Identifying non-additive multi-attribute value functions based on uncertain indifference statements

Omega 85 (2019) 49–67

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Identifying non-additive multi-attribute value functions based on uncertain indifference statements☆

Fridolin Haag[a,b,*], Judit Lienert[a], Nele Schuwirth[a], Peter Reichert[a,b]

[a] Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland
[b] ETH Zürich, Institute of Biogeochemistry and Pollutant Dynamics, Universitätstrasse 16, 8092 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 30 August 2017
Accepted 20 May 2018
Available online 26 May 2018

Keywords:
decision making/process
preference modeling
multi-attribute value theory
uncertainty
aggregation
environmental assessment

ABSTRACT

Multi-criteria decision analysis (MCDA) requires an accurate representation of the preferences of decision-makers, for instance in the form of a multi-attribute value function. Typically, additivity or other stringent assumptions about the preferences are made to facilitate elicitation by assuming a simple parametric form. When relaxing such assumptions, parameters cannot be elicited easily with standard methods. We present a novel approach for identifying multi-attribute value functions which can have any shape. As preference information indifference statements are used that can be elicited by trade-off questions. Instead of asking one indifference statement for each pair of attributes, we ask for multiple trade-offs at different points in the attribute space. This allows inferring parameters of complex value functions despite the simplicity of the preference statements. Parameters are estimated by taking into account preference and elicitation uncertainty with a probabilistic model. Statistical inference supports identifying the most adequate preference model out of several candidate models through quantifying the uncertainty and assessing the need for non-additivity. The approach is elaborated for determining value functions by hierarchical aggregation. We apply it to an assessment of the ecological state of rivers, which is used to support environmental management decisions in Switzerland. Preference models of four experts were quantified, confirming the feasibility of the approach and the relevance of considering non-additive functions. The method suggests a promising direction for improving the representation of preferences.

1. Introduction

Multi-criteria decision analysis requires evaluating decision alternatives across multiple objectives. A multi-attribute value function (MAVF) can be used to determine the overall degree of achievement across these objectives by evaluating this function at the predicted outcomes of the alternatives [1].[1] This kind of decision support relies on constructing a MAVF as a model of the decision-makers' preferences.

Several methods to specify a MAVF and estimate its parameters exist. Generally, they differ with regard to (a) which preference information is used, (b) which *a priori* restrictions are put on the preference model, i.e., what types of MAVF are considered, (c) how the MAVF is constructed or inferred, and (d) how uncertainty in preference information is handled.

Regarding (a) and (c), much progress has been made concerning the flexibility in used preference information and the robustness of methods to infer value functions [e.g., 4–6]. Regarding (b), in most cases the additive model is assumed [e.g., 7]. A fundamental property of additive value functions is the possibility for compensation (substitutability): If the weights of objectives are equal, a decrease in performance of one objective can be compensated by the same increase in performance of another objective – if this is possible within the ranges. If weights are not equal, an analogous statement holds when including a correction for the weights. Additivity requires preferential independence [8], i.e., the objectives' evaluations are not interacting. In practice, this does not necessarily hold [e.g., 9–12], which has created interest in alternative preference models. For modeling interacting objectives, fuzzy integrals, specifically the Choquet integral, have been a focus in the last two decades [e.g., 13–16]. Uncertainty in preferences (d) is often dealt with *a posteriori*, for instance by using stochastic multiobjective acceptability analysis [17] or sensitivity analysis [e.g., 18,19].

Without questioning existing methods, shortcomings exist in two regards. Firstly, preference information of decision-makers is

☆ This manuscript was processed by Associate Editor C. Chen.

* Corresponding author.
E-mail addresses: fridolin.haag@eawag.ch (F. Haag), judit.lienert@eawag.ch (J. Lienert), nele.schuwirth@eawag.ch (N. Schuwirth), peter.reichert@eawag.ch (P. Reichert).

[1] If decision outcomes are uncertain, a multi-attribute utility function should be used. Since utility functions can be derived from value functions by adding the risk attitude, identifying a value function is also useful for these cases [2,3].

https://doi.org/10.1016/j.omega.2018.05.011

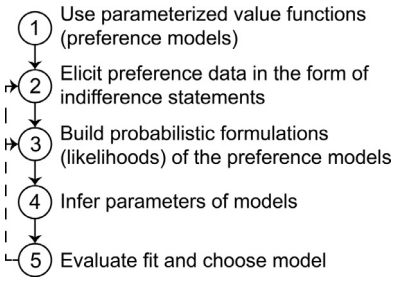

Fig. 1. Proposed procedure for identifying multi-attribute value functions of any shape.

always uncertain due to the inherent uncertainty of personal preferences and the uncertainty induced by the elicitation process. Yet, this uncertainty is often not dealt with explicitly and consistently in decision analysis. Secondly, preferences should not be forced to adhere to one specific form of the MAVF. Yet, often strict restrictions are placed on the shapes of MAVF to facilitate elicitation and estimation. We lack general methods for identifying MAVF that are not bound to specific functional forms and are feasible in practice regarding the elicitation process and the reliability of results.

In this paper we address these issues. We present a novel approach for constructing MAVF based on ideas from statistical learning. The approach is characterized by an explicit treatment of uncertainty and the usability with arbitrary forms of MAVF. Since practical feasibility of elicitation and inference are essential concerns, we illustrate the approach with a real-world assessment for river management.

In the following, we first introduce the general method for identifying MAVF based on uncertain indifference statements (Section 2). Subsequently, we specify this method for the hierarchical construction of a MAVF by aggregation (Section 3). We demonstrate its practicability with a case study (Section 4), before discussing implications (Section 5) and drawing conclusions (Section 6).

2. Method for identifying value functions in a probabilistic framework

A multi-attribute value function (MAVF) is a function that returns the degree of achievement of an objective as a function of attribute levels that characterize potential outcomes of decision alternatives. In this sense the MAVF is a model of the preferences of a decision-maker. If we know the MAVF and the consequences of decision alternatives, we can easily assess alternatives by evaluating the MAVF at the predicted attribute levels and solve ranking and choice problems.

More formally, a finite set of alternatives $\mathcal{A} = \{a,\ b,\ \ldots\}$ shall be evaluated with regard to an objective o. In an objectives hierarchy (see Section 3) this objective can be a higher-level objective with several sub-objectives. Each alternative $a \in \mathcal{A}$ is associated with predicted outcomes $\boldsymbol{x}_a = (x_{a,1}, \ldots, x_{a,n})$, with $x_{a,i}$ being the level of an attribute i that measures a consequence of alternative a. Let $X_i = [x_i^{\text{lb}},\ x_i^{\text{ub}}]$ be the set of potential outcomes for attribute i (interval bounded by a lower bound, lb, and an upper bound, ub). Note that intervals and even measurement units for different attributes usually differ as they quantify different aspects of the outcome. Let X denote the set of all possible outcomes for all attributes in a decision. We are interested in finding a MAVF $v: X \rightarrow [0, 1]$ that represents the evaluation – or degree of achievement – of the objective o for potential outcomes of the decision alternatives. To obtain such a MAVF, we propose a procedure based on five steps (Fig. 1), as detailed in Sections 2.1–2.5.

2.1. Using a parameterized function as value function

In our context, a value function can be any parameterized function $v(x_1, \ldots, x_n, \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$, if it maps decision outcomes to a value representing the decision-maker's preferences when choosing appropriate parameter values. In the following, only interval scale (measurable) value functions [8] are considered. A description of preferences by such a value function is possible if the preference for alternatives' outcomes and the preference for transitions between outcomes are complete and transitive orderings [20].

The value function can be additive or non-additive. An additive value function has the form $v(x_1, \ldots, x_n, \boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \cdot v_i(x_i)$ with parameters $\boldsymbol{\theta} = (w_1, \ldots, w_n)$. It only is a valid representation of preferences if mutual preferential independence of attributes is given and in the case of measurable value functions also difference independence [8]. If preferential independence is violated, i.e., there is interaction between objectives, non-additive value functions should be used. A non-additive value function can have any form except the additive one, as long as it adequately describes the preferences.

Commonly, one particular preference model is assumed *a priori* and subsequently its parameters are estimated based on preference information. We argue that this choice should not be made *a priori* to avoid restricting the expression of preferences. Selecting an appropriate model should be based on how well alternative models fit to the preference statements, balanced with model complexity. In this spirit, we explicitly start with several alternative preference models to then select the best.

2.2. Eliciting preference information

2.2.1. Indifference statements

Various types of preference information can be used for constructing value functions. These include holistic judgments or rankings of alternatives, pairwise comparisons of decision outcomes, or inter-objective comparisons such as rankings of importance, weights, or interaction between objectives [e.g., 4,5,14,21–23].

Indifference statements between pairs of decision outcomes (a, b) are a further possibility [2]. Each of the outcomes is characterized by a vector of attribute levels $\boldsymbol{x}_a = (x_{a,1}, \ldots, x_{a,n})$. In a classical trade-off question these levels are manipulated until indifference between the pair is reached: $a \sim b$ [1]. In comparison to statements about simple preference, such as $a \succ b$, indifference points provide more information [24]. This allows robust inference with fewer questions. A disadvantage is that trade-off questions are more difficult to answer than statements about simple preference and are thus more uncertain.

Trade-offs have a strong axiomatic foundation and the advantage of directly representing a property of the MAVF: When a decision-maker is indifferent between two potential decision outcomes, a measurable MAVF should have the same value for both. Trade-offs may change depending on the starting position in the value space [1]. This allows determining level-dependent interactions (e.g., non-constant substitution rates).

Behavioral aspects play a paramount role for eliciting meaningful preference statements. Two important factors are task complexity and task type. Task complexity here refers to the number of attributes that need to be considered when making an indifference statement. For choice experiments it has been found that with increasing numbers of attributes, non-attendance to (i.e., ignoring of) attributes increases [e.g., 25]. This is likely analogous for indifference statements and limits the dimensionality of meaningful trade-off questions to few attributes – or a large number of uncertain answers are needed for reliable parameter inference. However, with

a hierarchical approach (Section 3) larger decision problems can be tackled.

Repeated choice tasks (i.e., pairwise comparisons, deciding which outcome is preferable), matching tasks (i.e., adjusting one outcome until it is equally good as the other), or combinations – such as letting the decision-maker suggest an indifference state [26] – can be used for eliciting indifference statements. Since procedural invariance is not given, there is a long and ongoing debate which type of task leads to more reliable, unbiased, and valid preference statements [e.g., 27–29]. One phenomenon and potential bias is the prominence effect which suggests that in choices the more important dimension looms larger than in matching [24,27]. Loss aversion – "losses loom larger than corresponding gains" [30] – is another potential bias, but the dependence on the task type is still debated [e.g., 24,28,29,31].

Acknowledging the importance of behavior in elicitation, for the case study application we have decided to use matching tasks and compare a maximum of three attributes at once.

2.2.2. Determining an elicitation scheme

To be sensitive to non-additivity, we need to ask trade-offs at different points in the attribute space for each attribute combination. As we can ask a decision-maker only a limited number of questions, the choice of questions becomes a key issue. The elicitation scheme should allow estimating the parameters of the preference model well for a feasible number of questions and allow discriminating between alternative models.

The theory on experimental design makes it possible to formalize such requirements as optimality criteria, which can be used for finding an optimal elicitation scheme [see 32 for an overview]. Alternatively, an adaptive or flexible approach to determining elicitation questions is possible [e.g., 33–36].

Finding an optimal design for non-linear models which also allows model discrimination is a non-trivial problem, as the optimal design depends on model parameters (which are not known *a priori*) and maximizing model dissimilarity is an additional difficulty [37]. With many unresolved challenges remaining, e.g., concerning the formulation of optimality criteria, algorithms, and computation, we decided for a pragmatic approach: We determined our elicitation scheme in a simulation study where we tested different schemes with artificial answers (see Section 4.3.1).

2.3. Probabilistic framework for parameter estimation: Building an observation model

We propose to use a probabilistic framework for identifying a MAVF. Preference statements always contain uncertainty. Conceptually, this uncertainty can be separated into uncertainty of the decision-maker about her preferences and uncertainty due to the elicitation process ("observation error"), although empirically these are usually mixed. Therefore, the preference statements are treated as uncertain information rather than hard constraints. When formulating the uncertainties probabilistically, standard frequentist or Bayesian statistical inference techniques can be applied for estimating the parameters of the MAVF, their uncertainty, and the resulting uncertainty of the value function. Quantifying the uncertainty of the parameter estimates is important to assess the significance of aspects described by them, e.g., of non-additivity.

2.3.1. Formalizing indifference statements

In each question we ask for an indifference statement between two potential outcomes which are each characterized by a vector of attribute levels $(x_1, \ldots, x_n)$. One of these potential outcomes is completely specified and denoted by reference point, r: $\boldsymbol{x}^r = (x_1^r, \ldots, x_n^r)$. The other potential outcome is denoted by question point, q, and is missing the specification of the attribute i: $\boldsymbol{x}^q_{-i} = (x_1^q, \ldots, x_{i-1}^q, x_{i+1}^q, \ldots, x_n^q)$. The decision-maker is asked to provide the level of the attribute i, x_i^q, for which indifference in her preferences would be reached between the two potential outcomes: $(x_1^q, \ldots, x_{i-1}^q, x_i^q, x_{i+1}^q, \ldots, x_n^q) \sim \boldsymbol{x}^r$ – or to express the inability of reaching indifference within the range of the attribute i. Assuming a MAVF v with parameters $\boldsymbol{\theta}$, this translates to $v(x_1^q, \ldots, x_{i-1}^q, x_i^q, x_{i+1}^q, \ldots, x_n^q, \boldsymbol{\theta}) \stackrel{!}{=} v(\boldsymbol{x}^r, \boldsymbol{\theta})$. Solving for x_i^q we obtain the "correct" attribute level according to the model as $x_i^{q,v}(\boldsymbol{\theta}, \boldsymbol{x}^q_{-i}, \boldsymbol{x}^r)$. If the decision-maker were perfectly consistent with the preference model v and the parameter values $\boldsymbol{\theta}$, she would provide this answer for x_i^q. This model attribute level is thus defined by the implicit equation

$$x_i^{q,v}\left(\boldsymbol{\theta}, \boldsymbol{x}^q_{-i}, \boldsymbol{x}^r\right): \; v\left(x_1^q, \ldots, x_{i-1}^q, x_i^{q,v}\left(\boldsymbol{\theta}, \boldsymbol{x}^q_{-i}, \boldsymbol{x}^r\right), x_{i+1}^q, \ldots, x_n^q, \boldsymbol{\theta}\right) = v\left(\boldsymbol{x}^r, \boldsymbol{\theta}\right) \quad (1)$$

2.3.2. Assumptions about the error model

To account for the uncertainties in the preferences and the elicitation process, we assume the replies to be distributed normally around the deterministic solution provided by Eq. (1). We experimented with other distributions (e.g., beta distributions) but found them dissatisfying as they do not allow answers to fall outside the considered attribute ranges (Section 2.3.3). However, other distributions can be used by adapting the probabilistic model (Eq. (2)). Secondly, we assume the errors of the answers to different questions to be independent. Thirdly, we assume the variances ${\sigma_i}^2$ of the replies regarding the attribute i to be constant in the whole attribute space. It is important to allow answers to be outside the bounds $[x_i^{\text{lb}}, x_i^{\text{ub}}]$. This also eliminates the need for a decreasing variance at the interval bounds, as would be the case if the answers would be restricted to the interval. However, extension to non-constant variance is possible.

These assumptions seem very stringent. However, they lead to the simplest probability model of the "observation process" and we would need empirical evidence for its rejection to justify a more complex approach. In our case study, we statistically tested the assumptions (see Section 4.4). However, due to the small sample sizes, these tests do not have a high power.

2.3.3. Dealing with situations with no indifference point within given attribute ranges

Attribute levels are restricted to certain ranges, $X_i = [x_i^{\text{lb}}, x_i^{\text{ub}}]$. These are either defined by the decision context (alternatives under consideration will not lead to outcomes outside these ranges) or represent natural bounds of the attribute (e.g., substance concentrations cannot be negative). MAVF are defined over this range. However, it may happen that the decision-maker cannot provide an indifference point within the range. She would be indifferent when the attribute level is outside the bounds, or she cannot reach indifference at all when modifying this single attribute. Therefore, decision-makers were given the possibility to make such a statement for every question.

To use this preference information, we explicitly incorporated it in our probabilistic model (Eq. (2)). Depending on the parameter values, we distinguish two cases: If Eq. (1) can be solved – whether the model solution lies within or outside the range $[x_i^{\text{lb}}, x_i^{\text{ub}}]$ – we calculate either the probability density of the response within $[x_i^{\text{lb}}, x_i^{\text{ub}}]$ or the probability for the response being above or below this interval, depending on the observed value of the response i at which the probabilistic model is evaluated. If Eq. (1) cannot be solved, we reject the parameter values as inadequate.

2.3.4. Probabilistic model

To formulate the probabilistic model for multiple questions, we extend our notation by introducing an index j to refer to the question. Consequently, we denote the reference and question points

for question j by $\boldsymbol{x}^{r(j)}$ and $\boldsymbol{x}^{q(j)}_{-i(j)}$, and the sets of all reference and question points by $\boldsymbol{x}^{\boldsymbol{r}}$ and $\boldsymbol{x}^{\boldsymbol{q}}_{-\boldsymbol{i}}$, respectively. Finally, we denote the single reply to question j by $x^{q(j)}_{i(j)}$ and the set of all replies by $\boldsymbol{x}^{\boldsymbol{q}}_{\boldsymbol{i}}$.

Given our assumptions about the error model and given the preference model v, its parameters, the standard deviations σ_i, and the reference and question points, the probability distribution for receiving the answers that the decision-maker gave can be written as:

$$L_v\left(\boldsymbol{x}^{\boldsymbol{q}}_{\boldsymbol{i}}\middle|\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{x}^{\boldsymbol{q}}_{-\boldsymbol{i}},\boldsymbol{x}^{\boldsymbol{r}}\right)=\prod_j f\left(x^{q(j)}_{i(j)}\middle|\boldsymbol{\theta},\sigma_{i(j)},\boldsymbol{x}^{q(j)}_{-i(j)},\boldsymbol{x}^{r(j)}\right) \tag{2}$$

with

$$f\left(x^q_i\middle|\boldsymbol{\theta},\sigma_i,\boldsymbol{x}^q_{-i},\boldsymbol{x}^r\right)$$
$$=\begin{cases}\dfrac{1}{\sigma_i\sqrt{2\pi}}\cdot\exp\left[-\dfrac{\left(x^q_i-x^{q,v}_i\left(\boldsymbol{\theta},\boldsymbol{x}^q_{-i},\boldsymbol{x}^r\right)\right)^2}{2{\sigma_i}^2}\right] & \text{for } x^{\mathrm{lb}}_i\le x^q_i\le x^{\mathrm{ub}}_i\\ \dfrac{1}{\sigma_i\sqrt{2\pi}}\cdot\displaystyle\int_{-\infty}^{x^{\mathrm{lb}}_i}\exp\left[-\dfrac{\left(t-x^{q,v}_i\left(\boldsymbol{\theta},\boldsymbol{x}^q_{-i},\boldsymbol{x}^r\right)\right)^2}{2{\sigma_i}^2}\right]dt & \text{for } x^q_i<x^{\mathrm{lb}}_i\\ 1-\dfrac{1}{\sigma_i\sqrt{2\pi}}\cdot\displaystyle\int_{-\infty}^{x^{\mathrm{ub}}_i}\exp\left[-\dfrac{\left(t-x^{q,v}_i\left(\boldsymbol{\theta},\boldsymbol{x}^q_{-i},\boldsymbol{x}^r\right)\right)^2}{2{\sigma_i}^2}\right]dt & \text{for } x^q_i>x^{\mathrm{ub}}_i\end{cases}$$

This model depends on the value function v through the solutions $x^{q,v}_i$ of the implicit Eq. (1). The three cases are needed to account for answers outside the range, i.e., $x^q_{i(j)}<x^{\mathrm{lb}}_i$ or $>x^{\mathrm{ub}}_i$ (Section 2.3.3). The probability of the answer being outside the specified range is given by integrating the probability density function over the corresponding range (either $(-\infty,x^{\mathrm{lb}}_i)$ or $(x^{\mathrm{ub}}_i,\ \infty)$). Eq. (2) becomes the likelihood function of our model parameters once the replies, $\boldsymbol{x}^{\boldsymbol{q}}_{\boldsymbol{i}}$, have been substituted into the expression.

2.4. Inferring parameters of preference models

Parameter estimation can be formulated as an optimization problem under constraints. One important distinction between different approaches concerns the definition of the performance measure or optimality criterion. In Operations Research, such problems are classically formulated as minimizing a cost or error term [e.g., 14,22,23]. Within a probabilistic framework, we use a statistical measure as optimality objective [e.g., 22].

Substituting the replies of the decision-makers ("observations") into the probabilistic model given by Eq. (2) leads to the likelihood function that builds the basis for frequentist or Bayesian inference of the parameters of the MAVF, $\boldsymbol{\theta}$, and the error model, $\boldsymbol{\sigma}$. The strength of frequentist techniques is to get parameter estimates without the need of specifying prior knowledge ("prejudice"), but they require sufficient data to achieve parameter identifiability. In contrast, Bayesian techniques combine prior knowledge with the information gained from (new) data. Which technique is favorable depends on the availability and desire to use prior knowledge. In our application case, we chose to estimate the parameters in a frequentist context by maximum likelihood estimation.

In this technique, the best set of parameters is obtained by maximizing the likelihood function, i.e., the probability of the results evaluated at the elicited indifference levels. Parameter inference was implemented in R [38]. The problem was solved numerically, using the Rsolnp package [39] for optimization. It implements an augmented Lagrange barrier minimization algorithm [40] which was used to minimize the negative log-likelihood of our function under constraints. Constraints were that parameters were kept in the domain of their definition. Uncertainty of the parameter estimates can be inferred using bootstrapping [41,42], see Section 2.5.1. Parameter inference was conducted separately for each competing model.

2.5. Evaluating results

2.5.1. Quality of parameter estimates

A decisive advantage of the probabilistic framework is that we can assess the uncertainty of our parameter estimates. This makes it possible to evaluate the quality of parameter inference, using variance and bias in the estimate as measures of quality. Furthermore, we can determine whether the parameters of the preference model make it significantly different from the additive model.

In our application we used bootstrapping [42] for this purpose and repeated the parameter estimation with different answer samples. These answer samples were created by drawing with replacement from the empirical distribution of the residuals of the decision-maker's answers and the model values and then adding the drawn residuals to the model values.[2] Variance in the parameter estimate can be determined by the standard error of the bootstrap sample estimates (measure for random error in the estimate). Bias in the estimate can only be determined when the true value is known, as in a simulation study (Section 4.3.1). Here, we calculated bias of the parameter estimator as the deviation of the estimated mean of the bootstrap sample from the true mean (measure for systematic deviation in the estimate).

2.5.2. Model selection

Selecting the best preference model requires balancing model fit with model complexity. This can be captured by different selection criteria. Classical approaches are based on information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). These criteria penalize high parameter numbers, as more parameters will usually lead to a better fit. When a more precise representation of model complexity is of interest, more sophisticated criteria such as normalized maximum likelihood [43] should be used. The choice of a selection criterion will depend on the application case.

In our application, we have chosen non-additive value functions that have the same parameters as the additive model (weights) complemented by one additional parameter that characterizes the curvature of the iso-lines or iso-hyper-surfaces. Therefore, the model complexity penalty-term in the information criteria cancels out and we would choose the model with the best fit. Since the considered models have a similar shape of the iso-lines or iso-hyper-surfaces, this argument is also valid when using a more sophisticated model complexity quantification.

Thus, we decided to select the model based on the estimated standard deviation, σ, and check for the significance of non-additivity using the uncertainty range of the curvature parameter. As the models have similar complexity, the only remaining concern is overfitting. The simulation study helped determining the number of replies needed to avoid overfitting.

3. Hierarchical construction of MAVF by aggregation

The method presented in Section 2 is general and makes no assumptions about the problem structure. In this section we specify how a multi-attribute value function (MAVF) can be obtained when the problem is structured by an objectives hierarchy. A hierarchical structure has the advantage of requiring preference statements only for a subset of attributes at the same time.

[2] When answers are outside the range, no residual can be calculated. In this case there are fewer residuals than model values. We drew from the empirical distribution with replacement. If model values were outside the range, still a drawn residual was added to them.

3.1. Objectives hierarchies

A MAVF returns an evaluation of decision outcomes. In the context of value-focused thinking [44], we are interested in an evaluation with respect to concrete decision objectives. In multi-criteria problems, an overall goal is specified by several objectives which establish an objectives hierarchy. This hierarchy can be built by two approaches, which in practice are often mixed. In a top-down approach we start with a general objective which is further characterized by sub-objectives. In a bottom-up approach individual objectives are clustered to become sub-objectives of higher-level concepts. It is not uncommon to have an objectives hierarchy with only one single level; this case is sometimes called "non-hierarchical".

The decomposition of a decision problem with a hierarchy has practical advantages and is one solution to the "curse of dimensionality". The hierarchical structure represents independence conditions that help disassembling the identification of a MAVF into smaller parts and thus considerably simplifies its construction [2]. Interactions are not considered between all objectives simultaneously, but only within the same branch and level, which also makes elicitation more feasible.

3.2. Obtaining a MAVF by hierarchical aggregation

Given an objectives hierarchy, the construction of the value function for the overall objective can be divided into two steps. Firstly, value functions for the lowest-level sub-objectives are elicited as functions of the attributes relevant for these sub-objectives. Typically, due to the narrow definition of these objectives, they just need a single attribute or a small number of attributes. Secondly, value functions for higher-level objectives are constructed by using an aggregation function which depends on the values of the underlying sub-objectives and combining it with the value functions of these sub-objectives.

Formalizing this construction, elementary objectives (lowest-level objectives) o_i are directly evaluated with respect to their attributes $\{x_k, \ldots, x_l\}$ with a value function $v_i(x_k, \ldots, x_l, \boldsymbol{\theta}_i)$ with parameters $\boldsymbol{\theta}_i$. This value function can be linear or non-linear and can be single-attribute or multi-attribute. There are established methods for constructing single-attribute value functions, such as the mid-value splitting method, see [1]. For MAVF the procedure introduced in Section 2 can be used as presented to determine this value function.

Higher-level objectives are evaluated by aggregating the evaluations of their respective sub-objectives and thus depend on the attributes indirectly. Such a MAVF over objectives $\{o_p, \ldots, o_q\}$ on a specific hierarchical level may generally be written as:

$$v_{p,q}\left(x_1, ..., x_n, \boldsymbol{\theta}\right) = F\left(v_p\left(x_1, \ldots, x_n, \boldsymbol{\theta}_p\right), \ldots, v_q\left(x_1, \ldots, x_n, \boldsymbol{\theta}_q\right), \boldsymbol{\theta}_{pq}\right) \tag{3}$$

In practice, each value function v_i will only depend on a subset of the attributes $\{x_1, \ldots .x_n\}$. The function F is an aggregation function. Its form and parameters also depend on preferences of the decision-maker. The resulting MAVF over objectives $\{o_p, \ldots, o_q\}$, therefore, is composed of lower-level (single- or multi-attribute) value functions v_i and an aggregation function F. Value functions depend on attributes, whereas aggregation functions depend on the values of sub-objectives. Aggregation functions are not themselves value functions. Only their combination with lower-level value functions leads to a value function for the higher-level objective.

If the parameters $\{\boldsymbol{\theta}_p, \ldots, \boldsymbol{\theta}_q\}$ of the underlying value functions are known, we only need to determine the functional form and parameters $\boldsymbol{\theta}_{pq}$ of F to construct the MAVF. In this case, we can easily apply the parameter estimation procedure presented in Section 2 only to the aggregation step. We replace attributes, $(x_1, \ldots, x_n)$, in Eqs. (1) and (2) by values, $(v_p, \ldots v_q)$, of the underlying objectives and then estimate the parameters of the aggregation function. This can be done at arbitrary hierarchical levels, as long as the lower-level value functions are known.

For a multi-level hierarchy, evaluations are aggregated upwards along the hierarchy in a step-wise manner (i.e., following each branch) until an overall evaluation is reached. In this case, multiple aggregation functions are nested according to the hierarchical structure of the problem.

Since the MAVF in Eq. (3), $v_{p,q}(x_1, \ldots, x_n, \boldsymbol{\theta})$, is a parameterized function as introduced in Section 2, we could estimate all its parameters – both of the individual underlying value functions and the aggregation function – at once by the method presented in Section 2. In practice, however, it is often sensible to determine parameters of lowest-level value functions and aggregation functions separately.

3.3. Aggregation functions

An aggregation function maps an arbitrarily long list of input values on the same scale to one single representative output value. For a real interval $[0, 1]$ that contains the values to be aggregated, an aggregation function in $[0, 1]^n$ is a function $F^{(n)}$: $[0, 1]^n \rightarrow [0, 1]$ that is nondecreasing in each argument and fulfills the boundary conditions $F^{(n)}(0, 0, \ldots, 0) = 0$ and $F^{(n)}(1, 1, \ldots, 1) = 1$ [45]. In the context of MAVF, aggregation functions operate on the value space, not the attribute space.

For aggregation functions to be meaningful in preference modeling, we consider certain properties useful: Aggregation functions should be continuous and idempotent, meaning $F(v, \ldots, v) = v, \ \forall \ v \in [0, 1]$ (if all input values are equal, the overall value is identical to these). Furthermore, they need to be meaningful for the type of scales we operate on. For ordinal value functions, many aggregation functions are not meaningful [11]. Lastly, behavioral properties of the functions are relevant, as we would like to model different preference structures. Also, they should make it possible to incorporate varying sensitivities to different sub-objectives (for example by weighting). For an in-depth treatment of aggregation functions we refer to Grabisch et al. [46] and Beliakov et al. [45].

Aggregation functions with diverse behavior exist [45,46]. Basic behaviors are complementarity (synergy), independence (substitutability), and redundancy. These extreme cases can be modeled by the minimum (Min), (weighted) arithmetic mean (AM), and maximum (Max) aggregation functions, respectively. The aggregation functions differ in how sensitive the result is to unequal input values. If all inputs are equal, they give the same result due to idempotency. Complementarity means punishing unequal inputs, as the aggregated value is smaller than with the AM. Redundancy means favoring unequal inputs, as the aggregated value is higher than with the AM. This can be formalized by measures of *andness* and *orness* which express how similar to Min and Max an aggregation function behaves [46]. Aggregation functions can cover behaviors from complete *andness* (Min), over independence (AM), to complete *orness* (Max). Interactions may depend on the level of the values to be aggregated. For instance, there might be partial complementarity when values to be aggregated are low, independence when values are medium, and redundancy when values are high.

By choosing an appropriate aggregation function, we can model all kinds of interactions between objectives. In practice, most commonly the additive model with the weighted arithmetic mean as aggregation function is used. One often proposed alternative is multiplicative aggregation [1,8] which, however, is not idempotent. More recently, the Choquet integral has received increasing attention [e.g., 13,15,16]. Furthermore, minimum aggregation is applied

in practice to account for strict legal thresholds or restriction levels [47].

3.4. Eliciting preference statements in a hierarchy

For inferring the parameters of aggregation functions at different hierarchical levels, we need corresponding indifference statements. A hierarchical problem structure has the great advantage that preferences can be elicited for each subset of objectives separately. This reduces cognitively demanding multi-dimensional trade-offs, i.e., considering interactions between many objectives simultaneously.

However, there are drawbacks. Generally, decision-makers give indifference statements between states characterized by attributes. In hierarchical elicitation, communicating the attributes at higher hierarchical levels becomes challenging, because more and more attributes have to be considered and a certain value can be achieved by different combinations of the levels of attributes. This is a general drawback of a hierarchical elicitation procedure which also affects the simpler elicitation of weights of the additive model [e.g., 48]. In some cases, decision-makers are sufficiently familiar with states leading to a certain degree of fulfillment of sub-objectives so they can give indifference statements directly in the value space. This applies to the case discussed in Section 4, where the decision-makers routinely assess rivers with value classes in [0,1].

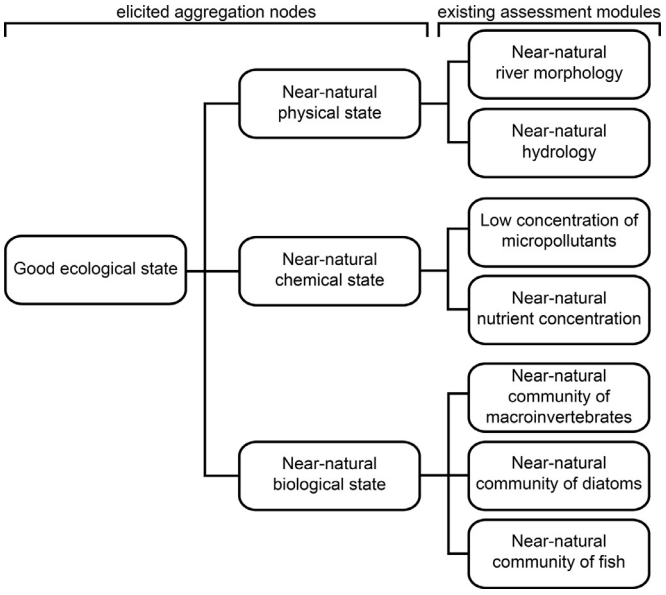

Fig. 2. Objectives hierarchy for assessing the ecological state of rivers. Each of the objectives on the lowest level shown here consists of an assessment module (e.g., near-natural community of fish) that can be described as a branch of an objectives-hierarchy and includes several sub-objectives at lower levels (see Appendix B, Fig. B1).

4. Application to river management

Assessing the ecological state of rivers is a key management issue for identifying and addressing perturbations and pressures on these systems and for supporting decisions about rehabilitation measures based on trade-offs with costs and ecosystem services [49]. To obtain a comprehensive overview, multiple quality elements are included covering biological, chemical, and physical aspects of rivers [e.g., 50].

For Switzerland, a modular concept for stream assessment has been established [51]. To allow an overall assessment based on several existing assessment modules (quality elements), an approach based on MAVT has been proposed [50,52]. However, aggregation of the results from the individual assessment modules to an overall state has been based on direct judgment [12], not explicitly on elicited preferences (but see [9]).

With this case study, we illustrate the use of the presented method for the step of identifying aggregation functions (as in Section 3.2) for the upper hierarchical levels of this assessment, as the value functions for the lower levels (assessment modules) are already known.[3] The preferences of experts working in river management were elicited and modeled.

4.1. Case description

The assessment of the ecological state of rivers can be structured in an objectives hierarchy covering the biological, physical, and chemical state. Each of these branches is described by several sub-objectives that cover the different assessment modules. For example, the biological state is described by assessment modules for the organism groups of fish, macroinvertebrates, and diatoms (Fig. 2).

Similarly to the European Water Framework Directive [49], the outcomes of all assessment modules are expressed by five color-coded quality classes. Furthermore, the outcomes can be presented in the form of a measurable value function on an absolute scale between 0 and 1 with equally distributed classes (poor state from 0 to 0.2 in red; unsatisfactory from 0.2 to 0.4 in orange, moderate from 0.4 to 0.6 in yellow, good from 0.6 to 0.8 in green; very good from 0.8 to 1 in blue) [53]. The value functions have three well-defined reference points on all hierarchical levels. A value of zero represents the worst possible state that is imaginable for a river in Switzerland. A value of one represents the near-natural state in the given anthropogenic influenced environment [51]. A value of 0.6 or above, i.e., good or very good quality class, means that there is no need for action while a value below 0.6 means that measures should be taken to improve the state.

The full objectives hierarchy including the lower levels is provided in Fig. B1 (Appendix B). Examples for two value functions on the lowest level are given for the nutrient assessment module (Appendix B, Fig. B2). The full hierarchy and value functions can be reproduced with the R package “ecoval” [54,55]. As they are not decisive for this application, not all lower-level value functions are presented.

The interviewed experts were familiar with the assessment modules as part of their daily work and can thus link the quality classes and values to real-world outcomes. Since some assessment modules are under development or in revision, we agreed with the interviewees to assume that a satisfying assessment module exists for each sub-objective.

Compared to the general approaches outlined in Sections 2 and 3, this case has specific characteristics which we could exploit. These are useful to ease the hierarchical elicitation:

- The interviewed experts were well familiar with the assessment modules and the value scales have a common and interpretable meaning for all objectives on all levels. Thus, it was not necessary to illustrate outcomes on a multi-dimensional attribute scale.
- Since we are only concerned with the upper part of the hierarchy, we can consistently operate on the level of values which are given by underlying value functions.
- The hierarchical structure requires maximally three objectives to be compared simultaneously.

[3] See http://www.modul-stufen-konzept.ch/index_EN.

Table 1
Aggregation functions considered in the river assessment application case.

Function name	Abbreviation	Generating function $f(v_i)$	$F_w^f(v)$	Additive for
Weighted arithmetic mean (additive model)	WAM	v	$\sum_{i=1}^{n} w_i v_i$	always
Weighted geometric mean with offset[a]	GEO-OFF	$\log(v+\delta)$	$\left(\prod_{i=1}^{n}(v_i+\delta)^{w_i}\right)-\delta$, with $\delta \in \bar{\mathbb{R}}_{\geq 0}$	$\delta=\infty$
Mixture between WAM and minimum[b]	WAM-Min		$(1-\gamma)\cdot\sum_{i=1}^{n} w_i v_i + \gamma\cdot\min(v)$, with $\gamma\in[0,1]$	$\gamma=0$
Weighted power mean (root-mean-power)	POW	$\begin{cases} v^{\gamma} & \text{if } \gamma\in\bar{\mathbb{R}}_{\neq 0} \\ \log(v) & \text{if } \gamma=0 \end{cases}$	$\begin{cases} (\sum_{i=1}^{n} w_i v_i^{\gamma})^{\frac{1}{\gamma}} & \text{if } \gamma\neq 0 \\ \prod_{i=1}^{n} v_i^{w_i} & \text{if } \gamma=0 \\ \min(v) & \text{if } \gamma=-\infty \\ \max(v) & \text{if } \gamma=\infty \end{cases}$	$\gamma=1$
Weighted exponential mean	EXPM	$\begin{cases} \gamma^{v} & \text{if } \gamma\in\mathbb{R}_{>0}\backslash 1 \\ v & \text{if } \gamma=1 \end{cases}$	$\begin{cases} \log_{\gamma}(\sum_{i=1}^{n} w_i\cdot\gamma^{v_i}) & \text{if } \gamma\in\mathbb{R}_{>0}\backslash 1 \\ \sum_{i=1}^{n} w_i v_i & \text{if } \gamma=1 \end{cases}$	$\gamma=1$

[a] The geometric mean has the sometimes undesirable property of being zero if one element is zero. Adding an offset eliminates this. For $\delta=0$, GEO-OFF reduces to the weighted geometric mean, for $\delta=\text{Inf}$, it reduces to the weighted arithmetic mean.
[b] The mixture between WAM and minimum is a composed aggregation function. Specifically, a weighted arithmetic mean of two quasi-arithmetic means. As it is a simple extension of the WAM that allows modeling veto effects [12], it was included.

- As values by definition have the same range $[0,1]$, we could further simplify the probabilistic model (Eq. (2)) by assuming all the σ_i to be equal per set of compared objectives.

4.2. Considered aggregation functions

In the spirit of working with multiple competing preference models, we tested several aggregation functions as part of the MAVF. Relying on prior knowledge for this assessment [12], we considered several functions of the family of weighted quasi-arithmetic means [46] to be suitable for this study (Table 1). In other applications, other functions may be more sensible. A well founded account of useful aggregation functions is provided by Grabisch et al. [46], Beliakov et al. [45], or Langhans et al. [12].

For aggregating a vector $\boldsymbol{v}=(v_1,\ldots,v_n)$ with n elements, quasi-arithmetic means are defined as $F^f(\boldsymbol{v},\boldsymbol{w}) := f^{-1}[\sum_{i=1}^{n} w_i\cdot f(v_i)]$, with f a continuous and strictly monotonic function (the generating function), f^{-1} its inverse, and $\boldsymbol{w}$ a vector of weights with $\sum_{i=1}^{n} w_i = 1$.

All functions in Table 1 contain the additive model (the WAM) as special case (last column). This allows us to address the question how "non-additive" preferences are. Apart from weights, $\boldsymbol{w}$, the functions have one additional parameter, which can be seen as indicative of the deviation from the additive model. In non-additive models weights can have a different meaning than in the additive model and can compensate or reinforce the effect of the additional parameter.

To make comparison easier, the functions were reparametrized: The additional parameter, now called α for all functions, is zero when the model coincides with the additive model, one when it is maximally different in direction of the minimum, and minus one when it is maximally different in direction of the maximum (Appendix A, Table A1). The effect of a certain change in α depends on the function. Therefore, the value of α is only indicative of the deviation from the additive model and not an absolute measure.

The chosen functions have a relatively simple form. However, considered together, they are able to represent a wide range of preferences between complementarity and redundancy and can show behaviors from complete *andness* to independence to strong *orness*, depending on the parameters. Level-dependent interactions between two and more objectives can also be modeled.

4.3. Preference elicitation

4.3.1. Determining an elicitation scheme

To test our method before conducting the elicitation interviews and to find elicitation schemes (sets of questions) with a high sensitivity to the parameters, we conducted a simulation study. The fundamental idea was to generate artificial, noisy answer data for different elicitation schemes based on one preference model and then use these artificial answers to estimate the parameters of this and other models.

The simulation allowed, firstly, determining how well the parameters of the model that generated the answer data could be recovered. Secondly, it allowed testing model discrimination between the additive model and others. Thirdly, by fitting other preference models to these data, it allowed assessing the distinguishability between models. Fourthly, it helped estimating the number of questions needed for reaching an acceptable parameter uncertainty and a reasonable model discrimination power as well as avoiding overfitting.

The ideal elicitation scheme depends on the goals of the preference modeling. In our case, important goals were discriminating between additive and non-additive preference models and estimating parameter uncertainty. The simulation helped us to find a suitable set of questions for these goals. For instance, as we wanted to determine a generic value function applicable to all midsized streams in Switzerland, our models should be accurate in the whole value space. Therefore, we opted for a general elicitation scheme which was independent of the compared objectives and alternatives. For other purposes, or when using preference models with a different number of parameters, the elicitation schemes and the required number of questions could be different.

The simulations study was subdivided into several parts. We describe only the case with three objectives, but it worked analogously for two objectives.

First, a full factorial design was determined, based on the factors that can be varied when asking a trade-off question. These are: (1) the values of objectives used as reference point, (2) which objective is varied in the question, (3) which is the response objective, (4) the magnitude of change of the values from the reference to the question situation, and (5) the direction of change, i.e., lowering or increasing the varied value. Depending on the number of factor levels, this leads to a large number of potential questions, here more than 11,000 (Table 2). Of these, fewer than 6000

Table 2
Factors for a full factorial design with three objectives.

Factor	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Number of factor levels
Value objective 1	0.1	0.2	0.3	0.5	0.7	0.9	6
Value objective 2	0.1	0.2	0.3	0.5	0.7	0.9	6
Value objective 3	0.1	0.2	0.3	0.5	0.7	0.9	6
Question variable	Obj. 1	Obj. 2	Obj. 3				3
Response variable	Obj. 1	Obj. 2	Obj. 3				3
Step size	0.2	0.3	0.4				3
Direction of change	Up	Down					2
							$\prod = 11,664$

Table 3
Criteria for reducing the full factorial design for the three objectives case.

Criterion	Description
i)	Focus on edges, as the differences between the models are smaller in the center due to the idempotency property: take out value levels of 0.3 and 0.5 for all objectives.
ii)	Have room for answers: If the level of the response variable is 0.1 or 0.2 the direction of change is downwards, if the level of the response variable is 0.9 the direction of change is upwards.
iii)	Minimize cognitive burden: If the question variable has a value of 0.2 or 0.7, the response variable has the same value.
iv)	Minimize cognitive burden: level of the third (neutral variable) is fixed. If the level of the question or response variable is 0.1 or 0.9 it also has to be either 0.1 or 0.9; if the level of the question variable is 0.7 or 0.2, it also is 0.7 or 0.2, respectively.
v)	Step size either 0.2 or 0.4.
vi)	Further limit magnitude and direction of change: if the level of the question variable is 0.9, it is decreased by 0.4; if it is 0.1, increased by 0.4; if 0.2, decreased by 0.2; if 0.7, the step size is always 0.2.

Table 4
Design sets to reduce the number of questions from a full factorial design for three objectives.

Elicitation layout	Applied reductions	Resulting number of questions
E1	None	5616
E2	i)	1440
E3	i), ii)	888
E4	i), ii), iii)	288
E5	i), ii), iii), iv)	108
E6	i), ii), iii), iv), v), vi)	42

are within the required interval and lead to different reference and question points (see top line in Table 4).

Secondly, a list of criteria (Table 3) was developed and iteratively refined to reduce the number of possible questions and the cognitive burden for interviewees. For example, by criterion iv) the value of the third variable, which is not directly involved in the trade-off, was fixed. Based on these criteria, the number of potential questions was narrowed down and more specific elicitation layouts were created (E2–E6, Table 4).

In a first simulation, we investigated how many questions were necessary for acceptable parameter estimation when we used these elicitation layouts. For this purpose we created concrete elicitation schemes with different numbers of questions (12, 18, 30, 42) based on the layouts in Table 4. As the desired number of questions is usually smaller than the number of possible questions, the questions were selected by random or stratified random sampling. To account for effects due to this selection, we created 10 instances of each scheme. The more selection criteria had been applied, the fewer questions are possible and the smaller is the variation in the elicitation schemes.

Next, artificial answers were calculated for the elicitation schemes using the aggregation functions (Table 1) and assumptions about likely parameter values. We used six different parametrizations with more equal and more unequal weights, and stronger and weaker non-additivity. We calculated the answers that would be given if these models were correct and added Gaussian noise with mean 0 and standard deviation 0.05 to represent uncertainty.

Subsequently, we performed parameter inference to see how well the generating function and other functional forms could fit these artificial answers. To analyze the estimates, 100 bootstrap samples were created and the inference was rerun with these answer samples. Results were analyzed by exploratory data analysis (Fig. 3).

As expected, the goodness of the parameter estimate, approximated by the standard error of the bootstrap sample, increased with the number of questions (decreasing lines in Fig. 3). The improvement from 12 to 18 questions was often considerable, in contrast to minor further improvement when asking 30 or 42 questions. We judged these further improvements not to outweigh the additional elicitation effort, especially since judgments of decision-makers might become more uncertain with fatigue. Therefore, we decided to use 18 questions.

The overall performance for those elicitation layouts in which there were fewer degrees of freedom for the question selection (layouts 5 and 6 in Fig. 3) was generally better than for random layouts (layouts 1 and 2). Therefore, we continued to work with the layout with most restrictions (E6) with an overall good performance across all cases for 18 questions.

As layout E6 contains 42 questions, there are different possibilities for drawing 18 questions out of these 42. In a second simulation study, we tested four possible subsets and a random layout against each other, using the same procedure as before. We examined the ability to recover parameters and to estimate other functional forms. Focusing on the σ parameter as an estimator for the goodness of fit, the differences between layouts 3–5 were generally small (Fig. 4). As expected, all functions were able to estimate the parameters of the additive model (panels in first column). Self-recovery worked well for all functions and question layouts 3–5 (diagonal panels). Functions could generally be discriminated from each other (other panels) – with some exceptions, e.g., the geometric mean with offset and the exponential mean. As it is completely deterministic, we decided to use variant 5 of elicitation layout 6 in the interviews, (Appendix B, Table B1). If we interpolated this elicitation scheme with the same logic to cases with more objectives, 32 questions would be necessary to compare four objectives and 50 questions for five objectives. However, the performance of these schemes cannot be deduced directly from the known performance with three objectives.

In a third simulation, we tested the influence of different degrees of uncertainty on the inference. For this purpose, we

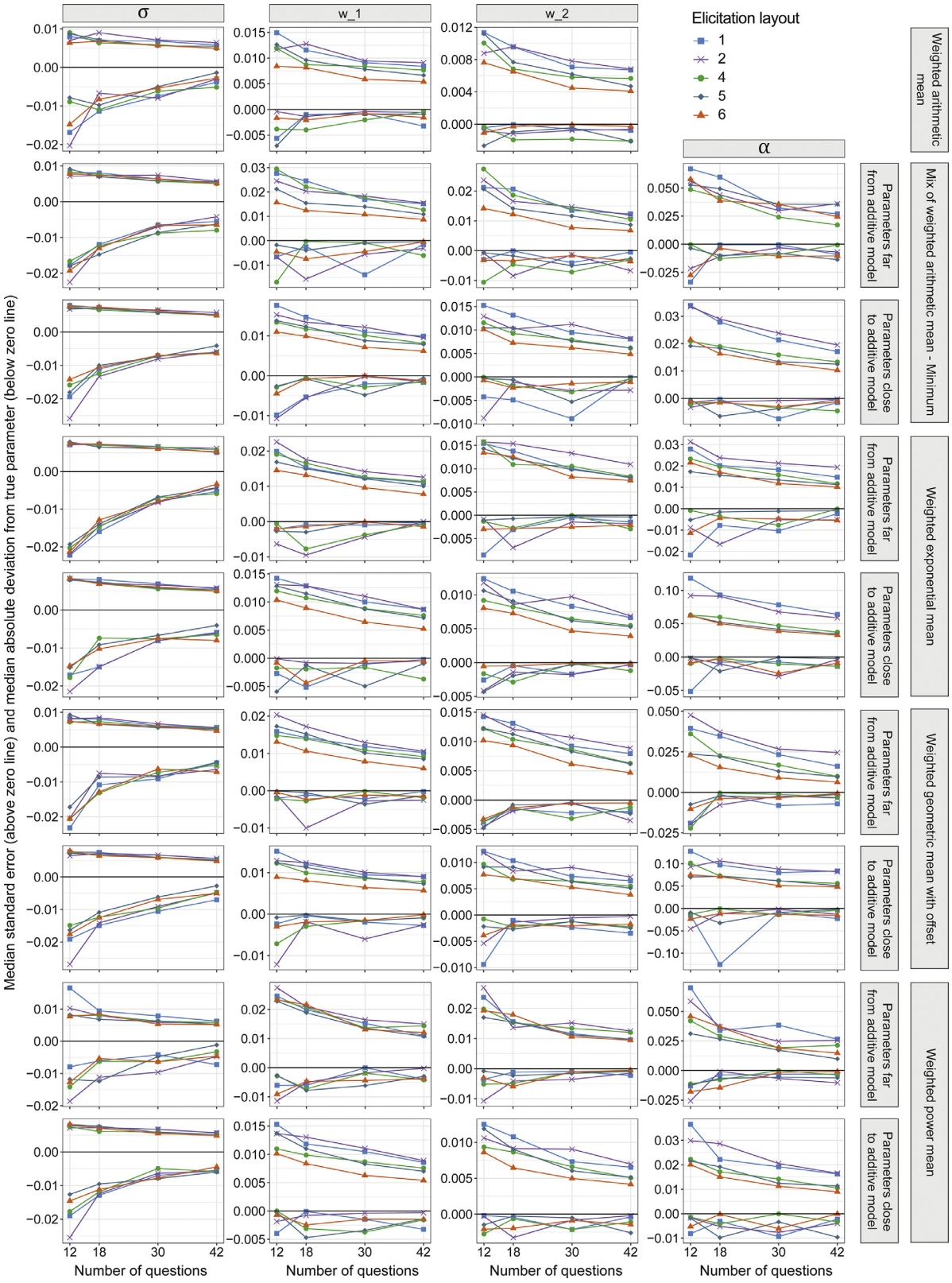

Fig. 3. Number of questions plotted against median standard error of bootstrap sample (above zero line) and median absolute deviation from true parameter value (below zero line) for self-recovery of parameters with data generated by the same function. Split by parameter, aggregation function, and parameter sets subdivided between those close to and far away from the additive model. Each point represents the median of 30 bootstrap runs with 100 samples each. The runs differ in the parameter values and the sampling of random elements in the elicitation schemes.

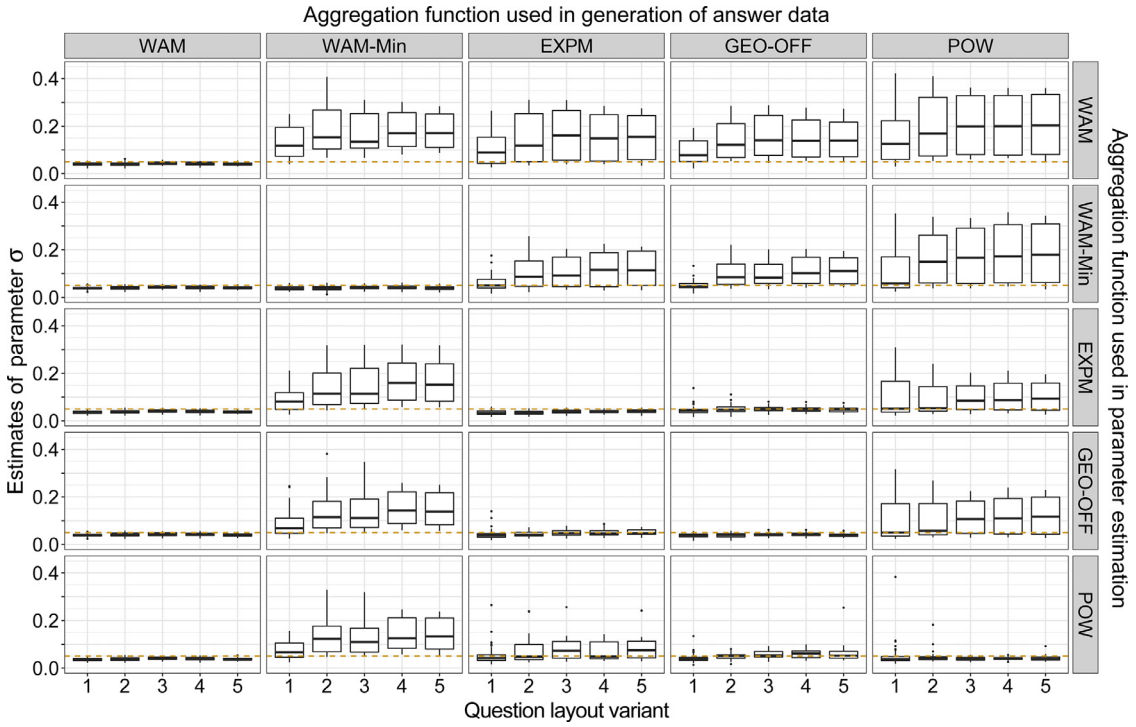

Fig. 4. Comparing the performance of different preference models to fit answer data generated by another model and the ability for self-recovery (diagonal panels). Results for five different variants of elicitation layouts with 18 questions (x-axis) are shown. The median bootstrap estimate of the σ parameter for different parameter combinations and 10 instantiations of each elicitation layout is shown. The dotted line is the value of sigma used in creating answer data (0.05).

created answers with different amounts of noise (standard deviation of 0.025, 0.05, 0.1, and 0.2) for 20 instantiations of a random layout and the layout used in interviews and performed inference. Focusing again on the estimates of the σ parameter, all functions were able to estimate the parameters of the additive model (Fig. 5, panels in first column), as expected. For self-recovery (Fig. 5, diagonal panels) there was little difference between the elicitation layouts. However, for discriminating between preference models (columns) the layout used in the interviews outperformed a random layout. The fit of functions that had not been used to generate

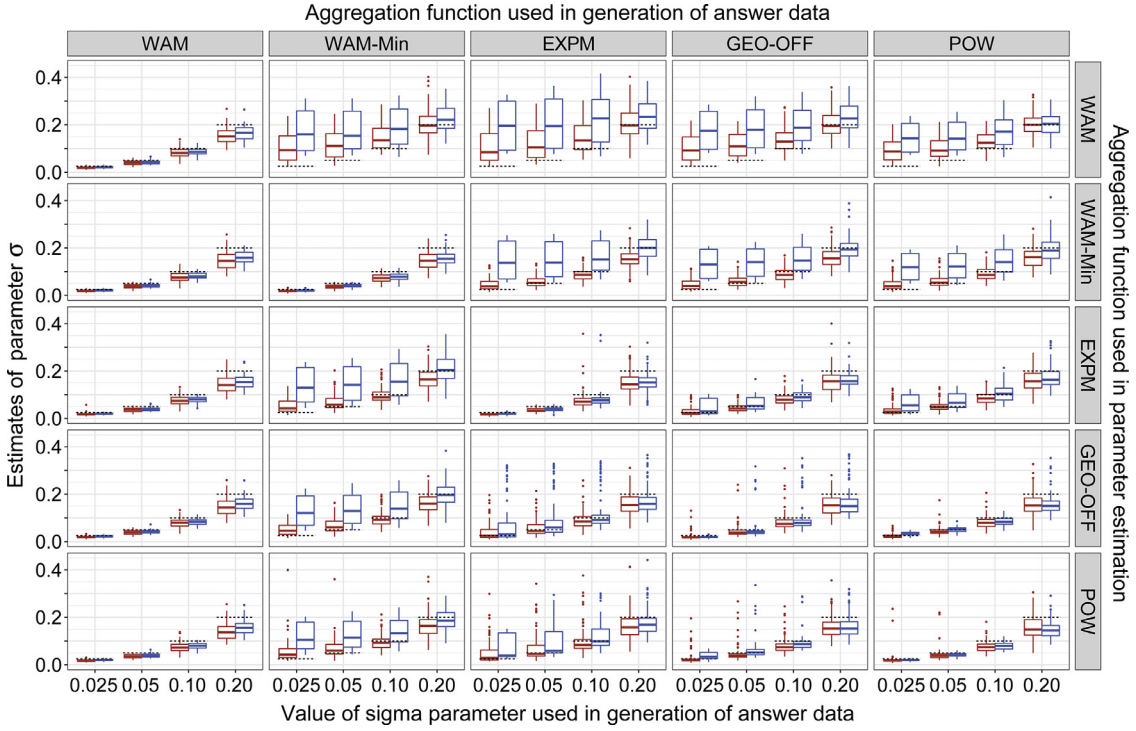

Fig. 5. Comparing the performance of different preference models to fit answer data generated by another model and the ability for self-recovery. Comparison of outcomes between a random question layout (left boxplots, red), and the layout used in the elicitation interviews (right boxplots, blue). Different values of sigma (noise) were used in data generation (x-axes). Boxplots of median bootstrap estimates of the σ parameter are shown, based on six different parameter combinations and 20 instantiations of both elicitation layouts. Dotted black lines depict the value of sigma used in creating answer data.

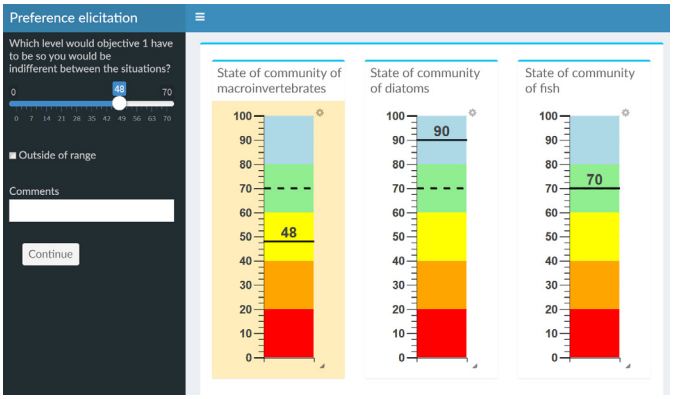

Fig. 6. Screenshot of the elicitation tool used in interviews. The text has been translated from German. A value scale [0,100] was implemented to allow experts to respond in integer numbers instead of [0,1]. The reference situation is indicated by dotted lines, the question situation by solid lines. The level of the question point for the objective with background shading can be adjusted. The question points of the other two objectives remain fixed. The value classes used in river management are visualized by color coding. For a reading example, see text.

the answers was much poorer for the interview layout (σ is high). This enhances the possibility for model discrimination. The value of the standard deviation of the observation error, σ, was generally underestimated by about 20%.

Overall, these results indicate that the models are identifiable with the used layouts and that the standard deviation, σ, can be used for discriminating among the non-additive models in our specific application (with models of the same complexity).

4.3.2. Eliciting indifference statements

We elicited preferences from three experts working in Swiss cantonal authorities responsible for surface water protection and one scientist involved in developing the Swiss stream assessment procedure who is co-author of this paper.

Preferences were elicited by computer aided personal interviews. We have developed an interactive tool to aid the task. It is implemented with shiny [56], a framework for building R based web applications. The tool displays attribute levels or values numerically and in bar plots, making explicit the ranges of attributes and the magnitude of trade-offs (Fig. 6).

Decision-makers were presented with a reference point and a question point. For one of the objectives, the question point had a different value than the reference point, i.e., it was either worsened or improved. The value of another objective was left open, so that it could be adjusted according to the preference of the respondent. If there was a third objective, it had the same value for both points. In a matching task, respondents were asked to adjust the value of the open objective of the question point until they were indifferent between the points. A possibility was to state that the indifference point was higher or lower than the given range and no trade-off was possible. To check for consistency, the reference point and adjusted question point were reversed. Decision-makers were asked how this change of the previously open objective could be compensated by now changing the other objective. If inconsistencies were detected, they were resolved by going back and forth between these directions of asking.

In the example depicted in Fig. 6, at the reference point (dotted line) all objectives are set to a good state (value = 70). For the question point, the objective on diatoms (middle panel) is improved from a good state (70) to a very good one (90), the state of fish (right panel) remains at 70. The respondent is asked to adjust the level of the objective on macroinvertebrates (left panel, shaded) in such a way that this counterbalances the improvement and the reference point and question point have the same overall value. In this example, the respondent is indifferent between a reference point with equal values (70, 70, 70) and a question point with values (48, 90, 70).

4.4. Results

From the indifference statements provided by the experts we inferred parameters of different aggregation functions for each higher-level objective. These are the weights for the sub-objectives and the α parameter. The model with the best representation of the preferences was selected based on the σ term. The probabilistic framework allowed us to determine the quality of our parameter estimates.

The results can be illustrated with a plot showing two sub-objectives on the x- and y-axis and the value of the main objective as iso-lines (Fig. 7). For the two objectives case, we can also plot the answers to trade-off questions as iso-lines (connected dots). As a reading example, consider the answers of expert 1 for the chemical state (left side, Fig. 7). In one question, both objectives were set to a value of 0.7 as reference point (square symbol). For the question point, the value of the assessment module micropollutants was improved to a value of 0.9 and the expert was asked by how much the objective nutrients would have to be lowered to counterbalance this improvement. The answer was 0.6, resulting in the value 0.9/0.6 (dot to the upper-left). In a second question, the nutrient assessment was improved to a value of 0.9, which had to be counterbalanced by worsening micropollutants to 0.63 (dot to the lower-right) according to the expert.

Since we treat the indifference statements as uncertain information and not as constraints, the preference statements are approximated rather than exactly matched by the models. In Fig. 7, the statements of expert 1 could best be approximated by a model which uses a mixture of the weighted arithmetic mean and minimum for aggregation. If the fit of the model were perfect, the iso-lines of the answers would be parallel to the iso-lines of the model.

As the non-additive models have more parameters, they lead to a lower estimate of the σ parameter and thus a better fit than the additive model (Table 5). The uncertainty range of the curvature parameter, α, can now be used to assess whether the preferences may nevertheless be indistinguishable from the additive model. Judged by the 98% interquantile range of the α estimates, this was the case only once (expert 3, physical state; Table 5). For all experts the resulting models were convex ($\alpha > 0$), pointing towards synergy or complementarity between objectives. Otherwise, there was considerable variation between experts in the resulting functions and parameters.

The magnitude of the error (σ) varied between experts and nodes from 0.02 to 0.14 (Table 5). This is the random error due to uncertainty in preferences, uncertainty due to elicitation, and deviation of the parameterized aggregation function from the “true” model. There are two lines of explanation which are indistinguishable by relying on the data alone. Firstly, the selected models might be more or less suitable for representing the preferences. Secondly, uncertainty in the experts’ preferences or bias in the elicited preferences may differ.

Such behavioral effects might, for instance, be related to loss-aversion [30]. When there was a loss in one objective in comparison to the reference situation and the experts were asked by how much this had to be compensated, we have found indications that this loss was overcompensated. Some experts compensated almost any loss by larger gains independent of the objective in question or levels of the reference points, leading to inconsistent statements. By asking consistency questions this could partly be mitigated.

Where possible given the small number of answers, we tested the assumptions of our error model based on the answers.

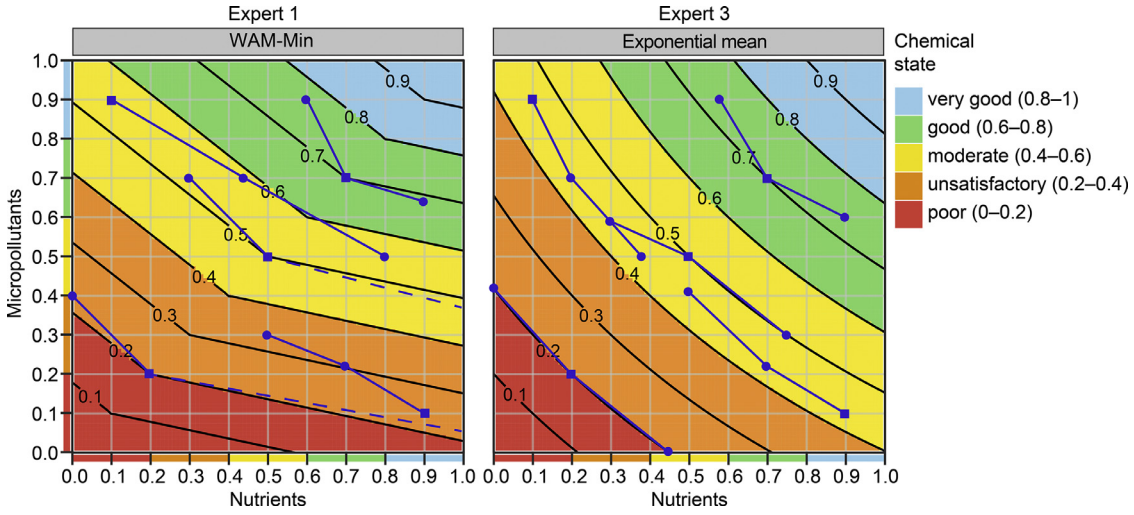

Fig. 7. Aggregation for the objective chemical state. Aggregation function on the left: mixture of weighted arithmetic mean and minimum (WAM-Min) for expert 1; on the right: weighted exponential mean for expert 4. The axes show the values for the objectives "good state of nutrients" and "good state of micropollutants". Color-coding on the panel indicates areas with "poor" to "very good" achievement of the objective chemical state, given the levels of the sub-objectives. Lines are iso-lines of equal value. Answers are given as connected dots: squares represent the reference points and dots the given answers. In case an iso-point was outside the range a dashed line is shown. For a reading example, see text.

Table 5
Parameters of the aggregation functions with best fit per aggregation node and expert. Standard deviation of bootstrap parameter estimation is given in brackets.

Objective	Parameter	Expert 1	Expert 2	Expert 3	Expert 4
Physical state	Aggregation function	POW	EXPM	POW[b]	POW
	$w_{morphology}$	0.54 (±0.03)	0.49 (±0.01)	0.53 (±0.03)	0.42 (±0.05)
	$w_{hydrology}$	0.46[a]	0.51[a]	0.47[a]	0.58[a]
	α	0.30 (±0.02)	0.66 (±0.02)	0.01 (±0.05)	0.24 (±0.04)
	σ	0.08 (±0.02)	0.02 (±0.00)	0.08 (±0.01)	0.14 (±0.04)
Chemical state	Aggregation function	WAM-Min	POW	EXPM	POW
	$w_{nutrients}$	0.24 (±0.05)	0.47 (±0.02)	0.48 (±0.02)	0.32 (±0.03)
	$w_{micropollutants}$	0.76[a]	0.53[a]	0.52[a]	0.68[a]
	α	0.26 (±0.06)	0.11 (±0.02)	0.43 (±0.10)	0.35 (±0.01)
	σ	0.08 (±0.02)	0.05 (±0.01)	0.05 (±0.01)	0.12 (±0.03)
Biological state	Aggregation function	WAM-Min	EXPM	WAM-Min	POW
	$w_{macroinvertebrates}$	0.35 (±0.02)	0.37 (±0.00)	0.34 (±0.01)	0.40 (±0.02)
	$w_{diatoms}$	0.14 (±0.01)	0.22 (±0.00)	0.29 (±0.01)	0.19 (±0.01)
	w_{fish}	0.50[a]	0.41[a]	0.38[a]	0.41[a]
	α	0.04 (±0.02)	0.33 (±0.02)	0.05 (±0.02)	0.28 (±0.03)
	σ	0.12 (±0.02)	0.02 (±0.00)	0.05 (±0.01)	0.14 (±0.02)
Ecological state	Aggregation function	POW	EXPM	EXPM	POW
	$w_{physical}$	0.24 (±0.01)	0.32 (±0.01)	0.21 (±0.01)	0.28 (±0.01)
	$w_{chemical}$	0.21 (±0.01)	0.32 (±0.01)	0.30 (±0.01)	0.28 (±0.01)
	$w_{biologal}$	0.55[a]	0.36[a]	0.49[a]	0.45[a]
	α	0.09 (±0.02)	0.43 (±0.03)	0.38 (±0.12)	0.12 (±0.02)
	σ	0.07 (±0.01)	0.03 (±0.01)	0.10 (±0.01)	0.07 (±0.01)

[a] As one of the weights is always given by normalization, it was not part of the parameter estimation and there is no estimate for its uncertainty.
[b] The 98% interquantile range of the bootstrap sample contains zero. Therefore this model was judged indistinguishable from the additive model.

Concerning normality, a Kolmogorov-Smirnov test on the residuals for every expert and node indicated that the null hypothesis of normality at the 0.05 significance level could never be rejected. Concerning the independence of errors, we calculated the autocorrelation of residuals. With a lag of one, we found autocorrelations larger than 0.5 in 3 out of 16 cases.

The results from the bootstrap method indicated that the parameter estimation led to acceptable uncertainties for the models that turned out to be best (Fig. 8 and Appendix B, Fig. B3). Usually, the models could be distinguished from the additive model, i.e., the uncertainty range for α was not overlapping with zero (Fig. 8, panels headed by α as example). Discrimination between the other models was sometimes difficult. The estimates for σ often overlapped, which indicates that we cannot be sure which non-additive model has the best fit (Fig. 8, panels headed by σ as example). As their functional forms can be alike in certain parameter regions, geometric mean with offset (GEO-OFF) and exponential mean (EXPM) are often similar in their ability to fit the data.

To explore the significance of the differences between the preference models of experts for the evaluation of real-world situations, we calculated how river sites would be evaluated by these models (Fig. 9). The situations were archetypes, but based on real monitoring data for the Swiss plateau. Mostly, the differences between experts were not huge and decreased with higher hierarchical level. For instance, for site 1 the maximum difference between the evaluations of the ecological state between experts 1–4 was only 0.015 (Fig. 9, top panel). However, for certain nodes differences were up to 0.1 and sometimes led to sites being

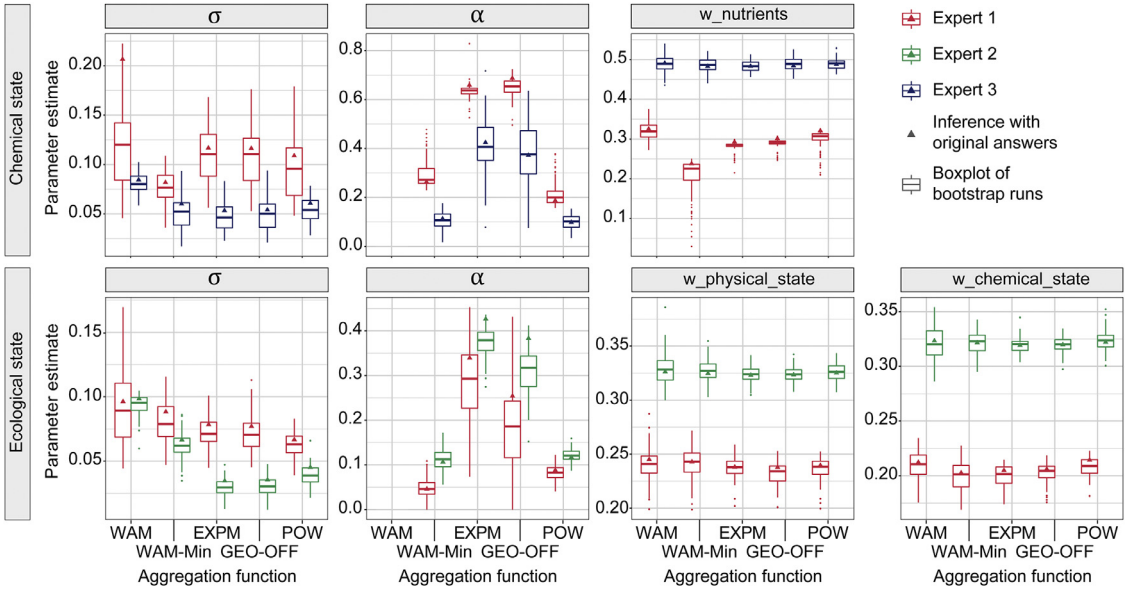

Fig. 8. Excerpt of the bootstrap results of estimating the parameters of different aggregation functions for the nodes chemical state and ecological state. Boxplots of the bootstrap estimates as well as the result without bootstrapping (triangles) are shown. The upper panels correspond to the experts and objectives in Fig. 7. The full results are given in Fig. B3 (Appendix B).

assigned to different river quality classes (e.g. sites 2, 3, and 4; Fig. 9). This might have implications for site management, e.g., regarding prioritization of measures. We also compared the results to two aggregation forms traditionally used in river management: minimum (Min, "one out all out" [47]) and arithmetic mean with equal weights (AM). In comparison to the experts' models, Min was very strict and often led to assignment of a different quality class (maximum difference to experts −0.35). The AM was closer to the evaluation of the experts. However, partly its evaluations were considerably higher (maximum difference to experts +0.12).

5. Discussion

5.1. Discussion of the case study

The goal of the case study was to identify value functions of experts that adequately represent their preferences about an integrated ecological assessment of river states. Together with cost and ecosystem services assessments, these valuations are intended to be applied to support river management in Switzerland.

The application showed that the suggested approach is indeed feasible and reliable. Multi-attribute value functions could be identified based on the indifference statements and their parameters could be estimated with acceptable error. In addition, the mean uncertainty in the replies, caused by the imprecision of the preference, the uncertainty of the elicitation process, and the structural error of the value function parameterization, could be estimated.

The interviews took about two hours to elicit almost 60 indifference statements. Clearly, this is more demanding than eliciting a set of weights for the additive model, but the resulting preference information is also much richer. If we just had wanted to infer parameters of one specific non-additive model over the objectives in our case study, ten trade-off statements would have been sufficient to estimate all α and weight parameters. However, this would have left us with no information about the uncertainty and no possibility of discriminating between models.

Preference statements are inherently uncertain, which was also mentioned by the interviewees. As with any elicitation, behavioral aspects are a crucial issue which deserves attention. While not systematically investigated, we suspect that loss aversion bias or criteria conflict might have been influencing factors [31]. Behavioral effects and possible de-biasing strategies for trade-off questions should be explored more systematically in further studies.

Despite the differences in the elicited preference models of the different experts, the evaluation of sites turned out similarly (Fig. 9). This indicates a good chance for finding a consensus solution that would be acceptable to all experts, e.g., by a group discussion and a joint fit to all answers.

In most cases, the elicited non-additive preference models were significantly different from the additive model. Sometimes, the experts' replies indicated level-dependent interactions. When the achievement of one objective was poor and of one objective good, a decrease in the degree of fulfillment of the well-fulfilled objective could be compensated by a small improvement in the degree of fulfillment of the other objective. However, when starting at a more balanced degree of fulfillment of both objectives, a larger improvement in the degree of fulfillment of the objective was needed to compensate for the same decrease in the other one (Fig. 7).

These findings add to a growing literature discussing the limitations of the additive model and the need to build MAVF for more complex preference structures [e.g., 9,10,12]. In practice, the accuracy of preference models, the effort for elicitation, and the ease of communication of results to decision-makers requires a delicate balance. In many real-world cases, relatively simple models and elicitation efforts may be sufficient to find a good compromise solution for a decision problem. However, we made the experience that in some cases also practitioners opt for non-additive models, for example in lake shore assessment [57].

5.2. Discussion of the method

With the development of our method, we address two important issues in preference modeling: uncertainty and non-additive preferences.

We propose to use a probabilistic framework for parameter inference for MAVF. The advantage of the approach is that uncertainty in elicitation and model fit can be explicitly considered and

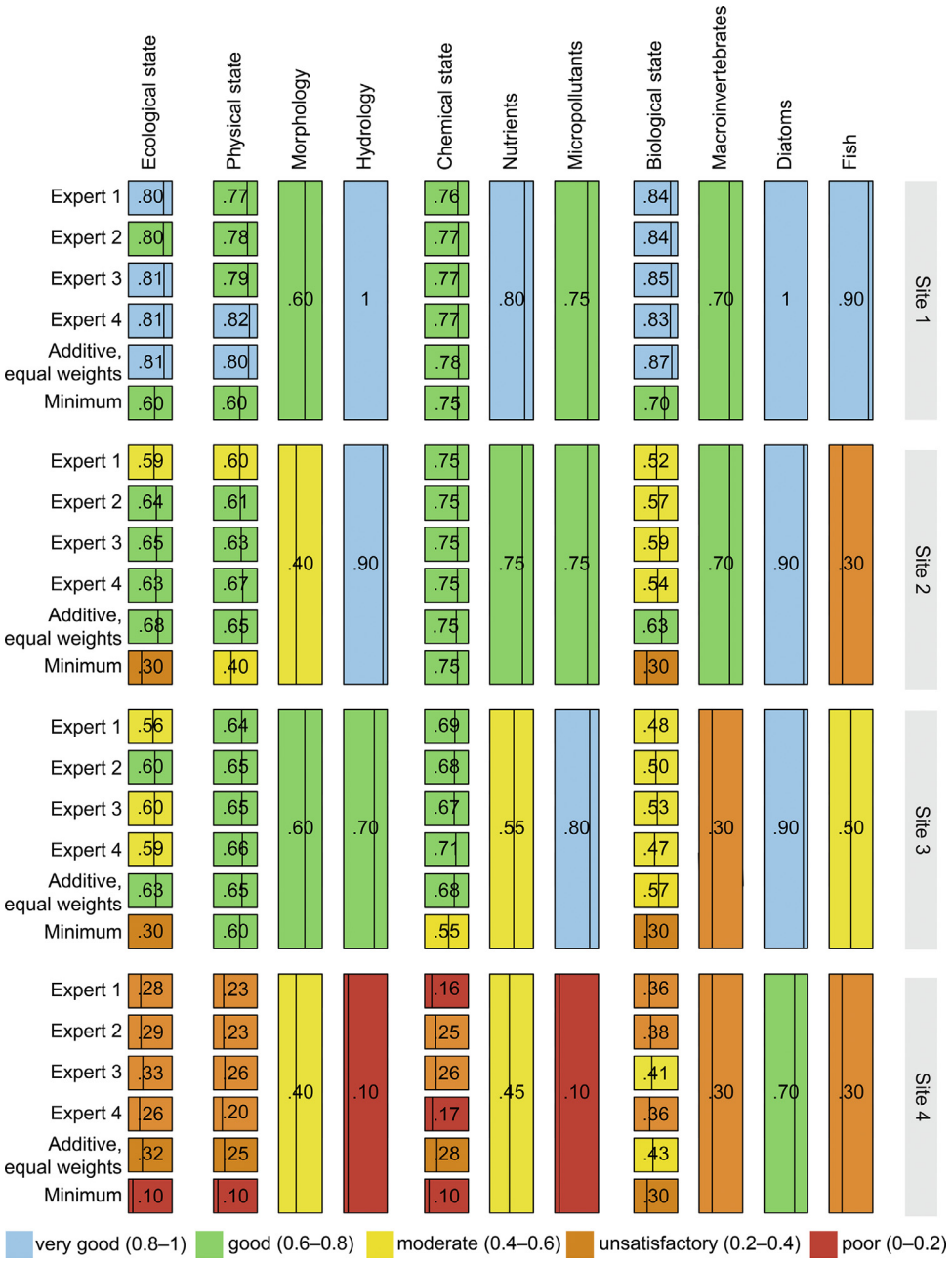

Fig. 9. Aggregation of evaluations along the hierarchy. We present four typical river situations (sites 1–4), based on real monitoring data from Switzerland. The evaluations according to the preference models of the four interviewed experts are shown. We also evaluated the data using two typical assumptions: Additive, equal weights: arithmetic mean using equal weights; Min: minimum aggregation.

quantified. It allows us to evaluate the uncertainty of the estimates of model parameters and to assess the importance of preference features described by these parameters (e.g., non-additivity). By knowing about the uncertainty we can thus be more deliberate in model selection and the interpretation of results.

The method does not require *a priori* restrictions on the shape of the MAVF. Thus, it can be used to infer parameters of a wide range of preference models. This includes models like the Choquet integral. However, we can go beyond these models and test new ones to represent diverse interactions between objectives, for instance level-dependent synergy and redundancy between the same objectives. A further advantage is that the elicitation procedure is independent of the form of the value function. This allows identifying the best fitting function *a posteriori* without a need for new elicitation. The presented method is a general approach for identifying MAVF and their parameters. As such it presents an alternative to other techniques suggested for identifying specific types of non-additive MAVF [e.g., 13,14,21,58,59].

The concepts can be further developed concerning the error model, the type of preference information, and the parameter inference. If there are empirical indications, we might want to relax the assumptions of our error model and, e.g., allow for non-constant variance. Furthermore, uncertainty in value functions on lower hierarchical levels could be propagated to higher levels when evaluating the MAVF. Instead of using indifference statements as preference information, the ideas can be adapted to other types of information, such as simple preference in pairwise comparisons of alternatives' outcomes. This merely requires a reformulation of the

probabilistic model of the responses and leads to a different likelihood function for the parameters. Lastly, if prior information exists, a Bayesian framework to parameter estimation may be more suitable than the frequentist approach we have chosen.

Decision models with interacting objectives require more information compared to the additive model to reliably estimate the interaction effects. Therefore, the development of feasible elicitation schemes is a key research direction. Instead of using a simulation study to identify question layouts, one could try to formally optimize elicitation schemes [32] or design an adaptive approach to elicitation. In an adaptive elicitation scheme, previous answers determine the questions to be asked next. Most approaches so far assume the additive model [e.g., 33–36], but the idea of adaptive elicitation has also been used for the Choquet integral [60]. Such an approach would also be well aligned with Bayesian parameter estimation. When the purpose of the MAVF is to differentiate between alternatives that are specified in advance (or even to select one best alternative), a promising direction is to take the alternatives more directly into account. When using an adaptive approach, we can stop elicitation when differentiation between the alternatives is possible. As the purpose of our case study was to develop a generic model that can be used to evaluate the ecological state of almost any midsized river in Switzerland – from very poor to very good state – we have not yet applied this idea.

6. Conclusion

We presented a novel method for identifying multi-attribute value functions (MAVF) based on indifference statements and using a probabilistic framework for parameter estimation. A real-world application to an ecological assessment procedure used for river management confirmed that the method is useful and feasible.

Instead of forcing preferences to fit an *a priori* specified model, we suggest to test how well different models fit the indifference statements made by the decision-makers. As additivity is a predominantly used *a priori* assumption, it is particularly interesting to test for non-additivity. Complex preference models can be inferred based on commonly used preference statements such as trade-offs. Ideas from statistical inference are promising for preference modeling because they allow explicit treatment of uncertainty in the elicitation and modeling phase.

This study contributes to a growing stream of literature on modeling interacting objectives by presenting a novel, practical, and well-grounded approach for doing so. In the presented case study, the most adequate preference models predominantly deviated from the additive model. We hope to stimulate further research in this direction, especially by testing the feasibility of the proposed approach in other cases.

Acknowledgments

We thank the Eawag Directorate for supporting this work with Eawag's Discretionary Funds. We are grateful to Pascal Bücheler for his work in programming the elicitation tool, to Christian Michel for testing, and to the experts for participating in interviews. We thank the editor and three anonymous reviewers for their valuable comments to an earlier version of the manuscript.

Appendix A: Reparametrization

See Table A1.

Appendix B: Expert elicitation and inference

Table B1
Elicitation scheme used for the interviews for three objectives.

Question	Reference point				Question point		
	Value.1	Value.2	Value.3		Value.1	Value.2	Value.3
1	0.7	0.7	0.7	∼	?	0.9	0.7
2	0.7	0.7	0.7	∼	0.7	?	0.9
3	0.7	0.7	0.7	∼	0.9	0.7	?
4	0.2	0.2	0.2	∼	0	0.2	?
5	0.2	0.2	0.2	∼	?	0	0.2
6	0.2	0.2	0.2	∼	0.2	?	0
7	0.1	0.9	0.9	∼	?	0.5	0.9
8	0.1	0.9	0.1	∼	?	0.5	0.1
9	0.9	0.1	0.9	∼	0.5	?	0.9
10	0.9	0.1	0.1	∼	0.5	?	0.1
11	0.9	0.9	0.1	∼	0.9	0.5	?
12	0.1	0.9	0.1	∼	0.1	0.5	?
13	0.9	0.1	0.9	∼	0.9	?	0.5
14	0.1	0.1	0.9	∼	0.1	?	0.5
15	0.9	0.9	0.1	∼	0.5	0.9	?
16	0.9	0.1	0.1	∼	0.5	0.1	?
17	0.1	0.9	0.9	∼	?	0.9	0.5
18	0.1	0.1	0.9	∼	?	0.1	0.5

Table A1
Parameters of aggregation functions expressed in terms of parameter $\boldsymbol{\alpha}$.

Function name	$F_w^f(v)$	Parameter expressed in terms of α[a]
Weighted geometric mean with offset	$\left(\prod_{i=1}^{n}(v_i+\delta)^{w_i}\right)-\delta,\ \text{with } \delta\in\bar{\mathbb{R}}_{\geq 0}$	$\delta=-\log(\alpha)$
Mixture between weighted arithmetic mean and minimum	$(1-\gamma)\cdot\sum_{i=1}^{n}w_iv_i+\gamma\cdot\min(v),\ \text{with } \gamma\in[0,1]$	$\gamma=\alpha$
Weighted power mean	$\begin{cases}(\sum_{i=1}^{n}w_iv_i^{\gamma})^{\frac{1}{\gamma}} & \text{if } \gamma\neq 0\\ \prod_{i=1}^{n}v_i^{w_i} & \text{if } \gamma=0\\ \min(v) & \text{if } \gamma=-\infty\\ \max(v) & \text{if } \gamma=\infty\end{cases}$	$\gamma=\ln(\frac{2}{\alpha+1}-1)+1$
Weighted exponential mean	$\begin{cases}\log_{\gamma}(\sum_{i=1}^{n}w_i\cdot\gamma^{v_i}) & \text{if } \gamma\in\mathbb{R}_{>0}\backslash 1\\ \sum_{i=1}^{n}w_iv_i & \text{if } \gamma=1\end{cases}$	$\gamma=-1\cdot\frac{\ln(\frac{\alpha+1}{2})}{\ln(2)}$

[a] α is zero when the model coincides with the additive model, one when it is maximally different in direction of the minimum. In addition, for the power mean and exponential mean α is minus one when the model is maximally different in direction of the maximum.

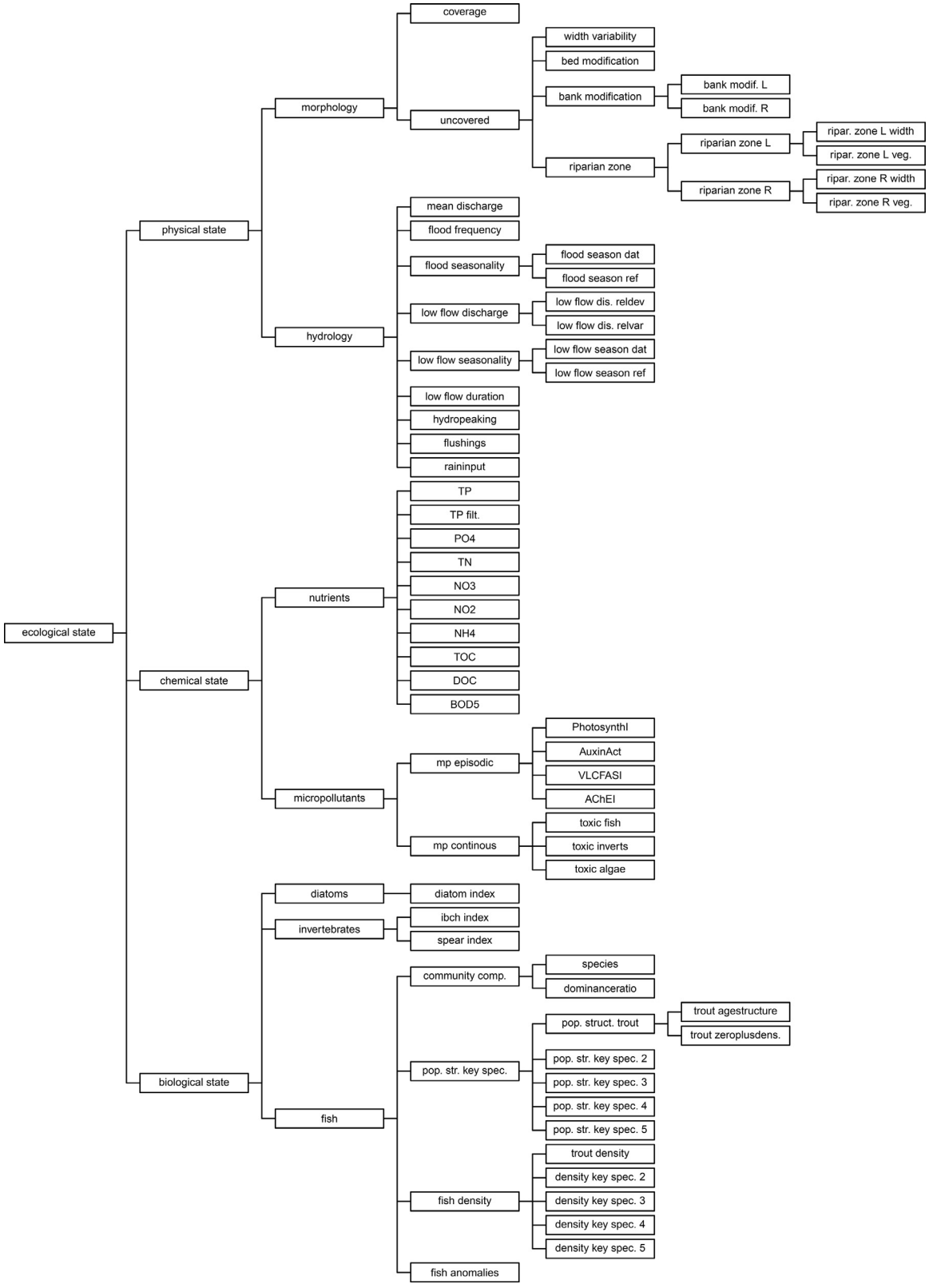

Fig. B1. Full objectives hierarchy for river management in Switzerland that includes existing assessment modules as well as modules that are in development or in revision.

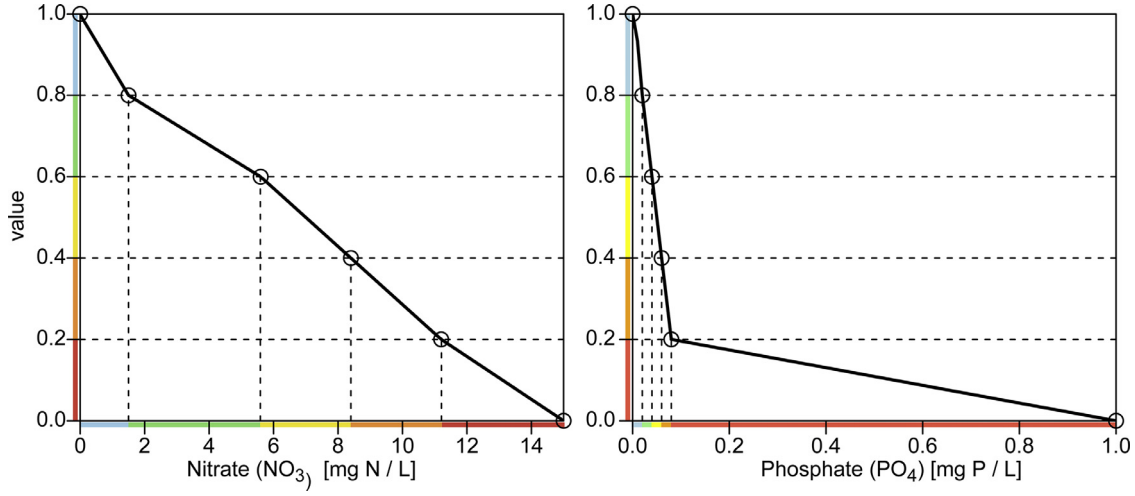

Fig. B2. Example of lowest level value functions for two sub-objectives of the nutrient assessment module. Chemical concentrations are mapped to a level of achievement on the value scale with color-coding for the five quality classes.

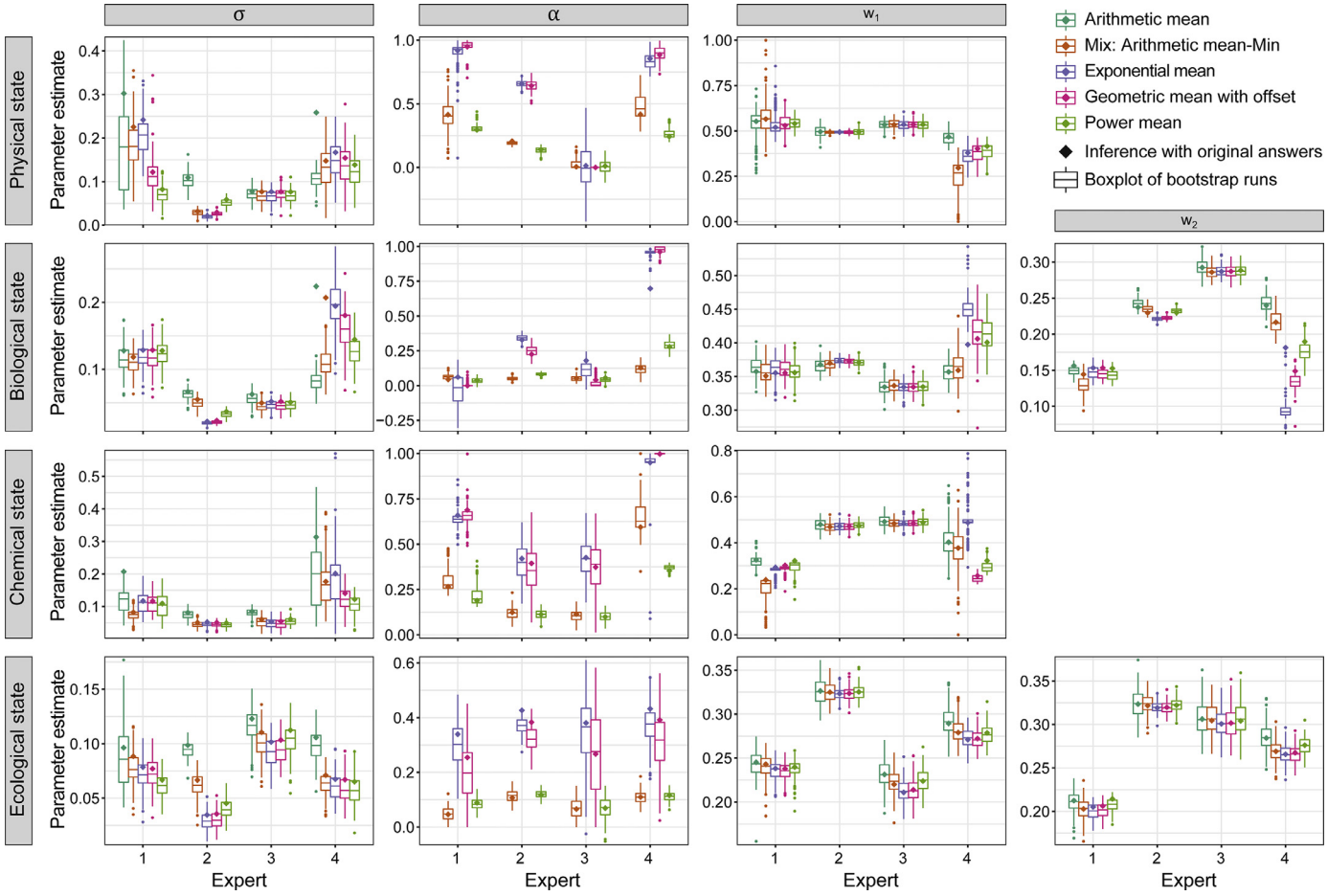

Fig. B3. Results of inference and bootstrapping with 200 bootstrap samples for all experts and aggregation functions.

Table B2
Elicitation scheme used for the interviews for two objectives.

Question	Reference point Value.1	Value.2		Question point Value.1	Value.2
1	0.7	0.7	∼	?	0.9
2	0.2	0.2	∼	?	0
3	0.7	0.7	∼	0.9	?
4	0.2	0.2	∼	0	?
5	0.9	0.1	∼	0.7	?
6	0.9	0.1	∼	0.5	?
7	0.1	0.9	∼	?	0.7
8	0.1	0.9	∼	?	0.5
9	0.5	0.5	∼	0.3	?
10	0.5	0.5	∼	?	0.3

References

[1] Keeney RL, Raiffa H. Decision with multiple objectives. New York: Wiley; 1976.
[2] Reichert P, Langhans SD, Lienert J, Schuwirth N. The conceptual foundation of environmental decision support. J Environ Manage 2015;154:316–32. http://dx.doi.org/10.1016/j.jenvman.2015.01.053.
[3] Dyer JS, Sarin RK. Relative risk-aversion. Manage Sci 1982;28:875–86. http://dx.doi.org/10.1287/mnsc.28.8.875.
[4] Greco S, Mousseau V, Slowiniski R. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. Eur J Oper Res 2008;191:416–35. http://dx.doi.org/10.1016/j.ejor.2007.08.013.
[5] Kadziński M, Greco S, Słowiński R. RUTA: a framework for assessing and selecting additive value functions on the basis of rank related requirements. Omega 2013;41:735–51. https://doi.org/10.1016/j.omega.2012.10.002.
[6] Figueira JR, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. Eur J Oper Res 2009;195:460–86. http://dx.doi.org/10.1016/j.ejor.2008.02.006.
[7] Belton V, Stewart TJ. Multiple criteria decision analysis: an integrated approach. Boston: Kluwer Academic Publishers; 2002.
[8] Dyer JS, Sarin RK. Measurable multiattribute value functions. Oper Res 1979;27:810–22. http://dx.doi.org/10.1287/opre.27.4.810.
[9] Langhans SD, Lienert J. Four common simplifications of multi-criteria decision analysis do not hold for river rehabilitation. PLoS One 2016;11:e0150695. https://dx.doi.org/10.1371/journal.pone.0150695.
[10] Rowley HV, Peters GM, Lundie S, Moore SJ. Aggregating sustainability indicators: beyond the weighted sum. J Environ Manage 2012;111:24–33. http://dx.doi.org/10.1016/j.jenvman.2012.05.004.
[11] Ebert U, Welsch H. Meaningful environmental indices: a social choice approach. J Environ Econ Manage 2004;47:270–83. http://dx.doi.org/10.1016/j.jeem.2003.09.001.
[12] Langhans SD, Reichert P, Schuwirth N. The method matters: a guide for indicator aggregation in ecological assessments. Ecol Indic 2014;45:494–507. https://dx.doi.org/10.1016/j.ecolind.2014.05.014.
[13] Angilella S, Greco S, Matarazzo B. Non-additive robust ordinal regression: a multiple criteria decision model based on the Choquet integral. Eur J Oper Res 2010;201:277–88. http://dx.doi.org/10.1016/j.ejor.2009.02.023.
[14] Grabisch M, Kojadinovic I, Meyer P. A review of methods for capacity identification in Choquet integral based multi-attribute utility theory applications of the Kappalab R package. Eur J Oper Res 2008;186:766–85. http://dx.doi.org/10.1016/j.ejor.2007.02.025.
[15] Grabisch M. The application of fuzzy integrals in multicriteria decision making. Eur J Oper Res 1996;89:445–56. http://dx.doi.org/10.1016/0377-2217(95)00176-X.
[16] Grabisch M, Labreuche C. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. Ann Oper Res 2010;175:247–86. https://doi.org/10.1007/s10479-009-0655-8.
[17] Lahdelma R, Hokkanen J, Salminen P. SMAA - stochastic multiobjective acceptability analysis. Eur J Oper Res 1998;106:137–43. https://doi.org/10.1016/S0377-2217(97)00163-X.
[18] Scholten L, Schuwirth N, Reichert P, Lienert J. Tackling uncertainty in multi-criteria decision analysis - an application to water supply infrastructure planning. Eur J Oper Res 2015;242:243–60. http://dx.doi.org/10.1016/j.ejor.2014.09.044.
[19] Zheng J, Egger C, Lienert J. A scenario-based MCDA framework for wastewater infrastructure planning under uncertainty. J Environ Manage 2016;183:895–908. Part 3 http://dx.doi.org/10.1016/j.jenvman.2016.09.027.
[20] Krantz DH, Suppes P, Luce RD. Foundations of measurement. Academic Press; 1971.
[21] Angilella S, Corrente S, Greco S, Słowiński R. Robust ordinal regression and stochastic multiobjective acceptability analysis in multiple criteria hierarchy process for the Choquet integral preference model. Omega 2016;63:154–69. http://dx.doi.org/10.1016/j.omega.2015.10.010.
[22] Corrente S, Greco S, Kadzinski M, Słowiński R. Robust ordinal regression in preference learning and ranking. Mach Learn 2013;93:381–422. http://dx.doi.org/10.1007/s10994-013-5365-4.
[23] Angilella S, Corrente S, Greco S, Słowiński R. MUSA-INT: multicriteria customer satisfaction analysis with interacting criteria. Omega 2014;42:189–200. http://dx.doi.org/10.1016/j.omega.2013.05.006.
[24] Carmon Z, Simonson I. Price–quality trade-offs in choice versus matching: new insights into the prominence effect. J Consum. Psychol 1998;7:323–43. https://doi.org/10.1207/s15327663jcp0704_02.
[25] Hensher DA. How do respondents process stated choice experiments? Attribute consideration under varying information load. J Appl Econometr 2006;21:861–78. https://doi.org/10.1002/jae.877.
[26] Branke J, Corrente S, Greco S, Gutjahr WJ. Using indifference information in robust ordinal regression. In: Gaspar-Cunha A, Henggeler Antunes C, Coello CC, editors. Evolutionary multi-criterion optimization: 8th international conference, EMO 2015, Guimarães, Portugal, March 29 –April 1, 2015 proceedings, part II. Springer International Publishing; 2015. p. 205–17. Cham.
[27] Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychol Rev 1988;95:371. https://doi.org/10.1037//0033-295x.95.3.371.
[28] Huber J, Ariely D, Fischer G. Expressing preferences in a principal-agent task: a comparison of choice, rating, and matching. Organ Behav Hum Decis Process 2002;87:66–90. http://dx.doi.org/10.1006/obhd.2001.2955.
[29] Attema AE, Brouwer WB. In search of a preferred preference elicitation method: a test of the internal consistency of choice and matching tasks. J Econ Psychol 2013;39:126–40. https://dx.doi.org/10.1016/j.joep.2013.07.009.
[30] Tversky A, Kahneman D. Loss aversion in riskless choice: a reference-dependent model. Q J Econ 1991;106:1039–61. https://dx.doi.org/10.2307/2937956.
[31] Deparis S, Mousseau V, Öztürk M, Huron C. The effect of bi-criteria conflict on matching-elicited preferences. Eur J Oper Res 2015;242:951–9. https://doi.org/10.1016/j.ejor.2014.11.001.
[32] Atkinson A, Donev A, Tobias R. Optimum experimental designs, with SAS. Oxford University Press; 2007.
[33] Ciomek K, Kadziński M, Tervonen T. Heuristics for prioritizing pair-wise elicitation questions with additive multi-attribute value models. Omega 2017;71:27–45. https://dx.doi.org/10.1016/j.omega.2016.08.012.
[34] Cavagnaro DR, Gonzalez R, Myung JI, Pitt MA. Optimal decision stimuli for risky choice experiments: an adaptive approach. Manage Sci 2013;59:358–75. http://dx.doi.org/10.1287/mnsc.1120.1558.
[35] van Valkenhoef G TervonenT. Entropy-optimal weight constraint elicitation with additive multi-attribute utility models. Omega 2016;64:1–12. http://dx.doi.org/10.1016/j.omega.2015.10.014.
[36] de Almeida AT, de Almeida JA, Costa APCS, de Almeida-Filho AT. A new method for elicitation of criteria weights in additive models: Flexible and interactive tradeoff. Eur J Oper Res 2016;250:179–91. http://dx.doi.org/10.1016/j.ejor.2015.08.058.
[37] Myung JI, Pitt MA. Optimal experimental design for model discrimination. Psychol Rev 2009;116:499–518. https://doi.org/10.1037/a0016104.
[38] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
[39] Ghalanos A, Theussl S. Rsolnp: general non-linear optimization using augmented lagrange multiplier method; 2015. R package version 1.16 https://CRAN.R-project.org/package=Rsolnp.
[40] Ye Y. Interior algorithms for linear, quadratic, and linearly constrained convex programming [PhD Thesis]. Stanford University; 1987.
[41] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994.
[42] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009.
[43] Myung JI, Navarro DJ, Pitt MA. Model selection by normalized maximum likelihood. J Math Psychol 2006;50:167–79. https://doi.org/10.1016/j.jmp.2005.06.008.
[44] Keeney RL. Value-focused thinking: a path to creative decisionmaking. Cambridge, Mass: Harvard University Press; 1992.
[45] Beliakov G, Pradera A, Calvo T. Aggregation functions: a guide for practitioners. Springer; 2007.
[46] Grabisch M, Marichal J-L, Mesiar R, Pap E. Aggregation functions. Cambridge: Cambridge University Press; 2009.
[47] European Communities. Guidance document no 13: Overall approach to the classification of ecological status and ecological potential. Common implementation strategy for the water framework directive (2000/60/EC). Luxembourg: Office for Official Publications of the European Communities; 2005.
[48] Marttunen M, Belton V, Lienert J. Are objectives hierarchy related biases observed in practice? A meta-analysis of environmental and energy applications of multi-criteria decision analysis. Eur J Oper Res 2018;265:178–94. https://doi.org/10.1016/j.ejor.2017.02.038.
[49] European Commission. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. Offic J Eur Communities 2000:43.
[50] Paillex A, Schuwirth N, Lorenz AW, Januschke K, Peter A, Reichert P. Integrating and extending ecological river assessment: Concept and test with two restoration projects. Ecol Indic 2017;72:131–41. https://doi.org/10.1016/j.ecolind.2016.07.048.
[51] Bundi U, Peter A, Frutiger A, Hutte M, Liechti P, Sieber U. Scientific base and modular concept for comprehensive assessment of streams in Switzerland. Hydrobiologia 2000;422:477–87. http://dx.doi.org/10.1023/a:1017071427716.
[52] Langhans SD, Lienert J, Schuwirth N, Reichert P. How to make river assessments comparable: a demonstration for hydromorphology. Ecol Indic 2013;32:264–75. http://dx.doi.org/10.1016/j.ecolind.2013.03.027.

[53] Schlosser JA, Haertel-Borer S, Liechti P, Reichert P. Konzept für die Untersuchung und Beurteilung der Seen in der Schweiz. Anleitung zur Entwicklung und Anwendung von Beurteilungsmethoden. Bundesamt für Umwelt; 2013. Umwelt-Wissen Nr. 1326 www.bafu.admin.ch/uw-1326-d.
[54] Schuwirth N, Reichert P, Langhans S. ecoval: procedures for ecological assessment of surface waters; 2017. R package version 1.1 https://CRAN.R-project.org/package=ecoval.
[55] Reichert P, Schuwirth N, Langhans S. Constructing, evaluating and visualizing value and utility functions for decision support. Environ Modell Software 2013;46:283–91. https://dx.doi.org/10.1016/j.envsoft.2013.01.017.
[56] Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: web application framework for R; 2016. R package version 0.14.2 https://CRAN.r-project.org/package=shiny .
[57] Niederberger K, Rey P, Reichert P, Schlosser J, Helg U, Haertel-Borer S, et al. Methoden zur Untersuchung und Beurteilung der Seen. Bundesamt für Umwelt (BAFU); 2016. Umwelt-Vollzug Nr. 1632. www.bafu.admin.ch/uv-1632-d.
[58] Beccacece F, Borgonovo E, Buzzard G, Cillo A, Zionts S. Elicitation of multiattribute value functions through high dimensional model representations: Monotonicity and interactions. Eur J Oper Res 2015;246:517–27. http://dx.doi.org/10.1016/j.ejor.2015.04.042.
[59] Greco S, Mousseau V, Słowiński R. Robust ordinal regression for value functions handling interacting criteria. Eur J Oper Res 2014;239:711–30. http://dx.doi.org/10.1016/j.ejor.2014.05.022.
[60] Benabbou N, Perny P, Viappiani P. Incremental elicitation of Choquet capacities for multicriteria choice, ranking and sorting problems. Artif Intell 2017;246:152–80. https://doi.org/10.1016/j.artint.2017.02.001.