most familiar to the initial workshop and score the condition of each on a scale of 1 to 10. We could then compare these overall expert assessments to the results of our models. Second, we scheduled another round of provincial workshops, in which we presented model results based on actual data to the experts. While these experts had designed the models, it was difficult for them to fully understand how the models would perform on actual data, especially given the novelty of this methodology for most of them.

The review was initiated and could have been done completely via e-mail, but we chose a workshop approach again for reasons stated above. Workshops were scheduled for a 1-day duration and were held in each region. Often, a year or more had passed since the original workshop, so a brief review of the modeling process was given again before delving into specific results. Review materials primarily consisted of the summary document from the first workshop, model diagrams, and a spreadsheet listing the raw values and evaluated scores for each indicator for each subwatershed. Our GIS specialist and an intern also assembled a comprehensive GIS database for each region, and the facilitators worked with each group to interactively delve into these data to answer questions and provide context. Because EMDS is integrated into a GIS, we also explored the models more interactively with the groups in this round.

After each validation workshop, we implemented suggested changes to the models and data and circulated a summary description of these changes, along with the amended model results, in spreadsheet format for final review.

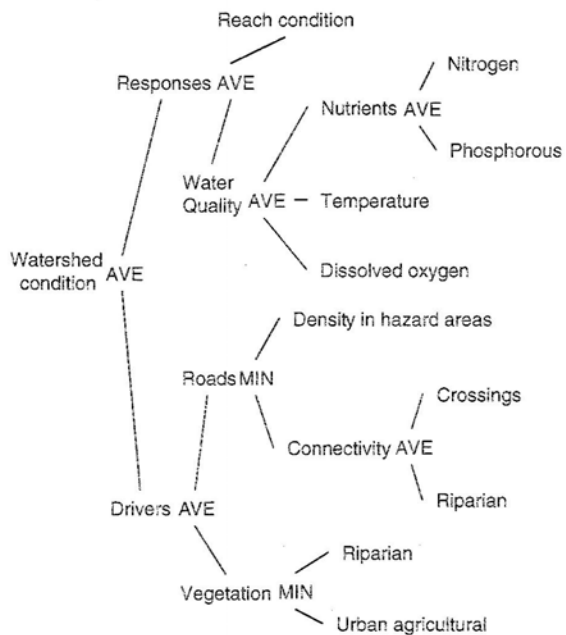
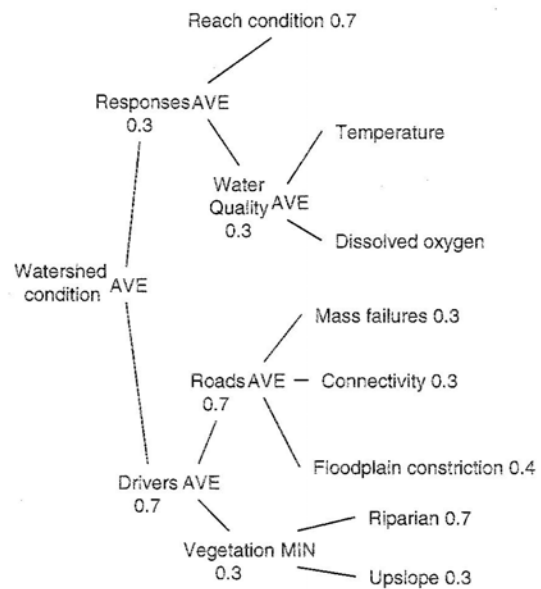
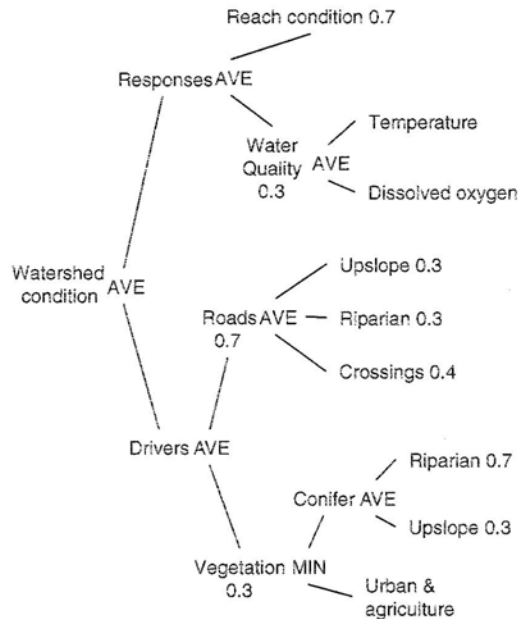
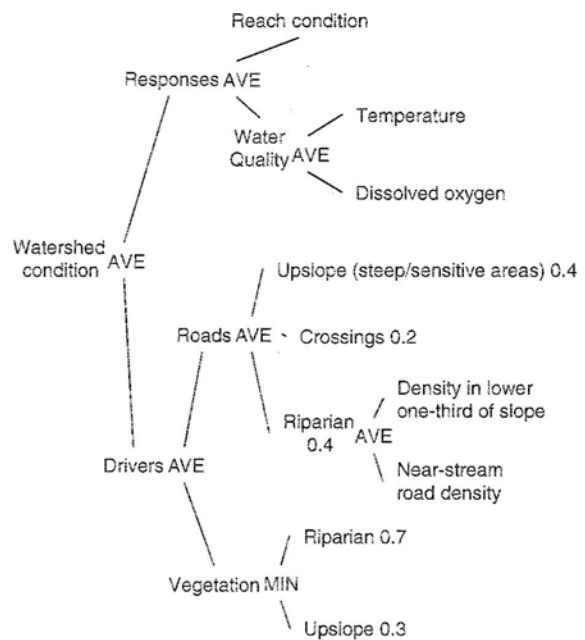
Results and discussion

In this section, we discuss results of our process for engaging subject-matter experts and lessons learned from our approach to knowledge-based modeling.

Choice of geographic scale

Organizing the modeling process by physiographic province had specific effects on the participation of experts and the resulting model structures. Holding separate workshops for the provinces seemed to promote both greater participation and greater attendance but also introduced differences to the structure of the models compared to a single, unified model. Organizing by province resulted in groups of six to ten experts attending each workshop, a number which provided some diversity of experience and perspective, and allowed each individual ample opportunity for input. We would anticipate less participation overall at a single modeling workshop, both in terms of overall attendance (due to less scheduling flexibility) and individual input (since there would be more participants).

Organizing the effort by physiographic province also had effects on the process of model development per se. Creating a separate model for each province kept the model structures and data inputs simpler as opposed to one large, unified model. While the model structures do not appear particularly complex (see Fig. 4), each required an average of 55 data inputs: 25 indicator measures and 30 evaluation function values. For the most part, these measures (e.g., percent area with mean tree diameter >20 in.) and values for evaluation criteria remained constant across a province, so provincial grouping reduced the time required for data preparation and input. However, three of the six models still required

OR/WA Coast and Olympic PeninsulaNorth CascadesWest and High CascadesKlamath-Siskiyou and Franciscan

^a Numbers next to model nodes represent node weights

Fig. 4 Final watershed model structure diagrams for the aquatic provinces. Numbers next to model nodes represent node weights

physiographic delineations at subprovincial scales: the North Cascades included an east/west division, High Cascades a north/central/south, and Klamath-Franciscan a wet/dry division. The subprovincial divisions were mainly implemented to provide different evaluation criteria (e.g., expected diameter, canopy cover) for accounted for expected vegetation differences across environmental gradients. Adjusting model behavior to account for physiographic differences at the subprovince level was easily implemented with logic switches, thus avoiding the need to further proliferate the number of models needed to represent the NWFP area.

Holding separate model-building workshops for each province did raise questions about model consistency. One of our participants posed this question in a workshop, "How much of this variation between provincial models is due to ecological differences and how much to the individuals involved?" Although we would expect a unified model to reflect some differences by physiographic province, it seems reasonable to assume that overall model variation would be less than for separate provincial models. We captured at least some of the ecological rationale for model differences in justifications provided by the experts for specific data measures and evaluation criteria, but we did not have a process to distinguish possible group effects, especially in the general structuring of the models.

In a similar vein, another participant asked about the consistency of interprovince comparisons given that different models were built for each province. Strictly speaking, the Monitoring Plan sampling design was not intended for interprovincial comparisons, but for an assessment of the overall plan area. But the question can also be interpreted more broadly as, "Can the overall evaluation be considered consistent if differently constructed models are used?" We have no definitive answer to this question. The models are congruent in the sense that they were all designed to answer the same question, and as stated above, they should differ to account for

ecological variation across provinces. However, when relying on a knowledge-based approach, these differences are confounded by differences in the mental models of the experts. A single workshop approach may have lessened this impact but, as stated above, would have faced numerous operational disadvantages.

Selection of experts

Selection of experts has been cited as one of the most important steps in the development of knowledge-based systems (Marakas). In our case, this selection was strongly based on organizational affiliation and how we structured the problem geographically, as well as their area of expertise. Overall, 33 experts participated in the first round of workshops (six to 10 per workshop), and they listed an additional 41 colleagues whom they had consulted on the process (six to 16 per workshop).

Fish biologists comprised 20 of the 33 attendees, with four each of hydrologists and aquatic ecologists, three watershed specialists, and one undefined. Nonattending contributors were more diverse in background, with 13 fish specialists, six hydrologists, five each of geologists and aquatic ecologists, four more general ecologists, and eight "other". The majority (26) of active participants were Forest Service employees, four were from the BLM, and only one each from the National Park Service, US Fish and Wildlife, and a state Fish and Wildlife agency representative. Indirect contributors were slightly more diverse, with 32 from the Forest Service, three from the BLM, four and two, respectively, from federal and state fish and wildlife agencies.

Upon reflection, we realize that the involvement of federal and state regulatory agencies should be improved since these agencies have considerable expertise in assessing stream and watershed conditions. The bias toward land management agencies stemmed from two reasons. The first reason was simply expediency; the Forest Service had the lead in the process and therefore a greater ability to identify its own experts and more authority to encourage them to participate. Second, historically, the watershed assessment process has evolved in a particular

regulatory context where land management agencies make an initial assessment of potential impacts of management activities (e.g., timber harvest), and this assessment is then reviewed independently by the regulatory agencies. However, since this assessment is not an evaluation of proposed activities (like an environmental impact statement) but rather a regional assessment of existing conditions, it is not subject to regulatory review. Therefore, to achieve greater interagency participation, it must be integrated into the workshop process of future assessments.

We found little mention of the effects of group size in the expert systems literature. A recent book on participatory systems dynamics modeling recommends sizes from 10 to 40 stakeholders, when the main concern is building sufficiently broad consensus to move a social decision-making process forward (van Den Belt 2004). The literature on focus groups recommends a considerably smaller range, from four to 12, to allow everyone satisfactory input and prevent the group from splintering (Krueger and Casey 2000). Because each of our participants had in-depth knowledge of the subject, and our modeling effort was more focused on capturing this expertise than achieving broad social consensus, the smaller group size appeared appropriate for this situation. A greater emphasis on interagency participation in future workshops will likely lead to larger workshops and a greater emphasis on consensus building.

Model construction

Model specification workshops

Each workshop lasted from 2 1/2 to 3 days. Because the knowledge-based modeling process was new to the participants, we found it advantageous to prescribe the sequence of activities, working first on the overall model structures then on the larger categories (e.g., roads, vegetation), proceeding down to specific indicators and evaluation functions. At the end, we iterated back to examine the overall model structure given the specific indicator criteria ultimately chosen. This last step proved to be important in a several cases; for example, when a group decided to set temperature curve limits at near lethal (rather than just un-

favorable) levels, they realized that this indicator should be integrated into the model using a minimum (limiting factor) rather average operator.

The four workshops produced model structures, which were similar in the broad sense but varied in the specific choice of indicators and weightings (see Fig. 4). Each group began by examining models made by the previous groups, or in the case of the first workshop, the pilot model produced by the monitoring team. Other authors have suggested that providing such a "straw man" can get groups started more quickly without unduly influencing the outcome, as evidenced by the difference between the initial and final documents (Schmoldt and Peterson 2000). We believed providing an example model was critical in this case because of the novelty of the task, and the structure of the final models is similar enough to surmise that the initial models influenced the outcome. Part of the similarity between model structures was due to the limited set of indicators available to the modeling process, but participants were innovative to an extent. Evaluation of individual indicators varied considerably between provinces. For example, an evaluation of riparian vegetation was included in each of the provincial models, but the width of the riparian buffer varied by province as did the size class of vegetation evaluated (Table 3).

Overall, the loosely structured workshop approach proved effective as a technique for knowledge elicitation. Using more time-consuming, single interview or verbal protocol analysis methods would have been impractical considering the number of experts involved. Additionally, the group context provided opportunities for discussion, which appeared to stimulate model formulation and production of associated judgments. Potential disadvantages include the time needed for such discussion and the possibility of failing to reach consensus on aspects of the model or, on the other hand, artificially reaching consensus because of "group think" or conflict avoidance. We did not experience such a deadlock, but we are not able to judge the extent to which the group setting homogenized the resulting judgments.

Other knowledge elicitation approaches are possible. For another module of the NWFP review, Marcot (2006) described a four-step process.

Table 3 Riparian buffer widths, tree size classes, and evaluation function thresholds for each aquatic province

Province (subprovince)	Buffer width (m)	Size class (in.) ^a	Evaluation function thresholds ^b			
			−1	+1	+1	0
OR/WA Coast	50	>20	25	50	75	85
Olympic	50	>20	25	50	NA	NA
Klamath, Siskiyou, Franciscan	50	>20	40	75	NA	NA
West Cascades	50	>20	60	100	NA	NA
North Cascades (West)	50	>20	65	88	NA	NA
North Cascades (East)	30	>20	65	88	NA	NA
High Cascades (N/S)	30	<10	80	30	NA	NA
High Cascades (Central)	30	<10	60	20	NA	NA

^aAverage conifer quadratic mean diameter (square root of the average squared diameter value; often used in forestry because of a closer relationship to stand volume than the arithmetic average)

^bPercentage of riparian area meeting the size class threshold. Note that the OR/WA Coast province used a four-point evaluation function, which reduced scores for an overabundance as well as underabundance of large trees, while all the other provinces used a two-point function. The High Cascades evaluated the percentage area in small trees

First, he worked with a single expert (or one small group) to produce an initial model. Second, this model was sent out to other experts for review, and the resulting suggestions were incorporated into the model as deemed appropriate by the original expert and modeler (similar to the scientific publication process). Third, the model results were compared to a dataset of known species occurrences. And fourth, when available, the model was compared to other independent occurrence data derived from randomized surveys. Although we had no independent dataset against which to compare (and our response variable of "watershed condition" is not directly observable), we could have used this independent peer review approach. However, such an approach does not provide as much opportunity for group consensus building, which we felt was crucial given the abstract nature of the concept of watershed condition. Additionally, we believe the workshop process was more amenable to broader participation and gave participants a feeling of joint ownership in the resulting models. Developing consensus among both managers and scientists was fundamental to the credibility of this assessment, and participation in the modeling process has been identified as an important factor in acceptance of model results in a number of cases (Gordon 2006; Manno et al. 2008; Shifley et al. 2008).

Keeping notes on the workshop discussion and decisions reached is a critical task that required

more expertise than we originally assumed. Our note takers at the early workshops were not involved in the model building and quickly became overwhelmed trying to transcribe the whole discussion among six to ten experts. A keen sense for the critical points of a model was essential, and we found that having the two facilitators able to trade off on this duty was optimal. The active moderator was also able to facilitate note taking by asking the group to summarize decisions and their justifications before moving on to the next topic. We were consistently able to capture the model structures, indicators, and evaluation criteria, but recording the complex discussions about justifications for these choices was more problematic. In the future, an even more structured approach to recording might be helpful, including agreeing on specific categories of information for the facilitator to summarize and the note taker to record (indicator, evaluation function, thresholds, and justifications) and providing the note taker with blank templates to fill in for the different evaluation function options.

Model building

Although the first draft models could be produced quite quickly (a matter of hours) from the workshop summaries, working out evaluation criteria that require more research and obtaining and processing the data needed (particularly the vegetation maps based on remote sensing) took

almost a year. Other modeling practitioners have commented on how data preparation is often the most time-intensive aspect of modeling (Johnson et al. 2007; Shifley et al. 2008).

Once the data were deemed as complete as possible, each model was run. In a number of cases, no data were available for an indicator in a model. The NetWeaver component of EMDS has a default method for handling missing data, which is to give it an evaluated score of zero, indicating no evidence either for or against the assessment objective. Such zero scores can sometimes affect the overall assessment score significantly, bringing up or down an otherwise negative or positive result. A dilemma we faced was whether to include or exclude assessment components lacking data. For example, the Northern Cascades group, as part of their roads evaluation, wished to evaluate floodplain impacts using two measures: (1) the percent of stream segments with a road within 20 m, and (2) the percent of floodplain restricted due to roads. Unfortunately, measuring the indicator for floodplain restriction proved to be beyond the available analytical resources, so this input was not available. In most of these cases, we chose to exclude such indicators from the evaluation rather than incorporate the default value of zero.

Model verification and validation

Verification of the models using test data designed to produce fully positive (+1) and fully negative (-1) results at the top level of a model's hierarchy was helpful in finding a number of model formulation errors. Although each of these datasets required a few hours to set up and test, the effort appeared to be worthwhile because the errors would have been more difficult to detect using actual field data.

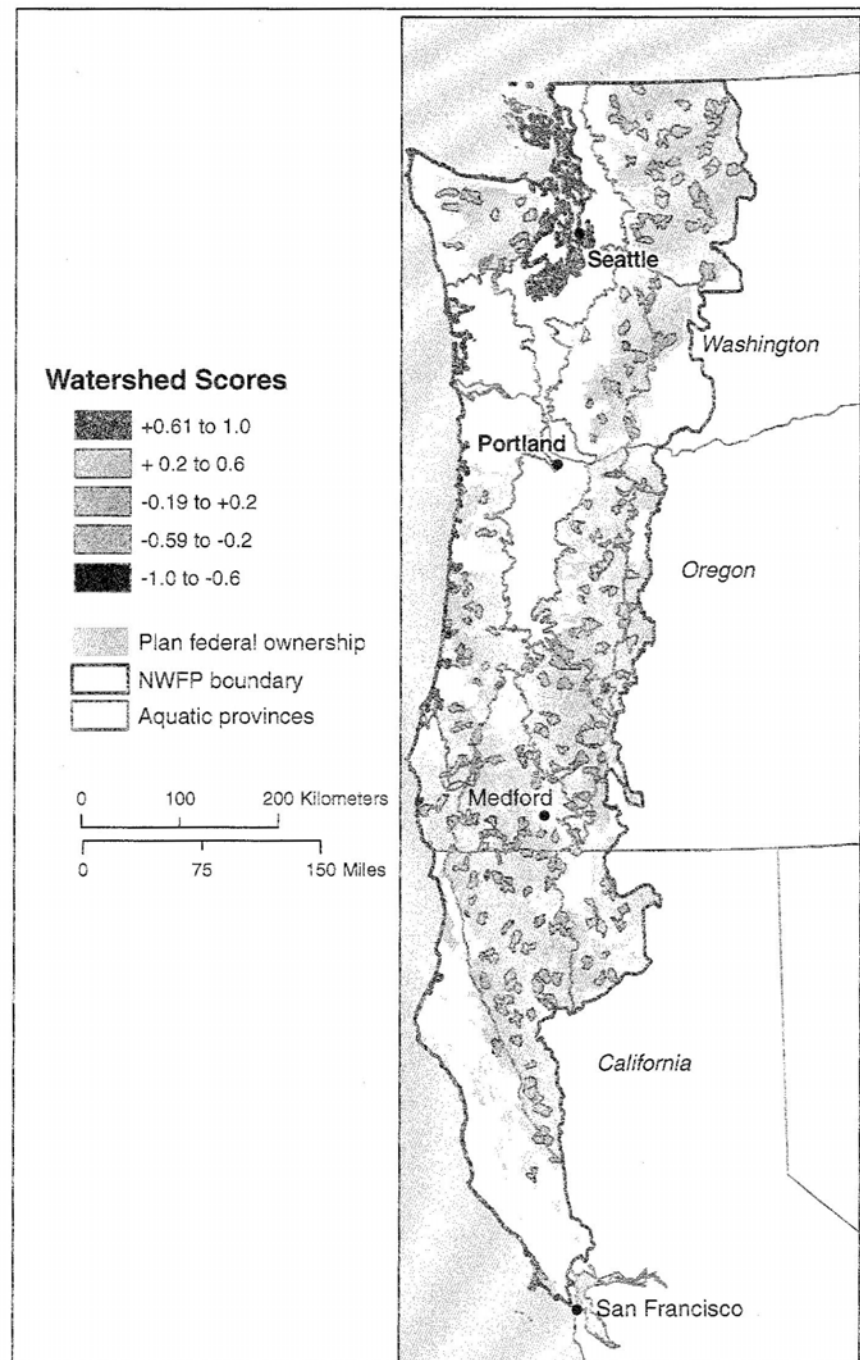
Our expert systems approach to validation (rechecking the model results with the experts) encountered an unanticipated difficulty in the first step: sending the model results out to the participants. It proved to be a challenge to format the model output in a way that it could be easily accessed and understood. We did not expect the participants to run GIS or other parts of the modeling software themselves. Exporting the results

to a spreadsheet resulted in a mass of acronyms and numbers, which were not ordered according to the model hierarchy. Through the design of output templates, manual formatting, and attached explanations, we were able to make the results considerably more tractable, but we still view this as a weakness in the EMDS modeling process that could be better addressed by the software. For future assessments, we also hope to export the color-coded GIS map outputs to a PowerPoint presentation file to make them accessible to participants. An example of these mapped outputs is presented in Fig. 5.

Our plan for comparing our models against holistic assessments provided by the experts was not successful. Very few participants completed this task, and the subwatershed boundaries used in the few lists we did receive did not coincide with the boundaries used by our modeling effort. Watersheds can be defined at a number of scales, and the federal delineation process for the scale we were modeling (sixth-field hydrologic units) was in flux during our assessment, resulting in the availability of different versions at different times. In retrospect, we should have discussed with the participants whether they could reasonably produce such a rating and, if so, provided maps of the subwatershed boundaries we were using for our model. This activity might have been accomplished as prework or as a first step in our workshops.

The validation workshops were more straightforward. Most of the participants were familiar with the modeling process, and we now had considerable data to consult. By projecting the modeling system on a screen, we were able to interactively step down through the model hierarchy and GIS data to see how the various levels of evaluation functioned. This process uncovered both criteria in the model that needed modification and occasional problems with the underlying data we were using. By the end of each validation workshop, consensus (in terms of no significant objections from any participant) was reached on the model structure. We did not have a method of quantifying the degree of agreement reached, but further exploration of this area would be useful in the future. This second round of workshops, focused on model validation, proved

Fig. 5 Watershed condition scores produced for the 250 randomly selected watersheds in the plan area



to be an essential step in the process, as evidenced by the numerous refinements in model structures and data suggested by participants.

Other knowledge-based-system projects have devised different validation strategies. Girvetz and Shilling (2003), in estimating the environmen-

tal impact of forest roads, found three independent measures with which to compare their model results: (1) a separate, field-based expert assessment of roads in one project area, (2) a record of decisions on road closures, and (3) a forest-wide record of road failures. None of these matched the objectives of their model exactly, but all provided a basis for comparison and reflection. Our attempt to elicit holistic watershed evaluations before the workshops was similar to their first validation measure. Our lesson learned is to devote more time (including some field time) to this aspect, including adjusting to different watershed delineations.

Finally, comparison against aquatic and riparian biota datasets could provide more empirical support for these models. As mentioned in the "Introduction," relationships between habitat and fish population numbers have been difficult to establish because of the complexity involved and the lack of extensive data. Pess et al. (2002) were able to find moderate correlations between adult salmon counts and a number of riparian and upslope attributes, including general land use classes. However, such fish counts from numerous sites over time are not widely available. Herger et al. (2003), using a more extensive one-time sampling approach for lower-order streams, found little correlation between simple species metrics (richness, percent tolerants) and human disturbance metrics. However, new biotic indices, which combine both abundance and diversity measures, show promise, and there are a number of available datasets in the Pacific Northwest with the necessary information (Hubler 2008; Hughes et al. 2004; Whittier et al. 2007). Using an extensive sample from the Oregon and Washington Coast Range region, Kaufmann and Hughes (2006) were able to factor out major natural landscape differences (basin size, stream gradient, lithology) and found numerous statistically significant relationships between a biotic index and in-stream, riparian, and upslope indicators. Other recent assessments have also included these biotic indices, although they have not attempted correlations with the other biophysical indicators (Hubler et al. 2009; Mulvey et al. 2009; Oregon DEQ 2004). Testing for relationships between our expert-based watershed

assessments and these biotic datasets would be a useful and feasible next step in our model validation methods.

Conclusions

Integrating decision support software into an expert-based assessment process provided a number of advantages. In particular, the modeling tool facilitated evaluation, documentation, and repeatability. While available software based on Bayesian belief networks or fuzzy logic modeling provides some basic structures, which can facilitate the assessment process, they are very generic systems that require the user to organize the domain-specific knowledge for any particular application. Hopefully, our discussion has highlighted some of the more important decisions needed to build specific assessment models, from the selection and elicitation of knowledge experts to the choice and integration of indicators.

A workshop-based, group knowledge elicitation approach was effective in the context of our modeling problem. Breaking the process into smaller regional workshops facilitated attendance and opportunities for individuals to actively contribute. Participation was lively and inclusive in the workshops, and this interaction enabled the participants to work out difficult issues as a group. However, our groups were relatively homogeneous in expertise and responsibility, and many of the individuals had a history of working together on other regional projects. Other authors have reflected on the challenges of participatory modeling with different types of stakeholder groups (Fall et al. 2001; Mendoza and Prabhu 2003; Rouwette et al. 2002; van Den Belt 2004). The new concepts involved in fuzzy logic modeling did not appear difficult for the experts to grasp, and the modeling process provided a much-needed framework for expressing and combining their knowledge. As the experts become more familiar with these tools, we expect them to further refine the models to better represent their understandings of the world. Many other expert knowledge elicitation approaches are available, and future

studies comparing their relative effectiveness and efficiency would be useful.

We also derived cautionary lessons on some of the more mundane (but nevertheless critically important) aspects of the assessment process. Recording the group process in notes is challenging. Scribes need to understand the desired knowledge structuring process and at least the basics of the subject matter; an explicit plan for coordination with the facilitator is recommended. Second, preparing the raw data for model input is likely to be time consuming and is a large potential source of errors. Special care should be taken with indicator metrics. Recall that the failure of a Mars orbiter was caused by miscommunication of English versus metric units (Lloyd 1999), and the units involved in environmental assessment are often complex (e.g., is the measure "miles of riparian road per square mile riparian area" or "miles of riparian road per mile of stream"?). The group will also need to decide how to handle missing data. Third, communicating model outputs back to the expert group for validation can be difficult because they are unlikely to be able to run the model themselves. Exporting model results and reformatting in a more universally accessible file format (Excel spreadsheets in our case) was time consuming and a less-than-ideal communication vehicle. Interactive displays of the model and GIS data in the validation workshops were much richer communication mechanisms but limited by the amount of time the group can be assembled. The increased availability of web conferencing systems could be used to ease this meeting bottleneck. Ultimately, adding software capabilities for users to interact individually with models over the world wide web should be possible.

In hindsight, one of the decisions with the most profound impacts turned out to be the geographic scale at which we engaged the experts. We expected some model differences between physiographic provinces due to ecological differences, but holding multiple knowledge elicitation workshops also introduced an unknown amount of variation based on the mental models of the different experts attending each session. Besides the unwieldy alternative of collapsing the elicitation into a single workshop, we see the following

three general approaches, which might enhance model consistency while maintaining the benefits of distributed input:

1. a *control* approach in which all groups are required to use the same model structure but allowed to vary weights and evaluation criteria according to provincial ecological importance;
2. an *informational* approach, which would provide participants with detailed information on the other models and encourage them to use similar structures as appropriate; and,
3. a *postprocessing* approach in which, after the workshops, the project team or a small group of regional experts (preferably who are able to attend all the workshops) would review all the models and propose changes to harmonize them, perhaps followed by an additional validation workshop.

Finally, validation (or at least confirmation) of knowledge-based systems remains a challenge because such modeling is often undertaken in situations for which statistical validation with biophysical data is not feasible. There are, however, other strategies that can be pursued, such as corroboration with related biophysical and sociopolitical indicators and comparison with independent expertise, either through field-based assessments or the peer review process.

Acknowledgements We would like to acknowledge S. Lanigan, K. Reynolds, P. Hessburg, D. Hohler, M. Jensen, T. Geier, and two anonymous reviewers for the helpful comments on earlier drafts.

References

- Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1986). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Middleton: Green Briar.
- Fall, A., Daust, D., & Morgan, D. (2001). A framework and software tool to support collaborative landscape analysis: Fitting square pegs into square holes. *Transactions in GIS*, 5, 67–86.
- Gallo, K., Moyer, C., & Lanigan, S. (2002). *Northwest Forest Plan Aquatic And Riparian Effectiveness Monitoring Program: 2001 pilot summary report*. USDA Forest Service Regional Ecosystem Office, Portland, OR. Retrieved from <http://www.reo.gov/monitoring/>

- reports/watershed/2001_Final_AREMP_Tech_Report.pdf.
- Gallo, K., Lanigan, S. H., Eldred, P., Gordon, S. N., & Moyer, C. (2005). *Northwest Forest Plan-The first 10 years (1994-2003): Preliminary assessment of the condition of watersheds*. Report PNW-GTR-647. Portland: USDA Forest Service Pacific Northwest Research Station.
- Girvetz, E., & Shilling, F. (2003). Decision support for road system analysis and modification on the Tahoe National Forest. *Environmental Management*, 32, 218–233.
- Gordon, S. N. (2006). *Decision support systems for forest biodiversity management: A review of tools and an analytical-deliberative framework for understanding their successful application*. Doctoral dissertation, Dept. of Forest Resources, Oregon State University, Corvallis, OR. Retrieved from <http://ir.library.oregonstate.edu/jspui/handle/1957/2592>.
- Herger, L. G., Weiss, A., Augustine, S., & Hayslip, G. (2003). *Modeling fish distributions in the Pacific Northwest Coast Range Ecoregion using EMAP data*. Report EPA/910/R-03/000. U.S. Environmental Protection Agency, Region 10, Seattle, WA.
- Hubler, S. (2008). *PREDATOR: Development and use of RIVPACS-type macroinvertebrae models to assess the biotic condition of Wadeable Oregon streams*. Hillsboro: Oregon Department of Environmental Quality, Laboratory and Environmental Assessment Division.
- Hubler, S., Miller, S., Merrick, L., Leferink, R., & Borisenko, A. (2009). *High level indicators of Oregon's forested streams*. Hillsboro: Oregon Department of Environmental Quality, Laboratory and Environmental Assessment Division.
- Hughes, R. M., Howlin, S., & Kaufmann, P. R. (2004). A biointegrity index (IBI) for coldwater streams of western Oregon and Washington. *Transactions of the American Fisheries Society*, 133, 1497–1515.
- Jensen, M. E., Reynolds, K., Andreasen, J., & Goodman, I. A. (2000). A knowledge-based approach to the assessment of watershed condition. *Environmental Monitoring and Assessment*, 64, 271–283.
- Johnson, K. N., Swanson, F., Herring, M., & Greene, S. (Eds.) (1990). *Bioregional assessments: Science at the crossroads management and policy*. Washington: Island.
- Johnson, K. N., Gordon, S. N., Duncan, S., Lach, D., McComb, B., & Reynolds, K. (2007). *Conserving creatures of the forest: A guide to decision making and decision models for forest biodiversity*. Corvallis: College of Forestry, Oregon State University. Retrieved from <http://www.ncseonline.org/ncssf/dss>.
- Kaufmann, P. R., & Hughes, R. M. (2006). Geomorphic and anthropogenic influences on fish and amphibians in Pacific Northwest coastal streams. In R. M. Hughes, L. Wang, & P. W. Seelbach (Eds.), *American fisheries society symposium: Landscape influences on stream habitats and biological assemblages*. Bethesda: American Fisheries Society.
- Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research*. Thousand Oaks: Sage.
- Lloyd, R. (1999). Metric mishap caused loss of NASA orbiter. Retrieved from <http://www.cnn.com/TECH/space/9909/30/mars.metric.02/>.
- Manno, I. P., Smardon, R., DePinto, J. V., Cloyd, E. T., & Del Granado, S. M. (2008). *The use of models in Great Lakes decision making: An interdisciplinary synthesis*. Report 16. Syracuse: Randolph O. Pack Environmental Institute.
- Marakas, G. M. (1999). *Decision support systems in the 21st century*. Upper Saddle River: Prentice Hall.
- Marcot, B. G. (2006). Characterizing species at risk I: Modeling rare species under the Northwest Forest Plan. *Ecology and Society*, 11, 10.
- Mendoza, G. A., & Prabhu, R. (2003). Qualitative multi-criteria approaches to assessing indicators of sustainable forest resource management. *Forest Ecology and Management*, 174, 329–343.
- Mitchell, R. B., Cark, W. C., Cash, D. W., & Dickson, N. M. (Eds.) (2006). *Global environmental assessments: Information, institutions, and influence*. Cambridge: MIT.
- Mulvey, M., Leferink, R., & Borisenko, A. (2009). *Willamette basin rivers and streams assessment*. Hillsboro: Oregon Department of Environmental Quality, Laboratory and Environmental Assessment Division.
- Oregon DEQ (2004). *Oregon's 2004 water quality assessment section 305(b) report*. Hillsboro: Oregon Department of Environmental Quality, Water Quality Division.
- Oreskes, N., Shrader-Frechette, K., & Beitz, K. (1994). Verification, validation, and the confirmation of numerical models in the earth sciences. *Science*, 263, 641–646.
- Pess, G. R., Montgomery, D. R., Steel, E. A., Bilby, R. E., Feist, B. E., & Greenberg, H. M. (2002). Landscape characteristics, land use, and coho salmon (*Oncorhynchus kisiuch.i*) abundance, Snohomish River, Wash., USA. *Canadian Journal of Fisheries and Aquatic Sciences*, 59, 613–623.
- Rauscher, H. M. (2000). Application scientific knowledge to decision-making in managing forest ecosystems: Forward. *Computers and Electronics in Agriculture*, 27, 1–6.
- Reeves, G. H., Hohler, D. B., Larsen, D. P., Busch, D. E., Kratz, K., Reynolds, K., et al. (2004). *Aquatic and riparian effectiveness monitoring plan for the northwest forest plan*. Report PNW-GTR-577. Portland: USDA Forest Service Pacific Northwest Research Station.
- Reeves, G. H., Williams, J. E., Burnett, K. M., & Gallo, K. (2006). The aquatic conservation strategy of the Northwest Forest Plan. *Conservation Biology*, 20, 319–329.
- Reynolds, K., & Peets, S. (2001). Integrated assessment and priorities for protection and restoration of watersheds. In: *Proceedings of the IUFRO 4.11 conference on forest biometry, modelling and information science*. UK: University of Greenwich.

- Reynolds, K. M., Jensen, M. E., Andreasen, J., & Goodman, I. (2000). Knowledge-based assessment of watershed condition. *Computers and Electronics in Agriculture*, 27, 315–333.
- Reynolds, K. M., Rodriguez, S., & Bevans, K. (2002). *Ecosystem management decision support 3.0 user guide*. Corvallis: USDA Forest Service Pacific Northwest Research Station.
- Rieman, B. E., Hessburg, P. F., Lee, D. C., Thurow, R. F., & Sedell, J. R. (2000). Toward an integrated classification of ecosystems: Defining opportunities for managing fish and forest health. *Environmental Management*, 25, 425–444.
- Rieman, B. E., Peterson, J. T., Clayton, J., Howell, P., Thurow, R., Thompson, W., et al. (2001). Evaluation of potential effects of federal land management alternatives on trends of salmonids and their habitats in the interior Columbia River basin. *Forest Ecology and Management*, 153, 43–62.
- Rouwette, E. A. J. A., Vermix, J. A. M., & van Mullekom, T. (2002). Group model building effectiveness: A review of assessment studies. *System Dynamics Review*, 18, 5–45.
- Schmoldt, D. L., & Peterson, D. L. (2000). Analytical group decision making in natural resources: Methodology and application. *Forest Science*, 46, 62–75.
- Sedell, J. R., Lee, D. C., Rieman, B. E., Thurow, R. F., & Williams, I. E. (1997). Effects of proposed alternatives on aquatic habitats and native fishes. In T. M. Quigley, K. M. Lee, & S. J. Arbelbide (Eds.), *Evaluation of EIS alternatives by the science integration team* (pp. 435–535). Portland: USDA Forest Service Pacific Northwest Research Station.
- Shifley, S. R., Thompson, Iii, F. R., Dijak, W. D., & Fan, Z. (2008). Forecasting landscape-scale, cumulative effects of forest management on vegetation and wildlife habitat: A case study of issues, limitations, and opportunities. *Forest Ecology and Management*, 254, 474–483.
- Turban, E., & Aronson, J. E. (2001). *Decision support systems and intelligent systems*. Upper Saddle River: Prentice Hall.
- USDA & USDI (1994). *Record of decision for amendments to Forest Service and Bureau of Land Management planning documents within the range of the northern spotted owl*. Washington: USDA Forest Service and Department of Interior Bureau of Land Management.
- van Den Belt, M. (2004). *Mediated modeling: A system dynamics approach to environmental consensus building*. Washington: Island.
- Whittier, T. R., Hughes, R. M., Lomnický, G. A., & Peck, D. V. (2007). Fish and amphibian tolerance values and an assemblage tolerance index for streams and rivers in the western USA. *Transactions of the American Fisheries Society*, 136, 254–271.