

Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie

Roger M Cooke,†‡§ Marion E Wittmann,‖ David M Lodge,‖ John D Rothlisberger,# Edward S Rutherford,†† Hongyan Zhang,†† and Doran M Mason‡‡

†Resources for the Future, Washington, DC, USA
‡University of Strathclyde, Glasgow, Scotland, United Kingdom
§Delft University of Technology, Delft, The Netherlands
‖Department of Biological Sciences, University of Notre Dame, Indiana, USA
#US Department of Agricultural Forest Service, Eastern Region, Milwaukee, Wisconsin
††Cooperative Institute for Limnology and Ecosystems Research, University of Michigan, Ann Arbor, Michigan, USA
‡‡National Oceanic and Atmospheric Administration, Great Lakes Environmental Research Laboratory, Ann Arbor, Michigan, USA

(Submitted 20 February 2014; Returned for Revision 23 April 2014; Accepted 30 June 2014)

Decision Analysis

ABSTRACT

Structured expert judgment (SEJ) is used to quantify the uncertainty of nonindigenous fish (bighead carp [*Hypophthalmichthys nobilis*] and silver carp [*H. molitrix*]) establishment in Lake Erie. The classical model for structured expert judgment model is applied. Forming a weighted combination (called a decision maker) of experts' distributions, with weights derived from performance on a set of calibration variables from the experts' field, exhibits greater statistical accuracy and greater informativeness than simple averaging with equal weights. New methods of cross validation are applied and suggest that performance characteristics relative to equal weighting could be predicted with a small number (1–2) of calibration variables. The performance-based decision maker is somewhat degraded on out-of-sample prediction, but remained superior to the equal weight decision maker in terms of statistical accuracy and informativeness. *Integr Environ Assess Manag* 2014;10:522–528. © 2014 The Authors. *Integrated Environmental Assessment and Management* published by Wiley Periodicals, Inc. on behalf of SETAC.

Keywords: Asian carp Classical model Cross validation Invasive species Structured expert judgment

INTRODUCTION

Motivation

Bighead (*Hypophthalmichthys nobilis*) and silver (*H. molitrix*) carp are cyprinid fishes native to eastern Asia and introduced in the early 1970s to the United States as biocontrol agents for nuisance algae in freshwater ponds and lakes (Fuller et al. 1999). Since these introductions, Asian carp have escaped into natural systems and caused unwanted ecological and economic impacts (Kolar et al. 2007; Garvey et al. 2010). Ongoing efforts to prevent the introduction of these species to the Great Lakes have incurred high costs in research, monitoring, and surveillance actions (ACRCC 2012). Restoration of the natural hydrological separation of the Great Lakes and Mississippi River basins (to prevent the passage of Asian carp and other nonnative species) has been proposed with an estimated cost of $4 to $10 billion (ACRCC 2012). As yet, there is no evidence that Asian carp populations have established self-sustaining populations in any of the Great Lakes. One important question relevant to future decisions about hydrological separation or other management strategies is: Will Asian carp successfully establish and cause ecological or economic damage in the Great Lakes? Structured expert judgment (SEJ) is used to address these questions.

Among the Great Lakes, Lake Erie is considered the most vulnerable to Asian carp invasion because of its close proximity to established Asian carp populations, habitat, and high value fisheries species under threat. Results relating to peak and equilibrium biomass of invasive Asian carp and native species are published in Wittmann, Cooke, Rothlisberger, Rutherford et al. (2014). Results relating to effectiveness of deterrent strategies are published in Wittmann, Cooke, Rothlisberger, Lodge (2014). The present study focuses on the SEJ methodology, with particular emphasis on cross validation.

Structured expert judgment is an established technique for probabilistic risk assessment (Apostolakis 1990; Cooke 1991; Aspinall 2010) and consequence analysis (Cooke and Goossens 2000), and it has previously been used for several environmental applications including assessments of the likelihood of natural disasters (volcanic eruption, dam failure) (Aspinall et al. 2003; Klugel 2011), consequences of nuclear accidents (Cooke and Goossens 2000), drivers of climate change (Morgan et al. 2001; Lenton et al. 2008), ice sheet dynamics (Bamber and Aspinall 2013) fisheries and ecosystems (Burgman 2005; Rothlisberger et al. 2010, 2012; Teck et al. 2010; Martin et al. 2011) and increases in human mortality attributable to air pollution (Evans et al. 2005; Tuomisto et al. 2005; Roman et al. 2008).

The "classical model" (Cooke 1991) for combining expert judgment was used in many of the above references, and is used here. This model is distinguished from other methods of

All Supplemental Data may be found in the online version of this article.

* To whom correspondence may be addressed: cooke@rff.org

Published online 9 July 2014 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/ieam.1559

combining expert judgments in that it treats experts' judgments as statistical hypotheses and measures performance in terms of statistical accuracy and informativeness (see below), based on assessments of calibration or "seed" variables. Calibration variables are taken from the experts' field, and their true values are known post hoc although unknown to the experts at the time of assessment. These measures (statistical accuracy and informativeness) are used to construct an optimal performance-based combination of the experts' assessments. The use of calibration variables serves the triple purpose of 1) validating expert performance, 2) enabling performance-based combination of experts' distributions, and 3) evaluating performance of various combinations of experts' judgments. These combinations are referred to as Decision Makers (DMs). This article extends the evaluation of DM performance by including "out-of-sample" cross-validation (e.g., validating model performance using variables not used to initialize the model).

Why out-of-sample validation for ecological SEJ studies?

Evaluating forecasting models, using "out-of-sample" variables is preferable to using "in-sample" variables (e.g., those that also serve to initialize the model). A forecasting method based on calibration variables assumes that performance on calibration variables predicts performance on the variables of interest, for which true values are not known. For applications with short temporal scales, such as leading economic indicators, true values of estimated variables of interest can be observed, enabling true out-of-sample validation (van Overbeek 1999; Qing 2002). However, in many applications, including ecological forecasting, out-of-sample validation based on observing the variables of interest is simply not possible due to temporal or spatial limitations associated with observing true outcomes of these variables. When variables of interest cannot be observed, out-of-sample validation resorts to cross validation: that is, on splitting calibration variable sets into complementary subsets, and predicting values in one subset (the test set) with a model initialized on the other subset (the training set). In ecological applications, out-of-sample validation must be attained through cross validation.

Summary

The focus of this article is to present and evaluate new out-of-sample validation methods to assess the performance of SEJ with ecologically motivated calibration variables. The following sections describe the expert judgment application to Asian carp invasion of Lake Erie, explores out-of-sample validation, and compares performance-based weighting and equal weighting on the variables of interest. A final section provides conclusions and suggestions for further work. Supplemental Data (SD) presents details of the expert judgment performance measures and the analysis of expert data in this study. SD_A gives details of the mathematical model and the analysis of the expert judgment data in this study, SD_B consists of the elicitation protocol, and SD_C is the briefing booklet sent to all experts before the elicitation.

Expert judgment application to Asian carp invasion

Eleven experts participated in this study (Table 1) and were offered compensation of $1000 (although some did not choose to receive compensation). A briefing booklet was prepared (SD_C) describing the scientific background of Asian carp biology in North America and the status of Lake Erie fisheries. This briefing booklet and the elicitation instrument were sent to the experts beforehand, and experts were encouraged to use all accessible sources of information to estimate their responses.

The variables of interest divide into 3 categories. First are quantities that concern the biomass of Asian carp both at peak and equilibrium conditions after establishment in Lake Erie,

Table 1. Scoring of individual experts, PW, and EW combinations

Expert	*p* Value[a]	Mean relative information caliber variables[b]	Assessed caliber variables (*n*)	Unnormalized weight[c]
1	0.1815	0.6121	15	0
2	0.1227	0.6648	15	0
3	0.005634	1.47	15	0
4	0.7606	0.8562	15	0.6513
5	0.666	0.84	15	0
6	1.93E-06	1.381	15	0
7	0.05946	1.158	15	0
8	0.615	1.086	11	0
9	0.5276	1.288	15	0
10	0.2587	0.8282	15	0
11	0.5276	0.8071	15	0
PW	0.7606	0.8562	15	0.6513
EW	0.3126	0.2943	15	0.09197

EW = equal-weighted; PW = performance-weighted.
[a]The *p* value of falsely rejecting the hypothesis that the realizations are independently drawn from a distribution complying with the expert's percentiles.
[b]Informativeness score.
[c]Combined score of the weighted experts used in forming the performance-based DM.

production and consumption of Asian carp at equilibrium conditions, and predation on Asian carp by other fish species currently present in Lake Erie at the predicted equilibrium condition. Second, some questions concerned equilibrium condition biomass of other fishes in Lake Erie (walleye, yellow perch, gizzard shad, rainbow smelt) following bighead and/or silver carp establishment. Finally, questions regarding the efficacy of different types of Asian carp deterrent strategies proposed for use to prevent Asian carp passage into Lake Michigan were asked.

There were 15 calibration variables. They included observed whole lake biomass measurements and average annual dietary fractions of Lake Erie fishes. These values are estimated annually by Task Groups of the Lake Erie Committee of the binational Great Lake Fisheries Commission and reported to the public in annual reports released each spring. The calibration variables were chosen for their relevance to the variables of interest. Elicitations were conducted with each expert individually. Experts gave 5, 50, and 95 percentiles of their subjective probability distribution for all uncertain quantities. At least 2 elicitors were present, one engaging the expert and encouraging him/her to verbalize his reasoning, and the other taking notes to record the rationale behind his/her uncertainty. A typical elicitation lasted 4 hours. The calibration variables and participating experts are listed in SD_A.

In the classical model, expert performance is measured in 2 dimensions: statistical accuracy and informativeness. Statistical accuracy (also called calibration) is the p value of falsely rejecting the hypothesis that the realizations are independently drawn from a distribution complying with the expert's stated percentiles. The words "calibration," "p value," and "statistical accuracy" are used interchangeably. Informativeness is defined as the Shannon relative information in the expert's distribution relative to a background measure chosen by the analyst. The information score does not depend on the realizations, and an expert can give him/herself a high information score by choosing percentiles very close together. The theory of strictly proper scoring rules is invoked to combine these measures as a "product with cutoff," whereby an expert is unweighted if statistical accuracy falls beneath a threshold value. This insures that, in the long run, an expert receives his/her highest expected weight by and only by stating percentiles corresponding to his/her true beliefs. The combined scores for each expert are normalized to provide weights for the performance weight (PW) DM. That is, the PW DM's distributions are weighted combinations of the experts' distributions. The threshold is chosen to optimize the combined score of the DM. The Shannon relative information score is a slow function, whereas the likelihood of observing realizations outside the 90% central confidence band goes decreases very quickly. The product of the calibration and information scores is thus dominated by the calibration score, and informativeness modulates between more or less equally accurate experts. Details on scoring are found in SD_A.

The performance of the individual experts and of the PW and EW DMs are compared in Table 1. PW is statistically more accurate and more informative than EW. In this case, expert 4 received weight 1, and in approximately one-third of all applications, one expert receives all of the weight. Expert 9 also shows very good statistical accuracy and a higher information score than expert 4. The combined score of expert 9 (0.6797) is higher than that of expert 4 (0.6513). However, because the optimization is based on the p value, ensuring the strictly proper scoring rule property, it is not possible to give weight 1 to expert 9 and weight zero to expert 4. We can either include expert 4 alone, or, by assigning the cutoff equal to the p value of expert 9 (0.5276), we combine experts 4, 5, 8, 9, and 11 using weights equal to their normalized combined scores. This would result in a p value of 0.7104 and an information score on calibration variables of 0.5115, for a combined score of 0.3634 (these numbers are not retrievable from Table 1 but require recalculation). This demonstrates the strong influence of the strictly proper scoring rule constraint.

It is significant that only 2 experts (3 and 6) have low statistical accuracy, whereby the corresponding statistical hypotheses would be rejected at the 5% level. This is among the best performing expert panels in this regard. Note that the statistical accuracy scores vary over 5 orders of magnitude whereas the informativeness scores vary within a factor 3. Robustness analysis of these results is found in SD_A.

IN- AND OUT-OF-SAMPLE VALIDATION

Background

Table 1 compares performance of the PW DM and EW DM on the calibration variables. Because these calibration variables are also used to derive the weights for performance-based weighting, this is "in-sample" validation. That PW should outperform EW in-sample is certainly not a mathematical theorem, and EW does occasionally outperform PW in-sample. Were there not a strong in-sample preference for PW over EW, there would be little motive for considering performance-based weighting at all. Cooke and Goossens (2008) summarized a TU Delft expert judgment database comprised of 45 studies completed by 2006, in which experts assessed calibration variables and showed that PW DM strongly outperforms the EW DM in-sample. Since then, the number of studies has nearly doubled.

Researchers have used the TU Delft database to explore new models and to study whether good performance on the calibration variables is linked with good performance on the variables of interest. In a few studies, variables of interest were later observed, enabling true out-of-sample validation. In most cases, like the present case, the variables of interest are not observable on time scales relevant for the decision problem. Therefore, various forms of cross validation have been applied. Clemen (2008) proposed a remove-one-at-a-time (ROAT) method according to which the calibration variables were removed one at a time and predicted by the model initialized on the remaining calibration variables. Clemen (2008) pooled these predictions, though originating from different DMs, and compared the resulting synthetic DM with the EQ DM. Of the 14 studies analyzed by Clemen (2008), the PW DM was superior to the equal weight DM on 9, which was not statistically significant in a sample size of 14.

The ROAT method is biased against the PW DM, because each calibration variable was predicted by a DM in which experts who assessed that particular item badly were up-weighted, and all variables were assessed in this manner. In many studies, removing 1 calibration variable can influence an individual expert's p value by a factor 3 or more (Cooke 2008), a feature explained by the fact that statistical accuracy is a very fast function. It is easy to under appreciate this effect, and Cooke (2012, 2014) gave a detailed example making this conspicuous. Other types of cross validation have been carried out by Lin and Cheng (2008, 2009) and Flandoli et al. (2010).

The most extensive study is the Eggstaff et al. (2013) analysis of 62 cases, which initialized the PW DM on all subsets of calibration variables and, in each case, predicted the complementary subset. In 45 of the 62 cases (73%) the PW DM outperformed the EW DM, and when PW DM was better, it tended to be much better. This provides an indication of the ROAT bias. Details may be found in SD_A.

Some researchers have applied other scoring rules to measure performance that apply to single variables rather than sets of variables, such as the quadratic or logarithmic rule. An extensive discussion in Cooke (1991) discourages this practice. A simple example clarifies the issue. Suppose an expert assess the probability of heads as 1/2 with a coin of unknown composition. On each toss with the coin, the score is the same for heads and tails. If these individual scores are added, then the sum score after 100 tosses is also independent of the actual sequence of outcomes, 50 heads and 50 tails gets the same score as 100 heads. A general conclusion of all this work is that the performance-based DM is degraded on out-of-sample prediction, but is superior to the equal weight DM.

When a cross validation study initializes the PW DM on a training set containing K of the N calibration variables, and measures performance on the remaining N–K variables, the following issues arise. First, if K is close to N, then the small number of out-of-sample predictions have low statistical power and are subject to the bias noted above. Second, if K is small, then the ability to distinguish experts' high and low statistical accuracy is low. If the experts' calibration scores on the K variables are similar, then weighting is driven by informativeness that is often negatively correlated with statistical accuracy (Cooke 1991). For intermediate values of K, statistical power is lost at both ends, though this may be partially compensated by averaging the scores over all training sets of size K. How much statistical power is "recovered" in this way is difficult to judge, as the training sets overlap. In short, we do not know the best way for performing out-of-sample validation at present.

In lieu of an optimal solution, the exhaustive approach of Eggstaff et al. (2013) at least balances the known biases. Its disadvantage is that it is cumbersome, especially for more than 10 calibration variables. Eggstaff et al. (2013) noted that with a small training set, the scores on the test set did not predict the scores on the larger set of calibration variables but did confirm the superiority of performance weighting against equal weighting. Unfortunately, this does not mean that future

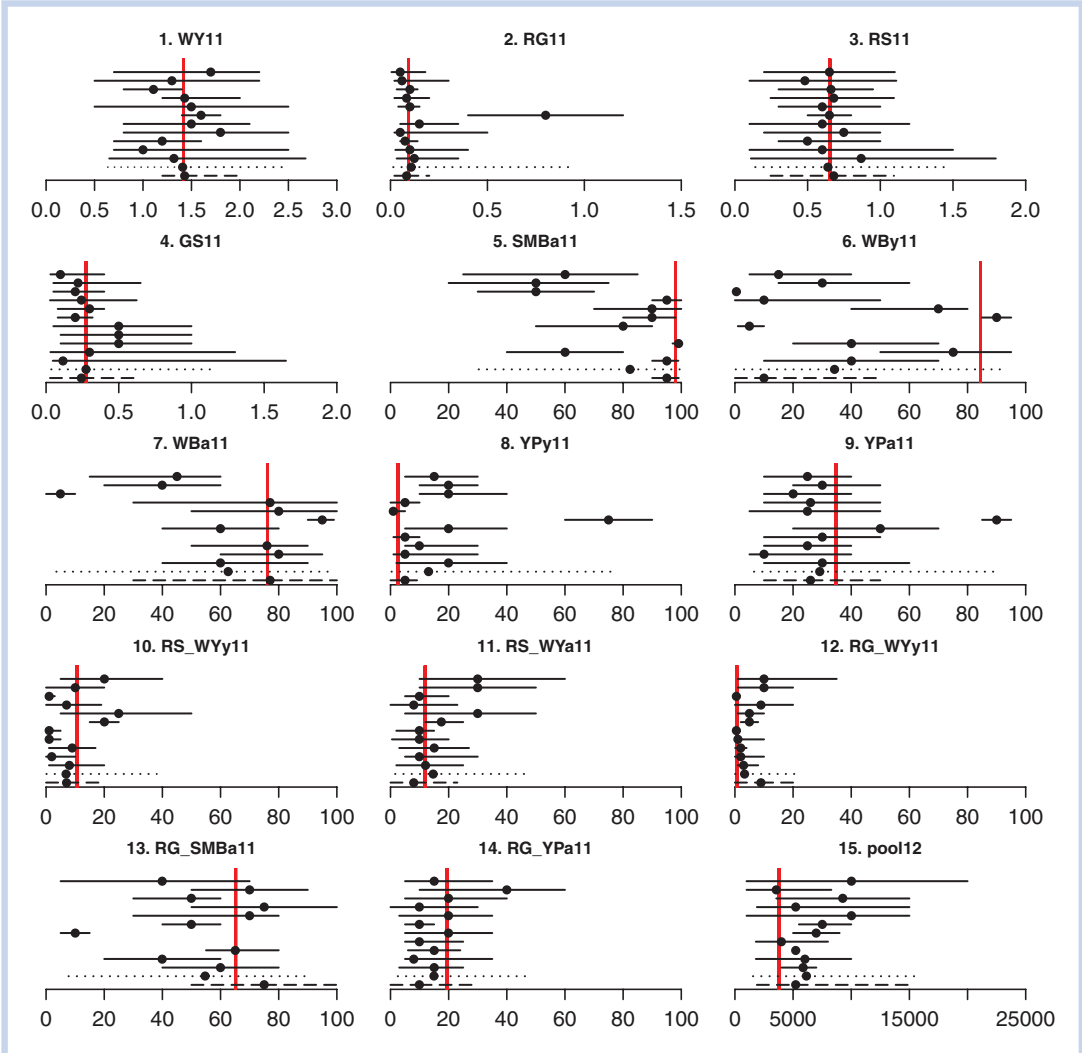

Figure 1. Range graphs for assessments of the calibration variables (variable numbering and associated acronyms are given in the SD_A). Solid lines denote the individual experts, endpoints are the 5 and 95 percentiles, dots denote the medians. The red vertical lines indicate the 2011 realizations of each calibration variable. Dotted lines denote the equal weight decision maker and dashed lines indicate the performance weight decision maker.

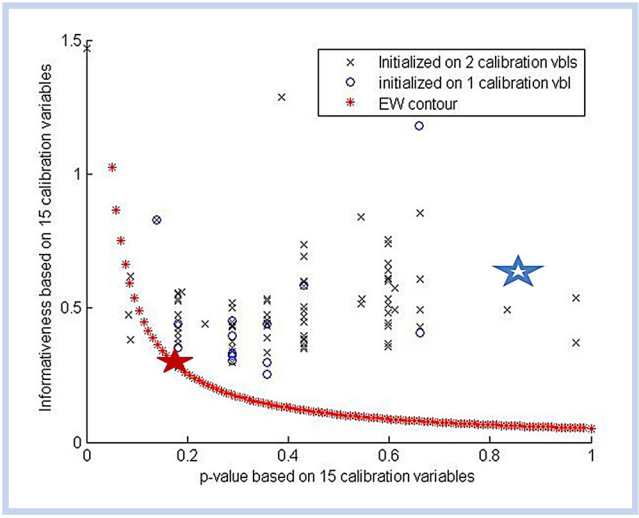

Figure 2. Cross validation results in which the global weight performance based DM is initialized on all subsets of 1 or 2 calibration variables and used to predict all 15 calibration variables. The solid star is the equal weight DM, and the curve consists of combinations of calibration and informativeness resulting in the same combined score as the equal weight DM. The hollow star is the performance DM for all calibration variables. Statistical accuracy and informativeness for EW DM are 0.1823 and 0.2808; for PW DM these are 0.6610 and 0.8562.

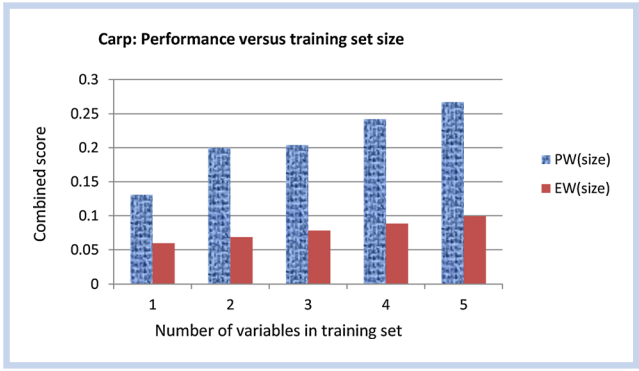

Figure 3. Combined scores of PW and EW averaged over size of training set, for sizes 1 to 6.

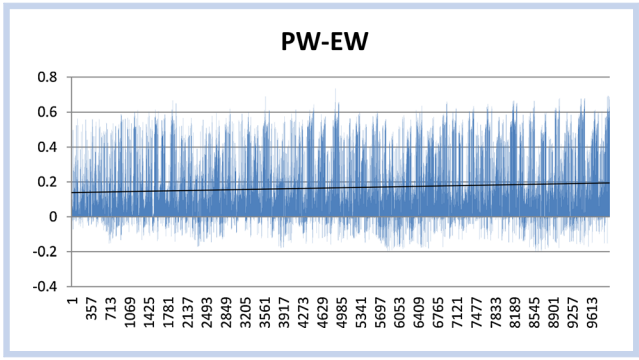

Figure 4. Differences of PW, EW scores for Asian carps study, based on 1 to 6 training variables out of 15 calibration variables, with trendline.

studies can make do with 1 or 2 calibration variables, as this result is attained by averaging over many sparse training sets. The present data set allows us to gain further insights in "sparsely trained PW DM's."

Out-of-sample validation for Asian carp study

For this exercise, expert 8 was removed, as he assessed only 11 calibration variables. All scores are now based on 15 calibration variables, and the scores are somewhat different than those in Table 1, where the statistical power of 11 calibration variables is used. Focusing first on training sets of size K = 1 and K = 2, Figure 2 compares the calibration and information scores of these very sparsely trained decision makers. Echoing results of Eggstaff et al. (2013) this shows 1) that the statistical accuracy and informativeness scores of the PW DM do not predict those in the original study, but 2) do outperform the EW DM, and 3) show considerable scatter. The Asian carp expert panel involved a relatively large number of high scoring experts; it remains to be seen if similar cross validation results emerge from other expert panels. To enable the comparison in Figure 1, the cross validation comparisons are also based on all 15 seed variables, not just 14 (initialized on 1 variable) or 13 (initialized on 2 variables).

Figure 3 shows the PW and EW combined scores averaged over each training set size from 1 to 6. There are 14 training sets of size 1, there are 15-choose-2 = 105 training sets of size 2, and so forth, going up to 5005 training sets of size 6.

Both the PW and EW scores increase with the training set size, reflecting the diminishing power of the test set. The EW DM is the same in all cases, as the weights do not depend on the training set. The rise in scores is purely a result of decreasing statistical power of the test set. However, PW increases faster than EW, suggesting a gain in performance from increased power of the training set over and above the power reduction of

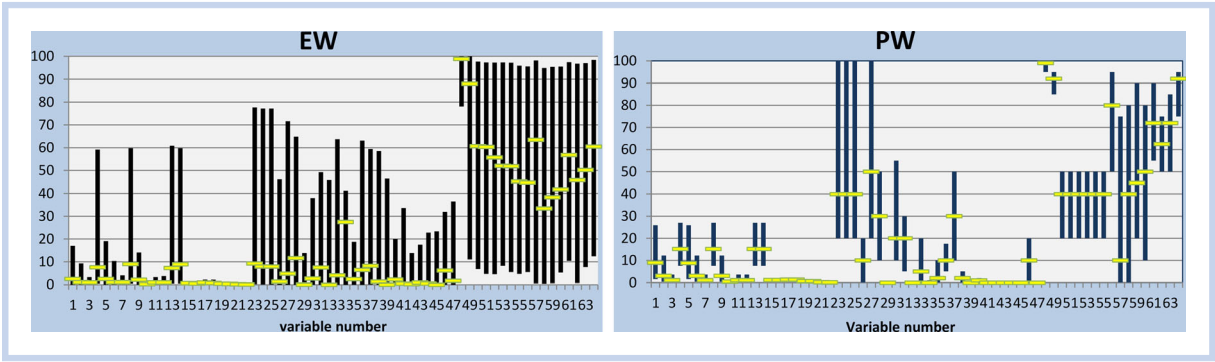

Figure 5. EW (left) and PW (right) quantiles for all variables of interest. Dots indicate median values, lines denote the 5% to 95% intervals. The physical dimensions vary between variables, but for a given variable, the dimensions for EW and PW are the same. Variables 1 through 14 concern biomass, consumption and production of Bighead and Silver carp, variables 15 through 66 concern biomass of native species after establishment, and dietary fractions.

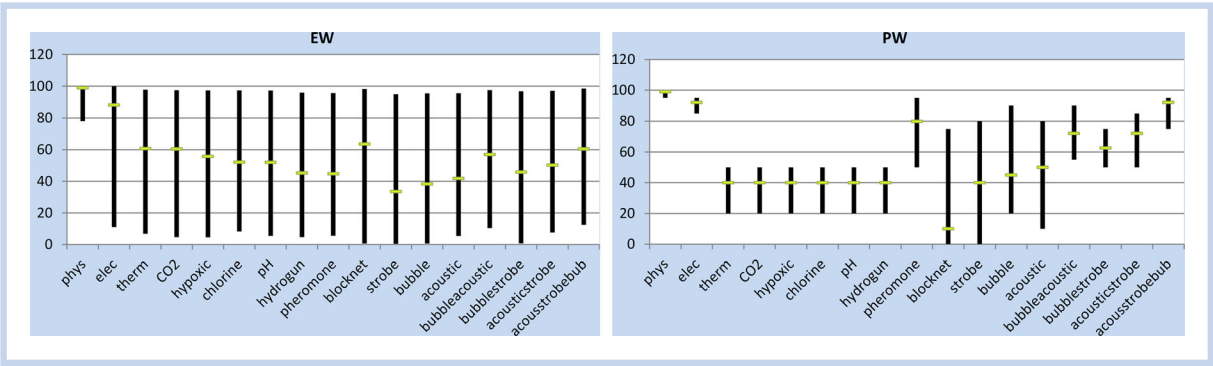

Figure 6. Equally weighted (EW; left) and performance-based (PW; right) expert assessments of the percentage of Asian carp prevented access to the Great Lakes as a result of implementing 17 proposed fish deterrent strategies in the Chicago Area Waterway System. Dots indicate median; lines denote the 5% and 95% intervals. For definitions of acronyms see SD_B or (Wittmann et al. 2014a).

the test set. Figure 4 compares the PW and EW differences, for each of the 9949 training sets of size 1 to 6.

The comparisons are indexed by size of the training set; with size 1 corresponding to indices 1 to 15 and size 6 corresponding to indices 4944 to 9949. Near the origin, values of PW-EW tend to be smaller, as is expected. The superiority of PW over EW in out-of-sample cross validation is evident. However, the differences in the cross validation comparisons tend to be smaller than the in-sample differences shown in Table 1, where the difference PW-EW is 0.56.

VARIABLES OF INTEREST

Variables relating to the biomass of Asian carps following establishment in Lake Erie are treated in depth in (Wittmann, Cooke, Rothlisberger, Rutherford 2014). Variables relating to the effectiveness of deterrent strategies are discussed in (Wittmann, Cooke, Rothlisberger, Lodge 2014). Present purposes are served by comparing the PW DM and EW DM for these 2 classes of variables of interest.

Figure 5 gives a schematic picture of the EW DM and PW DM. The variable numbering is the same as in the elicitation protocol in SD_B, from which the precise definitions may be retrieved. The scaling on the vertical axes is artificial in the sense that some variables are percentages, others are kilograms per square meter, etc. However, variables with the same number are scaled in the same way.

Figure 5 shows that the PW DM is usually—not always—more informative than the EW DM. Moreover, their median assessments can also vary substantially.

Figure 6 from Wittmann, Cooke, Rothlisberger, Lodge (2014) shows similar information for the 17 Asian carp deterrent strategies assessment. As with Figure 5, the PW DM is substantially more informative than the EW DM. For the most promising strategies, physical separation (SEP) and electric barrier (ELE), the median effectiveness assessments are nearly identical. For others, this is not the case. Pheromone attractant and/or repellant appears promising to the PW DM but not to the EW DM. For physical block net the reverse holds, the PW DM is highly skeptical whereas EW DM is not. The very large confidence bounds of the EW DM indicate that this DM has no pronounced opinion on any of these strategies except the first two.

CONCLUSIONS

This expert judgment study has demonstrated the experts' skill in performing probabilistic assessments and demonstrated the superiority of performance-based combinations of expert's judgments over equal weighting in this case. Although both performance-based weights and equal weights returned acceptable statistical performance, the performance-based combination was significantly more informative on both the calibration variables and on the variables of interest. Highly informative assessments are valuable only if these assessments are statistically accurate such that the narrower confidence bands are statistically defensible. The differences in the PW and EW DMs highlights the importance of validation.

Informative uncertainty distributions are very useful in making practical choices. For example, Figure 6 shows that the PW DM has a clear preference for 3 strategies (phys, elec, and acousstrobebulb), and this preference is based on the 90% confidence bands as well as on the median value. With the exception of SEP, the EW DM's confidence bands are so wide that they provide no practical value. Needless to say, narrow confidence bands are defensible only if their statistical accuracy is affirmed. In this case, based both on in-sample and out-of-sample validation, the statistical accuracy of the PW DM is actually better than that of the EW DM. Both forms of validation are important when structured expert judgment is used to quantify uncertainty.

A key question is whether the experts' performance on calibration variables will carry over to the variables of interest. These latter variables are not directly observable—otherwise we would not need expert judgment in the first place. Out-of-sample cross validation can be undertaken, whereby the performance weights are computed based on a subset of calibration variables (training set), and performance measured on the complementary set (test set). Using all nonempty subsets of calibration variables yields a super set of the relevant comparisons, but might be a very large set. With 25 calibration variables (not an unrealistic number) there are 33 554 431 training sets; computational advantages are achieved with smaller training sets. A further advantage would result if small training sets were generally sufficient to attest out-of-sample validity, as this would justify reducing the numbers of calibration variables. The present study provides some evidence for that conclusion but further assessments of other SEJ study outcomes will serve to more fully understand this issue.

SUPPLEMENTAL DATA

SD_A gives mathematical details.
SD_B is the elicitation protocol.
SD_C is the briefing booklet.

REFERENCES

[ACRCC] Asian Carp Regional Coordinating Committee. 2012. FY 2012 Asian carp control strategy framework. [cited 2013 February]. Available from: http://www.asiancarp.us/documents/2012Framework.pdf

Apostolakis G. 1990. The concept of probability in safety assessments of technological systems. *Science* 250:1359–1364.

Aspinall WP. 2010. A route to more tractable expert advice. *Nature* 463:294–295.

Bamber JL, Aspinall WP. 2013. An expert judgment assessment of future sea 1evel rise from the ice sheets. *Nature Clim Change* Available from: DOI: 10.1038/NCLIMATE1778

Burgman M. 2005. Risk and decisions for conservation and environmental management. Cambridge, UK: Cambridge Univ Press. pp 1–100.

Clemen RT. 2008. Comment on Cooke's classical method. *Reliab Eng Syst Safe* 93:760–765.

Cooke RM. 1991. Experts in uncertainty. Oxford, UK: Oxford Univ Press.

Cooke RM. 2008. Response to comments. Special issue on expert judgment. *Reliab Eng Syst Safe* 93:775–777.

Cooke RM. 2012. Pitfalls of ROAt cross validation comment on effects of overconfidence and dependence on aggregated probability judgments. *J Model Manage* 7:20–22.

Cooke RM. 2014. Validating expert judgments with the classical model in experts and consensus in social science—Critical perspectives from economics, sociology, politics, and philosophy. In: Martini C, Boumans M, editors. Ethical economy—Studies in economic ethics and philosophy. Springer.

Cooke RM, Goossens LHJ. 2000. Procedures guide for structured expert judgment in accident consequence modelling. *Radiat Prot Dosim* 90:303–309.

Cooke RM, Goossens LHJ. 2008. TU Delft expert judgment data base. Special issue on expert judgment. *Reliab Eng Syst Safe* 93:657–674.

Eggstaff JW, Mazzuchi TA, Sarkani S. 2013. The effect of the number of seed variables on the performance of Cooke's classical model. *Reliab Eng Syst Safe* 121:72–82.

Evans JS, Wilson A, Tuomisto JT, Tainio M, Cooke RM. 2005. What risk assessment can tell us about the mortality impacts of the Kuwaiti oil fires. *Epidemiology* 16:S137–S138.

Flandoli F, Giorgi E, Aspinall WP, Neri A. 2011. Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliab Eng Syst Safe* 96:1292–1310.

Fuller PL, Nico LG, Williams JD. 1999. Nonindigenous fishes introduced into inland waters of the United States. Special publication 27. Bethesda (MD): American Fisheries Society.

Garvey J, Ickes B, Zigler S. 2010. Challenges in merging fisheries research and management: The upper Mississippi River experience. *Hydrobiologia* 160:125–144.

Klugel JU. 2011. Uncertainty analysis and expert judgment in seismic hazard analysis. *Pure Appl Geophys* 168:27–53.

Kolar, CS, Chapman DC, Courtenay WR, Housel CM, Williams JD, Jennings DP. 2007. Bigheaded carps: A biological synopsis and environmental risk assessment. American Fisheries Society Special Publication 33, Bethesda, Maryland.

Lenton TM, Held H, Kriegler E, Hall JW, Lucht W, Rahmstorf S, Schellnhuber HJ. 2008. Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences* 105:1786–1793.

Lin S-W, Cheng C-H. 2008. "Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities?" Taiwan Proceedings of the 2008 IEEE IEEM, 8-11 Dec. 2008, p 425–429. Available from: DOI 10.1109/IEEM.2008. 4737904

Lin S-W, Cheng C-H. 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. *J Model Manage* 4:149–161.

Martin TG, Burgman MA, Fidler FF, Kuhnert PM, Low-Choy S, McBride M, Mengersen K. 2011. Eliciting expert knowledge in conservation science. *Conserv Biol* 26:29–38.

Morgan MG, Pitelka LF, Shevliakova E. 2001. Elicitation of expert judgments of climate change impacts on forest ecosystems. *Clim Change* 49:279–307.

Qing X. 2002. Risk analysis for real estate investment [PhD thesis]. Delft, the Netherlands: Delft Univ of Technology.

Roman HA, Walker KD, Walsh TL, Conner L, Richmond HM, Hubbell BJ, et al. 2008. Expert judgment assessment of the mortality impact of changes in ambient fine particulate matter in the US. *Environ Sci Technol* 42:2268–2274.

Rothlisberger JD, Lodge DM, Cooke RM, Finnoff DC. 2010. Future declines of the binational Laurentian Great Lakes fisheries, the importance of environmental and cultural change. *Front Ecol Environ* 8:239–244.

Rothlisberger JD, Finnoff DC, Cooke RM, Lodge DM. 2012. Ship-borne nonindigenous species diminish Great Lakes ecosystem services. *Ecosystems* 15:462–476.

Teck SJ, Halpern BS, Kappel CV, Micheli F, Selkoe KA, Crain CM, Martone R, Shearer C, Arvai J, Fischhoff B., et al. 2010. Using expert judgment to estimate marine ecosystem vulnerability in the California current. *Ecol Appl* 20:1402–1416.

Tuomisto JT, Wilson A, Cooke RM, Tainio M, Evans JS. 2005. Mortality in Kuwait due to PM from oil fires after the Gulf War. Combining expert elicitation assessments. *Epidemiology* 16:S74–S75.

Van Overbeek FNA. 1999. Financial experts in uncertainty [Masters thesis]. Delft, the Netherlands: Delft Univ of Technology.

Wittmann ME, Cooke RM, Rothlisberger JD, Lodge DM. 2014. Using structured expert judgment to assess invasive species prevention: Asian carp and the Mississippi-Great Lakes hydrologic connection. *Environ Sci Technol* 48:2150–2156.

Wittmann ME, Cooke RM, Rothlisberger JD, Rutherford ES, Zhang H, Mason D, Lodge DM. 2014. Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. *Conserv Biol* Available from: DOI: 10.1111/cobi.12369