A Summary of Existing Benthic Multi-Metric Biological Indices for Coastal and Estuarine Environments with Recommendations for Long Island Sound Embayments

Prepared for:

New England Interstate Water Pollution Control Commission (NEIWPCC)
650 Suffolk Street, Suite 410
Lowell, MA 01854

and

The Connecticut Department of Energy and Environmental Protection (CT DEEP)
79 Elm Street
Hartford, CT 06106-5127

Prepared by:

Tetra Tech
Center for Ecological Sciences
73 Main Street, #38
Montpelier, VT 05602

October 23, 2020

October 23,2020

Table of Contents

October 23,2020

1 Background/Introduction

The Long Island Sound Study's Comprehensive Conservation and Management Plan (CCMP, LISS 2015) calls for bolstering research, monitoring, and assessment to support effective management of the resources of Long Island Sound (LIS). Recently, the Long Island Sound Study (LISS) has placed additional emphasis on protecting and improving water quality, habitat, and biological diversity within the Sound's harbors and bays. In support of this focus, the program seeks additional scientific information to better assess the ecological condition of LIS embayments.

Assessments of benthic (bottom-dwelling) macroinvertebrates provide an important tool for evaluating the ecological condition of estuarine habitats. Metrics and indices based on the types, diversity, and abundances of benthic organisms can serve as useful indicators of ecological health with multiple applications. They can be employed to assess status and trends, serve as warning signs of degrading condition, or establish benchmarks to measure progress in restoration. Monitoring by the USEPA's National Coastal Condition Assessment (NCCA) is centered on assessments of benthic macroinvertebrates within the open waters of LIS, but not primarily within its coastal embayments (see map in Section 6.1). Precursors to the NCCA were the EMAP and NCA programs, which did sample embayments, data from which may be used to supplement the LISS embayment samples.

For the sampling year 2020, an intensification of the NCCA monitoring will provide for benthic macroinvertebrate sampling and assessment within LIS embayments. The monitoring will utilize the power of random statistical design and standard collection and analytical techniques of the NCCA to characterize the nutrients, sediments, and benthic macroinvertebrate community in Long Island Sound embayments. Embayment sampling in 2020 will include up to 60 stations in both New York and Connecticut waters (see map and further details in Section 6.2). USEPA contractors will conduct sampling following NCCA methods. Embayment monitoring will provide quality-assured data on water quality parameters (dissolved oxygen (DO), salinity, temperature, pH, water clarity), dissolved nutrients (nitrogen and phosphorus, chlorophyll a), sediment chemistry and toxicity (contaminant concentrations and bulk sediment toxicity), and benthic macroinvertebrate assemblages. The embayment sampling will be conducted in addition to the normal 2020 NCCA sampling conducted by Connecticut Department of Energy and Environmental Protection (CT DEEP) in the open waters of LIS (23 sites).

Currently, Tetra Tech is conducting an evaluation of existing estuarine and coastal benthic biological indices in support of development of a benthic macroinvertebrate index for the LIS embayments. To inform this effort, we reviewed a suite of papers from published literature to characterize methods and analytical approaches successfully employed in development of benthic indices in other estuarine and coastal systems. In the following sections, we present a review and summary of key elements of the literature reviewed. Included are an overview of previous studies, discussion of study design approaches, sampling methods, and indicator development analytical approaches. In the final section, key concepts from the literature review are integrated into recommendations for the LIS embayment index development. Tetra Tech's project is in support of the LISS, through a contract agreement with the New England Interstate Water Pollution Control Commission (NEIWPCC).

2 Estuarine and Coastal Benthic Macroinvertebrate Studies

The following estuarine and coastal benthic studies were reviewed and found to be relevant to benthic macroinvertebrate index development in embayments. They are the basis for information cited in the narrative of this report. Most studies were suggested by members of NEIPCC and the Tetra Tech project team based on prior work and familiarity with the subject matter. The study selection was augmented by studies found through journal searches using the terminology "estuarine benthic index united states", "estuarine benthic biological indices", and "estuarine benthic index long island", for instance. Particular attention was paid to studies whose index was developed for an area geographically similar to the Long Island Sound (e.g., Virginian Biogeographic Province, bodies of water with extensive embayments). These studies are presented in four general topic areas: 1) multi-metric index development (development of indices comprising several metrics of benthic community structure and function); 2) tolerance index development (development of indices based on rankings of species' tolerance/sensitivity to pollution); 3) index comparison and tests of application; and 4) selected literature that might be helpful in interpretation of results, including shifting trends in benthic community function within the northeast. The information presented below and additional details are included in Appendix A: 'Benthic Index_Literature Review Table.xlsx'.

2.1 Multi-metric Index Development

1. Mississippi Department of Environmental Quality (MDEQ). 2013. The Gulf Benthic Index for Mississippi (GBI-MS). Prepared for: Mississippi Department of Environmental Quality, Office of Pollution Control, Jackson, MS. Prepared by: Tetra Tech, Inc., Montpelier, VT. 30pp. plus 2 appendixes.
    a. Data/Source: USEPA National Coastal Assessment (NCA): 2000-2006; Mississippi Coastal Assessment (MCA): 2007-2011. No 2005 data due to Hurricane Katrina.
    b. Geography Covered: Mississippi estuarine and near-coastal areas, U.S.
2. Gulf of Mexico Alliance (GOMA). 2011. Benthic Index of Biological Integrity for Estuarine and Near-Coastal Waters of the Gulf of Mexico. Prepared for: The Gulf of Mexico Alliance (GOMA) and the Mississippi Department of Environmental Quality, Surface Waters Division, Jackson, MS. Prepared by: Tetra Tech, Inc., Center for Ecological Sciences, Owings Mills, MD. 79pp. plus 4 appendixes.
    a. Data/Source: USEPA NCA: 2000-2006
    b. Geography Covered: Estuarine and near-coastal waters of Texas, Louisiana, Mississippi, Alabama, Florida, Puerto Rico, and the Virgin Islands

3. Hale, S.S. and J.F. Heltshe. 2008. Signals from the benthos: Development and evaluation of a benthic index for the nearshore Gulf of Maine. Ecological Indicators 8:338-350.
   a. Data/Source: Calibration dataset –USEPA NCA: 2000-2001. Validation datasets – Massachusetts Water Resources Authority (MWRA) study of Boston Harbor and Massachusetts Bay (115 stations from 1991-1994), 101 stations from NCA 2002-2003 data, 37 stations in Casco Bay (a NOAA study conducted by Woods Hole, Larsen et al. 1983), used only to help select best index from list of candidates, rather than as validation
   b. Geography Covered: Gulf of Maine, Acadian Biogeographic Province, Northeast, United States. States include Maine, New Hampshire, and Massachusetts.
4. Malloy, K.J., D. Wade, A. Janicki, S.A. Grabe, and R. Nijbroek. 2007. Development of a benthic index to assess sediment quality in the Tampa Bay Estuary. Marine Pollution Bulletin 54:22-31.
   a. Data/Source: Tampa Bay Benthic Monitoring Program: 1993-2001
   b. Geography Covered: Tampa Bay Estuary, Florida, U.S.
5. Mebane, C.A., T.R. Maret, R.M. Hughes. 2003. An Index of Biological Integrity (IBI) for Pacific Northwest Rivers. American Fisheries Society 132, 239-261.
   a. Data/Source: Upper Snake River Basin Data – Maret 1997, O'Dell et al. 1998, USGS 2002a, Idaho Department of Fish and Game, Idaho Department of Environmental Quality, Idaho State University, and Idaho Power Company. Validation data - USGS NAWQA (Cuffney et al. 1997, Mullins 1997, Munn and Gruber 1997, USGS 2002b) and USEPA's EMAP (John Stoddard, USEPA, unpublished data).
   b. Geography Covered: Pacific Northwest, U.S.
6. Llanso, R.J. and M. Southerland. 2006. Hudson River estuary biocriteria application and validation. Submitted to: New York State Department of Environmental Conservation (NYSDEC), Albany, NY. Submitted by: Versar, Inc., Columbia, MD. 15pp. plus 3 appendixes.
   a. Data/Source: New York Department of Environmental Conservation: 2004
   b. Geography Covered: Hudson River Estuary, NY, U.S.
7. Llansó, R., M. Southerland, J. Vølstad, D. Strebel, and G. Mercurio. 2003. Hudson River estuary biocriteria final report. Submitted to: New York State Department of Environmental Conservation (NYSDEC), Albany, NY. Submitted by Versar, Inc., Columbia, MD, Tetra Tech, Inc., Owings Mills, MD. 110pp. Plus 3 appendixes and an executive summary.
   a. Data/Source: New York Department of Environmental Conservation sampling: 2000, 2001
   b. Geography Covered: Hudson River Estuary; Troy Dam downstream to the southern tip of Manhattan, U.S.
8. Llanso, R.J., L.C. Scott, J.L. Hyland, D.M. Dauer, D.E Russell, and F.W. Kutz. 2002. An estuarine benthic index of biotic integrity for the Mid-Atlantic region of the United States. II. Index Development. Estuaries 25(6A):1231-1242.

   a. Data/Source: Federal Sampling Programs – Mid-Atlantic Integrated Assessment (MAIA): 1997-1998, USEPA EMAP Carolinian Province: 1997-1998 and 1994-1997, NOAA National Status and Trends: 1997-1998. State Sampling Programs – Maryland Long-Term Benthic Monitoring: 1984-1998, Virginia Benthic Monitoring: 1985-1998, Environmental Monitoring and Assessment Virginian Province: 1990-1993, Coastal Bays Joint Assessment: 1993
   b. Geography Covered: Chesapeake Bay and Delaware estuaries, Chowan and Neuse Rivers, Pamlico-Albemarle Sounds and tributaries, Virginian Biogeographic Province, Delaware and Maryland Coastal Bays, U.S.

9. Paul, J.F., K.J. Scott, D.E. Campbell, J.H. Gentile, C.S. Strobel, R.M. Valente, S.B. Weisberg, A.F. Holland, and J.A. Ranasinghe. 2001. Developing and applying a benthic index of estuarine condition for the Virginian Biogeographic Province. Biological Indicators 1:83-99.
   a. Data/Source: USEPA EMAP: 1990-1993
   b. Geography Covered: Virginian Biogeographic Province, U.S.
10. Engle, V.D. and K. Summers. 1999. Refinement, validation, and application of a benthic condition index for northern Gulf of Mexico estuaries. Estuaries 22(3, Part A):624-635.
   a. Data/Source: USEPA EMAP-E: 1991, 1992
   b. Geographic Cover: Gulf of Mexico
11. Van Dolah, R.F., J.L. Hyland, A.F. Holland, J.S. Rosen, and T.R. Snoots. 1999. A benthic index of biological integrity for assessing habitat quality in estuaries of the southeastern USA. Marine Environmental Research 48:269-283.
   a. Data/Source: USEPA EMAP Carolinian Province: 1993-1995
   b. Geography Covered: Virginia, North Carolina, South Carolina, Georgia and Florida, U.S.
12. Weisberg, S.B., J.A. Ranasinghe, L.C. Schaffner, R.J. Diaz, D.M. Dauer, and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. Estuaries 20(1):149-158.
   a. Data/Source: Five CB sampling programs – CB Benthic Monitoring Program (MD): 1984-1994, CB Benthic Monitoring Program (VA): 1985-1994, USEPA EMAP: 1990-1993, James River Study: 1971-1972, Wolf Trap: 1987-1991
   b. Geography Covered: Chesapeake Bay, U.S.
13. Engle, V.D., K. Summers, and G.R. Gaston. 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. Estuaries 17(2):372-384.
   a. Data/Source: USEPA EMAP: 1991
   b. Geographic Cover: Gulf of Mexico states in the U.S., Florida to Texas
14. Whitlatch, R.B., R.N. Zajac. ~2011/Unpublished. Development of a Long Island Sound (LIS) Benthic Index for Assessing Environmental Conditions. Connecticut Sea Grant.
   a. Data/Source: USEPA EMAP, 1990-1993 (for index development); USEPA NCA 2000-2003 (index application)
   b. Geographic Cover: Long Island Sound, NY and CT, U.S.

2.2 Tolerance Index Development

1. Pelletier, M.C., D.J. Gillett, A. Hamilton, T. Grayson, V. Hansen, E.W. Leppo, S.B. Weisberg, and A. Borja. 2018. Adaptation and application of multivariate AMBI (M-AMBI) in US coastal waters. Ecological Indicators 89:818-827.
    a. Data/Source: USEPA NCA: 1999-2006. 3 validation datasets: Mid-Atlantic, Llanso et al. 2002; Southeast, Hyland et al. 1999b; Southern California, Weisberg et al. 2008
    b. Geography Covered: U.S. Coastal Waters (Pacific and Atlantic coasts, Gulf of Mexico)
2. Gillett, D.J., S.B. Weisberg, T. Grayson, A. Hamilton, V. Hansen, E.W. Leppo, M.C. Pelletier, A. Borja, D. Cadien, D. Dauer, R. Diaz, M. Dutch, J.L. Hyland, M. Kellogg, P.F. Larsen, J.S. Levinton, R. Llanso, L.L. Lovell, P.A. Montagna, D. Pasko, C.A. Phillips, C Rakocinski, J.A. Ranasinghe, D.M. Sanger, H. Teixeira, R.F. Van Dolah, R.G. Velarde, and K.I. Welch. 2015. Effect of ecological group classification schemes on performance of the AMBI benthic index in US coastal waters. Ecological Indicators 50:99-107.
    a. Data/Source: Ecological Group Assignment: USEPA NCA: 2000-2006, Regional Validation Datasets: Used in the creation of Southern California Bays and Estuaries Benthic Response Index, southeast U.S. (North Carolina to northern Florida) Benthic Index of Biotic Integrity, Mid-Atlantic Integrated Assessment (MAIA).
    b. Geographic Cover: Southern California (Southern California bight bays and estuaries. Point Conception to US-Mexico border), Southeast (US Southeast estuaries. North Carolina to northern Florida), and Mid-Atlantic (US Mid-Atlantic estuaries. Delaware Bay, DE to Pamlico Sound, NC), U.S.
3. Smith, R.W., M. Bergen, S.B. Weisberg, D. Cadien, A. Dalkey, D. 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.
    a. Data/Source: Six Southern California Bight (SCB) sampling programs: 1973-1994 – City of Los Angeles (1985, 1990), City of San Diego (1985, 1990), CSDLAC (1973, 1985, 1990), CSDOC (1985, 1990), Southern California Bight Pilot Project (1994), SCCWRP (1977, 1985, 1990)
    b. Geography Covered: Point Conception, CA and the U.S./Mexico border

2.3 Index Comparison and Tests of Application

1. Pelletier, M.C., A.J. Gold, I. Gonzalez, and C. Oviatt. 2012. Application of multiple index development approaches to benthic invertebrate data from the Virginian Biogeographic Province, USA. Ecological Indicators 23:176-188.
    a. Data/Source: USEPA EMAP. Encompasses 7 different monitoring efforts between 1990 and 2006
    b. Geography Covered: Virginian Biogeographic Province, U.S.

2. Rakocinski, C.F. 2012. Evaluating macrobenthic process indicators in relation to organic enrichment and hypoxia. Ecological Indicators 13:1-12.
    a. Data/Source: Sources unspecified: 2002-2005
    b. Geography Covered: North-central Gulf of Mexico – Grand Bay National Estuarine Reserve, Mobile Bay and Weeks Bay, East Bay within the larger Pensacola Bay system
3. Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Caiden, 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.
    a. Data/Source: Southern California Marine Bays – Bight (1998), Bight (2003), San Diego TMDL (2001-2002), SCCWRP and SPAWAR (2004), USEPA EMAP (1999). Polyhaline central San Francisco Bay: EMAP (2000), BADA (1994-1997), BPTCP (1994, 1997), RMP (1994-2000)
    b. Geography Covered: Southern California, San Francisco Bay, U.S.
4. Borja, A., D. Dauer, R. Diaz, R.J. Llanso, I. Muxika, J.G. Rodriguez, and L. Schaffner. 2008. Assessing estuarine benthic quality conditions in Chesapeake Bay: A comparison of three indices. Ecological Indicators 8:395-403.
    a. Data/Source: Chesapeake Bay Benthic Monitoring Program: 2003
    b. Geography Covered: Chesapeake Bay, Maryland and Virginia, U.S.
5. Borja, A. and D.M. Dauer. 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecological Indicators 8:331-337.
    a. Data/Source: Not applicable
    b. Geography Covered: Not applicable
6. Muxika, I., A. Borja, and W. Bonne. 2004. The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. Ecological Indicators 5:19-31.
    a. Data/Source: Data from six sources referenced in study
    b. Geography Covered: Incomplete – Spanish Basque coast, Spanish Mediterranean, Greece, Sweden, Belgian Coast
7. Nestlerode, J.A., M.C. Murrell, J.D. Hagy III, L. Hartwell, J.A. Lisa. 2019. Bioassessment of a Northwest Florida Estuary Using Benthic Macroinvertebrates. Integrated Environmental Assessment and Management 00, 1-12.
    a. Data/Source: Samples collected directly from the Pensacola Bay Estuary in 2016. Reference dataset from USEPA: 2000-2006
8. Dauer, D.M. and R.J. Llanso. 2003. Spatial scales and probability-based sampling in determining levels of benthic community degradation in the Chesapeake Bay. Environmental Monitoring and Assessment 81:175-186.
    a. Data/Source: CB Benthic Monitoring Program: 1999
    b. Geographic Cover: Chesapeake Bay, U.S.

9. Ranasinghe, J.A., J.B. Frithsen, F.W. Kutz, J.F. Paul, D.E. Russell, R.A. Batiuk, J.L. Hyland, J. Scott, and D. Dauer. 2001. Application of two indices of benthic community condition in Chesapeake Bay. Environmetrics 13:499-511.
   a. Data/Source: USEPA EMAP: 1990-1993
   b. Geography Covered: Chesapeake Bay, U.S.
10. Dauer, D.M., J.A. Ranasinghe, and J.B. Weisberg. 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in the Chesapeake Bay. Estuaries 23:80-96.
    a. Data/Source: Macrobenthic infaunal data from 5 Chesapeake Bay sampling programs: CB Benthic Monitoring Program (MD); 1984-1994, CB Benthic Monitoring Program (VA); 1985-1994, USEPA Environmental Monitoring and Assessment Program (EMAP); 1990-1993, James River Study; 1971-1972, Wolf Trap; 1987-1991. Only data processed with 0.5-mm mesh aperture and identified to the lowers possible taxonomic level were used.
    b. Geography Covered: Chesapeake Bay, U.S.

2.4 Other

1. Hale, S.S., M.M. Hughes, and H.W Buffum, H.W. 2018a. Historical trends of benthic invertebrate biodiversity spanning 182 years in a southern New England estuary. Estuaries and Coasts 41:1525-1538. Included because information on biodiversity trends in southern New England may be pertinent to interpretation of LISS embayments sampling results and index development.
   a. Data/Source: Data from 104 studies including: Joseph Totten's collections 1834-1835, August Gould's report on Mass. Invertebrates 1841, Joseph Leidy's 1885 collections, Addison Verrill/US Fish Commission surveys from 1870s-1890s, Newport Marine Zoological Laboratory at Castle Hill 1873-1910, Barnes 1906 list, Sumner's 1911 surveys, 1951-1986 Narragansett Bay benthic invertebrate study compilation, Invertebrate Zoology Collections of Yale Peabody Museum of Natural History database, Invertebrate Zoology Collections of the Smithsonian National Museum of Natural History database, MCZ Collections Database of the Harvard Museum of Comparative Zoology database, Ocean Biogeographic Information System, USEPA National Monitoring Program (EMAP, NCA, NCCA), Books and reports digitized by the Biodiversity Heritage Library
   b. Geographic Cover: Narragansett Bay, New England, U.S.
2. Hale, S.S., H.W. Buffum, and M.M. Hughes. 2018b. Six decades of change in pollution and benthic invertebrate biodiversity in a southern New England estuary. Marine Pollution Bulletin 133:77-87. Included because information on long-term trends in pollution and biodiversity trends in southern New England may be pertinent to interpretation of LISS embayments sampling results and index development.
   a. Data/Source: Combined USEPA National Coastal Program Survey Dataset (1990-2015) – EMAP: 1990-1993, NCA: 2000-2006, NCCA: 2010, 2015. Supplemented

with data from 2 references sites (North Jamestown and Spar Island) with comparable data from other studies.

b. Geographic Cover: Narragansett Bay, New England, U.S.

3. Hale, S., H.W. Buffum, J.A. Kiddon, and M.M. Hughes. 2017. Subtidal benthic invertebrates shifting northward along the US Atlantic coast. Estuaries and Coasts 40:1744-1756. Included because information on changes in benthic invertebrate distributions in coastal regions that could include New England may be pertinent to interpretation of LISS embayments sampling results and index development.
   a. Data/Source: Data from 3 USEPA coastal assessment programs that used similar sampling design and methods: EMAP-E, NCA, NCCA from time periods 1990-1993, 1994-1997, 2000-2006, and 2010.
   b. Geographic Cover: Data from the Carolinian and Virginian biogeographic regions of the U.S.
4. Hale, S.S., G. Cicchetti, and C.F. Deacutis. 2016. Eutrophication and hypoxia diminish ecosystem functions of benthic communities in a New England estuary. Frontiers in Marine Science 3:Article 249. Included because it is a region-specific evaluation of the effects of two major stressors on benthic community function and thus may be helpful interpreting LIS benthic sampling results and index development.
   a. Data/Source: USEPA NCA and National Coastal Condition Assessment (NCCA) 1990-2010 combined dataset
   b. Geographic Cover: Narragansett Bay, Rhode Island and Massachusetts, U.S.
5. Pelletier, M.C., A.J. Gold, J.F. Heltshe, H.W. Buffum, 2010. A method to identify estuarine macroinvertebrate pollution indicator species in the Virginian Biogrographic Province. Ecological Indicators 10, 1037-1048.
   a. Data/Source: USEPA's EMAP (five separate monitoring efforts between 1990-2001)
   b. Geographic Cover: Virginian Biogeographic Province

3 Study Design

3.1 Reference and Degraded Site Type Designation

In the context of environmental sampling and condition index development, reference sites are needed to provide an expectation for 'acceptable' biotic community conditions, to which biotic characteristics from other locations can be compared. In general, reference sites have been considered those in areas without anthropogenic influence; however, such 'pristine' (or 'natural') reference sites are rarely available in some areas like the highly developed LIS watershed. In order to utilize the best available conditions to set a 'good' baseline for comparison, the reference site concept has been expanded to include 'minimally disturbed conditions' (MDC), 'historical conditions' (HC), 'least disturbed conditions' (LDC), and 'best attainable conditions' (BAC) (Stoddard et al. 2006). While sampling programs, including the NCCA, sometimes tentatively identify certain sampling locations as 'reference' based on expert knowledge or readily available information such as surrounding landuse, it is often necessary to then use collected data to confirm appropriate reference classification. In studies that have done this in the context of developing a benthic condition index, many use habitat data (e.g., sediment contaminant concentrations, sediment toxicity testing, bottom water DO levels, etc.) as condition indicators that are independent of biological community data. Threshold values for these indicators can be chosen, potentially reflecting location-specific conditions as appropriate, in order to define a class of reference sites. The establishment of reference sites (and conversely, of degraded sites) is often one of the first steps in developing an index. Of the studies that included reference site information, most chose three criteria with which to establish their reference sites – high bottom DO, low sediment contamination, and high percent survival rate of the amphipod *Ampelisca* in toxicity tests. Three studies forewent the *Ampelisca* percent survival rate in favor of low total organic carbon (TOC) content of sediments. Three studies used four criteria: low metal and organic sediment contamination, high *Ampelisca* survival rate, low sediment TOC, and high bottom DO. In Hale et al. (2016), reference sites are classified as "normoxic sites" in that concentrations of bottom water DO were equal to or exceed 4.8 mg/L, as the goal was to better understand benthic community structure and function between sites in seasonally hypoxic areas and sites in normoxic areas in the absence of other stressors. Total organic carbon and sediment contaminants and toxicity were later used to eliminate sites with polluted sediments.

Three of the studies reviewed (GOMA 2011, Llanso et al. 2003, Ranasinghe et al. 2009) determined reference sites using stressor gradients. The quantification of a pollution gradient also was integral to the development of the Benthic Response Index (BRI) in Smith et al. (2001). Ordination analysis was used to quantify a pollution gradient within its sample calibration dataset and the pollution tolerance of each species was then determined based on distribution of abundance along the pollution gradient. Thresholds were established for reference condition and other levels of biological response: no effects range- median (ER-M, Long et al. 1995) exceedances, no more than one chemical exceeded effects range- low (ER-L), TOC concentration was what was expected based on sediment grain size, and the sample was distant from known contaminant sources.

The thresholds for reference site criteria vary from study to study, but are represented by the following ranges: bottom water DO – between >3 mg/L and >7 mg/L; sediment TOC – from <2 to 4%; *Ampelisca* survival – from $\geq$80 to 90% of control survival. Most studies that included sediment contaminant criteria used the ER-M and ER-L scalars (Long et al. 1995) to define sediment contamination levels to set thresholds that would indicate minimal or measurable biological responses in the benthic community. ER-M, for instance, represents the concentration at which 50% of collected data demonstrated adverse biological effects. No study that used ER-M and ER-L values allowed a contaminant to exceed an ER-M value and be considered a reference site. Pelletier et al. (2017) in addition allowed no ER-L exceedances for reference sites. Malloy et al. (2007) preferred the use of probable effects level (PEL) and threshold effects level (TEL) in the development of the Tampa Bay Estuary benthic index, a scalar approach developed specifically for Florida estuarine sediments. No TEL or PEL exceedances were allowed for reference sites. Most often, sites were required to meet each of the study-established criteria to be considered a reference site.

Both reference and degraded site groups must be defined to support testing of candidate metrics and indices. Similar stress indicators but with different thresholds were used in most reviewed studies to classify degraded sites: bottom water DO – between <2 and <5 mg/l; total organic carbon – between >3 and >5%; *Ampelisca* survival – between <50 and <80% of control survival. Most studies set the "degraded" site ER-M threshold to any (i.e. at least one) instance of value exceedance. ER-L thresholds varied between two or more, three or more, at least four or more, and over ten (Hale and Heltshe 2008, Paul et al. 2001, Pelletier et al. 2012) value exceedances. Malloy et al. (2007) established that the exceedance of any PEL resulted in "degraded" status while Van Dolah et al. (1999) established that three or greater ER-L/TEL exceedances or more than one ER-M/PEL exceedance resulted in "degraded" status. Most studies considered sites to be "stressed" or "degraded" if any criteria were failed. Pelletier et al. (2017) required that at least two of the criteria be met for a site to be considered "degraded".

Llanso et al. (2003) relied on a summed ER-M quotient and ER-M and ER-L values for 24 sediment contaminants as criteria for classifying reference and degraded sites as these criteria provide "a measure of the magnitude of sediment contamination in relation to biological effects." The study used ERM quotient values developed by Hyland et al. (1999a, 1999b) to describe the potential benthic impact due to the presence of sediment contamination. Summed ER-M quotients >2.352 "indicate presence of contaminants at concentrations likely to cause adverse effects on benthic assemblages", summed ER-M quotients between 2.352 and 11.352 represent "a range where high risks of benthic impacts are expected", and values above 11.352 represent "very high risk levels" of expected benthic impacts. Sites fell into either reference (no contaminants exceeded ER-M, no more than two contaminants exceeded ER-L, and sediment quotients were below the median value for the Virginian Province (associated with low risk of benthic impacts)), intermediate (didn't meet criteria for either reference or degraded), or degraded (any chemical exceeded ER-M and had sediment quotients above the median Virginian Province value) classifications.

3.1 Classification to Account for Natural Variation

The natural structure and composition of estuarine benthic assemblages is related to variation in habitat characteristics like salinity and sediment type. If such natural variation can be accounted for by partitioning habitats or other statistical means, assemblage responses to disturbance can potentially be detected with greater certainty (Smith et al. 2001). Of the studies reviewed that undertook site classification, most used cluster analyses of biological communities to define major habitat types (i.e. station classes or groups). The Mississippi Department of Environmental Quality (MDEQ) (2013) and Gulf of Mexico Alliance (GOMA) (2011) used cluster analysis and non-metric multidimensional scaling (NMS) ordination to determine classes. Smith et al. (2001) employed ordination analyses to quantify gradients of species change along environmental gradients, defining endmembers of the pollution gradient within ordination space using *a priori* defined unpolluted (reference) and polluted sites.

As an alternative to site classification, some studies accounted for natural variation in benthic metric values by investigating the presence of significant correlations between benthic metrics and habitat variables (e.g., salinity, sediment type, depth, etc.), and if present, using those relationships to 'correct' metric values for expectations based on the habitat variable (Paul et al. 2001, Engle et al. 1994). Pelletier et al. (2018) defined habitat by salinity zone based on the Venice Classification System to remove salinity bias seen in AZTI-Marine Biotic Index (AMBI) when applied in the U.S.

Most of the studies reviewed found salinity to be a key driver and utilized some analytical approach to account for it (MDEQ 2013, GOMA 2011, Borja et al 2008, Engle et al. 1994, Llanso et al. 2003, Llanso and Southerland 2006, Paul et al. 2001). A few studies (Pelletier et al. 2012, Ranasinghe et al. 2001, Weisberg et al. 1997) employed a salinity and sediment grain type habitat classification. Two studies (Engle and Summers 1999 and Llanso et al. 2002) included salinity, sediment grain type, and geographic location; and two additional studies (Pelletier et al. 2018 and Van Dolah et al. 1999) employed salinity and geographic location. Pelletier identified location categories because of differences in sampling gear among locations. As Dauer and Llanso (2003) sought to investigate the extent of benthic community degradation in the Chesapeake Bay at three different spatial scales, strata with which degradation was investigated were geographic and chosen based on managerially useful units for restoration.

3.3 Sampling Scheme

The NCCA is a probability-based survey of the Nation's coastal and estuarine waters (USEPA 2014). Using input from states and other partners, the USEPA used an unequal probability, stratified design to select 1,000 probabilistic sampling sites. Sample site stratification is based on major estuaries using the NOAA Coastal Assessment Framework and National Estuary Program (NEP).

Of the studies reviewed that provided sampling scheme information, most report using a probabilistic sampling design that selected stations randomly. Of those studies, four (Malloy et al. 2007, Paul et al. 2001, Van Dolah et al. 1999, and Llanso and Southerland 2006) note the specific method as stratified random sampling. Two studies, Engle and Summers (1999) and Engle et al. (1994), note hybrid probabilistic and deterministic site sampling designs. In Engle and Summers (1999), about two thirds of the sites sampled were randomly located while the remaining third of sites were specifically selected to answer sample design questions, perform quality control, and provide within- and between-year replications. Engle et al. (1994) notes a similar sampling scheme with near 60% of the locations randomly located, and near 40% of the sites selected for sediment deposition and as indicator testing and evaluation (ITE) sites due to anthropogenic impact. A few studies also mention the use of data from the Chesapeake Bay Monitoring Program which features two elements; a fixed-site monitoring sampling effort and a probability-based sampling effort, the latter of which randomly allocates 25 sampling sites among six strata within the estuary.

3.4 Disturbance Gradient Information

An alternative approach to categorizing sites according to whether they reflect reference or stressed conditions is to define stressor or disturbance gradients. One approach for defining a disturbance gradient is to assign environmental stress variables (e.g., sediment toxicity, dissolved oxygen, sediment contaminants) index scores ranging from 0-100 and average the scores for a site. Another approach is to define a stressor gradient in ordination space, using ‘good’ and ‘bad’ endmembers to establish endpoints for the gradient (Smith et al. 2001). Stressor gradients are often used to further evaluate ‘stressed’ sites, given the range of disturbances potentially represented within the category (GOMA 2011).

While fewer studies use a stressor gradient approach, GOMA (2011) developed a stressor gradient using three categories of stressors known to affect benthic assemblage health: sediment DO, the ‘good’ endmember was set as the 95th percentile of reference observations, which was then set to zero on 0-100 scale. Given the range of disturbances possibly represented within the ‘stressed’ category, sites were further evaluated using the defined stressor gradient.

Ranasinghe et al. (2009) explored the calibration of five benthic community condition indices in California bay and estuary habitats; for the Benthic Response Index (BRI) evaluation, a disturbance vector was used to account for species occurrence on the disturbance gradient, and IBI values were compared to the mean of ER-M quotients (mER-Mq = mean of ratios of individual contaminant concentrations relative to their corresponding ER-M values) to evaluate whether assessment results reflected significant differences in sediment contamination between disturbed and undisturbed samples.

4 Sampling Methods

4.1 Time of Year

Consistent with the 2015 USEPA National Coastal Condition Assessment (NCCA) field protocols, most studies (including USEPA National Coastal Assessment (NCA) for sampling prior to 2010) use a ‘summer’ sampling index period sometime between June and September (USEPA 2014), when it is assumed stresses to the benthic community will be high. Fewer studies use sampling data from outside of this summer period. No sampling data used began before May; however, sampling for some studies extended into September or October (e.g., Llanso et al. 2002, 2003, Malloy et al. 2007, Pelletier et al. 2012). One study, Rakocinski (2012), combined datasets that represented sampling periods from May-November (in 2003 and 2004), and June-December (in 2005).

4.2 Gear Type and Sampling Protocols

Several different methods have been used to collect benthic and sediment samples for data used in index studies. For the 2015 USEPA NCCA sampling program (from which many index development studies obtained their benthic and sediment sampling data), benthic samples (for community analysis) were collected using a 400 cm^2 Young modified Van Veen (or similar) grab; samples were then sieved through a 0.5 mm screen, and fixed with 10% buffered formalin solution (2 gallons of 100% Formalin (37% formaldehyde) solution buffered with 8 tablespoons

Borax and stained with ¼ teaspoon Rose Bengal crystals) (USEPA 2014). The Young modified Van Veen grab (400 cm^2), with samples sieved through a 0.5 mm screen, is the most commonly used sampling protocol, and has been adopted by some regional sampling programs, like the Tampa Bay Benthic Monitoring Program (from 1993-2001) and the NY Department of Environmental Conservation in the Hudson River Estuary (early 2000s). Uses of other gear types and screen sizes can reflect particular habitat requirements, such as Texas' shallow, sandy coastal lagoon habitats that required the use of PVC corers. Not all studies reviewed specified methods used to fix (preserve) benthic samples upon collection and sieving, though benthic samples collected from Alley Creek, a tidally-influenced, brackish water inlet from Great Neck Bay in Long Island Sound, Queens, NY, were preserved with a 10% buffered formalin solution (Hazen and Sawyer and Tetra Tech 2017).

Some sampling programs, like the Chesapeake Bay Benthic Monitoring Program collect samples using four kinds of gear. At "fixed stations", a 250 cm^2 hand operated box corer is used for the nearshore, shallow and largely sandy-bottom habitats of the mainstem bay and tributaries, while a 225 cm^2 Wildco box corers are used in water deeper than 4m. To maintain some continuity with previous sampling efforts dating back to the 1980s, a 250 cm^2 Petite Ponar grab is used at a Nanticoke River fixed site and a 440 cm^2 Young modified Van Veen grab is used for fixed stations first sampled in 1995, as well as at all probability sites (Versar, Inc. 2005).

Gear comparison studies have shown few effects on benthic macroinvertebrate sample characteristics using the standard dredge types recommended for sampling in the LIS embayments (e.g. the Young-Modified van Veen, Petite Ponar) (Caires and Chandra 2012; Dauer and Lane 2005; Elliot and Drake 1981a, 1981b; Howmiller 1971). Perhaps the most relevant study was conducted for the Virginia Department of Environmental Quality, showing that comparisons between a single Young grab and two combined Petite Ponar grabs indicated that benthic macroinvertebrate IBI and individual metric values were not significantly different (Dauer and Lane 2005). This comparison analysis was specifically designed to test gear and methods that would be applied for the NCA and other studies in the Chesapeake Bay.

4.3 Number of Grabs

Per the NCCA 2015 Field Methods manual (USEPA 2014), a single benthic sample is collected from each sampling location and submitted to the lab for species composition and abundance analysis. If a Young-modified Van Veen Grab (which samples 0.4 m^2) or similar sampler is used so that >0.03 m^2 of bottom substrate is sampled to a depth of at least 7 cm, then only one grab is collected. If <0.03 m^2 is collected with a single grab (e.g., a Petite Ponar which samples 0.0223 m^2), then two grabs are collected and composited into a single sample. Some other studies indicate taking replicate samples at each sampling location, though just under half of the studies reviewed provided specific information on number of grabs. Among the studies providing replicate grab information, most specified 3 grabs per site, although the numbers varied between 1 and 4 (e.g., Van Dolah et al. 1999).

4.4 Information Collected in Conjunction with Benthic Samples

Ancillary environmental/habitat data are typically collected in conjunction with benthic invertebrate samples. The 2015 NCCA field protocol (USEPA 2014) requires collection, when possible, of a hydrographic profile of dissolved oxygen, pH, salinity, and temperature using a multi-parameter water quality meter, starting just below the surface and taking measurements at specified depth intervals to 0.5 m from the bottom. Other parameters measured include water column transparency (using a Secchi disk), light attenuation using a PAR meter, water chemistry samples for analysis of ammonia (NH3), nitrate (NO3), nitrate-nitrite (NO3-NO2), total nitrogen (TN), total phosphorus (TP) and ortho-phosphate (PO4) (also called soluble reactive phosphorus (SRP), pH, conductivity and chlorophyll a (USEPA 2016) to determine nutrient enrichment and classification of trophic status, and algal toxin and microcystin water samples.

Most of the studies reviewed mention collecting information in conjunction with benthic samples (water quality and water chemistry data), but do not necessarily describe the specific methods. Studies that provide detail, though, most often mention collecting water column profiles of salinity, dissolved oxygen, pH, nutrient, water depth, and temperature and chlorophyll a via physical water samples, generally following protocols similar to the 2015 NCCA field protocols (i.e. taking measurements at 0.5 m intervals from the bottom to the surface). Rakocinski (2012) mentions taking measurements at 0.2m intervals between the bottom and surface. Four studies describe taking measurements for turbidity and only one, Paul et al. (2001), also measured light attenuation. Weisberg et al. (1997) is the only study that mentions measuring oxidation reduction potential (ORP).

4.5 Sediment Quality Measures

Beyond the grabs for benthic infauna, NCCA 2015 field protocols (USEPA 2014) specify that additional sediment grabs should be taken for chemical contaminant analysis (organics, metals and TOC), grain size determination, and acute whole sediment toxicity analysis. Multiple grabs are taken with the top 2 cm from each being removed and composited, until a minimum of 3 L of surface sediment is collected. The composite mixture is homogenized and split into specified sample containers, and analyzed for grain size, total organic carbon, moisture content, concentrations of metals, mercury, pesticides, and PCBs. Sediment toxicity testing is conducted using the amphipod, *Leptocherius plumulosus*, in 10-day tests with 5 replicates and 20 organisms per replicate for each sample and control.

Most of the studies reviewed mention collecting sediment grabs for sediment characteristics, chemical contaminants, and toxicity analysis, although many of the studies provide little detail regarding methods or rationale for the parameters measured. Parameters measured typically included silt-clay and total organic carbon (TOC) composition of the sediments. When chemical contaminant analyses were included, constituents analyzed typically included heavy metals, PAHs, PCBs, pesticides, and ammonia content of the sediment. Hale et al. (2018b), was additionally interested in analyzing the petroleum and synthetic organic compound content of sediments. Five studies describe the way their sediment samples were collected and prepared pre-analysis; two of which follow the 2015 NCCA method of collecting the top 2 cm of sediment, and three of which (Engle et al. 1994, Engle and Summers 1999, and Ranasinghe et al. 2001) removed and prepared 50 ml or 60 cc cores from each sediment grab for analysis. For sediment toxicity, three studies (GOMA 2011, Llanso et al. 2002, and Ranasinghe et al. 2001) used *Ampelisca abdita* % survival testing. Llanso et al. (2002) also used the Microtox bioassay to measure changes in bacterial luminescence as an indicator of acute sublethal effects in sediment elutriates.

4.6 Subsampling and Identifications

NCCA samples exhibit a wide range of organisms from under 100 to over 1000 andlaboratories are required to fully sort and identify all organisms in the sample to the lowest practical taxonomic level. Species is the target level except for Oligochaeta (Class), and Chironomidae (Family) in marine, polyhaline, and mesohaline region samples. NCCA excludes meiofauna from identification as it is under 0.5 mm in size. For each species or lowest identifiable taxonomic level, one representative organism is included in the laboratory's reference collection The U.S. EPA does not require or suggest the use of subsampling in the NCCA program. After laboratory processing of the 2015 NCCA data, a scientific panel working with data analysis recommended processing of the entire sample in all cases, without any subsampling (Pelletier, *personal communication*). There are no recommendations for subsampling in the NCCA 2020 laboratory operations manual (USEPA 2020). When subsampling was employed for the National Rivers and Streams Assessment (NRSA), subsampling procedures were as described in the 2015 Laboratory Operations manual (USEPA 2016). This subsampling method requires the sorter to evenly distribute each sample across sorting trays, placing an evenly divided grid over the sample, and randomly selecting three grids for sorting and picking organisms until all grids are sorted or the sorter has removed the maximum number of organisms from the subsample.

When noted, most studies identified all organisms to the lowest possible taxonomic level. An aspect of the utility of low taxonomic level identification is the ability to detect, through change in community composition, more subtle anthropogenic effects (e.g., change in temperature due to climate change). Several studies, though, noted that certain organisms were excluded prior to analysis, typically because they are not considered a component of the infaunal benthos (i.e. organisms living within or in close association with the substrate). Pelletier et al. (2012) excluded pelagic and epifaunal organisms to allow a better match with the Chesapeake Bay protocols to which comparisons were made. Gillett et al. (2014) excluded more than 300 taxa from consideration because they were epifaunal or identified only to a higher taxonomic level; Van Dolah et al. (1999) eliminated meiofauna (those smaller than 0.5 mm), pelagic (water column) fauna, terrestrial fauna, and any obvious epifauna from analysis; and GOMA (2011) removed taxa not classified as benthos, including terrestrial invertebrates, freshwater taxa, meiofauna, and planktonic or colonial organisms.

Few studies included information regarding subsampling methods. Mississippi Department of Environmental Quality (2013) used 200 organisms per sample as the maximum, and randomly resampled down to 200 organisms if a sample exceeded that amount in order to avoid bias in subsequent metric calculations. GOMA (2011) removed all samples from analysis below a minimum of individuals (<10 per sample). Llanso et al. (2003) and Llanso and Southerland (2006) used similar subsampling methods for their Hudson River biocriteria investigation; following the identification of oligochaetes and chironomids in a sample, if between 20 and 300 individuals were counted, "the sample was split and approximately 50% of the specimens were mounted for identification, while the remaining portion was used for biomass determinations." Samples were split by evenly spreading the specimens in a gridded tray and randomly selecting half of the total number of grids. If the number of individuals exceeded 300, grids were selected randomly until 150 specimens were selected for mounting and identification. Taxonomic counts for each oligochaete and chrionomid species were then adjusted proportionally.

5 Benthic Indices

5.1 Background

Benthic indices are tools used to monitor ecological condition through the interpretation of measures of benthic macroinvertebrate community composition. The interpretation of this complex information by benthic indices is achieved through easily communicated condition scores (Pelletier et al. 2018). Certain benthic indices can be applied not only to define ecological status and condition on a relative basis, but then can contribute to understanding the relationships between biological structure and function and the environmental stressors (contaminant exposure, organic enrichment, etc.) found to be prevalent in a region. Benthic indices are often developed using different approaches and result in different degrees of robustness, depending on the goals of the related bioassessment programs and management organizations (GOMA 2011). Some indices are developed for broad geographic areas spanning several state coastlines, some for single bodies of water, and some for entire continental estuarine waters (AMBI). There are trade-offs to consider between applicability of an index across a broader geographic range (i.e. its 'generality') and the potential 'precision' of an index in assessing site-specific conditions (Engle and Summers 1999, Hale and Heltsche 2008). For LIS embayments, there is the possibility of increasing precision through the development of an LIS embayments-specific index over the option of applying an existing index that was developed for a wider geographic area.

Benthic index development can be driven by different philosophies, some driven by ecological principal, some statistically driven, and some driven by a combination of both. The most common kinds of benthic indices used in the U.S. are multi-metric indices, ones that consider more than one component of benthic assemblages (richness, pollution tolerance, diversity, etc.) that reflect the effect of environmental stressors in order to differentiate between degraded and non-degraded sites. The multi-metric approach is emphasized in this review because that approach is the focus of anticipated analyses. Other approaches are described and will be considered in the anticipated analyses as needed.

Other benthic index approaches, like AMBI, use a species abundance-weighted tolerance approach to classifying ecological condition (*sensu* Hilsenhoff 1988), or use habitat characteristics to model the expected community for an area and compare that to the observed community as in RIVPACS (e.g., Hawkins et al. 2000). A RIVPACS approach was investigated for application in the Hudson River Estuary, but results were considered inconclusive, and the RIVPACS approach was not recommended (Llanso et al. 2003).

5.2 Multi-Metric Indices

In the development of a multi-metric benthic index, it is necessary to evaluate suites of benthic assemblage characteristics (richness, pollution tolerance, diversity, etc.), or metrics, that reflect the effects of environmental stressors in order to differentiate between degraded and non-degraded conditions. The incorporation of more than one metric in an index is expected to be more robust than an individual metric at detecting responses to a range of different potential stressors by capturing the distinctive responses of different structural and functional components of the benthic assemblages.

Per Van Dolah et al. (1999), several general approaches to developing multi-metric indices have been known to work well in U.S. estuarine environments. Most instances of index development begin with the compilation of a sometimes-large list of candidate benthic metrics to be tested for potential inclusion in the final index, gathered from literature and investigations of similar nature and/or based on accepted ecological principals. The steps used to further refine and select benthic metrics from the original candidates can comprise a large part of the index development process and are often done in varying ways. The various approaches used to evaluate and select metrics usually have the goal of identifying metrics that strongly differentiate *a priori* defined reference and degraded sites, although a more complex approach involving defining an integrated stressor gradient and testing the responses of benthic metrics along this gradient (e.g., MDEQ 2013) can also be used. Other metric selection criteria can include, for example, ecological meaningfulness, contribution of representative and unique information (i.e. lack of redundancy with other metrics) and having a sufficient range of values. The statistical methods used to achieve the goal of distinguishing between site types can range from simple t-tests or correlations to stepwise and canonical discriminant analysis or stepwise logistic regression.

In an estuarine setting, candidate metrics are also often ‘corrected’ to account for natural variations in metric values expected due to variations in salinity, sediment type, or other habitat characteristics, in order to increase the sensitivity of the metric to reflect stressor responses. Alternatively, the study area for which the index is being developed can be divided into regions that are similar in one or more of these key environmental drivers (e.g., salinity, with the study area separated into oligohaline, mesohaline, and polyhaline zones), for which separate indices are calibrated. Subsequent steps in the multi-metric index development process generally include assembly of an index (or several candidate index formulations) using the selected metrics, testing the efficacy of the index at distinguishing among reference and stressed sites, setting thresholds of index values that define good to poor conditions, and validation of the index using independent data sets. The application of these methods, in addition to others, through the development of several benthic indices are described below.

5.2.1 Candidate Metric List Compilation

There is little hard guidance on how extensive a list of candidate benthic metrics to compile for consideration in index development. Rationale for generating candidate metrics include their perceived utility in other index development efforts (e.g., Malloy et al. 2007 used the same list of candidate metrics employed by Engle and Summers 1999), accepted ecological principles, and observed distributions and patterns within the index calibration dataset. Benthic candidate metric lists tend to represent different categories of benthic assemblage components including species richness and diversity, taxonomic composition, productivity, trophic level abundance, species abundance, and pollution tolerance, as examples. While not every study provides the rationale behind their metric choices, some will. For instance, Weisberg et al. (1997) deliberately chose to include metrics less sensitive to collection gear type to make sure the index would be applicable to a wide array of datasets and avoided individual species metrics in favor of assemblage measures. Engle et al. (1994) chose metrics to represent "many of the major ecological attributes of the benthic assemblage". GOMA (2011) screened metrics with respect to the adequacy of data available for evaluation; for example, "metrics related to habit traits were dropped from consideration due to an insufficient number of taxa with associated trait designations". Additionally, the study made note of metric range sufficiency, stating that "if the range of values for a metric was very narrow, it would not function to illustrate response to stressor conditions". Studies will sometimes choose to include less common or novel metrics as well, as did Llanso et al. (2002) who included a recently (at the time) published metric, the North Carolina Sensitivity Index, and unpublished metrics, the Tolerance Score and the Tanypodinae-Chironomidae percent abundance ratio. Hale and Heltshe (2008) included a species tolerance value [ES(50)0.05] developed by Rosenberg et al. (2004) for a Benthic Quality Index (BQI). Candidate benthic metric lists often (but not always) begin larger and more encompassing, before the evaluation and selection process identifies the ones that best discriminate between degraded and non-degraded sites, and/or meet other criteria. For instance, Llanso et al. (2003) made it a point to "use as many metrics as possible initially", no matter the redundancy of some, to maximize potential metric combinations. When provided by the studies, initial candidate metric list sizes were: Mississippi Department of Environmental Quality (MDEQ) (2013) began with 14-19 metrics per habitat class; Van Dolah et al. (1999) tested 40; Weisberg et al. (1997) tested 17; GOMA (2011) tested 142; Hale and Heltshe (2008) tested 49; Llanso et al. (2002) tested 23; and Engle et al. (1994) tested 24.

5.2.2 Metric Manipulations

5.2.2.1 Corrections for Environmental Drivers.

Several studies (Malloy et al. 2007, Hale and Heltshe 2008, Engle et al. 1994, Paul et al. 2001, Engle and Summers 1999) tested the need to correct for influences of habitat attributes (salinity, grain size, etc.) on candidate metrics before evaluating the responses of these metrics to stressors, as these attributes are known to influence the abundance and diversity of resident benthic biota. Malloy et al. (2007) adjusted metrics if Pearson correlation coefficients were statistically significant at a 0.05 alpha level and had a r >0.25 value. In Hale and Heltshe (2008), each metric where the correlation with a habitat variable accounted for >25% of the variance was considered for normalization. Engle et al. (1994) used a method described by Weisberg et al. (1992) to correct for salinity in species richness and diversity metrics. Engle and Summers (1999) adjusted metrics to remove the effects of salinity based on modifications of the method used in Engle et al. (1994). Paul et al. (2001) evaluated relationships between individual characteristics and physical characteristics of habitat using Pearson correlation coefficients and normalized seven metrics using a polynomial regression for the specific habitat variable (salinity) described in Weisberg et al. (1993).

In some approaches, rather than correct metrics for salinity and/or sediment type, the study area was separated into habitat classes or strata, generally based on salinity (e.g., oligohaline, mesohaline, polyhaline, etc.), and indices were developed or calibrated separately for each class (e.g., Weisberg et al. 1997, Van Dolah et al. 1999, GOMA 2011, MDEQ 2013, Llanso et al. 2002, and Llanso et al. 2003).

5.2.2.2 Corrections for Normal Distribution

To assure normal distributions of benthic metrics, transformations may be used. For example, Engle et al. (1994) and Engle and Summers (1999) adjusted metrics based on proportions using arc-sine transformations and abundance metrics using $\log_{10}$ (value + 1) transformations. Malloy et al. (2007) adjusted proportion metrics using arc-sine transformations and abundance metrics using ln(abundance+1). Llanso and Southerland (2006) log(x+1) transformed its abundance-based metrics prior to standardization.

Metric adjustments can also account for non-linear responses using a “weighted” metric approach as demonstrated in a freshwater fish IBI (Mebane et al. 2003). In the weighted index approach, the major difference from the unweighted approach is that after metric selection, the metrics were scored on linear or curved scales, depending on responses perceived from cumulative distribution frequencies (see Section 5.2.4). This effectively accounts for non-normality of the metric distribution at the scoring step.

5.2.3 Metric Selection

Next in the process of multi-metric index development is the evaluation and selection of candidate metrics. Initial selection is commonly based on the ability of each metric to discriminate efficiently between degraded and non-degraded sites that are designated *a priori* as degraded or non-degraded (reference) based on independent criteria (often a combination of dissolve oxygen, sediment contaminant, and sediment toxicity criteria, and sometimes including degree of land use development or percent organic content of the sediments) (e.g., Weisberg et al. 1997, Van Dolah et al. 1999, GOMA 2011, MDEQ 2013, Llanso et al. 2002, Hale and Heltshe 2008, Malloy et al. 2007, Paul et al. 2001, Pelletier et al. 2012, 2017). Based on this, Llanso et al. (2002) retained as candidates metrics that, for each habitat, correctly classified at least 50% of the degraded sites in the calibration data set. For selection of candidate metrics, ability to discriminate >50% of reference and stressed sites is a good general threshold, which can be relaxed somewhat (e.g., to 40%) for some metric categories, depending on the quality of least-disturbed reference sites and the range of disturbance in the sampled sites. For individual candidate metrics, a high discrimination efficiency (e.g., 90%) is possible but not common. This resulted in twelve metrics being selected to determine the combination best suited for the final index. Selected metrics included measures of productivity (abundance), diversity (number of taxa, Shannon-Wiener diversity, percent dominance), species composition and life history (percent abundance of pollution-indicative taxa, percent abundance of pollution-sensitive taxa, percent abundance of Bivalvia, Tanypodinae-Chironomidae abundance ratio), and trophic composition (percent abundance of deep-deposit feeders).

Several studies used Mann-Whitney U-tests to compare differences in means of benthic metrics between reference and degraded sites (Weisberg et al. 1997, Van Dolah et al. 1999, Llanso et al. 2002, and Llanso et al. 2003,). In addition to the Mann-Whitney U-tests, Weisberg et al. (1997) applied the Kolmogorov-Smirnov test to abundance and biomass metrics as the “anticipated response at degraded sites could be higher or lower than at reference sites depending on the severity of the stress”. In Llanso et al. (2003), it was difficult to statistically distinguish among two groups of sites using tidal freshwater metrics, so reference and degraded site definitions were revised, and their metrics were tested using the Wilcoxon Rank test (due to smaller sample size). The results increased the number of statistically significant metrics but revealed very low values for some others and made apparent the redundancy of a few (e.g., the polychaete and spionid (Spionidae is a family of polychaete) metrics that were both represented by one species), limiting the number that could be used for index development. Few metrics responded in the expected direction in the oligohaline habitat including total abundance, total biomass, and most diversity measures. A subsequent Wilcoxon Rank test using the revised reference and degraded sites revealed additional significant metrics for that habitat. Generally, metrics that produced statically significant differences between degraded and reference sites were chosen for further indicator development. Ultimately six metrics were selected for the mesohaline index (as an example): Margalef’s diversity, total number of infaunal species, percent abundance of epifauna, number of species of polychaetes (or, alternatively, percent abundance), abundance of suspension feeding taxa, and percent biomass of carnivore-omnivore taxa.

GOMA (2011) used a combination of approaches to determine the ability of each candidate metric to differentiate between reference and stressed sites. These included discrimination efficiency (DE), z-scores, Wilcoxon U test, box-and-whisker plots, and quantile regression of scatter plots of metrics along the stressor gradient. Results of the metric evaluations were integrated using a weight of evidence approach with the most weight given to DE, z-score, and the quantile regression slope. MDEQ (2013) evaluated metric responsiveness within its habitat classes based on discrimination efficiency and z-scores. Metrics with scores that exhibited “strong and consistent responses along the stressor gradient” were kept for index evaluation. Metrics with low DE scores were considered in some cases if the z-score was high. Other selection criteria considered were ecological meaningfulness, contribution of representative and unique information, and sufficient range of values. Pearson correlations were performed among habitat classes in GOMA (2011) and MDEQ (2013) to evaluate redundancy amongst metrics, and redundant metrics were eliminated from analysis. Metrics that showed relatively strong responses along the stressor gradient were considered candidates for inclusion in site-class-specific multi-metric indices.

Iterations of stepwise and canonical discriminant analyses were applied to the candidate metrics in Engle et al. (1994) and Engle and Summers (1999) to evaluate metric ability to differentiate between degraded and reference sites. Stepwise discriminant analysis (SDA) was used to evaluate and choose subsets of candidate measures that best discriminated between degraded and reference sites. Canonical discriminant analysis (CDA) was then applied to that subset of metrics to test the classification efficiency of the set of metrics provided by SDA. Further metric selection was based on assessment of redundancy among some of the subsetted metrics.

In Malloy et al. (2007), SDA was performed to identify metrics that best discriminated between degraded and reference sites. In Paul et al. (2001), t-tests were applied first to all metrics to test the equality of means at reference and degraded sites. Then successive linear discriminant analyses (LDAs) were then used to find metrics that best discriminated between degraded and reference sites in the calibration data set; these were then tested in the validation data set. The LDA-driven metric selection process started with all candidate metrics but was iterative, subsequently adding or eliminating metrics, or limiting the number of variables in the discriminant function, to define the combination of metrics that best revealed differences between degraded and reference sites, using the SAS procedure for SDA.

To pare down their list of 49 candidate metrics that other studies had found to be useful, Hale and Heltshe (2008) used stepwise logistic regression and t-tests between the means of each metric at stations with high and low benthic environmental quality scores (BEQs), finding the best combinations of metrics through an iterative process.

5.2.4 Metric Scoring

Benthic metrics are measured in a wide range of scales, from abundances (e.g., number of individuals/m^2 that can have a range from zero to hundreds or thousands or more) to percent composition (obviously scaled from 0-100). To enable comparison across disparate metrics, they are often normalized, scaled or ‘scored’ before being combined into an index to give them (initially) even weights. In some studies, the distribution of values of each metric at reference sites is used to set percentile-based thresholds to establish a relative scoring scale that differentiates between reference and degraded sites. A scoring system for metrics based on these thresholds is then applied before final metric selection. Each metric is evaluated independently, and after scoring, selected metrics can be summed as a mode of combination in a final index.

For Llanso et al. (2003), metric thresholds were calculated based on the distribution of values at the reference sites. To address overlap between reference and degraded site distributions, the the 25th percentile of reference values for each metric-habitat combination was established as the threshold for differentiating reference conditions. Metric values above the 25th percentile received a score of 5 (on a 0 to 5 scale). Metric values falling below the 25th percentile threshold were considered to indicate impairment, and received a score of 0. For metrics expected to have higher values under degraded compared to reference conditions, the scoring system was reversed (i.e. upper thresholds of the 75th percentile were used).

Individual metric performance was evaluated using both calibration and validation datasets in separate steps. Metrics that correctly classified degraded sites using calibration and validation data at a 70% or higher were considered for inclusion in a final index. Criteria were lower for tidal freshwater mud and oligohaline habitats due to lower classification efficiencies in those habitats. Redundant metrics were excluded at this point. Effort was made to select metrics that represent different categories of biotic response.

In Llanso et al. (2002), the 10th and 50th percentile values of reference sites for each metric-habitat combination were established and the associated scores were 1 for metric values below the 10th percentile for a site, 3 for metric values between the 10th and 50th percentile, and 5 for values above the 50th percentile. Upper thresholds were established as the 90th percentile for some metrics (e.g., % abundance of pollution-indicative taxa) due to the expected response of higher percentages in degraded than reference sites and the scores are as follows: 1 for metric values above 90th percentile, 3 for metric values between 90th percentile and the median, and 5 for metric values below the median of corresponding reference values. Abundance and biomass respond bimodally to pollution (Weisberg et al. 1997) and received separate threshold values: an upper 90th percentile and a lower 10th percentile. The scoring system was established as follows: 1 for metric values below the 10th percentile or above the 90th percentile, 3 for values between the 10th and 25th percentiles or the 75th and 90th percentiles, and 5 for values between the 25th and 75th percentiles. Two scoring system modifications were made for certain cases: 1) no percentage or tolerance value metrics were scored when no fauna was present (as to not exaggerate the metric response at azoic sites); and 2) no pollution sensitive taxa related metrics that would score 3 or 5 when overall site abundance was low were scored to avoid an exaggerated response due to the presence of pollution-sensitive individuals among a small number of total individuals. Van Dolah et al. (1999) developed their scoring criteria separately for each metric based on the distribution of values at reference sites in its calibration dataset. The 10th and 50th percentiles of reference values were determined and scores of 1, 3, and 5 were given to metric values below the 10th percentile, between the 10th and 50th percentiles, or above the 50th percentile, respectively. Station index values were calculated by assigning scores to each component metric and averaging them. Index scores under three suggest "the presence of degraded benthic assemblages." Weisberg et al. (1997) established thresholds for metrics based on the distribution of values for metrics at reference sites at the 5th and 50th percentile values for reference sites in each habitat. Each metric was scored as 5, 3, or 1 depending on its degree of deviation from reference site conditions. Metric values below the 5th percentile were scored 1, values between the 5th and 50th percentiles were scored 3, and values above the 50th percentile were scored 5. As abundance and biomass respond bimodally, thresholds and scoring criteria were modified, such that metric values over the 95th and below the 5th percentile scored as 1, values between the 5th and 25th percentiles or between the 75th and 95th percentiles scored as 3, and values between the 25th and 75th percentiles scored as 5.

MDEQ (2013) scored metrics on a common scale of 0-100 prior to combination in an index as an average of all scores, with the optimal score determined by the distribution of data For metrics that decreased with increasing stress, the 95th percentile of all data within a habitat class was considered optimal and scored as 100 using the following equation: *MetricScore* = *MetricValue* – 5th Percentile divided by 95th Percentile – 5th Percentile. Metrics that increase with increasing stress were scored similarly but using the 5th Percentile of data as optimal. Percentiles other than 95th were sometimes used to reduce effects of skewed distributions. Llanso and Southerland (2006) similarly scored its metrics on a 0-100 scale prior to combination of metrics in an index by averaging scores. For metrics expected to decrease in value with increasing degradation, the 95th percentile metric value was assigned a score of 100. For metrics expected to increase with increasing degradation, the 5th percentile metric value was assigned a score of 100.Mebane et al. (2003) standardized metrics with a continuous scoring system 0-1 and weighted them as necessary to produce a 0-100 score IBI. The study implemented cumulative frequency distributions (CDFs) to "characterize the distribution of candidate metric values and to identify minimum and maximum score values". It set minimum scores to 0 and maximum scores to 1 for positive metrics, with maximum scores set "near the 95th percentile of scores for all sites, thereby reflecting a gradient of metric responses". Metric response curves, compared to a continuous disturbance gradient based on agricultural land use, were drawn. As the sites represented "wide ranges of anthropogenic disturbances", the slopes and shapes of the lines and curves reflect "the general pattern of values across the gradient of site conditions" and metric equations were fit to these responses. The response curves of metrics compared to a continuous disturbance gradient (agricultural land use) were drawn by eye and then equations were fit to the generalized response slopes or curves. This approach could be applied in an estuarine index development process if a limited number of index metrics could be pre-selected and a single disturbance gradient could be defined.

5.2.5 Index Assembly, Calibration, and Condition Thresholds

Index assembly refers to how the 'best metrics', selected through the various processes described above, are combined mathematically into a formula and then evaluated for performance. The process is often iterative; development of a formula could range from a simple additive (linear) combination of the normalized values or scores of the several final metrics selected, to a continuation of the more complicated statistical modeling (e.g., linear discriminant analysis, stepwise discriminant analysis (SDA)) that is used both to generate coefficients for each metric and a final, sometimes non-linear formulation. In some cases, multiple candidate index model formulations are generated and compared in terms of relative performance for final index model selection.

In Engle et al. (1994), the results of the first SDA suggested that one metric (proportion of expected number (i.e. salinity-adjusted) of polychaete species) alone explained 81% of the variation between degraded and non-degraded sites, and thus could be used to discriminate between these sites. It was determined that a model based on only a single metric would limit effective discrimination between sites and the SDA was run again without the polychaete richness metric, resulting in an eight-parameter model. Subsequent regression analysis showed that all but the first three metrics contributed minimally to the overall variance. CDA was performed using the first three metrics and revealed 90% of the variance was explained by the model with no misclassification of sites. The metrics chosen for index development included proportion of expected diversity, proportion of total abundance as tubificids, and proportion of total abundance as bivalves. Discriminant scores for each sample site were calculated and normalized to a scale of 0 to 10.

Engle and Summers (1999) employed SDA to find a seven-parameter model that explained 79% of the variance in the model. These seven metrics were tested for redundancy using Pearson correlations. Two of the metrics (total abundance and proportion of expected number of polychaete species) correlated significantly with the proportion of expected diversity and were removed, leaving five metrics for the final index (proportion of expected diversity, mean abundance of tubificids, % capitellids, % bivalves, and % amphipods). Metric values were normalized for inclusion in the index, using values from the EMAP-E probability-based sites in each year. The five metrics were combined into a composite index using coefficients from a CDA, with discriminant scores normalized to a range of 0 to 10 for ease of interpretation. Based on comparison of test-station index values to sites defined *a priori* as reference or degraded, index threshold values for categorizing reference and degraded sites were set as: values ≤ 3 indicate degraded sites, values >5 indicate reference sites, and values between 3 and 5 indicate sites with undetermined classification. Misclassification of sites using this 5-metric index was under 10% (9.1% of degraded sites and 4.2% of reference sites) and the model was accepted as final. The final index was significantly correlated with salinity and % silt-clay content of sediments, but it was determined that the relationships were driven by the large number of samples and were insignificant from an ecological perspective.

Malloy et al. (2007) applied a linear discriminant function to metrics selected by SDA to calculate discriminant scores, which were then used as metric coefficients in a final index. Three metrics (proportion of expected number of species, proportion of total abundance as spionid polychaetes, and proportion of total abundance as Capitellidae) accounted for a significant portion of the variation between healthy and degraded sites, and were used in the final Tampa Bay Benthic Index (TBBI) formula. Final index values were normalized using the formula presented in Engle and Summers (1999): Normalized TBBI Score = [(Raw TBBI score – M)/R)] X 100; where M = the minimum TBBI score; R = the range of TBBI scores.

Thresholds for TBBI scores to separate healthy and degraded conditions were set by comparing the cumulative distribution frequency of index values from the calibration data set, along with the frequency curves of false positives (incorrectly identified as degraded) and false negatives (failed to identify as degraded), with the intent of limiting the rate of false positive and false negative site assignments to 10%. Accordingly, index scores ≥ 83 were defined as healthy and <73 as degraded.

Based on their iterative LDA process of metric selection (above), Paul et al. (2001) generated and tested six candidate indices. Concerns raised with some of the indices included metrics that responded differently in the index (i.e. with a positive or negative sign) compared to expectation; and indices that did not meet targets for classification efficiency set at 90% and 80% in calibration and validation data sets. The chosen Virginian Biogeographic Province benthic index missed the goal for calibration by one site but met the targets for cross-validation and validation

– it included salinity normalized Gleason's D based upon infauna and epifauna, salinity-normalized tubificid abundance, and abundance of spionids.

The demarcation in the discriminant function score between reference and degraded sites was zero (because the calibration data contained the same number of sites in each category); and the index did not have to be scaled (linearly transformed) to set the demarcation at zero. Thus, the threshold for reference conditions is for the benthic index >0, and for degraded conditions is the benthic index ≤0.

The final phase of index development in Llanso et al. (2003) involved summing the scores of the metrics selected through the process described above. For the mesohaline index (as an example, with the 6 metrics indicated above), scores of 20 or higher indicated good benthic conditions and scores of 15 or lower indicated degraded benthic conditions. An index score of 10 or higher in oligohaline or tidal freshwater mud is considered good benthic condition, while a score of 5 or lower is considered degraded. Classification efficiencies (number of *a priori* reference or impaired sites correctly classified, expressed as a percentage of the total) were calculated for each condition group and the calibration (90% for the mesohaline index) and validation (83% for the mesohaline index) datasets.

In Llanso et al. (2002), developing a final index involved combining the metrics determined to be sufficiently sensitive into all possible combinations, and testing those combinations (candidate indices) for ability to discriminate reference and degraded sites using four criteria: correctly classifying sites in the calibration data set within 5% of the most efficient combination, correctly classifying sites in the validation data set within 10% of the most efficient combination, having the most number of metrics, and having a variety of functional categories. No combinations that considered biomass were considered due to sampling limitations in the MAIA region. Index values <3.0 indicated stressed benthic assemblages, suggesting the presence of degraded conditions. The index correctly classified 82% of all sites in an independent data set. Classification efficiencies of sites were higher in the mesohaline and polyhaline habitats (81-92%) than in the oligohaline (71%) and the tidal freshwater (61%).

Van Dolah et al. (1999) compiled combinations of the candidate metrics to further evaluate and determine the best combined index. Evaluations were made by calculating the rate at which each multi-metric index correctly classified degraded (index score <3) and reference (index score ≥3) stations; the metric combination of which that resulted in the highest percentage of correct classifications across various habitats was included in the final index. The index selected for use in the Carolinian Province was calculated using the average score of four metrics: mean abundance; mean number of taxa; 100 minus percent abundance of the top two numerical dominants; and percent abundance of pollution-sensitive taxa.

In Weisberg et al. (1997), metric scores were combined into an index by calculating a "mean score across all metrics for which thresholds were developed." Metrics included in the final index were limited to a single depth-distribution metric within each habitat, a single trophic composition metric, and included either an abundance- or a biomass-based metric of species composition but not both to avoid redundancy. A final index score <3 was defined as stressed assemblage conditions. The final B-IBI classified 93% of the validation sites correctly.

For final index assembly and assessment, Hale and Heltshe (2008) used several assessment criteria, including the Akaike Information Criterion (AIC) to select the best model, the area under a receiver operating characteristic (ROC) curve (where area under the curve gives a measure of separability in a classification model) to assess model discrimination between low and high quality sites, and the Hosmer-Lemeshow goodness-of-fit test to reject poorly fit models. The metrics accepted for the index model included the Shannon-Wiener diversity index, the station mean of species tolerance values (ES(50)0.05), and percent abundance of capitellid polychaetes. The benthic index was constructed as a formula that modeled the probability of a low BEQ and subtracted that from 1 to give an index where low values indicate low BEQ. The benthic index was then scaled to the range 0 to 10 by multiplying by 10. The logistic regression model that best fit the data, the Acadian Province Benthic Index ($APBI_1$), correctly classified 80% of the observations in the calibration dataset.

In GOMA (2011), metrics accepted from earlier statistical analyses were combined into candidate indices for each habitat class. The candidate indices selected for final assessment were those in each habitat class that best represented multiple aspects of benthic community structure and function, were not redundant in nature (Pearson correlation coefficients between metrics were calculated), and showed discrimination of reference and degraded sites (using DE and z-scores). The final index metric combinations and discrimination efficiencies for each site class were: Low Salinity – five metrics (% individuals as Bivalvia, % individuals as Spionidae, % individuals as predators, % individuals as tolerant, and Beck's biotic index); DE = 85.7%; High Salinity not Florida – five metrics (% individuals as Spionidae, % taxa as Polychaeta, % individuals as intolerant, % individuals as tolerant, and % taxa as intolerant), DE = 80.7%; and High Salinity south Florida – five metrics (# taxa as Bivalvia, % individuals as Polychaeta, % individuals as interface feeders, % individuals as tolerant, and Beck's biotic index), DE = 100%.

For MDEQ (2013), several index combinations (with up to 20 or more candidate indices) were calculated for each habitat class using "an iterative process of adding and removing metrics, calculating the index as an average of the metric scores, and evaluating index responsiveness." Indices were evaluated (using DE, z-scores, and representativeness and uniqueness) and those that performed best were further validated before approving one as final. Redundancy was evaluated using Pearson Product-Moment correlation analysis. Final indices are unweighted averages of metric scores per individual scoring formulas. The 25th percentile of reference calibration data for all habitat classes is suggested as the potential impairment threshold.

Llanso and Southerland (2006) standardized scores for each metric calculated from previous years' data and averaged them to develop a logistic regression model that could predict "the likelihood of contaminant effects as a function of the value of combined metrics at a site."

5.2.6 Index Validation

Typically, studies designate independent datasets with which to validate their final indices. Validation dataset discrimination efficiencies can then be compared to the calibration discrimination efficiencies; the expectation is often that there should be no significant difference. Indices can be validated in other ways as well, including the testing of other regional indices with the calibration dataset to compare performance.

The Engle and Summers (1999) index was computed for all sites sampled by EMAP-E in the Louisianan Biogeographic Province from 1991-1992. The index was validated using an independent sets of data from two subsequent years, 1993 and 1994, as well as data from special study sites representing between-year and within-year replicates. Validation involved three steps: 1) assessment of correct classification by index of an independent set of degraded and reference sites; 2) comparison of cumulative distribution function of the index among 4 years; and 3) assessment of correct classification of replicate sites by the index. Correct classification occurred when the benthic index was ≤3 at degraded sites or ≥5 at reference sites. Classification efficiency was adequate, but precision was likely sacrificed in favor or a more generalized index applicable across wide geographic area A cumulative distribution function (CDF) was computed from benthic index values weighted by surface area represented by base stations for each year of sampling. According to the Kolmogorov-Smirnov test, CDFs were not statistically different among the 4 years of samples. Within sites used to validate the consistency of classification by the BI, no sites were misclassified as degraded or reference. Correlation between the BI from temporal replicates was .83. Validation was determined successful.

Paul et al. (2001) developed an index and assessed several alternate models using 60 calibration sites and 52 validation sites. To compare the new index with the earlier 1990-1991 EMAP dataset index, the earlier index was applied to the 4-year (1990-1993) EMAP calibration and validation data. The authors determined that the newer index proved superior, scoring 86% correct classification of both reference and degraded sites in the combined data.

For the Llanso et al. (2003) index, classification efficiencies were relatively high for both the calibration (90%) and the validation (83%) datasets in the mesohaline habitat. In the oligohaline habitat, overall classification efficiencies were 88% for the calibration dataset and 80% for the validation dataset, although classification of individual condition groups (i.e. of reference and degraded sites) were not as high (75-79% was considered acceptable). In the tidal freshwater mud habitat, overall classification efficiency was 67% in the calibration dataset (and only 53% for degraded sites), but was higher (86% overall) in the validation dataset. Two other indices, the New York-New Jersey Harbor index and the EMAP Virginian Province Index were evaluated for the Hudson River sites. The New York-New Jersey Harbor index was evaluated for Hudson River mesohaline sites; most degraded sites were correctly classified by the index, but most of the reference sites were not. Many polyhaline species typically encountered in the NY-NJ Harbor complex were not present in the less saline samples. Applying the EMAP Virginian Province Index to Hudson River EMAP sites resulted in “negative benthic index values indicating degraded conditions and positive values indicating not degraded conditions disagreed with the sediment chemistry 54% of the time."

The final Llanso et al. (2002) index correctly classified 83% of the calibration dataset sites and 82% of the validation dataset sites. Classification efficiencies of sites were higher in the mesohaline and polyhaline habitats (81-92%) than in oligohaline (71%) and tidal freshwater (61%).

Van Dolah et al. (1999) used a combined 1993 and 1995 database as its validation dataset to confirm that the selected index "produced the highest correct classification efficiency (83%) of those considered for the overall study area." The Chesapeake Bay B-IBI developed by Weisberg et al. (1997) classified 93% of the validation sites correctly. Classification was poorer in low salinities; cumulative classification efficiency was 84% for habitats with salinity below 18 $^{o}/_{oo}$ and 97% for sites with salinity above 18 $^{o}/_{oo}$.

For Hale and Heltshe (2008), the selected index (APBI$_1$) correctly identified 26/37 low BEQ (i.e. degraded) stations and 70 out of 81 high BEQ (i.e. reference) stations for an overall classification accuracy of 81% using the calibration dataset. Using the validation dataset, the index correctly classified 23 of 28 low BEQ stations and 18 of 21 high BEQ stations, for an overall classification efficiency of 84%.

For each habitat class of GOMA (2011), the final indices were tested for their ability to discriminate between reference and stressed sites in an independent dataset; the 2005-2006 NCA dataset set aside as the validation dataset. DE scores were as follows: Low salinity (85.7%), High salinity-non-Florida (80.7%), High salinity-South Florida (100%), and North Florida (100%). The DE of habitat class indices were compared to that for the existing Engle and Summers (1999) GOM index.

Validation assessment of the MDEQ (2013) indices confirmed their ability to correctly classify reference and degraded sites based on biological conditions. DE in low salinity-high silt/clay habitat was 90%; in high salinity-high silt/clay habitat was 72.2%; in high salinity-low silt/clay habitat was 80%; in low salinity-low silt/clay habitat was 100%; and in low salinity-all substrates habitat was 93.3%.

5.3 Tolerance Indices

Though developed for freshwater streams, the Hilsenhoff Biotic Index (HBI; Hilsenhoff 1988) is a standard example of biotic tolerance indices. HBI estimates the overall tolerance of a biotic community to organic pollutants by assigning each taxon a tolerance number reflecting its sensitivity to organic pollutants. To calculate the index, each taxon tolerance value is weighted by the abundance of that taxon, the products are summed, and then the total is divided by the total abundance (i.e. the total number of specimens in the sample). Similarly, calculated tolerance indices have been adapted for estuarine benthic habitats (e.g., Smith et al. 2001), and the AZTI Marine Biotic Index (AMBI) (Borja et al. 2000) has been used in Europe for years to assess the degree of anthropogenic disturbance in coastal waters. Similar in concept to the HBI, AMBI is an abundance-weighted, tolerance value index that assesses habitat condition based on the relative abundances of taxa in different tolerance value groups (Gillett et al. 2015). Tolerance values of taxa (Ecological Groups (EG)) used to calculate the AMBI are assigned through consultation with subject matter experts and reference to the literature. EGs are labeled I (least tolerant taxa), II, III, IV, or V (most tolerant taxa). Applications of the European AMBI in the U.S. have typically performed poorly, disagreeing with locally calibrated indices, and thus limiting its widespread adoption (Gillet et al. 2014). Gillett et al. (2014) hypothesized that poor performance of AMBI in the U.S. may be "partially due to extrapolating EG values developed primarily from European waters" and that "using EG values in this manner may fail to account for potential differences in tolerance among habitats or across large geographies". Gillett et al. (2014) brought U.S. ecologists together to assign EG values based on their experience with benthic invertebrates in three coastal regions and for the entire U.S. Index performance was compared using region specific EG scores, national EG scores, a hybrid of national EG scores supplemented with standard international EG scores for taxa that U.S. experts didn't have sufficient expertise to assign, and standard international EG scores. Performance was evaluated by condition at pre-defined good and bad sites, concordance with existing local benthic indices, and independence from natural environmental gradients. AMBI performed best when using U.S. EG assignments augmented with standard international EG values, showcasing an over 80% correct classification rate in differentiating *a priori* good and bad sites and AMBI scores concordant and correlated with those of existing local indices. The AMBI, though, didn't perform well overall as it "tended to compress ratings away from the extremes and toward moderate condition and there was a salinity bias where high quality sites received increasingly poor condition scores with decreasing salinity" (Gillett et al. 2015).

Multivariate AMBI (M-AMBI) is an extension of AMBI, developed to address performance issues when there are few individuals and species present in a sample and to better tailor tolerance values to local settings. M-AMBI is a combined index of AMBI scores with habitat measures of richness and diversity (Shannon-Wiener). Metric values are standardized and combined using factor analysis and the factor scores are placed along orthogonal condition gradients defined for each habitat using reference and degraded anchor points. The position in Euclidean space represents the index score of the sample. Pelletier et al. (2017) evaluated M-AMBI in U.S. waters and compared its performance to that of U.S. AMBI in three ways: concordance with local indices used as management tools in three U.S. coastal geographic regions; classification accuracy for *a priori* defined good and bad sites; and insensitivity to natural environmental gradients. U.S. M-AMBI correlated highly with the local indices, accounted for issues in low diversity sites, and removed the compression response and salinity bias seen with U.S. AMBI.

Another tolerance index, the Benthic Response Index, was developed by Smith et al. (2001) for the Southern California coastal shelf environment. The index was developed using a two-step process in which ordination analysis was used to quantify a pollution gradient within a sample calibration dataset. The study established thresholds for reference conditions and four levels of possible biological response (values at which key community attributes were lost). The index is calculated as the abundance-weighted average pollution tolerance of species in a sample, determined based upon a species distribution of abundance along the pollution gradient.

5.4 RIVPACS Model

The River Invertebrate Prediction and Classification System (RIVPACS) is a method for assessing benthic macroinvertebrate communities that relies on predictions of taxa expected to occur in the absence of human disturbance. Originally developed for rivers, this concept can be applied to estuarine embayments. Comparison between the number of observed taxa and the number of expected taxa provides an assessment of biological condition at a site, assuming that the absence of the taxa expected in reference conditions is an indication of anthropogenic impacts. Llanso et al. (2003) compared the use of three indices (RIVPACS, IBI, and a discriminant analysis approach) in the Hudson River estuary to find a best fit for evaluating benthic condition. The RIVPACS analysis results were found inconclusive; it was weak in tidal-fresh habitats, worked well in mesohaline sites, but differences in sediment composition between degraded and reference mesohaline sites may have confounded the interpretation of results and it was therefore not recommended for use in the Hudson River estuary. Ranasinghe et al. (2009) calibrated and compared five indices, including RIVPACS, along the California coast. Experts ranking sample condition and evaluating benthic assemblage disturbance outperformed individual indices, but several index combinations outperformed the average expert, a few of which included RIVPACS.

6 Sample Sites in Long Island Sound and Its Embayments

6.1 Existing

Both USEPA coastal monitoring programs have sampled in the Long Island Sound: the NCA from 1990-2006 (as part of the Virginian Biogeographic Region and the Northeast Region surveys); and the NCCA in 2010 and 2015 (Figure 6-1). Note that the National Aquatic Resource Survey (NARS), of which NCCA is a component monitors all coastal, lake, river and stream, and wetland water resources, and therefor conducts nationwide surveys of each category once every five years on a rotating basis.

6.2 Planned

In addition to the planned NCCA 2020 sampling at Long Island Sound sites, as discussed in Section 1, up to 60 sites in Long Island Sound embayments (both in NY and CT) will be sampled using the random statistical design and the standard collection and analytical techniques of NCCA to evaluate benthic (e.g., nutrient, sediments, macrobenthic community) conditions for the Long Island Sound Study (LISS).

The NCAA 2020 survey follows a stratified probability survey design with two sampling site design components; 1) repeat sampling of sites sampled in 2010 and again in 2015, and 2) new site selection. LISS survey design is structured similarly: the first component consists of eight sites sampled previously through NCCA; and the second component selects new sites using a spatially balanced survey design with four strata: (CT_Bays), Connecticut Non-Bays

(CT_NonBays), New York Bays (NY_Bays), and New York Non-Bays (NY_NonBays). "Bays" sites refer to sites selected within the LIS embayments.

Combined survey designs for the Long Island Sound Study (LISS) follow these panels:

1. Base20_10RVT2: Sites from NCCA 2010 and 2015 that will be re-sampled twice in 2020 ('RVT' refers to 'revisit sites').
2. Base20_10RVT: Sites from NCCA 2010 that will be re-sampled once in 2020.
3. Base20_20RVT2: New sites that will be sampled twice in 2020.
4. Base20_20: New sites that will be sampled once in 2020.
5. Over20_20: New sites that are over-draw sample sites that will only be used if any Base20_20 site cannot be sampled in 2020.

In total, 94 sites were chosen for potential survey (Figure 6-2): 12 sites in NY_NonBays; 35 in NY_Bays; 12 in CT_NonBays; and 35 sites in CT_Bays. The LISS site selection summary, which describes the number of sites chosen by stratum and panel, is summarized in Table 6-1.

Table 6-1. LISS site selection summary describing the number of sites by stratum and panel.

Site Selection Summary						
	Base20_10RVT[1]	Base20_10RVT2[2]	Base20_20[3]	Base20_20RVT2[4]	Over20_20[5]	**Sum**
CT_Bays	0	0	29	1	5	**35**
CT_NonBays	4	1	5	0	2	**12**
NY_Bays	0	0	30	0	5	**35**
NY_NonBays	2	1	7	0	2	**12**
Sum	**6**	**2**	**71**	**1**	**14**	**94**

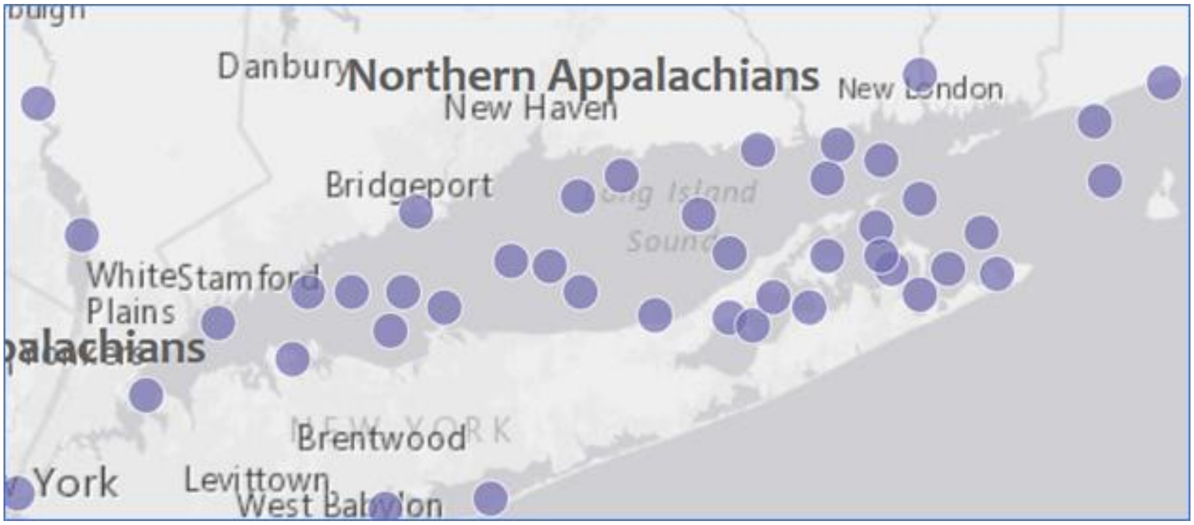

Figure 6-1. Map of NCCA 2010 Long Island Sound sample sites (from https://www.epa.gov/national-aquatic-resource-surveys/map-national-aquatic-resource-surveys-sampling-locations)

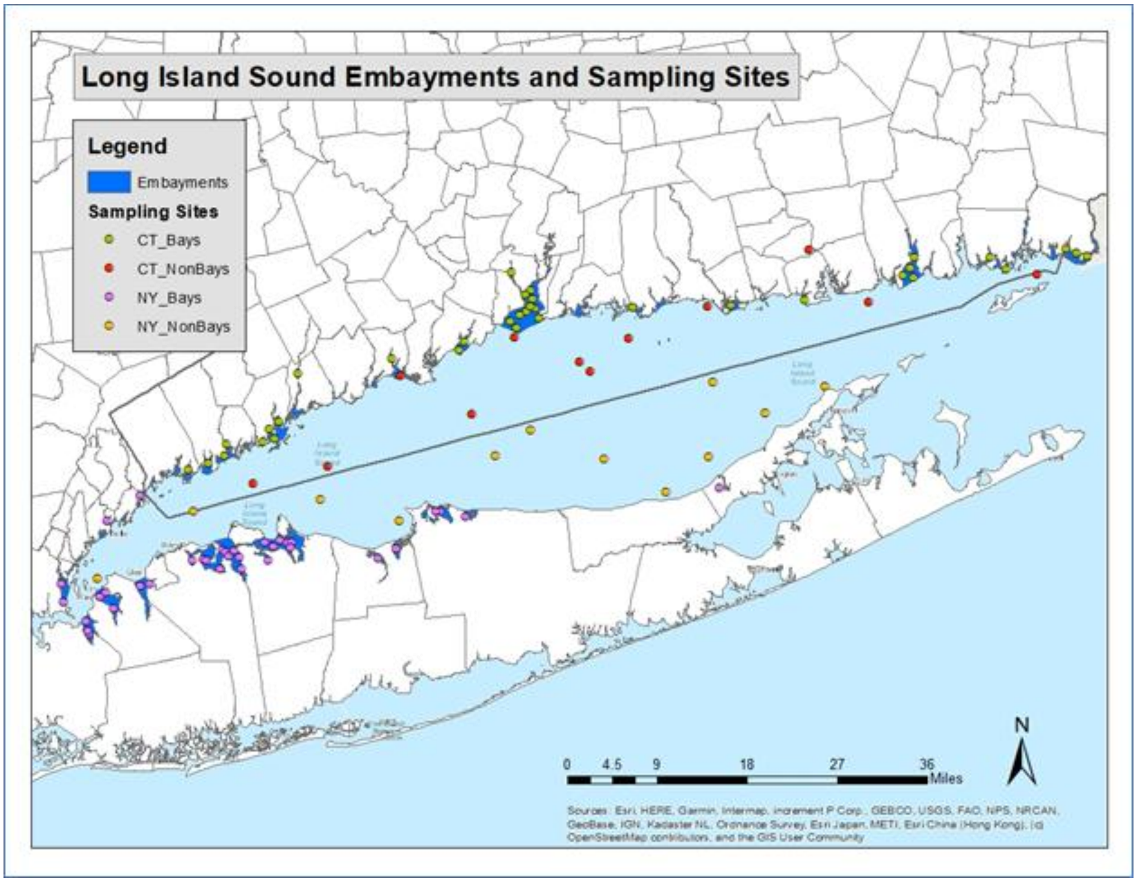
Long Island Sound Embayments and Sampling Sites
Legend
Embayments
Sampling Sites
CT_Bays
CT_NonBays
NY_Bays
NY_NonBays
0
4.5
9
18
27
36
Miles
N
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

Figure 6-2. Map of Long Island Sound embayments and 2020 planned LISS sampling sites.

6.3 Data Summary

As LISS 2020 sampling will follow standard NCCA sampling procedures (USEPA 2014 and USEPA 2016, see Section 4.4 above), it is expected that water quality parameters (e.g., dissolved oxygen, salinity, temperature, pH, water clarity), dissolved nutrients (e.g., nitrogen, phosphorus, chlorophyll-a), sediment chemistry and toxicity (contaminant concentrations and bulk sediment toxicity), and benthic macroinvertebrate community assemblage data will be collected at each site. Physical and chemical parameters will be measured and recorded at each site using a multi-parameter water quality sonde, a Secchi disk (for water clarity), and a PAR meter (for light attenuation). Water samples will be collected at each site using a water sampling device or pumping system for nutrient analysis. For benthic community analysis, one sediment grab sample (or two samples if grab size is less than 300 cm^2) will be collected to a depth of at least 7 cm using 400 cm^2 grab (Young modified Van Veen or other similar device) at each site. Samples will be sieved through a 0.5 mm mesh screen and preserved in a stained (e.g. rose Bengal) 100% percent buffered (e.g., buffered with borax) formalin solution. For sediment chemistry and toxicity testing, samples will be collected using the same method used for benthic community analysis. The top 2 cm of sediment samples will be removed and composted into a bowl and subsequent grabs will be taken until 3L of sediment is collected. The samples will be sent to laboratories and analyzed using several methods (summarized in USEPA 2016) to characterize benthic conditions.

Per NCCA SOPs, sediment toxicity will be tested using 10-day amphipod survival testing; silt-clay, total organic carbon (TOC), heavy metals, PAHs, pesticides, PCBs, and ammonia composition of sediments will be analyzed using documented testing methods (USEPA 2016). These data can then be used to define disturbance gradients or thresholds for reference and degraded site definition. Habitat characteristic data may also be used to support classification of habitat types, and correlation analysis to potentially better detect benthic assemblage responses to disturbance. The laboratory will have sorted and identified organisms in the benthic assemblage samples to the lowest practical taxonomic level. Benthic community analysis data will allow for the summarization of organisms likely to be seen in the LISS embayments. In combination, these data will support benthic macroinvertebrate multi-metric index development.

7 Recommendations

It should be noted at the outset that complete development of a benthic multi-metric condition index is predicated on having sufficient and appropriate data to complete the suite of analyses needed to calibrate (i.e. initial development of) an index, and then validate the index on an independent set of data. It is anticipated that about 60 LISS embayment samples will be available for this effort. The overarching approach will be to proceed with data compilation and review, and through analysis steps that will help define whether there is sufficient data to complete the index development process. These analyses should include review of physical and chemical habitat data to support setting criteria and thresholds for defining reference and degraded sites; application of these to sampled sites to determine how many sites can be categorized as reference (and degraded); classification to determine how stations should be grouped for index development; and review of biological data and preliminary calculation of metrics. Together, these efforts will support a review of the adequacy of the 2020 LISS embayment data to proceed through index calibration and validation. If initial analysis finds it appropriate, a different condition index approach (e.g., tolerance index) may be pursued. For context, Table 7-1 summarizes numbers of samples used by other researchers in their benthic index development (calibration and validation).

Table 7-1. Numbers of samples used by selected studies in coastal and estuarine benthic index development.

Citation	Index Name/ Location	# Samples	
		Calibration	Validation
Multi-Metric Indices			
MDEQ 2013	Gulf of Mexico IBI	415 (cal. & val.)	--
GOMA 2011	GOMA IBI	1,281 (yrs 2000-04)	~298 (yrs 2005-06)
Hale and Heltshe 2008	Benthic index, nearshore Gulf of Maine	182	~66
Malloy et al. 2007	Benthic index, Tampa Bay	578 (cal. & val.)	--
Llanso and Southerland 2006	Hudson R. estuary benthic index	90 (cal. & val.)	--
Llanso et al. 2003	Hudson R. estuary benthic index	215	63
Llanso et al. 2002	Mid-Atlantic estuary benthic index	2/3 of 2,083 total	1/3 of 2,083 total
Paul et al. 2001	Virginian Province, benthic index	60	52
Engle and Summers 1999	Gulf of Mexico benthic index	1,023 total (subset from 1991 & 1992 for calibration)	Subset of 195 sites from 1993-94
Van Dolah et al. 1999	Southeastern US	75	96
Weisberg et al. 1997	Chesapeake Bay B-IBI	2,319 (cal. & val.)	-- (not specified)
Engle et al. 1994	Gulf of Mexico benthic index	546	None available for validation
Tolerance Indices			
Pelletier et al. 2018	M-AMBI for US coastal	4,061 (cal. & val.)	--(not specified)
Gillett et al. 2015	AMBI for US coastal	Not specified	--
Smith et al. 2001	Southern California shelf	717	-- (not specified)

7.1 Data Compilation – Applicability, Continuity

The first step in the integrated process of developing a multi-metric index of biotic integrity for Long Island Sound embayments will be to assemble pertinent data sets. Pooling any available benthic invertebrate and ancillary environmental data has the benefit of potentially increasing the number of data points and the spatial scope of data upon which an index can be based. However, there must be some defined continuity in the data in order to support credible analyses. This can best be assured by applying some review/selection criteria, recommended to include:

1) Similar sample collection methods (e.g., a particular concern would be comparing samples collected by different gear types that sample very different areas, as the more area is sampled the more likely it is to collect a higher number of taxa which would impact richness metrics);

2) Similar sample processing methods, including (but not necessarily limited to) adequate levels of taxonomic identification similar to the NCCA methods that will be employed in the LIS, an adequate number of individuals in each sample, and similar inclusion/exclusion rules for all taxa groups; and

3) Concurrent measurements of physical-chemical habitat parameters, minimally those needed to correct for expected natural variations in abundance and distribution of taxa across estuarine gradients (e.g., salinity, sediment type), and other data that can help

explain organism responses (e.g., DO, as well as sediment contaminants and toxicology, though these latter data types are often not available).

The core data set will be the Summer 2020 embayment samples that will be collected in association with the 'Long Island Sound Study' (LISS), as an enhancement of and in conjunction with the planned 2020 NCCA Long Island Sound sampling (described in Section 6 above). This data set will be comprised of approximately 60 benthic samples plus ancillary data (see Section 6 for a list).

Additional data that will be reviewed as a potential supplement to this core set includes benthic data from the open-water area of Long Island Sound both the previously collected LIS NCCA samples from 2010 and 2015 (also mentioned in Section 6) and the 2020 open-water LIS NCCA samples, due its proximity to the LIS embayments. Also included in the review will be earlier LIS open-water and embayment EMAP/NCA data and, potentially, northeastern and Mid-Atlantic U.S. regional data from states (e.g., Maryland, New Jersey) and other estuarine sampling ventures (MAIA). If initial analysis proves it necessary to collect more data, sources of regional data comparable to LIS embayments will be screened for through sorting by factors including salinity, sediment type, and reference condition status. Other screening considerations include sampling method, sample processing methods (e.g., subsampling and level of taxonomic identification). Because NCCA sample collection and processing protocols will be used in the LISS, continuity among these data sets will be assured.

Any historic benthic data sets within the embayments of LIS could be reviewed, potentially including sample results from an Alley Creek study (Hazen and Sawyer and Tetra Tech 2017). Alley Creek is a tidally influenced, brackish water inlet from Great Neck Bay in Long Island Sound, Queens, NY. Benthic invertebrate, water quality, and vegetation sampling was conducted to develop a baseline of the surrounding habitat prior to a planned chlorination program for the local combined sewer overflow (CSO) tank. Over the course of three sampling events corresponding with spring, summer, and fall conditions in 2016, samples were collected upstream and downstream of the CSO. The monitoring plan included the use of petite ponar grabs to collect creek bottom sediment at 4 different sampling sites. In addition, other sets of ancillary environmental data could be obtained, reviewed, and incorporated for context (e.g., salinity, substrate, DO, etc.).

7.2 Methods – precedent or strength of analysis

All data reviewed and deemed comparable should be compiled in a database. Tetra Tech has developed and maintains options for data management and metric calculation, sensitivity analysis, and index reporting. Based on consultation with CT DEEP, we recommend using a package in R called BioMonTools. This will provide the efficiency of having existing code for calculating numerous common estuarine benthic metrics, while also easily accommodating the introduction of additional metrics, all of which can then be calculated repeatedly and reliably. Once an integrated database is established, a progression of analyses can be undertaken that will ultimately lead to testing and selection of a final index. In brief summary, the steps include:

1) organization of data and calculation of metrics;
2) characterization of reference and disturbed sites, and/or of a disturbance gradient;
3) classification of sites to account for natural variability, or alternatively evaluation of metrics for correlation with major environmental drivers (e.g., salinity, sediment type) with associated correction of metrics as needed;
4) assessment of metric sensitivity to stressors; and
5) assembly, testing, and selection of a multi-metric index.

Calculation of metrics. There are a large number of measures of benthic community structural and functional characteristics that can be calculated and evaluated as candidate metrics for potential inclusion in a final index. They fall into categories such as taxonomic composition, diversity, pollution tolerance/sensitivity, functional feeding group designations, and other trait categorizations. The number of metrics recommended for calculation at the beginning of the index development process is often a matter of best professional judgement (BPJ), for which guiding principles can include representing a spectrum of ecological structural and functional categories, sensitivity to a diverse array of stressor types, and starting with a sufficient number of metrics within such categories to support good statistical choices. For example, Nestlerode et al. (2019) compared two indices that were calibrated to Gulf of Mexico data sets, the GOM B-IBI (GOMA 2011) and the EMAP-E benthic index (Engle and Summers 1999) of which the results disagreed. Nestlerode et al. (2019) concluded that the index correlated with diversity measures (EMAP-E) met their expectations best, though the GOM B-IBI was not recommended for application in the higher salinity sites where the index was tested and the lower salinity GOM B-IBI index (which was applicable in 3 test sites) showed expected patterns with diversity. GOM B-IBI was calibrated using the best performing metrics relative to an objectively defined disturbance gradient and other selection criteria, as was the EMAP-E benthic index, though using different data sets. If more confidence in certain metrics was expressed during metric selection, those metrics could have been included. For example, if diversity was an important theoretical response to the index developers, diversity metrics could have been included regardless of performance in the calibration data set. Shannon diversity was included in the EMAP-E index and not in the GOM B-IBI. Therefore, the EMAP-E index correlates well with Shannon diversity and other diversity measures. The selection process for metrics in an index should be scrutinized by the developers and end-users to ensure that they are confident in the ecological meaning of the metrics and that the metrics address relevant stressor-response patterns.

Data limitations also must be taken into account. For example, metrics can be calculated based on richness (number and percent of taxa), abundance (number and percent individuals) or biomass, but while richness and abundance usually are measured, biomass often is not. The types of information available also can limit ability to calculate some metrics. For example, some trait-based metrics such as habitat preferences, feeding types, or temperature sensitivities, cannot be calculated if there are insufficient data or literature information to classify enough of the taxa present in the data set to support robust spatial comparisons. This type of limitation prevented, for example, GOMA (2011) from considering habit trait metrics, as well as those based on biomass or depth of occurrence in the sediment in its index development. Because these organisms are adapted to a highly variable environment, there is a high natural variation in estuarine benthic metrics, which makes the detection of differences among different condition site classes more challenging. For LIS, it is recommended that a relatively wide range of structural and functional metrics be calculated based on abundance and richness; biomass metrics will not be supported. In addition, trait-based metrics should be calculated as possible, including (but not necessarily limited to) pollution sensitivity/tolerance, temperature preference, feeding and habitat types.

Characterization of reference/disturbed sites or a disturbance gradient. To support the assessment of metric responses to stressors in the environment to which an index is going to be applied, there must be an independent way to identify locations that are good, or minimally impacted by those stressors, and locations that are impacted. A typical approach is to select several metrics that meaningfully reflect the stressed (i.e. degraded, impaired) conditions in the study area, and then establish criteria for each of those stress metrics that are sensitive to those stressors (i.e. reliably distinguish reference and degraded conditions). Commonly applied indicators of stressed conditions are bottom dissolved oxygen (DO) concentration, sediment contaminant concentrations, and sediment toxicity (e.g., MDEQ 2013, GOMA 2013, Engle et al. 1994, Engle and Summers 1999, Gillett et al. 2015, Llanso et al. 2002, 2003, Malloy et al. 2007, Paul et al. 2001). Hale and Heltshe (2008) used a combination of these three stressors plus the percent of organic carbon in the sediments. Table 7-2 summarizes examples of the stress metrics and criteria established by several of the estuarine benthic index studies reviewed.

Table 7-2. Examples of stress metrics and criteria established by several estuarine benthic index studies.

Study	Station Type	DO	Sed. Chem.	Sed. Tox.	Other
Engle et al. 1994	reference	>3 mg/l	No contaminants > ER-M[1]	% survival for *Ampelisca abdita* (10 d) or *Mysidopsis bahia* (96 h) (acute sediment bioassays) not different from controls	--
	degraded	≤2 mg/l	At least 1 contaminant > ER-M	--	--
Engle and Summers (1999)	reference	>3 mg/l	No contaminants > ER-M and not >3 contaminants > ER-L[2]	control-corrected percent survival for *Ampelisca abdita* (10 d) and *Mysidopsis bahia* (96 h) (acute sediment bioassays) > 85%.	--
	degraded	≤2 mg/l	At least 1 contaminant > ER-M or >3 contaminants > ERL	control-corrected percent survival for *Ampelisca abdita* (10 d) and *Mysidopsis bahia* (96 h) (acute sediment bioassays) < 80%.	--
Hale and Heltshe 2008	reference (meet all of these criteria)	≥ 5 mg/l	No exceedances of ER-Ms and ≤3 exceedances of ER-L values	*Ampelisca* survival >80% of controls	TOC < 4% dry wt
	degraded (meet at least one of these criteria)	< 5 mg/l	At least one exceedance of ER-M or ≥10 exceedances of ER-Ls	*Ampelisca* survival <80% of controls	TOC > 5% dry wt
Llanso et al. 2002	reference (meet all of these criteria)	> 3 ppm (=mg/l)	No exceedances of ER-Ms and ≤2 exceedances of ER-L values	Sediments not toxic in *Ampelisca* (i.e. not significantly different from and not <80% of control survival) or Microtox bioassays (i.e. where toxicity defined as the EC50 of test sediments (sediment concentration that reduces bacterial light production by 50% relative to water controls) was ≤ 0.2% for sediments with silt-clay content ≥ 20%, the EC50 was ≤ 0.5% for sediments with silt-clay content <20%, or the EC50 of test sediments was significantly different from controls)	--
	degraded (meet at least one of	< 2 ppm (= mg/l)	Any contaminant exceeded ER-M	sediments were toxic in the Ampelisca or Microtox bioassays	--

Study	Station Type	DO	Sed. Chem.	Sed. Tox.	Other
	these criteria)				
Llanso et al 2003	reference	--	No chemical contaminant concentrations exceeded ER-Ms, and ≤ 2 chemical contaminants exceeded ERLs	--	Sediment quotients (mean of ratios of individual contaminant concentrations relative to their corresponding ER-M values) were below the median value for the Virginian Province
	degraded (both criteria met)	--	Any chemical contaminant concentration exceeded the ER-M	--	Sediment quotients above the median Virginian Province value
Malloy et al. 2007	reference (both criteria met)	>4.5 mg/l and	--	--	No TEL[4] or PEL exceedances
	degraded	<2.5 mg/L	--	--	1 or more PEL exceedance
Paul et al. 2001	reference (meet all criteria)	≥ 7 mg/l	No contaminant in sediment exceeded an ER-M, and no more than three contaminants exceeded ER-Ls	Sediment was not toxic (i.e. survival rates exceeded 80% of controls and did not differ significantly from them).	--
	degraded (at least one of these)	≤ 2 mg/l	At least one exceedance of ER-Ms or >10 exceedances of effects range-low (ER-L) values	survival <80% of control and significantly different from controls	--

[1] ER-M = Effects Range Median the concentration at which 50% of collected data demonstrated adverse biological effects (based on Long et al. 1995).

[2] ER-L = Effects Range Low, the concentration at which 10% of collected data demonstrated adverse biological effects (based on Long et al. 1995).

[3] TOC = total organic carbon (in sediments, typically as a percent based on dry weight).

[4] TEL = Threshold Effects Level; PEL = Probable Effects Levels (MacDonald Environmental Services 1994)

For the LIS embayments index development effort, the commonly used stress indicators of DO, sediment contaminant concentrations, and toxicology will be evaluated (noting that natural habitat determinants of estuarine community distributions, such as salinity and sediment type, will be tested and accounted for separately, see Classification of Sites, the next subsection). In addition, sediment organic carbon and land use, as well as nutrient loading modeling outputs and nutrient isotopes results will be considered. The contribution of these indicators to defining reference and degraded conditions will be explored by site and by embayment.

In addition to the development of categorical disturbance categories, it is possible to characterize disturbance in terms of a continuous gradient. GOMA (2011) did this for Gulf of Mexico locations, using this approach to calculate relative disturbance (pollution) tolerance/sensitivity values for benthic invertebrate taxa. However, initially there will probably not be sufficient data to independently calculate such values for LIS embayment taxa. Instead, the tolerance of benthic organisms to disturbance/pollution will be derived from values available in published literature or databases (GOMA 2011, Gillett et al. 2015). Derivation of organism tolerance values using the disturbance gradient within LIS embayments would probably be difficult due to limited sample sizes for sites, organism occurrence, and disturbance conditions.

Classification of sites. The composition of estuarine benthic assemblages is strongly influenced by patterns of salinity, sediment type, latitude, and depth (e.g., Boesch 1973, 1977a, Dauer et al. 1987, Holland et al. 1987, Schaffner et al. 1987). Characterizing and accounting for such natural variability to the extent possible, allows assemblage responses to disturbances to be detected with greater certainty (Smith et al. 2001). There are two approaches that have been applied successfully to the development of estuarine benthic indices. One is to group sites into classes based on similarity of benthic assemblages. Benthic community composition varies in a continuum along estuarine gradients of salinity (and other factors) (e.g., Boesch 1977a, Holland et al. 1987, Pearson and Rosenberg 1978) making classification boundaries somewhat arbitrary. Nevertheless, this approach can be successful at defining regions within which an effective index can be calibrated. Typically, an index is developed and calibrated for each class identified.

Site classification should be done using reference stations that are minimally disturbed or least disturbed (*sensu* Stoddard et al. 2006) and pass the indicator thresholds established for separating reference and disturbed stations. Cluster analysis is a recommended technique for grouping samples into categories based on taxonomic composition. The process would compare all pair-wise sample combinations using the Bray-Curtis (also known as the Sorenson) dissimilarity metric, which can be calculated based on taxa abundance (log-transformed to satisfy assumptions of normality) or presence/absence. A clustering routine (e.g., the one in PC-Ord [McCune and Mefford 2006]) then progressively joins samples into groups, resulting in a dendogram that can be examined for ecologically meaningful groupings. The association of recognized clusters with salinity, sediment type, or other environmental factors (e.g., distance from shore, longitude, embayment size, etc.) can be evaluated, and further assessment of how strongly or uniquely various taxa are associated with clusters can be pursued using ordination and indicator species analysis. Indicator species analysis can, for example, help define expectations for the class and suggest metrics that are responsive to stressors within a class. Like cluster analysis, non-metric multi-dimensional ordination (NMS) starts with a Bray-Curtis matrix, but then plots the points in multiple dimensions, plotting similar points closer to each other and dissimilar points further apart, showing site groups. Environmental variables and taxa abundances can be correlated to the ordination axes, showing how site groups separate or overlap and their relationships to the environmental gradients.

Site classification will need to accommodate sample sizes for analysis. Recognition of multiple site classes, each with a unique reference condition, will depend on having more than a few least-disturbed reference sites to characterize the reference conditions. For the proposed study, which should have approximately 60 sites (samples), there is a likelihood that 20 – 25 reference sites will be identified. This would allow for one site class with sufficient data for calibration and validation of the index. If additional site classes are recognized, giving fewer reference sites per class, calibration might be less robust and validation with a reserved data set might not be possible. Rigorous index development with separated classes may not be possible with just 60 samples and would necessitate the use of additional sample sources.

Another option for accounting for natural variation in benthic assemblages due to salinity and other environmental factors is to evaluate the particular relationship between each metric and environmental factor using correlation or regression analysis. For any relationships that are significant, the metric can be 'corrected' for the amount of variation in the metric associated with variation in the environmental parameter. That is, the significant equation between a benthic metric and a particular environmental parameter generates an expected value for the metric at that point in the environmental gradient, which can then be used in an index. This type of continuous adjustment of individual metrics to individual environmental factors may be employed if sampling across the LIS embayments covers a sufficient range of salinity (or other) values to impact expectations for various benthic metrics. It may be a preferred approach if a particular site group (for example, the lowest salinity (oligohaline to tidal freshwater) is under-represented in number of samples, making it impossible to calibrate a separate index for that group.

Assessment of metric sensitivity. A series of screening and analysis steps contribute to the selection of candidate metrics for the index (Flotemersch, et al. 2006). One metric screening test is for an adequate range of observed metric values. If the number of taxa within a metric category is less than 2 or 3, the range of response would be considered inadequate to consistently identify a difference between reference and impacted sites. Subsequent analysis would focus on the ability of each candidate metric to discriminate between reference and stressed sites. This can be evaluated using discrimination efficiency (DE), z-scores, Wilcoxon U test, box-and-whisker plots comparing metrics between reference and stressed site categories, and quantile regression of scatter plots of the metrics along the stressor gradient. For DE, the distribution statistics of a metric are compared between reference and degraded sites. For example, for metrics that decrease with increasing stress, the DE would be the percent (%) of stressed sites with a metric value falling below the 25th percentile of the reference distribution of metric values. For metrics that increase with increasing stress, the DE would be the % of stressed sites with a metric value above the 75th percentile of the reference distribution. An acceptance criterion for DE values calculated in this way might then be set at, say, >50% for each site class.

The z-score, a common approach for standardizing data, is another method for determining how well a metric separates reference from stressed sites. Calculated as the difference between the mean metric value at reference sites and the mean at stressed sites, divided by the standard deviation of values at reference sites, this "standardized" statistic is comparable across all metrics. When normal distributions of metrics cannot be assumed or achieved by an appropriate data transformation, the Wilcoxon U score, a non-parametric comparison of mean metric values, can be applied to determine whether the mean values for a metric at reference and stressed sites are significantly different. Box-and-whisker plots can be used to display the mean, the 25th to 75th percentiles, and the ranges of the metric values for comparison between reference and stressed sites within a site class, as a good visualization of how differently a metric responds between reference and stressed sites. Quantile regression can be applied to examine and compare the two upper quantiles (75th, 90th) for a consistent and interpretable response to a stressor gradient (if a stressor gradient is estimated).

Because benthic metrics are measured in a wide range of scales, they must be normalized, scaled or 'scored' before being combined into an index to give them (initially) even weights. An option for accomplishing this is to set the range of values for a metric at reference sites (potentially eliminating high-end outliers) to a uniform scale, say 0-100, that would be used across all metrics. Another option is, again using the range of values for a metric at reference sites, is to set percentile-based thresholds to define ranges of values to score into a limited number of scores (e.g., 1, 3, 5) as a relative scoring scale. Each candidate metric should be scored to a similar scale before being combined into candidate indices.

Assembly, testing, and selection of a multi-metric index. For metrics that pass the filters of sufficient information and range of response, a weight of evidence approach can be used to integrate results of the metric sensitivity evaluations in order to identify the subset of candidate metrics that should be retained and evaluated as components within an index. It is recommended that greatest weight given to DE, z-score, and quantile regression slope (if a stressor gradient is estimated).

Once candidate metrics are picked for each site class, a series of metric combinations can be tested as candidate indices. The simplest approach for assembling a multi-metric index is a simple linear (additive) combination of candidate metrics. An implicit assumption of this approach is that all metrics included in the index have the same weight. The 'order' of such an additive equation reflects the number of metrics included; it is recommended that a sufficient number of metrics be included to potentially represent a spectrum of different community structural and functional characteristics and types of sensitivities/tolerances. There is potential variety in the functional components and modes of interacting with the environment that can be captured by different categories of biological metrics. On the other hand, too many metrics included in a linear model can add complexity without substantially improving discrimination efficiency. Candidate index models with >1 to 8 component metrics could be considered. Other factors can be considered to limit the number of candidate metrics included and contribute to the representativeness and robust functioning of an index.

Metrics that are strongly correlated are potentially contributing redundant information on station condition. Pearson correlation coefficients can be calculated to identify redundant metrics being considered for inclusion in a single index, in which case a decision can be made to include only one of a redundant pair in a candidate index. Consideration also can be given to conceptual redundancy in the types of ecological information provided by metrics, which might be expected to be higher between metrics from similar functional categories. Maximizing the range of metric types included in an index has more potential to function robustly in terms of representing community condition. Therefore, candidate indices with two or more metrics from the same category (e.g., functional type, pollution tolerance, taxonomic richness, etc.) can be considered less favorable than those with a greater functional diversity of metrics.

The goal is to generate a multi-metric model that is better able to discriminate between reference and stressed sites than an individual metric. Testing of candidate multi-metric indices would again involve calculation of discrimination efficiencies (DE's) as described for metric evaluation, as the percent of stressed sites with an index value falling below the 25th percentile of the reference distribution of index values. Similarly, z-scores for each candidate index would be calculated as the difference between the mean index value at reference sites and the mean at stressed sites, divided by the standard deviation of values at reference sites. Index values at stressed sites should be lower than at reference sites, so z-scores are expected to be negative. The z-score represents how different the stressed site index results are from reference index results in units of reference index standard deviations, so that results among all candidate indices can be compared. The smaller the z-score, the better the candidate index differentiates stressed and reference sites.

If either site classes or individual correction of metrics for any correlations with salinity, sediment type or other environmental factors are applied in index development, final tested and selected indices should be tested for any significant residual correlation with environmental variables. If significant residual correlations are found, judgement will have to be made whether the magnitude of the relationship is sufficient to warrant any further correction efforts.

For each site class (should different site classes be defined for separate index development), the candidate index selected as optimal based on the above testing should then be validated by testing for ability to discriminate between reference and stressed sites in an independent data set. Typically this is done by subsetting the available data prior to index development, so that the portion of data to be used for validation remains separate from (i.e. is independent of) the portion of data used in index development (i.e. 'calibration'). As a rough estimate, a minimum of 10-20 sites might be needed for validation, in order to cover a range of reference to degraded sites, and in addition to represent variation across any observed estuarine gradient (e.g., due to gradients in salinity and potentially other habitat variables). This may be a challenge in this initial development of an LIS embayments index, given the relatively small size of the data set anticipated to be available.

For application of the index in determining biological condition, index thresholds will be selected, typically separating good-fair conditions from poor conditions. The thresholds will be decided through analysis and considerations of the regulatory agency. The considerations should include the rates of Type I and Type II error in the index. Targets for minimal Type II error, should be no more than 30%, corresponding to a DE of 70%. If the DE is based on the 25th percentile of reference, then the Type I error is 25%. Type I and Type II index errors can be balanced against each other when selecting an index threshold. This would recognize that the index is not inherently biased towards or against protective measures. The errors can be unbalanced if justified. For example, if reference sites were selected based on a sliding scale of disturbance then there might be higher Type I error rates to recognize that disturbance was not entirely absent from reference sites and the reference biological condition was formed based on sub-optimal sites. The regulatory agency's protective intent might also be expressed when selecting condition thresholds.

Should an additive index model not achieve an acceptable level of discrimination between reference and stressed sites, alternative approaches could be used to estimate non-linear index models. For example, linear discriminant analysis (LDA) or stepwise discriminant analysis (SDA) can be used to generate coefficients for each metric and a final formulation that can then be tested for performance as described above. Pelletier and others (2012) found that non-linear metric scoring on a continuous scale (Mebane et al. 2003) resulted in better discrimination than a discrete scoring approach (Van Dolah et al. 1999) in a comparison study using EMAP data from the Virginian Province. In addition, a logistic regression model performed as well as the non-linear continuous approach in a comparison with the same data set (Pelletier et al. 2012). These alternative approaches will be considered if needed based on results of the simpler approach.

7.3 New Approaches

In addition to the widely used metrics, classification factors, or analysis approaches described above, potential new approaches (for example recently published novel approaches), or data or study results unique to the LIS can be considered. For the pending LIS sampling season, sediment stable isotopes of nitrogen and carbon will be analyzed. This analysis will not only identify nutrient concentrations, but will imply nutrient sources, especially differences between human or natural sources (Smucker et al. 2018). Sediment isotope analysis might be useful for associating nutrient disturbance with local embayment inputs or natural processes.

In addition, there are study results of N loading per embayment for the LIS (Vaudrey et al. 2016). It may be possible to relate observed patterns of N-loading and identified N and C sources to LIS land use patterns and use this information to help define and distinguish reference and disturbed sites or define a disturbance gradient.

A few studies have used a less commonly applied metric, percentage of organic carbon in sediments, to distinguish reference and disturbed sites, and this metric may have application in LIS. For example, Hale and Heltshe (2016) used %C_{org} in sediments >5% TOC dry weight as one criterion identifying degraded sites, and Pelletier et al. (2017) defined reference sites including a TOC criterion <2%. This criterion might also be evaluated for the LIS index.

7.4 Potential Additional Methods

The NCCA sampling intensification in LIS embayments will include macroinvertebrate community analysis and sediment stable isotope analysis. From the macroinvertebrate analysis, community metrics will be calculated, and a biological assessment index will be developed. From the stable isotope analysis, sediment N and C will be attributed to natural or human sources, which will inform the disturbance gradient for the biological index calibration. Should the results of sediment stable isotope analysis prove useful as a disturbance indicator, it would be recommended that it be considered for continuation.

There is an additional analysis that might be considered to tie the two analyses together and allow interpretation of resource-consumer interactions in benthic communities along an environmental gradient. Stable isotope analysis for biota is a relatively new technique. It has been done in streams (Smucker et al. 2018) but may not be practical in the LIS due to questions of sampling feasibility (e.g., holding times, collecting sufficient biomass for analysis, etc.). Measuring stable isotopes (δ15N and δ13C) in aquatic organisms among different trophic levels (e.g., primary and secondary consumers) can provide information regarding how food webs change, because of trophic fractionation in biota. Stable isotopes of biota integrate the exposure to and effects of urbanization and nitrogen from human sources reaching aquatic ecosystems over time. The responsiveness of biota δ15N to urban development in watersheds and anthropogenic increases in nitrogen has been established in riverine systems and support their use as indicators in monitoring programs. The responses of sediment and macroinvertebrate stable isotope analyses across a range of estuarine conditions can inform management and protection goals and be used as one way to quantify the effectiveness of future nitrogen and land-based management efforts in coastal embayments.

Should stable isotope analysis for biota be considered, efforts in addition to those already planned for the NCCA sampling would include three significant procedures:

- Collection of a second macroinvertebrate sample targeted to contain at least 1.0 g dry weight of benthic organisms
- Preservation of that sample by chilling and freezing either before or after identification
- Identification into categories of primary and secondary consumers
- Analysis of stable isotopes (δ15N and δ13C) in the laboratory

The second macroinvertebrate sample would be in addition to the one already planned for community analysis. It is uncertain whether the planned methods would yield 1.0 g dry weight of organisms. The isotope analysis requires approximately 0.5 g of dry matter for each trophic level. A minimum of 1.0 g would be enough if the primary and secondary consumers were equally distributed in the sample. It might be flawed to assume that the trophic levels would be equally distributed, so an ample sample size should be targeted.

While the benthic community sample can be preserved with buffered formalin with rose bengal stain, there is some evidence that such preservative would interfere with analysis of C and N isotopes. The preferred preservation method would be to chill the sample on ice from the boat to the lab, with a holding time of 24-48 hours. Timely delivery to the lab might be a hurdle for logistics in applying this preservation technique. Alternatives that would be less cumbersome are being investigated.

At the lab, organisms from the chilled sample would be separated into primary and secondary consumers, and then frozen. The frozen samples can be held weeks or months before stable isotope analysis.

8 References

Boesch, D. E 1973. Classification and community structure of macrobenthos in the Hampton Roads area, Virginia. Marine Biology 21:226-244.

Boesch, D. E 1977a. A New Look at the Zonation of Benthos along the Estuarine Gradient. Pages 245-266 In B. C. Coull, editor. Ecology of Marine Benthos. University of South Carolina Press, Columbia, South Carolina, USA.

Boesch, D.F. 1977b. Application of Numerical Classification in Ecological Investigations of Water Pollution. EPA-600/3-77-033. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Corvallis, Oregon. Special Scientific Report No. 77, Virginia Institute of Marine Sciences (VIMS).

Borja, A. and D.M. Dauer. 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecological Indicators 8:331-337.

Borja, A., Franco, J., Pérez, V., 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, A., D. Dauer, R. Diaz, R.J. Llanso, I. Muxika, J.G. Rodriguez, and L. Schaffner. 2008. Assessing estuarine benthic quality conditions in Chesapeake Bay: A comparison of three indices. Ecological Indicators 8:395-403.

Dauer, D.M., R.M. Ewing, and A.J. Rodi, Jr. 1987. Macrobenthic distribution within the sediment along an estuarine salinity gradient. Internationale Revue der Gesamten Hydriobiologie 72:529-538.

Dauer, D.M., J.A. Ranasinghe, and J.B. Weisberg. 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in the Chesapeake Bay. Estuaries 23:80-96.

Dauer, D.M. and R.J. Llanso. 2003. Spatial scales and probability-based sampling in determining levels of benthic community degradation in the Chesapeake Bay. Environmental Monitoring and Assessment 81:175-186.

Engle, V.D., K. Summers, and G.R. Gaston. 1994. A benthic index of environmental condition of Gulf of Mexico estuaries. Estuaries 17(2):372-384.

Engle, V.D. and K. Summers. 1999. Refinement, validation, and application of a benthic condition index for northern Gulf of Mexico estuaries. Estuaries 22(3, Part A):624-635.

Flotemersch, J. E., J. B. Stribling, and M. J. Paul. 2006. Concepts and Approaches for the Bioassessment of Non-wadeable Streams and Rivers. EPA 600-R-06-127. US Environmental Protection Agency, Cincinnati, Ohio.

Gillett, D.J., S.B. Weisberg, T. Grayson, A. Hamilton, V. Hansen, E.W. Leppo, M.C. Pelletier, A. Borja, D. Cadien, D. Dauer, R. Diaz, M. Dutch, J.L. Hyland, M. Kellogg, P.F. Larsen, J.S. Levinton, R. Llanso, L.L. Lovell, P.A. Montagna, D. Pasko, C.A. Phillips, C Rakocinski, J.A. Ranasinghe, D.M. Sanger, H. Teixeira, R.F. Van Dolah, R.G. Velarde, and K.I. Welch. 2015. Effect of ecological group classification schemes on performance of the AMBI benthic index in US coastal waters. Ecological Indicators 50:99-107.

Gulf of Mexico Alliance (GOMA). 2011. Benthic Index of Biological Integrity for Estuarine and Near-Coastal Waters of the Gulf of Mexico. Prepared for: The Gulf of Mexico Alliance (GOMA) and the Mississippi Department of Environmental Quality, Surface Waters Division, Jackson, MS. Prepared by: Tetra Tech, Inc., Center for Ecological Sciences, Owings Mills, MD. 79pp. plus 4 appendixes.

Hale, S.S. and J.F. Heltshe. 2008. Signals from the benthos: Development and evaluation of a benthic index for the nearshore Gulf of Maine. Ecological Indicators 8:338-350.

Hale, S.S., G. Cicchetti, and C.F. Deacutis. 2016. Eutrophication and hypoxia diminish ecosystem functions of benthic communities in a New England estuary. Frontiers in Marine Science 3:Article 249.

Hale, S., H.W. Buffum, J.A. Kiddon, and M.M. Hughes. 2017. Subtidal benthic invertebrates shifting northward along the US Atlantic coast. Estuaries and Coasts 40:1744-1756.

Hale, S.S., M.M. Hughes, and H.W Buffum, H.W. 2018a. Historical trends of benthic invertebrate biodiversity spanning 182 years in a southern New England estuary. Estuaries and Coasts 41:1525-1538.

Hale, S.S., H.W. Buffum, and M.M. Hughes. 2018b. Six decades of change in pollution and benthic invertebrate biodiversity in a southern New England estuary. Marine Pollution Bulletin 133:77-87.

Hawkins, C.P., R.H. Norris, J.N. Hogue, and J.W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456–1477.

Hazen and Sawyer and Tetra Tech. 2017. Alley Creek Benthic Invertebrate, Water Quality, and Vegetation Sampling; Trip Report 3: November 14 and 23, 2016 and Final Report: Discussion of Results of the Overall Sampling Program. Prepared for The City of New York Department of Environmental Protection.

Hilsenhoff, W.L. 1988. Rapid Field Assessment of Organic Pollution with a Family-Level Biotic Index. Journal of the North American Benthological Society 7(1):65-68.

Holland, A.F., A. Shaughnessy, and M.H. Heigel. 1987. Long-term variation in mesohaline Chesapeake Bay benthos: spatial and temporal patterns. Estuaries 10:227-245.

Hyland, J. L., R. F. Van Dolah, J. F. Paul, J. K. Summers, W. L. Balthis, and V. D. Engle. 1999a. Predicting benthic stress from sediment contamination along the U.S. Atlantic and Gulf of Mexico Coasts. Presentation No. 684 at ERF'99, September 25-30, 1999, New Orleans, LA.

Hyland, J. F., R. F. Van Dolah, and T. R. Snoots. 1999b. Predicting stress in benthic communities of Southeastern U.S. estuaries in relation to chemical contamination of sediments. Environmental Toxicology and Chemistry 18:2557-2564.

Larsen, P.F., Johnson, A.C., Doggett, L.F., 1983. Environmental benchmark studies in Casco Bay–Portland Harbor, Maine, April 1980. NOAA Technical Memorandum NMFS-F/NEC-19. NOAA, NMFS, Northeast Fisheries Center, Woods Hole, MA.

Llanso, R.J. and M. Southerland. 2006. Hudson River estuary biocriteria application and validation. Submitted to: New York State Department of Environmental Conservation (NYSDEC), Albany, NY. Submitted by: Versar, Inc., Columbia, MD. 15pp. plus 3 appendixes.

Llanso, R.J., L.C. Scott, J.L. Hyland, D.M. Dauer, D.E Russell, and F.W. Kutz. 2002. An estuarine benthic index of biotic integrity for the Mid-Atlantic region of the United States. II. Index Development. Estuaries 25(6A):1231-1242.

Llansó, R., M. Southerland, J. Vølstad, D. Strebel, and G. Mercurio. 2003. Hudson River estuary biocriteria final report. Submitted to: New York State Department of Environmental Conservation (NYSDEC), Albany, NY. Submitted by Versar, Inc., Columbia, MD, Tetra Tech, Inc., Owings Mills, MD. 110pp. Plus 3 appendixes and an executive summary.

MacDonald Environmental Services, Ltd. 1994. Approach to the assessment of sediment quality in Florida coastal waters, vol. 1. Development and Evaluation of Sediment Quality Assessment Guidelines. Prepared for Florida Department of Environmental Protection.

Malloy, K.J., D. Wade, A. Janicki, S.A. Grabe, and R. Nijbroek. 2007. Development of a benthic index to assess sediment quality in the Tampa Bay Estuary. Marine Pollution Bulletin 54:22-31.

McCune, B. and M. J. Mefford. 2006. PC-ORD. Multivariate Analysis of Ecological Data. Version 5.18. MjM Software, Gleneden Beach, Oregon, U.S.A.

Mebane, C.A., T.R. Maret, R.M. Hughes 2003. An Index of Biological Integrity (IBI) for Pacific Northwest Rivers. American Fisheries Society 132, 239-261.

Mississippi Department of Environmental Quality (MDEQ). 2013. The Gulf Benthic Index for Mississippi (GBI-MS). Prepared for: Mississippi Department of Environmental Quality, Office of Pollution Control, Jackson, MS. Prepared by: Tetra Tech, Inc., Montpelier, VT. 30pp. plus 2 appendixes.

Muxika, I., A. Borja, and W. Bonne. 2004. The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. Ecological Indicators 5:19-31.

Nestlerode, J.A., M.C. Murrell, J.D. Hagy III, L. Hartwell, J.A. Lisa. 2019. Bioassessment of a Northwest Florida Estuary Using Benthic Macroinvertebrates. Integrated Environmental Assessment and Management 00, 1-12.

Paul, J.F., K.J. Scott, D.E. Campbell, J.H. Gentile, C.S. Strobel, R.M. Valente, S.B. Weisberg, A.F. Holland, and J.A. Ranasinghe. 2001. Developing and applying a benthic index of estuarine condition for the Virginian Biogeographic Province. Biological Indicators 1:83-99.

Pearson, T.H., and R. Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology: An Annual Review 16:229-311.

Pelletier, M.C., A.J. Gold, J.F. Heltshe, H.W. Buffum, 2010. A method to identify estuarine macroinvertebrate pollution indicator species in the Virginian Biogeographic Province. Ecological Indicators 10, 1037-1048.

Pelletier, M.C., A.J. Gold, I. Gonzalez, and C. Oviatt. 2012. Application of multiple index development approaches to benthic invertebrate data from the Virginian Biogeographic Province, USA. Ecological Indicators 23:176-188.

Pelletier, M.C., D.J. Gillett, A. Hamilton, T. Grayson, V. Hansen, E.W. Leppo, S.B. Weisberg, and A. Borja. 2018. Adaptation and application of multivariate AMBI (M-AMBI) in US coastal waters. Ecological Indicators 89:818-827.

Rakocinski, C.F. 2012. Evaluating macrobenthic process indicators in relation to organic enrichment and hypoxia. Ecological Indicators 13:1-12.

Ranasinghe, J.A., J.B. Frithsen, F.W. Kutz, J.F. Paul, D.E. Russell, R.A. Batiuk, J.L. Hyland, J. Scott, and D. Dauer. 2001. Application of two indices of benthic community condition in Chesapeake Bay. Environmetrics 13:499-511.

Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Caiden, 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.

Rosenberg, R., M. Blomqvist, H.C. Nilsson, H. Cederwall, and A. Dimming. 2004. Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. Marine Pollution Bulletin 49:728–739.

Schaffner, L. C., R.J. Diaz, C.R. Olson, and I.L. Larsen. 1987. Faunal characteristics and sediment accumulation processes in the James River Estuary, Virginia. Estuarine, Coastal and Shelf Science 25:211-226.

Smith, R.W., M. Bergen, S.B. Weisberg, D. Cadien, A. Dalkey, D. Montagne, J.K. Stull, and R.G. Verardes. 2001. Benthic response index for assessing infaunal communities on the southern California mainland shelf. Ecological Applications 11(4): 1073-1087.

Smucker, N.J., Kuhn, A., Cruz-Quinones, C.J., Serbst, J.R. and Lake, J.L., 2018. Stable isotopes of algae and macroinvertebrates in streams respond to watershed urbanization, inform management goals, and indicate food web relationships. Ecological indicators, 90, pp.295-304.

Stoddard, J.L., D.P. Larsen, C.P. Hawkins, R.K. Johnson, and R.H. Norris. 2006. Setting expectations for the ecological conditions of streams: the concept of reference condition. Ecological Applications 16(4):1267–1276.

USEPA. 2014. National Coastal Condition Assessment: Field Operations Manual. EPA-841-R-14-007. U.S. Environmental Protection Agency, Washington, DC.

USEPA. 2016. National Coastal Condition Assessment 2015: Laboratory Operations Manual. EPA-841-R-14-008. U.S. Environmental Protection Agency, Office of Water, Washington, DC. 2016.

USEPA. 2020. National Coastal Condition Assessment 2020: Laboratory Operations Manual. U.S. Environmental Protection Agency, Office of Water, Washington, DC. 2020. EPA # 841F19004

Van Dolah, R.F., J.L. Hyland, A.F. Holland, J.S. Rosen, and T.R. Snoots. 1999. A benthic index of biological integrity for assessing habitat quality in estuaries of the southeastern USA. Marine Environmental Research 48:269-283.

Vaudrey, J.M.P, C. Yarish, J.K. Kim, C. Pickerell, L. Brousseau, J. Eddings, and M. Sautkulis. 2016. Connecticut Sea Grant Project Report: Comparative Analysis and Model Development for Determining the Susceptibility to Eutrophication of Long Island Sound Embayments. Project number R/CE-34-CTNY. 46 p. Accessed February 2017. http://longislandsoundstudy.net/wpcontent/uploads/2013/08/Vaudrey_R-CE-34-CTNY_FinalReport_2016.pdf.

Versar, Inc. 2005. Chesapeake Bay Benthic Monitoring Program - Data Collection and Processing. http://www.baybenthos.versar.com/DsgnMeth/FieldLab.htm

Weisberg, S.L., J.B. Frithsen, A.F. Holland, J. Paul, K.J. Scott, J.K. Summers, H. Wilson, R. Valente, D. Heimbuch, J. Gerritsen, S. Schimmel, R. Latimer. 1992. EMAP-Estuaries Virginian Province 1990 Demonstration Project Report. United States Environmental Protection Agency, Office of Research and Development, Environmental Research Laboratory, Narragansett, Rhode Island. EPA/600/01-92/XXX.

Weisberg, S.B., J.B. Frithsen, A.F. Holland, J.F. Paul, K.J. Scott, J.K. Summers, H.T. Wilson, D.G. Heimbuch, J. Gerritsen, S.C. Schimmel, and R.W. Latimer. 1993. Virginian Province

Demonstration Project Report, EMAP-Estuaries, 1990. EPA/620/R-93/006. US Environmental Protection Agency, Office of Research and Development, Washington, DC.

Weisberg, S.B., J.A. Ranasinghe, L.C. Schaffner, R.J. Diaz, D.M. Dauer, and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. Estuaries 20(1):149-158.

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

Whitlatch, R.B., R.N. Zajac. ~2011/Unpublished. Development of a Long Island Sound (LIS) Benthic Index for Assessing Environmental Conditions. Connecticut Sea Grant.