Calibration and validation of the AZTI's Marine Biotic Index (AMBI) for southern California marine bays

Heliana Teixeira[1], Stephen B. Weisberg, Angel Borja[2], J. Ananda Ranasinghe, Donald B. Cadien[3], Ronald G. Velarde[4], Lawrence L. Lovell[3], Dean Pasko[5], Charles A. Phillips[6], David E. Montagne[7], Kerry J. Ritter, Fuensanta Salas[1] and João C. Marques[1]

Abstract

Benthic indices are useful indicators of sediment condition, but many indices are difficult to employ because they require large calibration datasets. The AZTI's Marine Biotic Index (AMBI) requires minimal local calibration, but it was developed in Europe and the validity of its extension to distant regions is unclear. Here we compare its performance in southern California's marine bays with that of the Benthic Response Index, a locally derived data-intensive index. AMBI was calibrated in four ways: 1) using the original AMBI species' classifications developed in Europe; 2) augmenting the original classifications with closely-related taxa, following AMBI guidelines; 3) using local expertise to independently classify taxa; and 4) revision of the local expert classifications by European developers of the index. These approaches were applied to a 685 sample data set and assessed relative to the BRI by comparing samples' classification from best to worst and by evaluating the level of agreement in assigning samples into four condition categories. The AMBI was validated against environmental proxies of disturbance and expert judgement, using consensus agreement about sample condition developed by nine benthic ecologists. The first AMBI approach did not work well, as only 24% of the 928 taxa were on the original AMBI species list, resulting in only 11% of the samples meeting the required 20% of classified individuals for AMBI application. The other approaches classified substantially more taxa, allowing application to 75 to 98% of the samples, respectively. Both of these approaches were significantly correlated with the BRI, though the correlations were lower than between the AMBI runs. None of the AMBI approaches, though, compared well with either the BRI or the validation data when placing samples into perturbation categories, with the AMBI having a greater central tendency. AMBI categorized less than 5% of the samples as reference compared to almost one-third of the samples by the

[1]*IMAR-CMA, Marine and Environmental Research Centre, Department of Life Sciences, University of Coimbra, Portugal*

[2]*AZTI – Tecnalia, Marine Research Division, Herrera Kaia Portualdea n/s, Pasaia, Spain*

[3]*Sanitation Districts of Los Angeles County, Whittier, CA*

[4]*City of San Diego, Marine Biology Laboratory, San Diego, CA*

[5]*Orange County Sanitation District, Fountain Valley, CA*

[6]*City of Los Angeles, Environmental Monitoring Division, Playa del Rey, CA*

[7]*Independent Researcher, Penn Valley, CA*

experts or BRI, and substantially underestimating the number of severely affected samples. Species most responsible for disagreements between BRI and AMBI approaches were identified. Four modifications to enhance AMBI performance were identified: 1) incorporate local expertise in assigning ecological classifications; 2) use transformed abundance weighting to reduce the effect of dominant species; 3) calibrate the categorization scaling using expert judgement; and 4) use the AMBI in combination with other measures, such as the M-AMBI. The success of these modifications is specific to this study, but they are likely to enhance AMBI's performance worldwide.

Introduction

Benthic macrofauna have been used extensively over the past three decades to assess environmental impacts from discharges and outfalls (Dauer, 1993, Hunt *et al.* 2001, Borja *et al.* 2003, Diaz *et al.* 2004, Hall *et al.* 2005). Benthic macrofauna are used because they are sensitive and relatively immobile residents in sediments, where contaminants accumulate. One factor enhancing their use has been development of benthic indices that summarize the complex species composition information into a linear scale from good to bad condition that facilitates interpretation in a management context (Marques *et al.* 2009).

Many index approaches have been developed over the last decade and they vary considerably in their data requirements for calibration. A good example of such contrasts in data requirements is the southern California Benthic Response Index (BRI; Smith *et al.* 2001, Ranasinghe *et al.* 2009) and the AZTI's Marine Biotic Index (AMBI; Borja *et al.* 2000). Both indices are based on abundance-weighted pollution tolerance of the species present at the site, but the BRI requires large amounts of habitat specific data covering the entire disturbance gradient to develop species' pollution tolerance scores. In contrast, the AMBI assigns tolerance classifications to species based on consensus expert judgement, with those assignments transferable among geographies.

The AMBI fills a valuable need for indices that can be applied in geographies or habitats with little previous benthic sampling that lack the large amount of data needed for more complex index development. As a result, AMBI has been widely used in European estuarine and coastal habitats from the Northern Sea to the Mediterranean (Borja *et al.* 2003, 2009; Salas *et al.* 2004; Muxika *et al.* 2005; Carvalho *et al.* 2006; Grémare *et al.* 2009) and has had successful adaptations to other geographic regions in North and South America, Greenland, North Africa, Southeast Asia or Southwest Indian Ocean (Cai *et al.* 2003, Muniz *et al.* 2005, Afli *et al.* 2008, Bigot *et al.* 2008, Borja *et al.* 2008, Callier *et al.* 2008, Josefson *et al.* 2008, Bakalem *et al.* 2009, Borja and Tunberg, 2011).

However, these applications assume that ecological group assignments are universal and AMBI can be applied in any region where there are organisms with previous ecological group assignments. It is unclear how well those ecological group assignments apply to habitats or geographies distant from those for which they were originally developed. Here we apply the AMBI in southern California, testing different criteria of classifying species, and assess its performance relative to the locally validated BRI approach (Smith *et al.* 2003, Ranasinghe *et al.* 2009).

Methods

The AMBI was applied in southern California using several methods for species classifications and its outcome compared to the local BRI approach with regard to both scores classification ordering of sites and categorical classification of site condition. AMBI performance was also evaluated using two data sets that were originally used to validate the BRI: a) a best expert judgment classification scale (Weisberg *et al.* 2008) representing a full gradient of disturbance from non-impacted to highly impacted sites, and b) comparing site classifications against physicochemical proxies of disturbance level.

Index Description

AZTI's Marine Biotic Index (AMBI)

The AMBI (Eq. 1) is based on the abundance-weighted average disturbance sensitivity of the macroinvertebrate species in a sample (Borja *et al.* 2000). Each species is *a priori* assigned to one of five ecological groups (EG) based on expert opinion, as summarized by Grall and Glémarec (1997):

- Group I (EGI), species very sensitive to organic enrichment and disturbance, usually present only under unpolluted conditions. They include specialist carnivores, some deposit-feeding tubicolous polychaetes and species that structure communities. Most have long-life cycles;

- Group II (EGII), species indifferent to enrichment or disturbance, always present in low densities with non-significant variations over time. These include suspension feeders, less selective carnivores and scavengers;
- Group III (EGIII), species tolerant of excess organic matter enrichment, that may occur under normal conditions, but their populations are stimulated by organic enrichment. They are surface deposit-feeding species, such as tubicolous spionids;
- Group IV (EGIV), second-order opportunistic species. They are mainly small subsurface deposit-feeding polychaetes, such as cirratulids; and
- Group V (EGV), first-order opportunistic species, able to resist high disturbance. These are deposit-feeders, which proliferate in reduced sediments.

AMBI = [(0)(%EGI)+(1.5)(%EGII)+(3)(%EGIII)+(4.5)(%EGIV)+(6)(%EGV)]/100 Eq. 1

The index produces a final score on a continuous scale from 1 to 6 (7 in azoic sediments) and five categories define benthic community health (Borja *et al.* 2000): 'Undisturbed' (<1.2), 'Slightly Disturbed' (1.2-3.3), 'Moderately Disturbed' (3.3-5), 'Heavily Disturbed' (5-6) and 'Extremely Disturbed' (>6).

A multivariate extension of AMBI (M-AMBI; Muxika *et al*. 2007), which incorporates species richness and Shannon-Wiener ($\log_2$) diversity, was additionally applied in this study. Since the dataset covered a wide range of environmental conditions in a single habitat, the reference values adopted for the M-AMBI calculation were the best values found for its three metrics (no. sp = 108; H' $\log_2$ = 5.77; AMBI 4 = 1.14). The result varies between 0 and 1, with 1 indicating the best quality. Four thresholds define five categories on this M-AMBI scale: 'High'>0.77, 'Good' 0.77-0.53, 'Moderate' 0.53-0.38, 'Poor' 0.38-0.20, and 'Bad' <0.20, identified by intercalibration with other methods during the European Water Framework Directive intercalibration exercise (Carletti and Heiskanen 2009).

Benthic Response Index (BRI)

Benthic Response Index values are also abundance-weighted average disturbance tolerance scores for species in a sample (Eq. 2), but the tolerance score for each species is calculated statistically by assessing its average position along a pollution or disturbance gradient in multivariate ordination space (Smith *et al.* 2001). For this study, we used the optimized southern California marine bay benthos tolerance scores of Ranasinghe *et al.* (2009). The BRI also differs from AMBI in that transformed abundance weighting is used, rather than a simple abundance-weighting, decreasing the effect of the highest density species. This procedure was found to increase the impacted signal detection by the BRI in offshore benthic communities of southern California (Smith *et al.* 2001). For this study we used a fourth-root abundance data transformation from the calibration of the BRI to southern California marine bays (Ranasinghe *et al.* 2009). The general index formula is:

$$BRI_s = \frac{\sum_{i=1} a_{si}^f \; p_i}{\sum_{i=1}^{n} a_{si}^f} \qquad \text{Eq. 2}$$

where BRI_s is the BRI value for sampling unit s, n is the number of species with tolerance scores in s, p_i is the pollution tolerance of species i, a_{si} is the abundance of species i in s, and f is an exponent used to transform the abundance values.

Benthic Response Index values range from 0 to 100, from less to more disturbed infaunal communities. Across that range, four condition categories ('Reference/Undisturbed', 'Low Disturbance', 'Moderate Disturbance' and 'High Disturbance') were established at index values of 39.96, 49.15 and 73.27, based on loss of 25, 75, and 95% of the reference species pool (Smith *et al.* 2001, 2003; Ranasinghe *et al.* 2009).

Data Description

Data for this study comes from within a single habitat for index development purposes, the "Southern California Marine Bays" described in Ranasinghe *et al.* (In Press). The AMBI, M-AMBI and BRI were applied to 685 samples collected in southern California marine bays by eight sampling programs from 1998 to 2005 (Table 1). At each sampling site, sediments were collected with a 0.1 m^2 Van Veen grab and sieved through 1 mm mesh screen. Materials retained on the screen were placed in a relaxant solution of 1 kg $MgSO_4$ or 30 ml propylene phenoxytol *per* 20 L of seawater for at least 30 minutes and then fixed in buffered 10% formalin. In the laboratory, organisms retained on the screens were sorted from debris, identified to the lowest practical taxon (most often species), and counted.

The samples were selected to span a wide range of environmental conditions, from undisturbed to impacted. Five measures were used to distinguish potentially polluted sites: 1) Long *et al.* (1995) effects range median (ERM) quotient; 2) number of ERM exceedences; 3) amphipod acute toxicity tests (% of control *Eohaustorius* survival); 4) percent total organic carbon (TOC); and 5) total nitrogen (TN).

AMBI Calibration for Southern California Marine Bays

Four versions of the AMBI were applied to the data. In the first application (RUN1), we applied ecological group assignments only to those taxa for which there were exact matches in the AMBI database (available at www.azti.es) as of December 2007. In the second application (RUN2), we expanded the ecological group assignments to closely related species, a common practice in AMBI applications, following the guidelines of Borja *et al.* (2008). In the third application (RUN3), we had six experienced southern California benthic ecologists develop EG classifications for southern California species following AMBI criteria for species classification (Borja *et al.* 2000, 2008). The list of EG values based on local expert consensus is available in electronic format as Supplemental Information (SI). The fourth application of AMBI (RUN4) revised species EG assignments in a compromise between local expert opinion and previous AMBI species assignments (see SI). The list is also available from http://ambi.azti.es (February 2010 list).

The AMBI was applied to all 685 samples collected in southern California marine bays (Table 1), using software version 4.1 (available at http://ambi.azti.es) following author guidelines (Borja and Muxika 2005). Samples with more than 20% of the individuals not assigned to an EG were not included in subsequent analyses, following AMBI guidelines (Borja and Muxika 2005). Before calculating AMBI, taxa above genus level in the southern California data set were deleted (70 taxa), with the exception of the following taxonomic groups, which provide meaningful ecological information at the considered taxonomical level: Actiniaria, Cerianthiaria, Chironomidae, Diptera, Dolichopodidae, Halcampidae, Hirudinea, Lineidae, Nemertea, Oligochaeta, Pennatulacea, Phoronida, Runcinidae, Sipuncula, Tubulanidae, Tubulariidae and Turbellaria.

Table 1. Data sources and ranges of environmental characteristics.

Program	Location	Period	Samples
Bight'98	Southern California marine bays	1998	113
Marina Del Rey	Marina Del Rey	1998-2003	60
WEMAP'99	Southern California marine bays	1999	24
SPAWAR	Chollas and Paleta Creeks, San Diego Bay	2001-2002	61
NAASCO	San Diego Bay	2001	175
Huntington Harbor	Los Alamitos Bay, Huntington Harbor	2001-2003	118
Bight'03	Southern California marine bays	2003	119
WEMAP'05	Southern California marine bays	2005	15
Total			**685**

	Latitude (decimal degrees)	Salinity (psu)	Depth (m)	Sediment Fines (%)
Southern California marine bays (n=685)	32.5 - 34.5°N	27.2 - 39.4	0.4 - 30	1 - 100

The local experts removed 64 additional taxa in their classification (RUN3), which they considered out of habitat. In total, dropped taxa constituted less than 2% of the abundance in the dataset.

Comparison of Assessments

Index comparison was done in two ways: a) ordinal, examining the scores ordering of site condition, and b) categorical, examining the site condition category in which the index placed the site. The first analysis was conducted using Pearson correlations among the AMBI (four runs), M-AMBI and BRI scores. The level of agreement on condition categories was evaluated using Kappa analysis (Cohen 1960, Landis and Koch 1977) by establishing 'moderate', 'good', 'very good', and 'almost perfect' levels of agreement using the equivalence table of Monserud and Leemans (1992). Fleiss–Cohen weights were applied (Fleiss and Cohen 1973) because misclassifications between distant categories (e.g., between 'Undisturbed' and 'High Disturbance') are more important than misclassifications between closer categories (e.g., between 'Undisturbed' and 'Low Disturbance'). The BRI establishes correspondence to four categories while AMBI and M-AMBI contain five. Therefore, to allow categorical comparison, these indices two most affected categories (AMBI 'Heavily Disturbed' and 'Extremely Disturbed' and M-AMBI 'poor' and 'bad') were merged to match the BRI 'High Disturbance' category.

The BRI uses weighted abundances while the AMBI uses raw abundance data. For the southern California marine bays, a fourth root transformation of the abundance data was found to provide the best correlation of the BRI with a disturbance gradient (Ranasinghe *et al.* 2009), and hence was adopted for applying the index in such habitats. To evaluate AMBI responses under the same conditions as the BRI, the AMBI was tested with abundance data similarly transformed. The transformation effect was tested by repeating AMBI RUN2, RUN3 and RUN4 with transformed abundances and comparing the results to BRI scores as described above.

Evaluation of Index Performance

The performance of the AMBI, the M-AMBI, and the BRI were evaluated in two ways: 1) by comparing index assessments with independent best professional judgement assessments; and 2) by correlating the index assessments with potential indicators of environmental disturbance.

Validation against Best Professional Judgement (BPJ)

The best professional judgement data set included 21 samples that were originally used to validate the BRI (Weisberg *et al.* 2008). These 21 samples were part of a data set for which nine benthic ecologists developed consensus agreement about sample condition using best professional judgment (details in Weisberg *et al.* 2008). Three of the 24 original Weisberg *et al.* (2008) samples were eliminated, two because of poor agreement among the experts as to their condition and one due to limited interest because only a single species was present in the sample. To maintain independence, the group of nine experts participating in the development of this expert judgement scale shared only three experts with the group of six experts undertaking the species classification exercise in the present study.

Index success was assessed by correlating the BPJ ranks of the 21 samples from best to worst with the indices classifications and also by evaluating the level of agreement among approaches regarding the categorization of such samples. Spearman rank correlation was used to compare indices' scores with BPJ ranks and Kappa analysis used for categories comparison (as described above).

Validation against Environmental Pollution Proxies

The performance of these indices in southern California marine bays was also evaluated using physicochemical and pollutant information; by calculating Pearson correlations between index values and 1) ERM quotient; 2) number of ERM exceedences; 3) acute toxicity tests survival (% of control *Eohaustorius* survival); 4) % TOC; and 5) TN.

Comparison between the Two Methods for Species Ecological Classification

The agreement in ecological classification of species between the two approaches was evaluated by comparing BRI tolerance scores and AMBI EG assignments, for the 370 taxa with tolerance scores with the correspondent EG classification provided by each of the four AMBI runs. Equivalence between the AMBI ecological groups (EG) and the BRI tolerance scores was achieved by selecting BRI condition category thresholds to maximize the weighted kappa statistic between AMBI and BRI categorizations. Because the number of taxa in EG categories IV and V were low, they were merged into a single category. Standard linear weights for the kappa statistic were used (Cicchetti and Allison 1971) to allow greater flexibility for categorizations when optimizing partial agreement. The 'optimal' set of BRI tolerance thresholds was selected by computing weighted kappa statistics for a large set of possible candidates. These candidates were selected by choosing all permutations of three thresholds, taken at 5% increments of the tolerance range. In addition, distances between individual thresholds within each set were constrained to be no less than 10% of the range. These conditions ensured that optimization converged and thresholds within a set were not too close to one another. The following thresholds, yielding the largest weighted kappa value for EG categories (using 294 taxa below Family level), were selected as optimal: -41.7 < EG I < 8.8 > EG II < 48.1 > EG III < 66.8 > EG IV/V > 176.7, with a weighted kappa of 0.15 ('poor' agreement).

Results

Classification of Taxa into AMBI EGs

Only 219 of the 928 taxa in the southern California data set were on the original AMBI list, and only 11% of the RUN1 samples met the 80% of individuals classified into an EG requirement for AMBI application (Table 2). The number of classified taxa increased to 630 in RUN2, when species not on the AMBI list were assigned to an EG following AMBI criteria for closely-related taxa, meeting the AMBI application requirement for 75% of the samples. When local expertise was used to assign EGs, almost all species were classified and 98% of the samples had a high enough percentage of classified individuals for AMBI application.

The distribution of taxa in the five EG categories was similar between RUN1 and RUN2 (Table 3). However, the distribution changed when taxa were classified by outhern California experts, with the percentage of taxa assigned to EG II increasing from 36% to 65%. Despite this change, 56% of the taxa (n = 552) were classified the same by AMBI and local experts, with the disagreements not associated with any particular taxonomic group

Table 2. Summary of AMBI runs: percentage of samples classified (with more than 80% of abundance assigned to ecological groups); number (n) of taxa assigned an ecological group and ignored.

AMBI	Samples Classified	*n* Taxa Assigned	*n* Taxa Ignored
RUN1 (exact taxa)	10.9%	219	70
RUN2 (AMBI authors)	75.0%	630	70
RUN3 (local experts)	98.0%	748	134
	(Total Samples: 685)	(Total Taxa: 928)	

(SI): Annelida accounted for 37% of the taxa and 40% of the disagreements; Arthropoda accounted for 32% of the taxa and 25% of the disagreements; Mollusca accounted for 23% of the taxa and 26% of the disagreements.

Correlations among Index Scores

Application of the AMBI using any of the four methods for species EG assignment resulted in significantly correlated assessment scores (Table 4). However, RUN1 was applicable to only a limited number of samples. In contrast, RUN2, RUN3 and RUN4 were applicable to a meaningful proportion of the southern Californian samples and were strongly correlated. Therefore, the latter were thoroughly analysed.

AMBI scores for the four runs were all significantly correlated with the BRI scores (Table 4; Figure 1), but the correlations were much lower than among the AMBI runs. The correlation between the AMBI and BRI was similar using either local experts' assignments (RUN3) or AMBI authors' assignments (RUN2), and slightly higher when using a combination of both (RUN4). Using abundance transformed data in AMBI calculations contributed to increase substantially the correlation of all AMBI runs with the local BRI index scores (Table 4).

The multivariate AMBI approach (M-AMBI) yielded considerably higher correlations with the BRI than the AMBI, regardless of the criteria used to classify species (RUN 2, 3 or 4) (Table 4). Figure 1 shows lower dispersion of samples between the approaches when using the M-AMBI instead of the AMBI.

Table 3. Numbers (n) and percentages (%) of taxa in ecological groups (EG) using the three classification approaches.

	AMBI RUN1 (exact taxa matches to the AMBI database)		AMBI RUN2 (classification of closely related taxa)		AMBI RUN3 (classification by southern California experts)	
Ecological Groups	*n*	%	*n*	%	*n*	%
EG I	70	32.0	235	37.2	162	21.7
EG II	79	36.1	235	37.3	485	64.8
EG III	39	17.7	98	15.6	77	10.3
EG IV	28	12.8	56	8.9	19	2.5
EGV	3	1.4	6	1.0	5	0.7
Total *n* taxa assigned	**219**		**630**		**748**	

Level of Agreement among Category Assessments

There was great similarity in how the different AMBI runs classified samples into four assessment categories (Table 5). However, these classifications differed considerably from classification of the same samples by the BRI. The low agreement (Table 5) is related to a greater central tendency for the AMBI than the BRI (Figure 1). Whereas the AMBI runs placed only about 1% of the samples into the unaffected category, the BRI placed almost one-third of the samples into that category. Similarly, the BRI categorized about 30% of the samples as affected or severely affected, but less than 10% of the same samples were categorized similarly by AMBI (Table 5).

On the other hand, the M-AMBI category classifications presented a much higher agreement with the BRI than the AMBI alone (RUN2: 'Moderate' agreement Kappa = 0.50, p <0.0001, n = 514; RUN3: 'Good' agreement Kappa = 0.57, p <0.0001, n = 671; RUN4: 'Good' agreement Kappa = 0.62, p <0.0001, n = 671). The M-AMBI distribution of samples among the four condition categories is more balanced than for AMBI (Figure 1). M-AMBI RUN2 showed an opposite trend to the BRI, classifying the majority (59%) of the samples at the disturbed end of the scale against only 38% for the BRI. The other M-AMBI runs showed patterns similar to the BRI, with the majority of samples in the 'Undisturbed' category (Figure 1). Nevertheless, a tendency for M-AMBI to concentrate classifications on the 'Marginal Deviation from Affected' category is still observed as with the AMBI alone (Figure 1).

Agreement between AMBI and BRI in category classification improved when transformed abundance data were used in AMBI calculations (RUN2 Kappa = 0.39 'Low', p <0.0001, n = 447; RUN3 Kappa = 0.43 'Moderate', p <0.0001, n = 682; RUN4 Kappa = 0.55 'Moderate', p <0.0001, n = 684).

Index Validation

Best Professional Judgment Validation

Rank order correlations between the 21 samples judged by experts and the AMBI were comparable to those for the BRI (Table 4). AMBI RUN3 correlated best with expert judgement, although all correlations were high. Data transformation did not improve AMBI correlation with BPJ rankings for these samples, in contrast to the improvement in rank order correlation with the BRI observed for the entire dataset. The best correlations with the experts (BPJ) rank of samples were observed with AMBI RUN3 and all runs of the M-AMBI multivariate approach (Table 4).

On the other hand, categorical comparisons of AMBI with the expert judgement classifications (Table 6) were similar to the pattern observed when comparing AMBI to BRI for all the data. Neither of the AMBI runs identified any 'Unaffected' samples, and more than half of the samples were classified as 'Marginal Deviation from Reference.' Nevertheless, 'Very Good' agreement was achieved according to the kappa scale (Table 6). The BRI had 'Almost Perfect' agreement with the expert classifications. If

Table 4. Pearson correlations between AMBI, M-AMBI and BRI scores and Spearman rank correlations between the three methods and expert judgement ranks (BPJ). For each pair the correlation coefficients are presented in the first line, and the number of pairs of data values used to compute each coefficient is presented in the second line. All correlations were statistically significant with p-values <0.001.

	AMBI			BRI	BPJ
	RUN 1	RUN 2	RUN 3		
AMBI					
RUN 1	---	---	---	0.59	0.81
				75	21
RUN 2	0.99	---	---	0.33	0.81
	75			514	21
RUN 2 *(abundance 4th root transformed)*	---	---	---	0.53	0.79
				447	21
RUN 3	0.87	0.74	---	0.37	0.93
	75	514		670	21
RUN 3 *(abundance 4th root transformed)*	---	---	---	0.61	0.90
				682	21
RUN 4	0.99	0.87	0.88	0.45	0.88
	72	501	656	670	21
RUN 4 *(abundance 4th root transformed)*	---	---	---	0.70	0.77
				684	21
M-AMBI					
RUN 2				0.66	-0.95
				514	21
RUN 3				0.67	-0.97
				670	21
RUN 4				0.74	-0.92
				671	21
BRI					0.89
					21

the classifications are collapsed into two categories ('Unaffected' and 'Marginal Deviation from Reference' into 'Undisturbed', and 'Affected' and 'Severely Affected' into 'Disturbed'), there was agreement on condition for 17 (out of 21) samples for AMBI RUN2, 19 samples for AMBI RUN3, 17 samples for AMBI RUN4, and 20 samples for the BRI (Table 6).

For the M-AMBI, substantially different classifications of the validation samples were produced depending on the criteria (AMBI RUN) used to classify species (Table 6), with the level of agreement with expert classification varying from 'Low' to 'Almost Perfect'. Also, as observed for the AMBI alone, almost no samples were classified as 'Unaffected' (the equivalent of M-AMBI 'High' in the BPJ categories; Table 6).

To explore category mismatches between AMBI and the BPJ, we further examined the species driving the AMBI scores for RUN3 (based on local expertise assignments; Figure 2). This approach classified 13 samples as 'Marginal Deviation from Reference', eight of which Expert Judgment evaluated as either 'Unaffected' or 'Affected' (Table 5). These eight

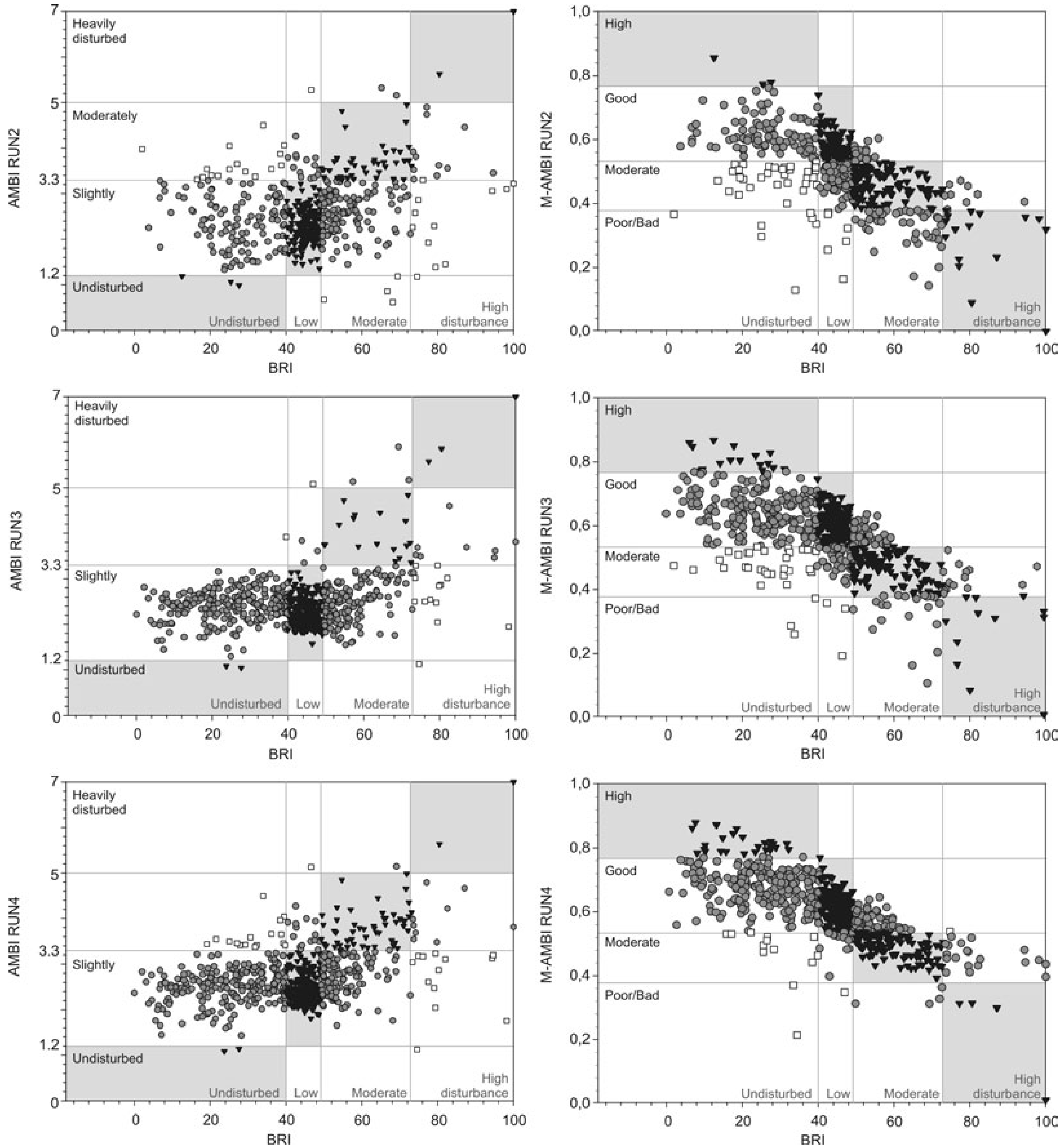

Figure 1. Comparison of sample classification (scores and categories) between AMBI and M-AMBI (RUN2, RUN3, RUN4) and the BRI. The number of samples in each pair is presented in Table 4. Thresholds (grey lines) across index scales show samples that fall into the different quality categories: perfect matches between the two approaches (black triangles samples) fall in grey boxes; the remaining samples with mismatch of one category (grey circles) or more (white squares) fall in white boxes.

mismatched samples were all essentially dominated or co-dominated by individuals of species classified in EG II and III of AMBI (Figure 2). However, the six 'Unaffected' samples according to Expert Judgement all had individuals from sensitive taxa *sensu* AMBI (EG I: between 1 and 15.4%), which were not present in the two 'Affected' samples. On the other hand, these two 'Affected' samples included higher percentages of individuals classified as opportunistic species (EG IV and V: between 8 to 16%), while the 'Unaffected' samples either had no or lower percentages of opportunistic taxa (EG IV and V: between 0 and 12%). The range of AMBI values for the Expert Judgement 'Unaffected'

Table 5. Percentages of samples assessed into four condition categories by the AMBI and BRI for the entire data set (RUN2: n = 514; RUN3 n = 670; RUN4 n = 670). Each RUN adds up to 100%. The level of agreement on category assignments between each pair of methods is indicated, as measured by Kappa analysis.

	BRI				Agreement Level
	Unaffected	Marginal Deviation from Reference	Affected	Severely Affected	
AMBI RUN2					
Unaffected	0.6%	0.0%	0.8%	0.2%	
Marginal Deviation from Reference	22.8%	32.3%	23.7%	2.1%	Low Kappa = 0.25 p <0.0001
Affected	3.1%	2.7%	8.2%	2.1%	
Severely Affected	0.0%	0.2%	0.4%	0.8%	
AMBI RUN3					
Unaffected	0.3%	0.0%	0.0%	0.1%	
Marginal Deviation from Reference	34.8%	34.0%	23.3%	1.8%	Low Kappa = 0.32 p <0.0001
Affected	0.1%	0.5%	2.5%	1.2%	
Severely Affected	0.0%	0.1%	0.5%	0.8%	
AMBI RUN4					
Unaffected	0.3%	0.0%	0.0%	0.2%	
Marginal Deviation from Reference	32.5%	32.2%	18.2%	1.8%	Low Kappa = 0.32 p <0.0001
Affected	2.4%	2.2%	7.9%	1.3%	
Severely Affected	0.0%	0.2%	0.2%	0.6%	

samples (from 1.8 to 2.2; Figure 2) was lower than the range of values observed for the Expert Judgement 'Marginal Deviation from Reference' samples (ranging from 2.2 to 3.2). AMBI values for the Expert Judgement 'Affected' samples were within the range of values observed in the Expert Judgement 'Marginal Deviation from Reference' samples (2.7 and 3.1; Figure 2).

Validation using Environmental Proxies for Disturbance

The three methods (AMBI, M-AMBI, and BRI) were significantly correlated with environmental parameters measured to capture the level of disturbance at the study sites, in spite of the low correlation coefficients observed (Table 7). The best correlations with disturbance proxies were for sediment toxicity measured as percentage of *Eohaustorius* survival and total organic carbon (TOC). The mean values of the indices increased (AMBI and BRI) or decreased (M-AMBI) as expected from non-toxic, to toxic and highly toxic samples (Figure 3). Despite high dispersion of the data, the mean BRI values for each toxicity class pointed progressively to 'Undisturbed', 'Low,' and 'Moderate Disturbance' as survival decreased. AMBI and M-AMBI mean values for all toxicity classes corresponded to 'Slightly Disturbed' or 'Good' status.

Healthier communities were detected at lower TOC values (Figure 3). High dispersion

Table 6. Percentages of the 21 expert judgement (BPJ) samples assessed by the AMBI, M-AMBI, and BRI into four condition categories. Each RUN adds up to 100%. The level of agreement on category assignments between each pair of methods is indicated, as measured by Kappa analysis.

	Expert Judgement (BPJ)				Agreement Level
	Unaffected	Marginal Deviation from Reference	Affected	Severely Affected	
AMBI RUN2					
Unaffected	0.0%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	28.6%	14.3%	9.5%	0.0%	Very Good
Affected	0.0%	9.5%	14.3%	14.3%	Kappa = 0.71
Severely Affected	0.0%	0.0%	0.0%	9.5%	*p* <0.0001
AMBI RUN3					
Unaffected	0.0%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	28.6%	23.8%	9.5%	0.0%	Very Good
Affected	0.0%	0.0%	14.3%	9.5%	Kappa = 0.84
Severely Affected	0.0%	0.0%	0.0%	14.3%	*p* <0.0001
AMBI RUN4					
Unaffected	0.0%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	28.6%	14.3%	9.5%	0.0%	Very Good
Affected	0.0%	9.5%	14.3%	14.3%	Kappa = 0.71
Severely Affected	0.0%	0.0%	0.0%	9.5%	*p* <0.0001
M-AMBI RUN2					
Unaffected	0.0%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	19.0%	0.0%	0.0%	0.0%	Low
Affected	0.0%	4.8%	19.1%	0.0%	Kappa = 0.32
Severely Affected	9.5%	19.0%	4.8%	23.8%	*p* <0.0176
M-AMBI RUN3					
Unaffected	0.0%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	28.6%	9.5%	0.0%	0.0%	Very Good
Affected	0.0%	14.3%	14.3%	0.0%	Kappa = 0.82
Severely Affected	0.0%	0.0%	9.5%	23.8%	*p* <0.0001
M-AMBI RUN4					
Unaffected	4.8%	0.0%	0.0%	0.0%	
Marginal Deviation from Reference	23.8%	19.0%	4.8%	0.0%	Almost Perfect
Affected	0.0%	4.8%	19.0%	4.8%	Kappa = 0.87
Severely Affected	0.0%	0.0%	0.0%	19.0%	*p* <0.0001
BRI					
Unaffected	23.8%	9.5%	0.0%	0.0%	
Marginal Deviation from Reference	4.8%	9.5%	0.0%	0.0%	Almost Perfect
Affected	0.0%	4.8%	19.1%	14.3%	Kappa = 0.89
Severely Affected	0.0%	0.0%	4.8%	9.5%	*p* <0.0001

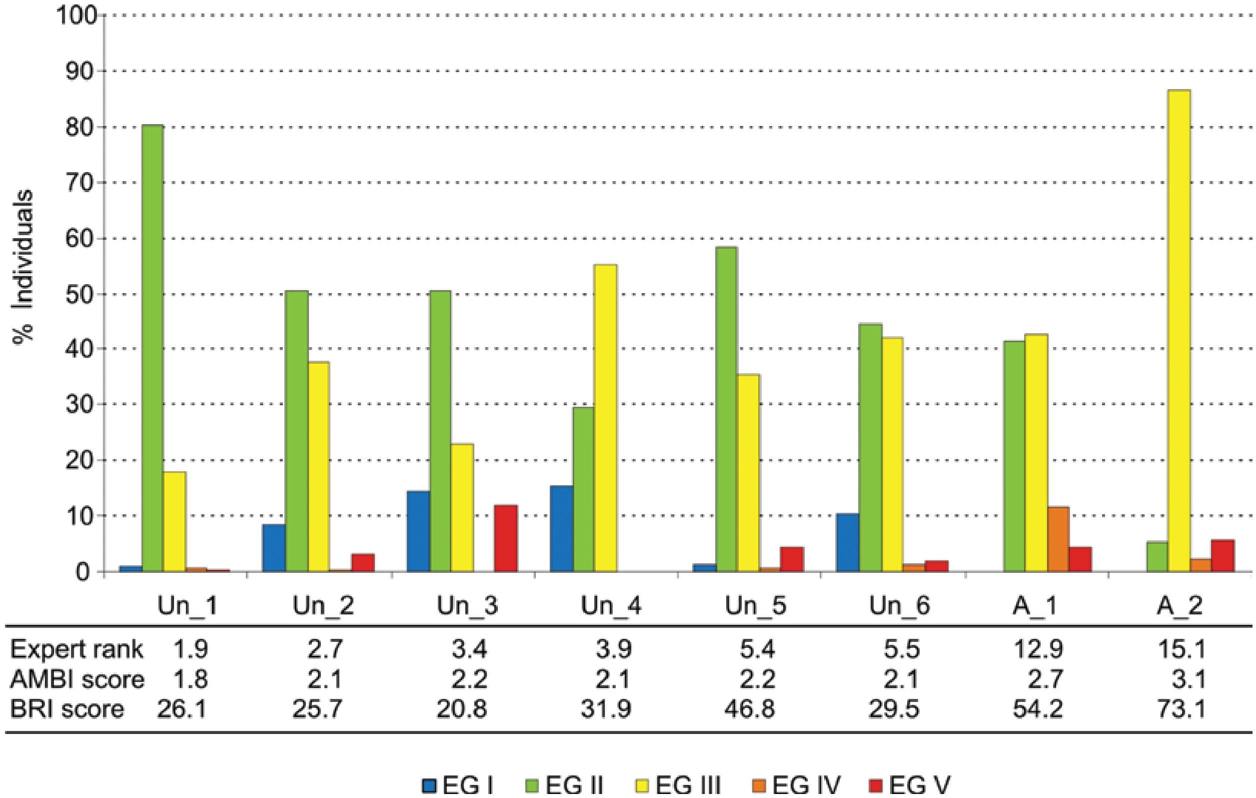

Figure 2. Distribution of individuals (%) among the five AMBI ecological groups (EG I to V) for the eight 'Marginal Deviation from Reference' samples of AMBI RUN3, classified as 'Unaffected' (Un) or 'Affected' (A) by Expert Judgement. For each sample, the Expert Judgment rank, the AMBI RUN3 score, and the BRI score are shown.

was observed, especially for the BRI, with many samples in poor condition with low levels of TOC. The indices also were correlated with other habitat measures, such as depth and salinity for the BRI, and percentage of fine particles in the sediment for the AMBI approaches (Table 7).

Using fourth root transformed data in AMBI calculations, only slightly increased the strength of correlations with environmental proxies for disturbance. AMBI RUN3 registered the biggest increase in correlations for percent fines, toxicity survival, number of ERM exceedences and ERM quotient to reaching values for AMBI RUN 4. However, correlations with TOC (for RUN2 and RUN3) and TN (for RUN2) decreased slightly when fourth root transformations were applied.

Comparison of the Methods for Species Classification

There was poor agreement between the two methods of classifying species by tolerance and sensitivity to disturbance. BRI tolerance score ranges for the five AMBI EGs overlapped greatly, independently of the AMBI criteria (RUN1 to RUN4) deriving them (Figure 4). The distribution of tolerance scores for the five EGs was not clear distinct sets of values, although mean tolerance scores increased gradually from sensitive species EGs (I and II) to opportunistic species groups (IV and V). Nonetheless, the most sensitive species according to the BRI (lowest tolerance scores) were consistently classified in EG II (species indifferent to enrichment or disturbance) by all AMBI criteria, with this ecological group presenting the lowest mean BRI tolerance scores rather than EG I (species very sensitive to organic enrichment and disturbance). It was also evident that the number of species classified as opportunistic in AMBI (EG IV and V) were extremely rare compared to the number of species with considerably high tolerance scores in the BRI. The same overall pattern was observed when only the 193 species with tolerance scores present in the 21 samples validation subset were used for comparison of BRI and AMBI.

Table 7. Pearson correlations between indices and environmental parameters. For each pair, correlation coefficients are presented in the first line, number of pairs of data values used to compute each coefficient is presented in the second line, and p-value <0.05 in the third line. Significant correlations are in bold font.

	Depth (m)	Salinity	% Fines	TOC	TN	ToxEoSurv	N_Erm	Erm_Q
BRI	**-0.55**	**-0.24**	**0.19**	**0.18**	**-0.17**	**-0.23**	**0.21**	**0.16**
	685	685	685	670	268	604	669	669
	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.000
AMBI RUN2	**-0.19**	**-0.12**	**0.19**	**0.25**	**-0.21**	**-0.22**	0.03	0.05
	514	514	514	504	191	443	503	503
	0.000	0.006	0.000	0.000	0.003	0.000	0.481	0.252
AMBI RUN3	**-0.20**	**-0.25**	**0.13**	**0.24**	**-0.17**	**-0.31**	-0.02	-0.01
	671	671	671	656	260	595	655	655
	0.000	0.000	0.001	0.000	0.007	0.000	0.598	0.825
AMBI RUN4	**-0.19**	**-0.14**	**0.31**	**0.25**	-0.11	**-0.35**	**0.15**	**0.11**
	671	671	671	656	260	595	655	655
	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.005
Richness	**0.19**	**0.10**	**-0.32**	**-0.18**	0.04	**0.11**	**-0.19**	**-0.14**
	685	685	685	670	268	604	669	669
	0.000	0.009	0.000	0.000	0.470	0.005	0.000	0.000
Shannon-Wiener Diversity	**0.35**	**0.18**	**-0.09**	**-0.10**	**0.17**	**0.17**	0.02	-0.03
	685	685	685	670	268	604	669	669
	0.000	0.000	0.022	0.013	0.005	0.000	0.587	0.373
M-AMBI2	**0.26**	**0.15**	**-0.29**	**-0.24**	0.11	**0.19**	-0.07	-0.08
	514	514	514	504	191	443	503	503
	0.000	0.000	0.000	0.000	0.136	0.000	0.098	0.087
M-AMBI3	**0.32**	**0.21**	**-0.22**	**-0.21**	**0.14**	**0.25**	-0.06	-0.07
	671	671	671	656	260	595	655	655
	0.000	0.000	0.000	0.000	0.023	0.000	0.127	0.056
M-AMBI4	**0.29**	**0.16**	**-0.29**	**-0.21**	0.11	**0.24**	**-0.13**	**-0.12**
	671	671	671	656	260	595	655	655
	0.000	0.000	0.000	0.000	0.067	0.000	0.001	0.002

There were a substantial number of observations for which the AMBI and BRI approaches were contradictory with regard to a species tolerance and sensitivity to disturbance (Table 8; SI). The highest BRI tolerance score (176.74) was for the polychaete *Paraonella platybranchia,* which was classified in AMBI EG I by local experts (RUN 3). Consequential mismatches were detected between species classifications for approximately 40% of the 34 taxa contributing 80% of the abundance in the dataset (Table 8).

Discussion

AMBI validation in southern California revealed an excellent agreement with expert judgement for ranking samples, similar to that found for the local BRI approach. However, the AMBI did not perform as well as the BRI for the categorical assessments. The use of M-AMBI yielded a better correlation with BRI assessments, but it barely increased the agreement of AMBI with the BPJ validation. The three main approaches tested revealed similar capacity for detecting environmental pollution effects on benthic invertebrate communities.

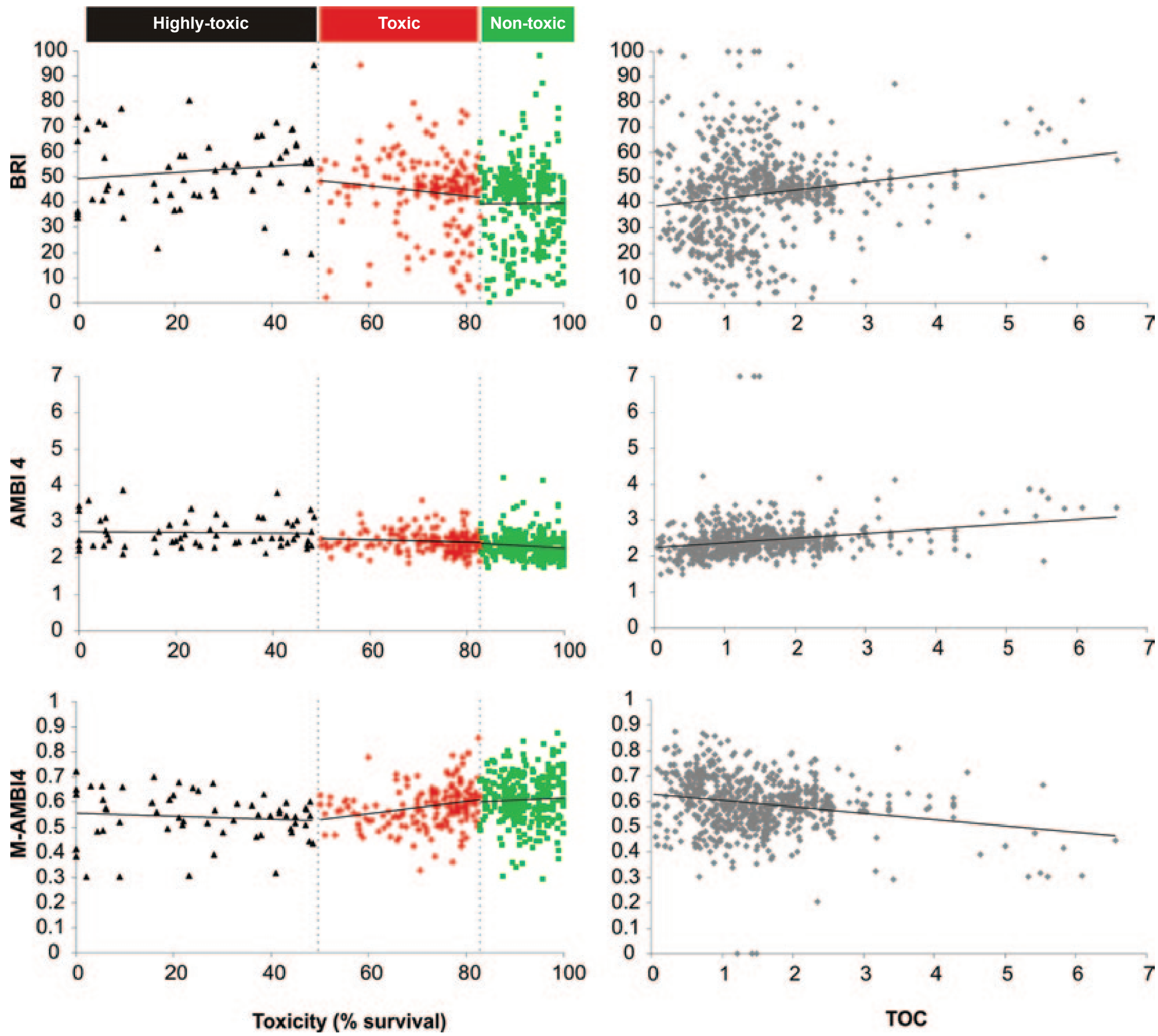

Figure 3. BRI, AMBI and M-AMBI variability with environmental pollution proxies including increasing levels of toxicity (non-toxic to highly toxic) as measured by the percentage of survival of *Eohaustorius* (left) and total organic carbon (TOC) measured as % dry weight (right).

The major weakness of AMBI in the new habitat was found to be its central tendency resulting in less discriminatory power among samples. The results indicated that this could be improved with adjustments to the index calibration by a) improving species classification; b) applying a data transformation; c) adjusting thresholds for condition categories; and d) using a multivariate AMBI approach (M-AMBI).

Increasing AMBI Discriminatory Power

Much of the disparity in the assessments between the two indices could be attributable to differences in species ecological classification. Despite differences among AMBI criteria, the distribution of species across the gradient of sensitivity/tolerance showed a systematically higher concentration of assignments on the sensitive end of the tolerance gradient for all AMBI runs compared to a more balanced distribution of species tolerance scores for the BRI. Especially for species considered opportunistic or highly tolerant to disturbance (Figure 4), the low proportion observed in AMBI can explain the low variance found for the index scores across different levels of disturbance (Figure 3) and hence the concentration of assessments in the 'Marginal Deviation from Reference' category.

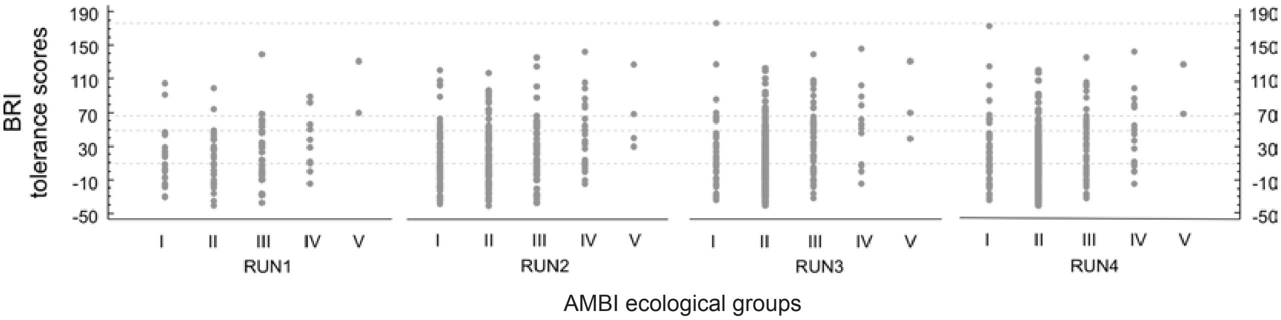

Figure 4. Distribution of BRI species tolerance scores among the five AMBI ecological groups (EG I to V) according to different criteria of classifying species: exact taxa (RUN1, n = 97); authors' criteria (RUN2, n = 280); local experts (RUN3, n = 337); revised (RUN4, n = 339). Dashed lines indicate equivalence between tolerance scores and ecological groups after application of a kappa optimization procedure.

Another factor affecting AMBI performance is the difference in abundance weighting strategies between the two indices. The BRI uses a fourth root transformation, which lessens the effect of the dominant species in the index calculation. The AMBI based on fourth root transformed data significantly improved correlation with the BRI while maintaining a high correlation with expert condition rankings (Table 4). A slight overall improvement was also detected for AMBI (all runs) correlation with environmental proxies of disturbance, although no substantial differences were observed. It is unclear whether a transformation as radical as fourth root is necessary, as Smith *et al.* (2001) observed that less severe transformations were also successful. This is consistent with Warwick *et al.* (2010), who recently tested several transformations for AMBI and concluded that index performance was enhanced with modestly-transformed data, declining with the increase of the severity of transformations.

The third factor is threshold calibration. The AMBI compared better with the BRI in correlations than it did in the categorical assignments, where it had more central tendency and lesser discriminatory power. This was particularly apparent in comparison to the expert judgement index validation sites. This suggests that the AMBI works, but that the default AMBI quality thresholds (Borja *et al.* 2000) used in this study may need adjustment to meet local ecological expectations. The BPJ scale used to validate the indices indicates a greater need on the good side of the gradient (Figure 2). Such BPJ scales could be powerful independent tools to guide future threshold definition and ecological index intercalibration efforts.

Instead of redefining the boundaries of AMBI categories, the index authors (Muxika *et al.* 2007) suggested that an alternative way to overcome this would be the complementary use of several indices, as in M-AMBI, and setting the reference conditions for each index according to the expectations for specific habitats, as already done in other parts of the US, such as the Chesapeake Bay (Borja *et al.* 2008) and a Florida estuary (Borja and Tunberg 2011). The quality thresholds would then be defined on the final score for the samples, after a multivariate procedure, reflecting a departure from reference conditions. Indeed the M-AMBI performed slightly better than BRI or AMBI alone when compared to the expert judgment validation data set (Tables 4 and 6). The inclusion of species richness and Shannon-Wiener diversity index contributed to increase the range of index values across the gradient of disturbance (Figures 1 and 3). Nevertheless a central tendency was still observed for the category assessment which indicates that the process of defining quality thresholds is not solved solely by the incorporation of new metrics. On the other hand, M-AMBI performance was dependent on the AMBI RUN used (Table 6), indicating that the criteria to classify species is crucial for any multimetric using the AMBI.

Species Ecological Classification

The AMBI assumes that species ecological behaviour is intrinsic and each species is bonded to a single ecological group (EG) classification worldwide. However, for 308 of 630 taxa the local experts did not agree with previous AMBI classifications and for 42 of the taxa the disparity went beyond one level of EG (SI). Such mismatch is either a consequence of different interpretation of

Table 8. Comparison of AMBI Ecological Group (EG) classifications (by local experts as used in AMBI RUN3) and BRI tolerance scores for 34 taxa representing 80% of the abundance in the data. Species for which there is a clear mismatch between classifications are in bold font. Total abundance of the taxa and percentage of samples in which it occured are also presented. (n.a. / --: not assigned).

Taxa	Total Abundance (Ind. m^{-2})	Presence (% of samples)	AMBI EG (Local experts)	BRI Tolerance Scores
Pseudopolydora paucibranchiata	77992	74.6	**III**	**81.7**
Mediomastus sp.	53115	85.8	III	57.8
Kalliapseudes crassus	32578	6.4	III	--
Exogone lourei	31208	63.8	II	41.9
Euchone limnicola	20569	58.1	**III**	**76.3**
Musculista senhousia	20186	57.2	**II**	**68.0**
Scoletoma sp. A	15337	44.5	II	19.7
Leitoscoloplos pugettensis	13511	84.2	III	64.4
Grandidierella japonica	13340	38.7	**III**	**106.0**
Streblospio benedicti	13167	13.3	**IV**	**61.8**
Capitella capitata Cmplx	12801	19.1	V	130.8
Oligochaeta	12066	41.9	V	70.0
Scoletoma sp.	11014	48.5	II	--
Theora lubrica	10496	63.6	**III**	**46.6**
Prionospio (Prionospio) heterobranchia	10078	72.6	II	37.5
Pista percyi	9142	58.0	II	42.0
Synaptotanais notabilis	8209	30.7	**III**	**75.3**
Fabricinuda limnicola	7935	17.7	II	50.0
Cossura sp.	7785	23.8	III	--
Monocorophium acherusicum	7678	9.5	III	61.7
Amphideutopus oculatus	7150	49.8	**III**	**-0.4**
Phoronida	5464	35.9	n.a.	99.3
Tagelus subteres	4738	35.5	II	37.3
Dorvillea (Schistomeringos) sp.	4663	40.9	**V**	**38.1**
Scyphoproctus oculatus	4562	15.9	**III**	**25.4**
Spiophanes duplex	4249	41.9	**III**	**31.8**
Neanthes acuminata Cmplx	3942	30.9	III	58.9
Dipolydora sp.	3779	11.5	n.a.	56.6
Euphilomedes carcharodonta	3228	47.2	III	57.2
Polyophthalmus pictus	3007	6.1	I	0.9
Diplocirrus sp. SD1	2921	33.7	**III**	**46.6**
Armandia brevis	2822	26.3	**IV**	**7.1**
Podocerus fulanus	2563	23.4	II	21.6
Acteocina inculta	2516	27.0	**II**	**110.1**

the EG concept or from distinct local perception of species behaviour in the new geography.

Interestingly, the use of local expertise to classify species did not result in substantially higher agreement between AMBI and the local validated index. Even using local experts, there were still important disagreements on species ecological classifications between the EG assignments and the BRI tolerance scores (Figure 4). This was particularly apparent for species classified into EG II by local experts that corresponded to low BRI tolerance scores (Figure 1). Some of the local experts suggested that the differences could have resulted from ambiguity in the EG definitions, but the prevailing feeling was that local knowledge on species ecological strategies might be insufficient and the data intensive BRI approach may provide more accurate information for some taxa than expert perceptions. Certainly those species where there was the greatest disagreement among the BRI tolerance values, AMBI classifications and local expert classifications (Table 8; SI) should be the focal point for additional investigation, both locally and internationally.

There is another important difference in how the BRI assigns tolerance values that remains to be resolved. Unlike AMBI, BRI assumes that species behaviour is habitat dependent, with the numbers and kinds of benthic animals that occur in reference areas varying naturally by habitat (Smith *et al.* 2003). Different pollution tolerance values are empirically developed for each species in each habitat according to the distribution of species along gradients of disturbance (Ranasinghe *et al.* 2009). Accuracy of BRI tolerance values depends on inclusion of data from a disturbance gradient covering a broad range of conditions, as well as accurate definition of the gradient, which may not be easy in habitats such as bays and estuaries where disturbance and pollution from multiple sources is distributed by complex circulation patterns. In contrast, AMBI applies a single classification for each species across all habitats and geographies. One of the aims of a pan-European scale study (Grémare *et al.* 2009) was to test the validity of the use of a single list of sensitivity/tolerance levels by comparing BQI $E(S_{50})_{0.05}$ between subareas, covering both marine and estuarine habitats. Corroborating the Smith *et al.* (2001) studies, they provided evidence suggesting that the species sensitivity/tolerance levels change with geographical location.

Findings from this study, though, may improve AMBI's performance not only for the new geography but also worldwide. The two approaches for classifying species according to their sensitivity or tolerance to disturbance presented obvious inconsistencies that go beyond methodological subtleties. This reinforces that species ecological strategies might be geographically or habitat dependent, feeding the controversy around the plasticity of species ecological behaviour. On the other hand, local endemics, that are similar to European species are frequently discriminated upon close morphological examination, or based on molecular data. In consequence, the number of proven "cosmopolitan" species has shrunken considerably in recent years. This could be a problem, depending on how the ecological performance of these "ghost" taxa differs from the nominal species. The critical species pointed out by the results should be the object of future study, not only to clarify their sensitivity/tolerance level and allow for more robust ecological assessments, but to better understand which factors most affect population dynamics of benthic species.

Literature Cited

Afli, A., R. Ayari and S. Zaabi. 2008. Ecological quality of some Tunisian coast and lagoon locations, by using benthic community parameters and biotic indices. *Estuarine, Coastal and Shelf Science* 80:269-280.

Bakalem, A., T. Ruellet and J.C. Dauvin. 2009. Benthic indices and ecological quality of shallow Algeria fine sand community. *Ecological Indicators* 9:395-408.

Bigot, L., A. Grémare, J. Amouroux, P. Frouin, O. Maire and J.C. Gaertner. 2008. Assessment of the ecological quality status of soft-bottoms in Reunion Island (tropical Southwest Indian Ocean) using AZTI marine biotic indices. *Marine Pollution Bulletin* 56:704-722.

Borja, Á., J. Franco and V. Perez. 2000. A marine Biotic Index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. *Marine Pollution Bulletin* 40:1100-1114.

Borja, Á., I. Muxika and J. Franco. 2003. The application of a Marine Biotic Index to different impact sources affecting soft-bottom benthic communities along European coasts. *Marine Pollution Bulletin* 46:835-845.

Borja, Á. and I.Muxika. 2005. Guidelines for the use of AMBI (AZTI's Marine Biotic Index) in the assessment of the benthic ecological quality. *Marine Pollution Bulletin* 50:787-789.

Borja, Á., D.M. Dauer, R. Díaz, R.J. Llansó, I. Muxika, J.G. Rodriguéz and L. Schaffner. 2008. Assessing estuarine benthic quality conditions in Chesapeake Bay: A comparison of three indices. *Ecological Indicators* 8:395-403.

Borja, Á., J.G. Rodríguez, K. Black, A. Bodoy, C. Emblow, T.F. Fernandes, J. Forte, I. Karakassis, I. Muxika, T.D. Nickell, N. Papageorgiou, F. Pranovi, K. Sevastou, P. Tomassetti and D. Angel. 2009. Assessing the suitability of a range of benthic indices in the evaluation of environmental impact of fin and shellfish aquaculture located in sites across Europe. *Aquaculture* 293:231-240.

Borja, Á. and B.G. Tunberg. 2011. Assessing benthic health in stressed subtropical estuaries, eastern Florida, USA using AMBI and M-AMBI. *Ecological Indicators* 11:295-303.

Cai, L., N.F.Y. Tam, T.W.Y. Wong, L. Ma, Y. Gao and Y.S. Wong. 2003. Using benthic macrofauna to assess environmental quality of four intertidal mudflats in Hong Kong and Shenzhen Coast. *Acta Oceanologica Sinica* 22:309-319.

Callier, M.D., C.W. McKindsey and G. Desrosiers. 2008. Evaluation of indicators used to detect mussel farm influence on the benthos: Two case studies in the Magdalen Islands, Eastern Canada. *Aquaculture* 278:77-88.

Carvalho, S., M.B. Gaspar, A. Moura, C. Vale, P. Antunes, O. Gil, L.C. Fonseca and M. Falcão. 2006. The use of the marine biotic index AMBI in the assessment of the ecological status of the Obidos lagoon (Portugal). *Marine Pollution Bulletin* 52:1414-1424.

Carletti, A. and A.S. Heiskanen (eds). 2009. Water Framework Directive Intercalibration Technical Report. Part 3: Coastal and Transitional Waters. JRC Scientific and Technical Reports, EUR 23838 EN/3.

Cicchetti, D.V. and T. Allison. 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 11:101-109.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37-46.

Dauer, D.M. 1993. Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin* 26:249-257.

de Paz, L., J. Patrício, J.C. Marques, Á. Borja and A.J. Laborda. 2008. Ecological status assessment in the lower Eo estuary (Spain). The challenge of habitat heterogeneity integration: A benthic perspective. *Marine Pollution Bulletin* 56:1275-1283.

Diaz, R.J., M. Solan and R.M. Valente. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. *Journal of Environmental Management* 73:165-181.

Fleiss, J.L. and J. Cohen. 1973. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33:613-619.

Grall, J. and M. Glémarec. 1997. Using biotic indices to estimate macrobenthic community perturbations in the Bay of Brest. *Estuarine, Coastal and Shelf Science* 44:43-53.

Grémare, A., C. Labrune, E. Vanden Berghe, J.M. Amouroux, G. Bachelet, M.L. Zettler, J. Vanaverbeke, D. Fleischer, L. Bigot, O. Maire, B. Deflandre, J. Craeymeersch, S. Degraer, C. Dounas, G. Duineveld, C. Heip, M. Herrmann, H. Hummel, I. Karakassis, M. Kedra, M. Kendall, P. Kingston, J. Laudien, A. Occhipinti-Ambrogi, E. Rachor, R. Sardá, J. Speybroeck, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, M. Wlodarska-Kowalczuk and A. Zenetos. 2009. Comparison of the performances of two biotic indices based on the MacroBen database. *Marine Ecology Progress Series* 382:297-311.

Hall, Jr., L.W., D.M. Dauer, R.W. Alden III, A.D. Uhler, J. DiLorenzo, D.T. Burton and R.D. Anderson. 2005. An integrated case study for evaluating the impacts of an oil refinery effluent on aquatic biota in the Delaware River: Sediment quality triad studies. *Human and Ecological Risk Assessment* 11:657-770.

Hunt, J.W., B.S. Anderson, B.M. Phillips, R.S. Tjeerdema, K.M. Taberski, C.J. Wilson, H.M. Puckett, M. Stephenson, R. Fairey and J. Oakden. 2001. A large-scale categorization of sites in San Francisco Bay, USA, based on the sediment quality triad, toxicity identification evaluations, and gradient studies. *Environmental Toxicology and Chemistry* 20:1252-1265.

Josefson, A.B., J.L.S. Hansen, G. Asmund and P. Johansen. 2008. Threshold response of benthic macrofauna integrity to metal contamination in West Greenland. *Marine Pollution Bulletin* 56:1265-1274.

Labrune, C. J.M. Amouroux, R. Sarda, E. Dutrieux, S. Thorin, R. Rosenberg and A. Grémare. 2006. Characterization of the ecological quality of the coastal Gulf of Lions (NW Mediterranean). A comparative approach based on three biotic indices. *Marine Pollution Bulletin* 52:34-47.

Landis, J.R. and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159-174.

Long, E.R., D.D. MacDonald, S.L. Smith and F.D. Calder. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sedmients. *Environmental Management* 19:81-97.

Marques, J.C., F. Salas, J. Patrício, H. Teixeira and J.M. Neto. 2009. Ecological Indicators for Coastal and Estuarine Environmental Assessment: A User Guide. Wit Press. Southampton. Boston, MA.

Monserud, R. and R. Leemans. 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling* 62:275-293.

Muniz, P., N. Venturini, A.M.S. Pires-Vanin, L.R. Tommasi and Á. Borja. 2005. Testing the applicability of a Marine Biotic Index (AMBI) for assessing the ecological quality of soft-bottom benthic communities in the South America Atlantic region. *Marine Pollution Bulletin* 50:624-637.

Muxika, I., Á. Borja and W. Bonne. 2005. The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. *Ecological Indicators* 5:19-31.

Muxika, I., Á. Borja and J. Bald. 2007. Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. *Marine Pollution Bulletin* 55:16-29.

Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Cadien, R.G. Velarde and K.J. Ritter. 2009. Calibration and evaluation of five indicators of benthic community condition in two California bay and estuary habitats. *Marine Pollution Bulletin* 59:5-13.

Ranasinghe, J.A., K.I. Welch, P.N. Slattery, D.E. Montagne, D.D. Huff, H. Lee II, J.L. Hyland, B. Thompson, S.B. Weisberg, J.M. Oakden, D.B. Cadien, Velarde. *In Press*. Habitat-related benthic macrofaunal assemblages of bays and estuaries of the Western United States. *Integrated Environmental Assessment and Management*.

Salas, F., J.M. Neto, Á. Borja and J.C. Marques. 2004. Evaluation of the applicability of a marine biotic index to characterise the status of estuarine ecosystems: the case of Mondego estuary (Portugal). *Ecological Indicators* 4:215-225.

Smith, R.W., M. Bergen, S.B. Weisberg, D.B. Cadien, A. Dalkey, D.E. Montagne, J.K. Stull and R.G. Velarde. 2001. Benthic response index for assessing infaunal communities on the southern California mainland shelf. *Ecological Applications* 11:1073-1087.

Smith, R.W., J.A. Ranasinghe, S.B. Weisberg, D.E. Montagne, D.B. Cadien, T.K. Mikel, R.G. Velarde and A. Dalkey. 2003. Extending the southern California Benthic Response Index to assess benthic condition in bays. Technical Report No 410. Southern California Coastal Water Research Project. Westminster, CA.

Teixeira, H., F. Salas, Á. Borja, J.M. Neto, J. Patrício, R. Pinto, H. Veríssimo, H., J.Á. García-Charton, C. Marcos, Á. Pérez-Ruzafa and J.C. Marques. 2008. Ecological indices tracking distinct impacts along disturbance-recovery gradients in a temperate NE Atlantic Estuary – Guidance on reference values. *Estuarine, Coastal and Shelf Science* 80:130-140.

Warwick, R.M., K.R. Clarke and P.J. Somerfield. 2010. Exploring the marine biotic index (AMBI): Variations on a theme by Ángel Borja. *Marine Pollution Bulletin* 60:554-559.

Weisberg, S.B., B. Thompson, J.A. Ranasinghe, D.E. Montagne, D.B. Cadien, D.M. Dauer, D.R. Diener, J.S. Oliver, D.J. Reish, R.G. Velarde and J.Q. Word. 2008. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. *Ecological Indicators* 8:389-394.

Acknowledgments

This study was supported by a Ph.D. Grant (SFRH/BD/24430/2005) awarded by the Portuguese Foundation for Science and Technology (FCT), and also by the research projects WISER (FP7-ENV-2008-226273; funding by European Union) and RECONNECT (PTDC/MAR/64627/2006; funding by the FCT).

Supplemental Information

Supplemental information is available online at ftp://ftp.sccwrp.org/pub/download/DOCUMENTS/AnnualReports/2011AnnualReport/ar11_SupplementalInfo_AZTI_AMBI.xls.pdf