JOURNAL OF QUATERNARY SCIENCE (2013) **28**(6) 617–629 ISSN 0267-8179. DOI: 10.1002/jqs.2657

Reconstructing Holocene sea level using salt-marsh foraminifera and transfer functions: lessons from New Jersey, USA

ANDREW C. KEMP,[1]*,† RICHARD J. TELFORD,[2,3] BENJAMIN P. HORTON,[4,5] SHIMON C. ANISFELD[1] and CHRISTOPHER K. SOMMERFIELD[6]

[1]School of Forestry and Environmental Studies, Yale University, New Haven, CT 06511, USA
[2]Department of Biology, University of Bergen, Bergen, Norway
[3]Bjerknes Centre for Climate Research, Bergen, Norway
[4]Sea Level Research, Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ, USA
[5]Division of Earth Sciences and Earth Observatory of Singapore, Nanyang Technological University, 639798, Singapore
[6]College of Earth, Ocean, and Environment, University of Delaware, Lewes, DE, USA

Received 13 March 2013; Revised 4 June 2013; Accepted 22 July 2013

ABSTRACT: We present an expanded training set of salt-marsh foraminifera for reconstructing Holocene relative sea-level change from 12 sites in New Jersey that represent varied physiographic environments. Seven groups of foraminifera are recognized, including four high- or transitional-marsh assemblages and a low-salinity assemblage. A weighted-averaging transfer function trained on this dataset was applied to a dated core from Barnegat Bay to reconstruct sea level with uncertainties of ± 14% of tidal range. We evaluate the transfer function using seven tests. (1) Leave-one-site-out cross validation suggests that training sets of salt-marsh foraminifera are robust to spatial autocorrelation caused by sampling along transects. (2) Segment-wise analysis shows that the transfer function performs best at densely sampled elevations and overall estimates of model performance are over optimistic. (3) Dissimilarity and (4) non-metric multi-dimensional scaling evaluated the analogy between modern and core samples. The closest modern analogues for core samples were drawn from six sites demonstrating the necessity of a multi-site training set. (5) Goodness-of-fit statistics assessed the validity of reconstructions. (6) The transfer function failed a test of significance because of the unusual properties of some cores selected for sea-level reconstruction. (7) Agreement between reconstructed sea level and tide-gauge measurements demonstrates the transfer function's utility. Copyright © 2013 John Wiley & Sons, Ltd.

KEYWORDS: analog matching; Barnegat Bay; leave-one-site-out cross validation; partitioning; sea-level indicators; weighted averaging.

Introduction

Salt-marsh foraminifera are a valuable tool for reconstructing relative sea level (RSL) changes. Their use as a proxy for sea level is underpinned by a robust relationship to tidal elevation (e.g. Scott and Medioli, 1978; Gehrels, 1994; Horton and Edwards, 2006). Foraminifera on salt marshes have differing ecological associations with the balance between inundation by salt water and subaerial exposure that corresponds closely to tidal elevation. Accordingly, salt-marsh sub-environments (low, high, and transitional marsh) can be distinguished from one another using assemblages of foraminifera characterized by the relative abundance of particular species. Recognition of these assemblages where they are preserved in coastal sediment allows RSL to be reconstructed from an understanding of their modern distribution and ecological preferences. The composition and elevational range of foraminiferal assemblages varies among sites and regions because of the secondary influence of environmental variables other than tidal elevation, such as climate and salinity (de Rijk, 1995; Hayward *et al.*, 1999; Kemp *et al.*, 2009a; Wright *et al.*, 2011). Therefore, it is necessary to document modern distributions of salt-marsh foraminifera from the sites and regions where RSL will be reconstructed (Edwards *et al.*, 2004; Horton and Edwards, 2006; Callard *et al.*, 2011). The spatial and physiographic scope of datasets describing the modern distribution of salt-marsh foraminifera should reflect, and provide analogues for, paleoenvironmental conditions that are represented in the sedimentary archives used for RSL reconstruction.

Current research into reconstructing RSL changes can be divided into two prevailing strategies. The first strategy is to generate discrete RSL reconstructions from individual core samples (termed sea-level index points) collected at a suite of sites and compiled into a regional RSL history (e.g. Shennan and Horton, 2002; Törnqvist *et al.*, 2004). Investigations that produce sea-level index points typically focus on larger RSL changes over longer (e.g. the Holocene) timescales and require a relatively large and diverse dataset of modern salt-marsh foraminifera to accurately interpret assemblages potentially deposited in varied sedimentary environments and climatic conditions. The second strategy produces continuous records of recent (approximately the last 100–3000 years) RSL using ordered samples from a single core of high-marsh sediment selected because it exhibited minimal changes in stratigraphy (e.g. Gehrels *et al.*, 2005; Kemp *et al.*, 2009b). As past environmental variability at a single, well-chosen, site is relatively low, a less diverse and smaller dataset of modern foraminifera is usually needed to provide adequate analogy between modern and core material. Emphasis on recognizing subtle RSL changes has led to transfer functions being used to provide quantitative reconstructions with the best possible precision.

This study serves three purposes. First, we provide an expanded dataset of modern foraminifera from the US mid-Atlantic region. This dataset is intended for generating both Holocene sea-level index points and continuous reconstructions of late Holocene RSL. Salt-marsh foraminifera were described in 175 samples from 12 sites selected to represent a

*Correspondence: A. C. Kemp, at †present address below.
†Present Address: Department of Earth and Ocean Sciences, Tufts University, Medford, MA 02155, USA. E-mail: andrew.kemp@tufts.edu

diversity of sedimentary environments, including 56 samples from three sites presented by Kemp *et al*. (2012a). Secondly, we investigate the value and necessity of developing sub-regional and regional training sets from multiple sites that include varied physiographic settings. Thirdly, we test if a weighted-averaging transfer function trained on this dataset can accurately reconstruct RSL and use seven techniques to evaluate model performance. These are a rigorous and up-to-date series of tests that are widely applicable in paleoenvironmental reconstruction, but some of which have been under utilized in RSL reconstruction. Specifically we test:

1. The effect of spatial autocorrelation introduced by sampling along transects using leave-one-site-out cross validation.
2. The effect of uneven sampling of the environmental gradient using segment-wise analysis.
3. The analogy between core and modern samples using dissimilarity and
4. non-metric multi-dimensional scaling.
5. The goodness-of-fit between core samples and elevation.
6. The statistical significance of the reconstruction by comparing it with alternative models trained on random environmental data.
7. How well the transfer function can reproduce the sea-level history recorded at a nearby tide gauge (The Battery in New York City) when applied to a dated core of salt-marsh sediment spanning the last ~200 years.

Regional Setting

Southern New Jersey is part of the wider US mid-Atlantic region and the 12 study sites share a common climate and oceanographic regime. Tidal marshes in the mid-Atlantic are ecologically distinct from the *Juncus roemerianus*-dominated systems of the south-eastern US (Eleuterius, 1976) that experience a warmer climate and closer proximity to the Gulf Stream. They are also differentiated from tidal marshes in New England that have an appreciably different prevailing climate, coastal geomorphology and Quaternary geologic history, but are dominated by many of the same plant species. The extended training set is broadly representative of tidal marshes and foraminiferal assemblages at the regional (mid-Atlantic) scale. The 12 sites were selected to include examples of the diverse types of tidal marsh (including low-salinity sites with strong fluvial influence) that are present in southern New Jersey and the mid-Atlantic region. Therefore, individual sites and groups of sites sharing similar physiographic traits represent local and sub-regional conditions, respectively. Eleven of the sites are located on the Atlantic coast between Great Bay and Cape May (Fig. 1). This coast is characterized by a lagoon system comprising open bays and salt marshes lying inland of a barrier-island chain. Nine inlets allow direct exchange of water between the lagoons and open Atlantic Ocean. In contrast, there is no barrier island–lagoon system along the coast of the Delaware Bay where the Sea Breeze site is located. The region has a semidiurnal tidal regime. Great diurnal tidal ranges [mean lower low water (MLLW) to mean higher high water (MHHW)] are larger on the ocean side of the barrier islands (e.g. 1.4 m at Atlantic City) than in the lagoons (typically ~1.1–1.2 m). Tidal influence extends up to 25 km from the coast into bays and brackish river systems.

Modern salt marshes in the US mid-Atlantic region (including southern New Jersey) form extensive platforms dissected by tidal channels that are particularly large and sinuous on Delaware Bay marshes (Ferland, 1990). Tidal-flat environments are rare and low-marsh settings are typically vegetated by *Spartina alterniflora* (tall form). They are flooded by tides at least once a day and exist between approximately mean tide level (MTL) and mean high water (MHW; Tiner, 1985). The high-marsh floral zone is dominated by *Spartina patens*, *Spartina alterniflora* (short form) and *Distichlis spicata* (Daddario, 1961). This zone is inundated less frequently and occurs above MHW (Tiner, 1985). The narrow border (transitional zone) between high salt marsh and freshwater upland exists above MHHW and is characterized by *Phragmites australis*, *Iva fructescens* and *Baccharis halimmifolia*. At sites with stronger freshwater influence, *Typha angustifolia* and *Schoenoplectus americanus* are common (Daddario, 1961; Stuckey and Gould, 2000). Salt marshes are replaced upstream by brackish marshes dominated at all vegetated intertidal elevations by *Phragmites australis*, *Typha angustifolia*, *Spartina cynosuroides* and *Schoenoplectus americanus* (Tiner, 1985); examples of these environments are particularly common around Great Egg Harbor (Fig. 1).

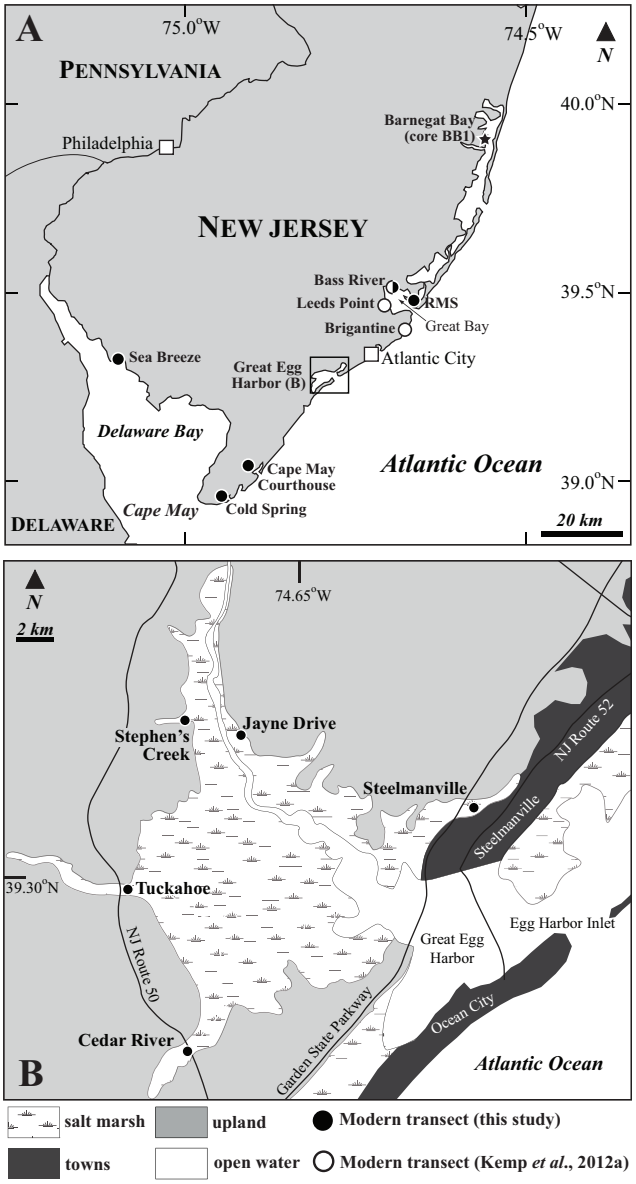

Figure 1. Location of study sites in New Jersey. (A) Sites at which transects were collected for this investigation are indicated by filled circles, and open circles represent sites with existing data from Kemp *et al*. (2012a). The Barnegat Bay site where core BB1 was collected is denoted by a star symbol. RMS = Rutgers Marine Station. (B) Sites in Great Egg Harbor.

A core of salt-marsh sediment (BB1) was collected from Barnegat Bay (Fig. 1) in a short-form *Spartina alterniflora* vegetation zone with a surface elevation of 0.10 m above MTL. The upper 41 cm of the core spans approximately the last 200 years and was previously dated by recognizing chronohorizons of lead concentrations, ratios of lead isotopes and ^{137}Cs activity (Kemp *et al.*, 2012b). The great diurnal tidal range at this site is small (0.17 m) because of its back-barrier location distal to an inlet and the restriction of tidal flow by shallow water, shoals and salt-marsh islands (Kennish, 2001).

Materials and Methods

Foraminiferal analysis

A transect was established at each site across the prevailing environmental and elevational gradient. Sampling points (stations) were positioned along transects to include principal plant communities with emphasis on high-marsh plant communities that are commonly the basis for sea-level reconstruction. This linear, within-site sampling is necessary (and consequently widespread) in sea-level research using salt-marsh microfossils because the primary environmental gradient is elevation within a single site rather than elevation among multiple sites. However, this sampling regime contradicts the assumptions about spatial autocorrelation and the independence of modern samples that are implicit in many numerical techniques. At each station a surface (0–1 cm) sample was collected, preserved in buffered ethanol and stained with Rose Bengal to differentiate between individuals that were living and dead at the time of collection (Murray and Bowser, 2000; Figueira *et al.*, 2012). Samples were sieved under running water to isolate the foraminifera-bearing fraction between 63 and 500 μm. A wet-splitter divided the retained sediment into representative sub-samples for counting. All samples were counted wet under a binocular microscope. A minimum of 100 individuals were enumerated from a known volume of sample, which is adequate to describe low-diversity assemblages typical of salt-marsh foraminifera (Fatela and Taborda, 2002). Identifications of foraminifera were confirmed by comparison with type and figured specimens lodged at the Smithsonian Institution, Washington, DC. All species of *Ammobaculites* were combined into a generic group because identification was frequently hindered by broken individuals. Calcareous taxa were merged into a single group after counting because relatively few individuals were present and the most common genus varied among sites despite occupying the same position in the tidal frame. The foraminiferal data presented and analysed are dead assemblages that represent a population averaged over several years, making them a suitable analog for paleoenvironmental interpretation (Horton, 1999). Foraminifera in core BB1 were counted from 1-cm-thick slices of sediment prepared and analysed in the same way as modern samples, with the exception of Rose Bengal staining. All taxa were included in the analysis of modern and core assemblages.

Sample elevations and tidal data

Sample elevations were established using one of three methods (Table 1). At six sites, samples were directly leveled to National Oceanic and Atmospheric Administration (NOAA) tidal benchmarks. At four sites, Real Time Kinematic (RTK) satellite navigation (Leica GPS 1200+) was used to establish a temporary benchmark to which individual samples were leveled. The accuracy of RTK measurements was confirmed by leveling to nearby National Geodetic Survey (NGS) benchmarks when possible. At Sea Breeze and Cold Spring Harbor, sample elevations were established by directly leveling to an NGS benchmark. Measured sample elevations reported to orthometric datums were converted to tidal datums using NOAA's VDatum transformation tool (v.2.3.5 with the New Jersey coastal embayment dataset). To combine data from individual sites into a single regional dataset, a standardized water level index (SWLI) modified from Horton and Edwards (2006) was used:

$$SWLI = \frac{(Alt_{ab} - MLLW_b)}{MHHW_b - MLLW_b}$$

where Alt_{ab} is the measured altitude of sample *a* collected at site *b* and $MHHW_b - MLLW_b$ is the great diurnal tidal range at site *b*.

Statistical analysis

Groups of foraminiferal assemblages were recognized and described using Partitioning Around Medoids (PAM) with Euclidean distances, where the appropriate number of partitions was determined by maximum average silhouette width (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990). Analysis was performed on square root transformed abundance data.

Table 1. Summary of site data.

Site	Samples	No. of taxa	Leveling	Range (m MTL)	Range (SWLI)
Bass River*	29	14	RTK, VDatum	−0.44 to 0.59	8–100
Brigantine Barrier*	15	12	RTK, VDatum	−0.48 to 0.79	11–108
Cape May Courthouse	9	11	RTK, VDatum	0.45 to 0.69	85–102
Cedar River	14	12	Tidal benchmark	0.41 to 0.63	89–113
Cold Spring	7	11	Benchmark, VDatum	0.42 to 0.88	82–114
Jayne Drive	5	11	Tidal benchmark	0.96 to 1.02	129–134
Leeds Point*	26	12	RTK, VDatum	−0.13 to 0.76	35–115
Rutgers Marine Station	30	16	Tidal benchmark	−0.13 to 1.04	35–142
Sea Breeze	14	13	Benchmark, VDatum	−0.03 to 0.95	47–119
Steelmanville	8	13	Tidal benchmark	0.31 to 0.55	72–91
Stephen's Creek	9	10	Tidal benchmark	−0.52 to 0.56	5–96
Tuckahoe	9	12	Tidal benchmark	0.33 to 0.57	77–98

Modern salt-marsh foraminifera were described from 12 sites in southern New Jersey. Sites marked with * are counts fully or partially reported in Kemp *et al.* (2012a). Sample elevation in the tidal frame was established by leveling directly to tidal benchmarks or by converting from orthometric (RTK or benchmark) to tidal heights using the VDatum transformation tool. Multiple transects were sampled at Bass River, Leeds Point and Rutgers Marine Station.

Gradient analysis and ordination was performed using the Vegan package (v.2.0-5; Oksanen *et al.*, 2012) for the R statistical language v.2.15.1 (R Core Development Team, 2011). Each analysis included all species and a square root transformation of abundance data. Detrended correspondence analysis (DCA) estimated the length of the environmental gradient in the training set of modern salt-marsh foraminifera (Birks, 1995). Canonical correspondence analysis (CCA) estimated the importance of tidal elevation in explaining species distributions in the modern training set. The ratio of the first constrained eigenvalue (λ_1) to the second, unconstrained eigenvalue (λ_2) indicated the importance of tidal elevation as an explanatory variable. When the ratio λ_1/λ_2 exceeds 1, tidal elevation is an important environmental variable.

Weighted-averaging (WA) transfer functions were developed in the rioja package (v.0.7-3; Juggins, 2009) for R with inverse (WA-inv), classical (WA-cla) and monotonic (WA-mono) deshrinking. WA was selected because the assumed unimodal response of species to elevation is ecologically reasonable and there is strong evidence that it generates reliable paleoenvironmental reconstructions in real and simulated datasets (Juggins and Birks, 2012). Extension of the model to include partial least squares regression (WA-PLS) showed that component 2 of a WA-PLS transfer function did not improve performance sufficiently to warrant the use of the more complicated model. All taxa were included and abundance was expressed as square root transformed percentages to stabilize variance in the dataset. The WA-inv transfer function was applied to core BB1 to reconstruct sea level. All taxa in core samples were included and abundance was expressed as square root transformed percentage data. Transfer function output was the estimated elevation (in SWLI units) at which each assemblage formed with a sample-specific uncertainty derived by bootstrapping ($n = 1000$) that includes both the S_1 and the S_2 components. This uncertainty reflects an approximately 66% confidence interval, which is close to the environmental tolerance of a species (Birks, 1995). Using a ~95% interval would make the uncertainty too large to discern paleoenvironmental changes from reconstruction uncertainties in almost all instances (Juggins and Birks, 2012).

To investigate how the spatial composition of the training set used in transfer function development influences reconstructions of sea level we developed alternative weighted-averaging transfer functions from subsets of the expanded dataset. Single-site (local) transfer functions were trained on data from Leeds Point, Bass River and Rutgers Marine Station (there were too few samples from the remaining sites to develop additional models). Transfer functions based on samples from Great Bay (Leeds Point, Bass River, Rutgers Marine Station) and Great Egg Harbor (Tuckahoe, Cedar River, Stephen's Creek, Jayne Drive, Steelmanville) provided sub-regional models from distinct physiographic marsh types.

The proficiency of the regional transfer function was assessed using seven methods that together provide a comprehensive and up-to-date suite of techniques for evaluating paleoenvironmental reconstructions from transfer functions:

i. Model performance was assessed using leave-one-site-out (LOSO) cross-validation, where all samples from a single site are omitted from the training set and data from the remaining sites are used to predict them (Payne *et al.*, 2012). This technique is a more appropriate estimate of transfer function performance than leave one out (LOO) for clustered datasets such as those used in sea-level research because it takes into account the effect of spatial autocorrelation from sampling along transects when quantifying transfer function performance.
ii. The effect of uneven sampling of the environmental gradient (elevation) on model performance was analysed by segment-wise division of the gradient into 10 equal parts for each of which a root mean squared error of prediction (RMSEP) and r^2 was calculated (Telford and Birks, 2011a). This analysis gauged the influence of sample distribution along the environmental gradient in estimating transfer function performance because most training sets used in sea-level research (and other paleoenvironmental fields) rarely satisfy the assumption of even sampling that is implicit when cross validating the entire training set in one step.
iii. The analogy between modern and core samples was analysed using the analog package (v.0.8-2; Simpson, 2007) for R. Measured dissimilarity estimated the degree of analogy between core and modern samples to judge if each core sample has an adequate modern counterpart in the training set on which to base a paleoenvironmental interpretation. To establish critical thresholds (at the 2, 5, 10 and 20% level) from the modern training set, the Bray–Curtis distance metric was used to calculate the dissimilarity between all possible pairs of modern samples. The choice of thresholds is subjective, but these values have been widely and effectively used in a broad range of paleoenvironmental reconstructions (e.g. Overpeck *et al.*, 1985; Jackson and Williams, 2004; Simpson, 2012).
iv. Non-metric multi-dimensional scaling (NMDS) with Bray–Curtis dissimilarity was used to graphically display the taxonomic distance among modern and core samples. It complements measures of analogy by passively showing the trajectory of core samples through an ordination space populated by the training set. The trajectory also provides a way to confirm reconstructions of elevation produced by the transfer function.
v. Goodness-of-fit statistics for core samples were derived by passively fitting samples from BB1 into a constrained ordination (CCA) of the New Jersey modern dataset with tidal elevation as the only constraint following the approach of Simpson and Hall (2012). The squared residual length between core samples and their fitted positions on the first constrained axis is compared with residual differences in the modern dataset. Thresholds at 90, 95 and 99% were established from the modern dataset for progressively worse fits (weak, poor and very poor, respectively). These thresholds are subjective, but recommended by Simpson and Hall (2012). This technique is a principal means to evaluate transfer function reconstructions, but is under utilized and rarely presented (Juggins and Birks, 2012; Simpson and Hall, 2012). The analysis was conducted in the analog package (v.0.8-2; Simpson, 2007) for R on square-root-transformed species data.
vi. The statistical significance of the reconstructions was tested following the procedure described by Telford and Birks (2011b) and performed in the palaeoSig package (v.1.1-1; Telford, 2012) for R. To be deemed statistically significant, the proportion of variance in the fossil data explained by the reconstruction should exceed that explained by 95% of 999 alternative reconstructions trained on random environmental data. Telford and Birks (2011b) encouraged the application of this test to all transfer functions and concluded that models failing the test should be interpreted with caution. It has not been applied to transfer functions developed from salt-marsh foraminifera.

vii. Comparison of reconstructed RSL from core BB1 with historical measurements at the nearby (~75 km) tide gauge at The Battery in New York City. This approach has often been used to assess and validate the utility of sea-level reconstructions from salt marshes (Gehrels *et al.*, 2002). If the transfer function is able to replicate the instrumental record the reconstruction (rather than the model) is deemed to be reliable.

Regional Groups of Foraminifera

Foraminifera were present in 175 samples collected from 12 sites, including three (56 samples) described in Kemp *et al.* (2012a). Twenty species were recognized of which 16 made up 10% or more of the dead assemblage in at least one sample. Table 1 provides details for each transect and the distribution of foraminifera along new transects is summarized in supplementary Figs S1–S10. Appendix 1 includes all foraminiferal data and sample elevations and distinguishes samples newly described in this study from those in Kemp *et al.* (2012a).

Samples of modern foraminifera were combined into a single dataset from which seven groups were identified using PAM (Fig. 2). These groups represent distinctive assemblages of foraminifera that exist on modern tidal marshes in southern New Jersey and probably represent the principal assemblages of salt-marsh foraminifera in the wider mid-Atlantic region. Low-marsh and tidal-flat environments throughout the study region are characterized by two groups (F dominated by *Ammobaculites* species and G dominated by *Miliammina fusca*) that frequently occurred in close proximity to one another. One or both of these groups was present at eight of the sites and were probably not recorded at the other sites because of the skewed distribution of samples towards high-marsh settings (Fig. 3A). The ubiquitous nature of these groups has been recognized in low salt-marsh and tidal-flat environments in the south-eastern and mid-Atlantic US (Ellison *et al.*, 1965; Ellison and Nichols, 1976; Goldstein *et al.*, 1995; Hippensteel *et al.*, 2000), New England (Scott and Leckie, 1990; Gehrels, 1994; de Rijk, 1995; Edwards *et al.*, 2004) and Atlantic Canada (Scott and Medioli, 1978; Scott *et al.*, 1981; Smith *et al.*, 1984).

Group A includes samples from transitional environments (often above MHHW) and is characterized by *Haplophragmoides manilaensis*. High salt-marsh environments typically between MHW and MHHW are represented by three groups (B, C and D). The dominant species of foraminifera in group B is *Trochammina inflata*. Group C consists of samples in which *Arenoparrella mexicana* was the characteristic species. Group D is made up of samples with high proportions of *Tiphotrocha comprimata* and includes samples from low-salinity and high-marsh environments. At lower salinity sites in Great Egg Harbor (Fig. 1), high-marsh samples are dominated by *Ammoastuta inepta* (Group E; Fig. 2). Variability in high-marsh assemblages among sites with different physiographic characteristics is typical of datasets that include multiple sites and probably reflects the secondary and sub-regional influence of environmental factors such as climate or salinity (Kemp *et al.*, 2009a; Wright *et al.*, 2011). For example, 19 of the 20 samples in Group E were situated

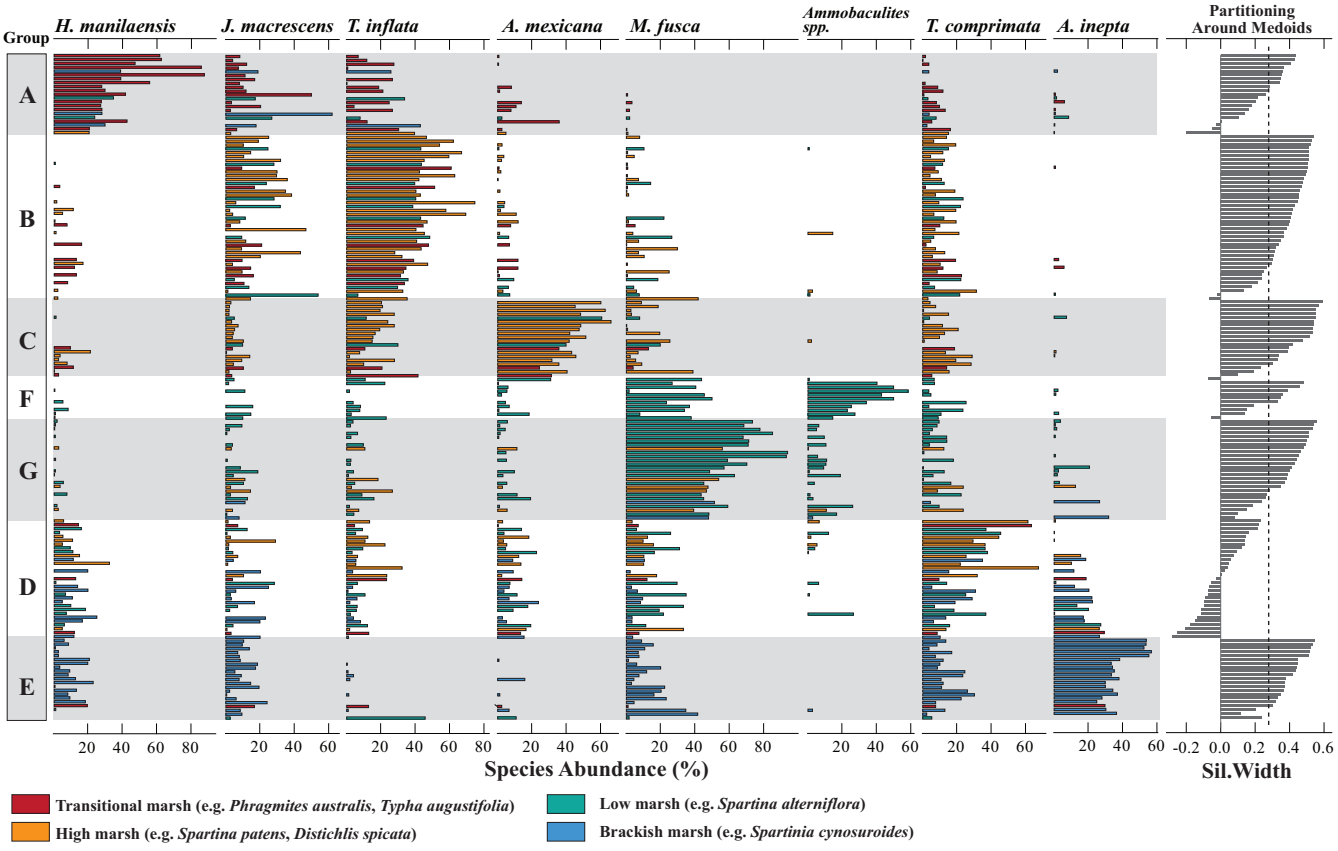

Figure 2. Combined dataset of 175 samples of modern salt-marsh foraminifera from 12 sites in New Jersey, including data from Kemp *et al.* (2012a). Samples are divided into the seven groups (A–G) identified using PAM. Bar color denotes the dominant type of vegetation at the sampling station. Low-marsh samples include a small number of tidal-flat samples and those vegetated by *Spartina alterniflora*. High-marsh samples were vegetated by *Spartina patens* and *Distichlis spicata*. The transitional samples represent environments above MHHW on the fringe of freshwater upland ecosystems. The low-salinity samples are largely from sites around Great Egg Harbor. Silhouette width (sil. width) for each sample is presented to the right; the average for the complete dataset was 0.29 (dashed vertical line). This figure is available in colour online at wileyonlinelibrary.com.

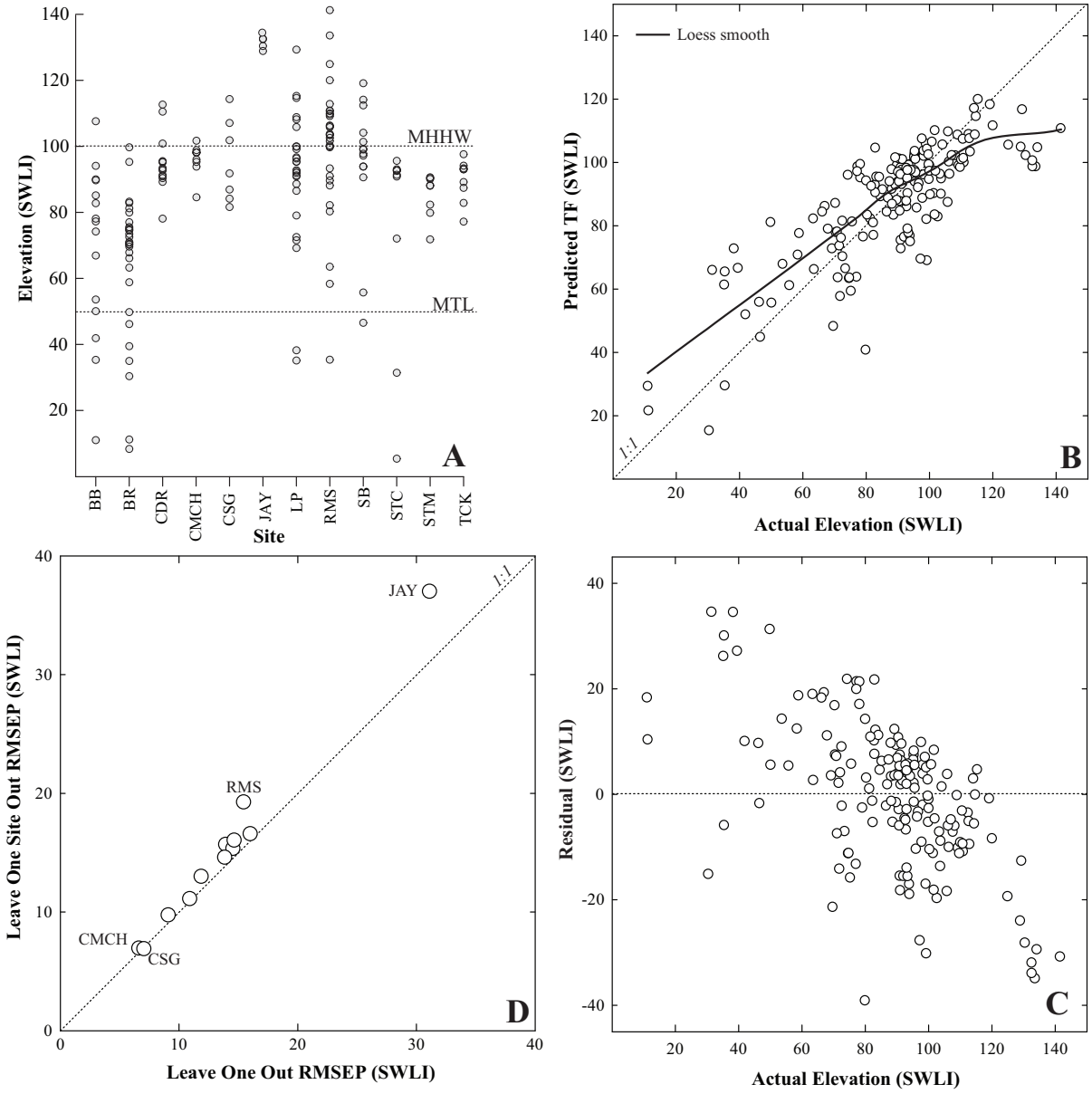

Figure 3. Transfer function development and performance. (A) Distribution of samples over the elevational gradient at each of the 12 sites. (B) Transfer function performance of actual and predicted elevations for 172 samples. The transfer function is a weighted averaging with inverse deshrinking (WA-inv) model under leave-one-site-out (LOSO) cross validation. Solid line is a loess smooth of the data points. (C) Residuals between elevation predicted by the transfer function and actual sample elevation measured at the time of collection. (D) Comparison of root mean squared error of prediction (RMSEP) from leave-one-out (LOO) and leave-one-site-out (LOSO) cross validation for each site. BB = Brigantine Barrier, BR = Bass River, CDR = Cedar River, CMCH = Cape May Courthouse, CSG = Cold Spring, JAY = Jayne's Drive, LP = Leeds Point, RMS = Rutgers Marine Station, SB = Sea Breeze, STC = Stephen's Creek, STM = Steelmanville, TCK = Tuckahoe.

at lower salinity sites in Great Egg Harbor (Fig. 1B; Cedar River, Tuckahoe, Stephen's Creek, Jayne Drive and Steelmanville). This group dominated by *Ammoastuta inepta* is typical of low or brackish salinity regimes (Scott *et al.*, 2001), Great Egg River is tidal for its lower 23 km and salinities decrease up river as the relative influence of freshwater input increases. Wright *et al.* (2011) demonstrated that high-marsh assemblages of foraminifera varied among regions and recognized the dominance of *Haplophragmoides* species in North Carolina, *Jadammina macrescens* in Connecticut and *Balticammina pseudomacrescens* in Newfoundland. In North Carolina, Kemp *et al.* (2009a) sampled a diverse range of marshes in the Albemarle–Pamlico estuarine system and recognized seven sub-regional groups of foraminifera, including five from high salt-marsh environments. The presence of sub-regional groups of high-marsh foraminifera in New Jersey and other regions suggests that accurate reconstructions of RSL from high-marsh sediment require a training set drawn from varied physiographic environments to capture this sub-regional diversity. In a space-for-time substitution the proximity of these assemblages to one another indicates that even over short time scales it is reasonable to expect the dominant group of foraminifera to change within a single core even where the marsh kept pace with sea-level rise and maintained a near-constant elevation in the tidal frame.

Transfer Function Development and Evaluation

DCA (Hill and Gauch, 1980) of the New Jersey training set returned a first axis length of 3.1 standard deviation units, indicating high turnover of species and supporting use of unimodal methods (Birks, 1995). CCA showed that tidal elevation explained 12.6% of variance in the training set of foraminifera. The ratio between the constrained CCA axis and the first unconstrained axis (λ_1/λ_2) was 0.76, which is less

than the value of 1 that is often used to indicate if the environmental variable under consideration (elevation) is an important one for explaining species distributions. However, transfer functions developed from datasets where $\lambda_1/\lambda_2 < 1$ may have utility (Dixit *et al.*, 1993). WA transfer functions were developed for reconstructing RSL from assemblages of foraminifera preserved in buried salt-marsh sediment.

Cross validation of a clustered dataset

Transfer functions assume that training sets are composed of independent samples (Birks, 1995). RSL reconstructions from salt-marsh foraminifera (and diatoms) normally employ training sets where multiple samples were collected at individual sites separated by many kilometers (e.g. Fig. 1). These clustered datasets are an appropriate way to sample a prevailing environmental gradient (tidal elevation) that is more pronounced within a single site than it is among sites. However, the methods most commonly used in sea-level research (and other fields) to evaluate transfer function performance (e.g. LOO) may consequently return overly optimistic results for clustered datasets. LOSO cross validation offers an alternative means to evaluate the performance of transfer functions trained on clustered modern datasets (Payne *et al.*, 2012).

A WA-inv transfer function trained on the expanded New Jersey modern dataset was evaluated with LOO and LOSO cross validation (Table 2; Fig. 3). Transfer function performance under both methods indicates that there was a good relationship between measured and transfer function-predicted elevations. The cross-validated precision is similar to reported values for other transfer functions of salt-marsh foraminifera from eastern North America that varied from $\pm 3.7\%$ of tidal range in Nova Scotia (Gehrels *et al.*, 2005) to $\pm 25\%$ in North Carolina (Kemp *et al.*, 2009b). The increased RMSEP (difference of 1.7 SWLI units) under LOSO is unlikely to cause a discernible difference to RSL reconstructions except at sites with very large tidal ranges, and may in part be caused by using a smaller sample size than LOO during cross validation (Payne *et al.*, 2012). For example, the increase equates to an additional uncertainty of $< \pm 0.01$ m at Barnegat Bay or ± 0.03 m at Sea Breeze, which is the site with the largest tidal range in the training set.

Table 2. Transfer function performance.

	RMSEP	r^2	Average bias	Maximum bias	Mean segment RMSEP
LOO					
WA-inv	13.96	0.62	0.01	29.48	18.43
WA-cla	17.21	0.62	0.07	21.13	19.46
WA-mono	13.98	0.61	0.00	29.48	18.55
LOSO					
WA-inv	15.66	0.52	0.81	34.87	20.06
WA-cla	19.64	0.54	1.69	29.25	23.00
WA-mono	15.83	0.51	0.87	34.78	20.56

Results for weighted-averaging (WA) transfer functions developed using the regional training set of 172 samples of salt-marsh foraminifera. Inverse deshrinking (WA-inv), classical deshrinking (WA-cla) and monotonic spline deshrinking (WA-mono) were applied. Two types of cross validation assessed model performance. Leave-one-out (LOO) excluded individual samples, while leave-one-site-out (LOSO) excluded all samples from one site. Root mean square error of prediction (RMSEP), average bias and maximum bias are reported in SWLI units. Segment-wise RMSEP was calculated by dividing the elevational gradient into 10 equal parts; value shown is the mean across all 10 segments.

Comparison of RMSEP for individual sites under the two cross-validation schemes indicates that the decrease in model performance with LOSO cross validation was largely caused by assemblages from Jayne Drive and Rutgers Marine Station (Fig. 3D). For the remaining sites there was little difference between RMSEP estimated by LOO and LOSO (average 0.75 SWLI). The Jayne Drive site is unique in the expanded training set because it includes samples from very high elevations with unusually high abundances of *Miliammina petila*. As such the transfer function based on the remaining sites had difficulty predicting those from Jayne Drive when they were excluded during LOSO cross validation. Model performance as evaluated by LOSO would therefore be improved by adding sites similar to Jayne Drive to the training set.

Comparison of model performance under LOO and LOSO cross validation indicates that spatial autocorrelation in clustered training sets causes slightly over-optimistic estimates of RMSEP. The difference (1.7 SWLI units) is relatively less pronounced for training sets of salt-marsh foraminifera than has been documented for other paleoenvironmental proxies (Payne *et al.*, 2012) and probably reflects the combination of a robust choice of transfer function method, a long environmental gradient at most sites (3.1 standard deviation units), limited differences in secondary gradients among sites and low-diversity assemblages that exist at multiple sites throughout the study region. Clustered training sets of salt-marsh foraminifera (in New Jersey at least) appear robust to the influence of spatial autocorrelation introduced by sampling along transects.

Effect of uneven sampling of the environmental gradient

Even distribution of samples along the environmental gradient is not a requirement for transfer function development, but estimation of the environmental optima and tolerance of taxa is most efficient when it is evenly sampled (Telford and Birks, 2011a). Modern training sets of salt-marsh foraminifera often include a disproportionately large number of high-marsh samples and relatively few low-marsh samples (e.g. Horton and Edwards, 2006; Kemp *et al.*, 2009a; Wright *et al.*, 2011). This is caused by sampling regimes emphasizing collection of modern analogues for core material (usually high-marsh peat) or practical reasons preventing sampling of some portion of the environmental gradient. The expanded New Jersey training set is not atypical in having a distribution of samples biased towards higher tidal elevations between 60 and 120 SWLI units (Fig. 4). Uneven sampling along an environmental gradient can cause bias in transfer function performance estimated during cross validation (Telford and Birks, 2011a). Performance will be better in the heavily sampled part of the environmental gradient than the less frequently sampled parts because there are more analogs that are retained more frequently in procedures such as LOO. Telford and Birks (2011a) presented an alternative (segment-wise) procedure for quantifying the influence of uneven sample distribution on model performance by dividing the environmental gradient into segments of equal length.

The influence of uneven sample distribution on the regional WA-inv transfer function was investigated by estimating RMSEP (LOSO cross validation) for 10 equal intervals of elevation (Fig. 4). Under segment-wise analysis, RMSEP (20.1 SWLI units) exceeded the value estimated when sample distribution was not considered (15.7 SWLI units), indicating that assessing transfer function performance across all elevations in a single step is overly optimistic when the environmental gradient was sampled unevenly. Individual segments had RMSEP of 10.3–35.9 SWLI units and the over-represented

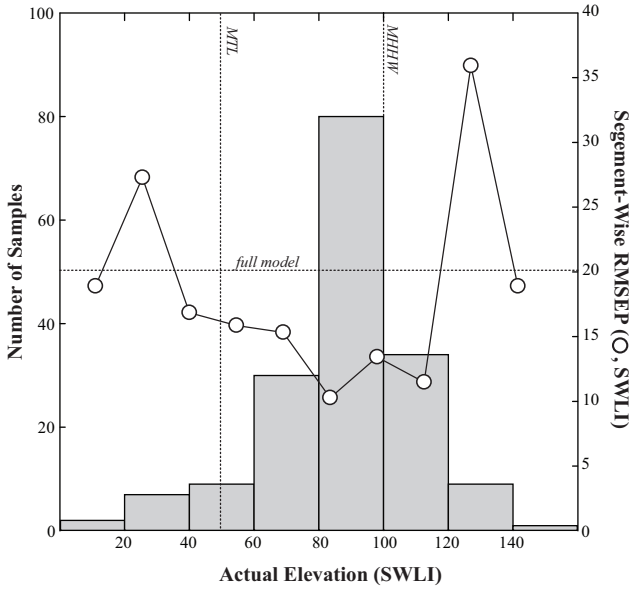

Figure 4. Influence of uneven sampling of the elevational gradient on transfer function performance. Grey histogram is the number of samples (left axis) within groups of 20 SWLI units. Open circles are the root mean square error of prediction (RMSEP) estimated for 10 segments of equal elevational range using the weighted averaging transfer function with inverse deshrinking and leave-one-site-out (LOSO) cross validation. Under segment-wise analysis the full LOSO model had an RMSEP of 20.1 SLWI units (horizontal dashed line). The transfer function performs most strongly at elevations that were best represented in the modern dataset. MTL = mean tide level, MHHW = mean higher high water.

part of the elevational gradient (high marsh) had the smallest RSMEP, while the largest RMSEP values were from segments with fewer samples (low marsh). The same pattern was observed under LOO cross validation and with different deshrinking techniques (Table 2).

The segment-wise results indicate that model performance is variable across the gradient of tidal elevation and that overall model performance is not always a good predictor of actual performance in reconstructing sea level. Reconstructions from assemblages of foraminifera that formed in environments similar to the high-marsh settings that were heavily sampled in the training set would have smaller errors than those more similar to under-sampled, low-marsh assemblages despite a decrease in full model performance. Therefore, decreased overall model performance from uneven sampling does not necessarily lead to more uncertain sea-level reconstructions.

Transfer Function Application and Evaluation

Foraminifera were enumerated from 27 samples in core BB1 between the surface and a depth of 45 cm (Fig. 5). The regional WA-inv transfer function was applied to core BB1 to reconstruct the elevation (in SWLI units) at which each sample formed with a sample-specific error derived from bootstrapping. Reconstructed elevations for samples between depths of 45 cm and 5 cm indicate that these samples formed close to MHHW (SWLI from 92.7 to 111.5) as indicated by the dominance of high-marsh taxa. The uppermost 3 cm had increased abundances of *Miliammina fusca* and *Ammobaculites* spp. and the transfer function estimated that these samples were deposited at lower tidal elevations between 58.5 and 74.3 SWLI units. The average, sample-specific uncertainty (~66% confidence interval) for these reconstructed elevations was ± 14.2 SWLI units, equating to ± 2.4 cm given the small great diurnal tidal range in Barnegat Bay and assuming a constant tidal range for the period under consideration.

Analogy between modern and core samples

To evaluate the ecological plausibility of RSL reconstructions from the transfer function, minimum dissimilarity between core and modern samples (distance to closest single modern analog) was calculated using the Bray–Curtis distance metric and compared with thresholds established by pairwise comparison of modern samples (Fig. 5). Five samples in BB1 exceeded the 20% threshold and the closest analogs were drawn from six different sites in the modern dataset, including two low-salinity sites at Tuckahoe and Cedar River. The need for analogs from multiple sites and varied physiographic settings to reconstruct sea level from a core of high salt-marsh sediment representing only ~200 years illustrates the necessity for, and value of, a diverse training set to capture sub-regional variability in assemblages (Horton and Edwards, 2005).

The choice of a 20% threshold is higher than that employed in many studies (e.g. Overpeck *et al.*, 1985; Anderson *et al.*, 1989) because the low diversity of foraminiferal assemblages coupled with over-sampling of high-marsh environments (Figs 3A and 4) produced absolute threshold values that are very small because of the high degree of similarity among modern samples. In homogenous data sets, such as the New Jersey training set, the absolute dissimilarity between two samples may be small despite representing a relatively high percentile of ranked dissimilarities. Conversely, in heterogeneous datasets, two samples at a low threshold of dissimilarity may actually have very different taxonomic compositions. The weak analogy for core samples at 20, 23, 27 and 28 cm was partly caused by the presence of *Miliammina petila* in abundances exceeding its contribution to the modern dataset. *Miliammina petila* had a maximum occurrence in modern samples of 19%, but contributed 27–34% of individuals in the four core samples. This problem was also highlighted by Kemp *et al.* (2012a) at another core location (Leeds Point, NJ; Fig. 1) and despite considerable expansion, it is apparent that the modern training set still lacks a modern equivalent of this assemblage, suggesting that additional modern sampling may be necessary.

NMDS complements the use of dissimilarity measures to judge the similarity in species composition of modern and core samples. It also shows if reconstructed elevations agree with the observed elevations of training set samples composed of similar assemblages. Scatter of modern samples shows separation of Great Egg Harbor samples from the others on the second axis (NMDS 2), probably reflecting differences in assemblages found in low-salinity settings with stronger fluvial influence (Fig. 6A). The trajectory of core samples moved towards a less densely populated region of the NMDS plot at depths below 20 cm, reflecting a change in dominant assemblage from *Tiphotrocha comprimata* to *Jadammina macrescens* with *Miliammina petila*. Samples from BB1 fell within the scatter of modern samples, indicating that core samples are adequately represented by modern equivalents. The trajectory towards lower elevations agrees with transfer function reconstructions for samples in the upper 3 cm that included higher abundances of the low-marsh taxa *Miliammina fusca* (Group G).

Goodness of fit

Goodness of fit measures how well modern and core samples are fitted to the first axis of an ordination constrained by tidal elevation. Well-fitted samples have smaller residual distances to the axis than those with larger residual distances, which

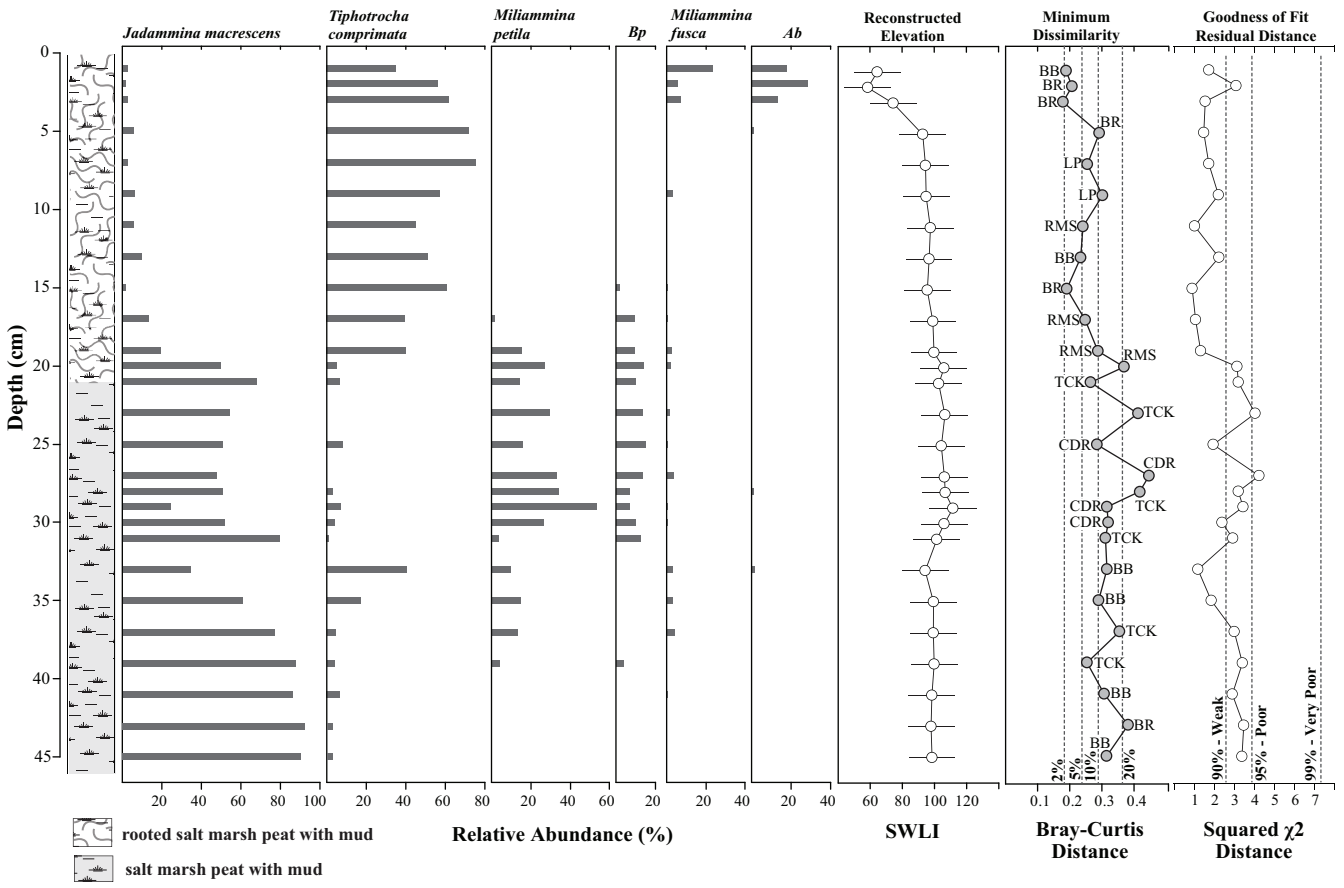

Figure 5. Barnegat Bay core analysis (BB1). Application of the weighted averaging transfer function with inverse deshrinking transfer function to core samples generated reconstructions of elevation. Sample-specific error was derived from bootstrapping ($n = 1000$) and expressed in SWLI units. Bray Curtis distance to the single closest modern analog estimates the degree of analogy between core and modern samples. The site from which the closest analog was drawn is listed next to each datapoint (BB = Brigantine Barrier, BR = Bass River, LP = Leeds Point, RMS = Rutgers Marine Station, TCK = Tuckahoe, CDR = Cedar River). Thresholds at 2, 5, 10 and 20% (vertical, dashed lines) were derived from pairwise analysis of the modern dataset. Goodness of fit was measured as the squared residual fit of core samples in comparison with thresholds (vertical dashed lines) established from the modern dataset. Bp = Balticammina pseudomacrescens; Ab = Ammobaculites spp.

have a poor fit to elevation. This method for evaluating transfer function performance has been under utilized in many paleoenvironmental fields (Simpson and Hall, 2012) and is largely absent from sea-level reconstructions. Thresholds for assessing goodness of fit were established from the New Jersey training set. In core BB1 eleven samples exceeded the 90% threshold for a weak fit to elevation and two (at 23 and 27 cm) exceeded the 95% threshold for a poor fit (Fig. 5). The weak fit of samples below 20 cm in BB1 is coincident with a shift to fossil assemblages dominated by *Jadammina macrescens* and with *Miliammina petila*, which PAM showed are assemblages poorly represented in the regional training set (Fig. 2).

Test of model significance

Telford and Birks (2011b) proposed a method to test the statistical significance of transfer functions and recommended that all reconstructions are evaluated in this manner. The reconstruction is compared with 999 alternative models trained on random environmental variables. It is deemed significant if it explains more of the variance observed in a core than 95% of the randomized models. The regional WA-inv transfer function developed from the New Jersey training set and applied to core BB1 was subjected to this test (Fig. 6B). It accounted for 28% of downcore variance, but did not explain more variance than 95% of the alternative models and therefore failed the significance test of Telford and Birks (2011b).

This apparent failure probably reflects an unusual property of some sea-level reconstructions rather than indicating a true failure of the model. For a salt-marsh in equilibrium with sea-level rise, the elevation reconstructed by applying a transfer function to downcore assemblages of foraminifera will be unchanged. This is apparent in core BB1 that was dominated by high salt-marsh foraminifera resulting in transfer function reconstructions close to 100 SWLI (Fig. 5). The stability of reconstructed elevation (particularly in comparison with reconstruction errors of $\sim \pm 14$ SWLI units) results in there being minimal variability for any model to explain and the reconstruction appears to vary only by chance. Under this circumstance a random model would perform equally as well as the WA-inv model in explaining downcore variability. This is an important caveat to the test of statistical significance as sea-level reconstructions often target unbroken sequences of high salt-marsh sediment to deliberately minimize downcore variability. An additional caveat is that the low number of effective species of foraminifera in the core make it unlikely that any WA reconstruction would be significant under this test (Telford and Birks, 2011b).

Spatial scale

The spatial extent of training sets used in reconstructing sea level continues to be debated in the context of model performance and the diversity of sites needed to provide adequate analogy between modern and core samples (Horton and Edwards, 2005; Watcham *et al.*, 2013). To investigate the

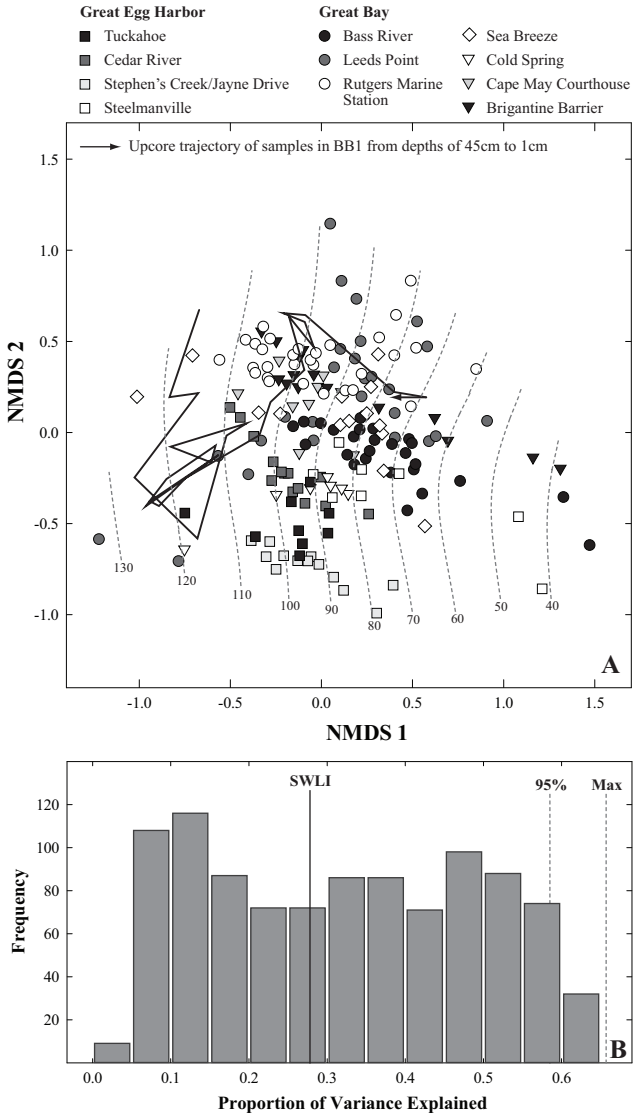

Figure 6. (A) Non-metric multidimensional scaling (NMDS) of modern samples (symbols) with core samples from BB1 presented as trajectories (solid line). Symbols distinguish sites from Great Egg Harbor, Great Bay and other sites. Analysis performed on Wisconsin transformed square root abundance data. Dashed lines mark the position of standardized water level index (SWLI) elevations at labeled intervals. (B) Histogram of the proportion of variance in core BB1 explained by 1000 WA-inv reconstructions, including 999 random environmental data and one on the regional New Jersey dataset of modern foraminifera. Solid black line marks the proportion of variance explained by the regional reconstruction of SWLI elevation, which is lower than the 95th percentile of the null distribution (dashed line). The dashed line (Max) marks the proportion of variance explained by the first axis of a principal components analysis of the fossil data, the maximum that it is possible to explain with a single reconstruction.

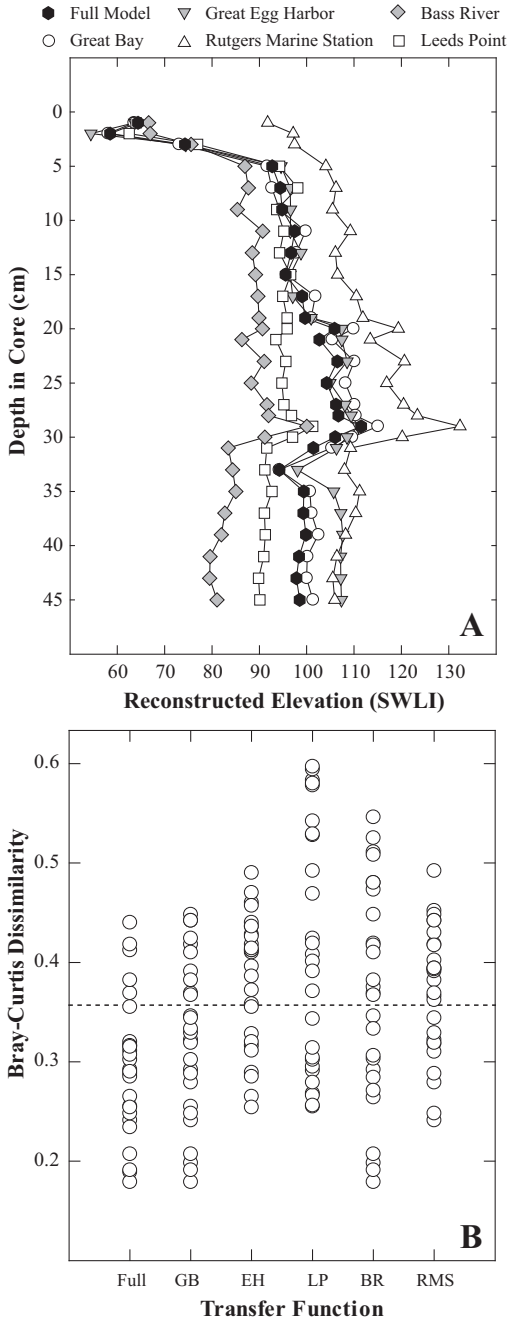

Figure 7. (A) Reconstructed elevation from transfer functions (WA-inv) trained on individual sites, two groups of sites and all sites (full model). Elevations are reported in standardized water level index units (SWLI). (B) Dissimilarity between modern and core (BB1) samples measured by Bray–Curtis distance for transfer functions trained on all sites (full), two groups of sites and individual sites. GB = Great Bay, EH = Egg Harbor, LP = Leeds Point, BR = Bass River, RMS = Rutgers Marine Station. The 20th percentile of dissimilarity (0.357) is marked by a dashed horizontal line and was calculated for the full, regional training set.

effect of developing local, sub-regional, and regional transfer functions we developed five additional WA-inv models trained on single sites (Leeds Point, Bass River and Rutgers Marine Station) and also using sub-regional groups of samples from sites sharing broad physiographic conditions in Great Egg Harbor and Great Bay (Fig. 7A).

Each of the six transfer functions (including the regional model of all New Jersey sites) individually displays the same pattern of paleomarsh elevation predictions when applied to assemblages in core BB1. The difference among transfer functions however was large, with models trained on single sites predicting the highest (Rutgers Marine Station) and lowest (Leeds Point) elevation reconstructions that differed by an average of 24.9 SWLI units. Models trained on multiple sites (Great Egg Harbor, Great Bay and the full New Jersey dataset) reconstructed elevation between the extremes from single site models. Differences among single site (local-scale) transfer functions reflect the inability to capture the complete relationship between high-marsh species and elevation because of the distribution of samples and/or because particular taxa are absent or rare at some sites. For example, the Bass River model generated low reconstructions because no modern samples were collected above 100 SWLI (MHHW), whereas the Rutgers Marine Station model included 19 samples above 100 SWLI (MHHW) and produced

correspondingly high reconstructions. For both of these models some important species in BB1 were poorly represented in the training set. *Jadammina macrescens* was rare at Bass River, while at Rutgers Marine Station *Miliammina fusca* was rare, causing that model to fail in reconstructing the lower elevations near the top of the core that is a feature of the other models (Fig. 7A). The development and application of local, sub-regional and regional transfer functions illustrates that transfer functions trained on multiple modern sites that span a range of physiographic conditions generate less extreme reconstructions of elevation.

In dissimilarity analysis, core samples less frequently exceeded the 20% threshold when the regional and sub-regional transfer functions were applied compared with the local models (Fig. 7B). Five samples in the regional model exceeded the 20th percentile threshold established from dissimilarity in the training set. The sub-regional Great Bay model had the second fewest samples exceeding the threshold (nine), while the other models included 14–16 samples with minimum dissimilarity exceeding the same threshold. Regional transfer functions are less likely to return no analog results than models based on individual sites and demonstrate that it was necessary to include sites spanning a range of physiographic conditions (in particular the inclusion of sites with strong fluvial influence) to reconstruct sea level in BB1.

Comparison of sea-level reconstructions with instrumental records

Tide-gauge records of RSL change during the historical period present an opportunity to test if the transfer function can accurately reconstruct RSL. A meaningful comparison between reconstructed and observed RSL is most robust when using a site with small tidal range such as Barnegat Bay (great diurnal range of 0.17 m) to ensure that the reconstructed vertical uncertainty is not larger than the historical rise in RSL. The Battery tide gauge in New York City is approximately 75 km north of Barnegat Bay and has an almost continuous history of sea-level measurements since 1854 AD, providing a long time series for comparison. This combination of a sea-level reconstruction from a microtidal site and a long, near-continuous tide-gauge record is ideal for testing the accuracy of reconstructions from the transfer function.

We reconstructed RSL (Fig. 8B) by subtracting transfer function (WA-inv, full New Jersey dataset) estimates of elevation from measured sample altitude (derived from depth in core). Sample age estimates were from an age–depth model developed using lead concentrations, stable lead isotopes and peak ^{137}Cs concentration (Fig. 8A; Kemp *et al.*, 2012b). The reconstruction was restricted to the upper 41 cm of BB1 (the lowest dated level) and therefore spans the period since approximately 1820 AD. It shows approximately 0.45 m of rise since the start of the 19th century, with sample-specific vertical uncertainties of about ± 2.5 cm because of the small tidal range in Barnegat Bay. This great diurnal tidal range is unusual and most locations on the US mid-Atlantic coast have a considerably larger range. While the uncertainty estimated by the transfer function at Barnegat Bay does not offer considerable improvement over lithostratigraphic interpretation of a high-marsh peat, in most cases the vertical error in reconstructing RSL using the transfer function would be less than classifying samples as being from high-marsh or low-marsh environments.

There is close agreement between the reconstructed and instrumental records of RSL change as evidenced by the tide-gauge data largely falling within the uncertainties of reconstructed RSL (Fig. 8B). There is no apparent divergence between the instrumental and reconstructed records caused by samples lacking close modern analogs or with a poor fit to tidal elevation. This agreement demonstrates that the expanded dataset of New Jersey salt-marsh foraminifera and WA-inv transfer function reliably reconstructed RSL from high-marsh sediment. However, Juggins and Birks (2012) caution that this agreement validates the reconstruction, but not necessarily other reconstructions that use the same transfer function.

Conclusions

We described the distribution of modern foraminifera from 12 salt marshes in southern New Jersey to provide an expanded training set for reconstructing Holocene sea level (index points and continuous records) in the US mid-Atlantic region from a range of sedimentary environments. Seven groups of foraminifera were identified, including a characteristic low-salinity group, four high- and transitional-marsh groups indicating variability among sites, and two widespread low-marsh groups. The presence of these groups and comparisons of local, sub-regional and regional transfer functions indicated the necessity of compiling training sets from multiple sites to capture spatial variability in assemblages caused by factors such as salinity.

A weighted-averaging transfer function trained on the regional dataset was developed for reconstructing sea level with an estimated precision of ± 14% of great diurnal tidal

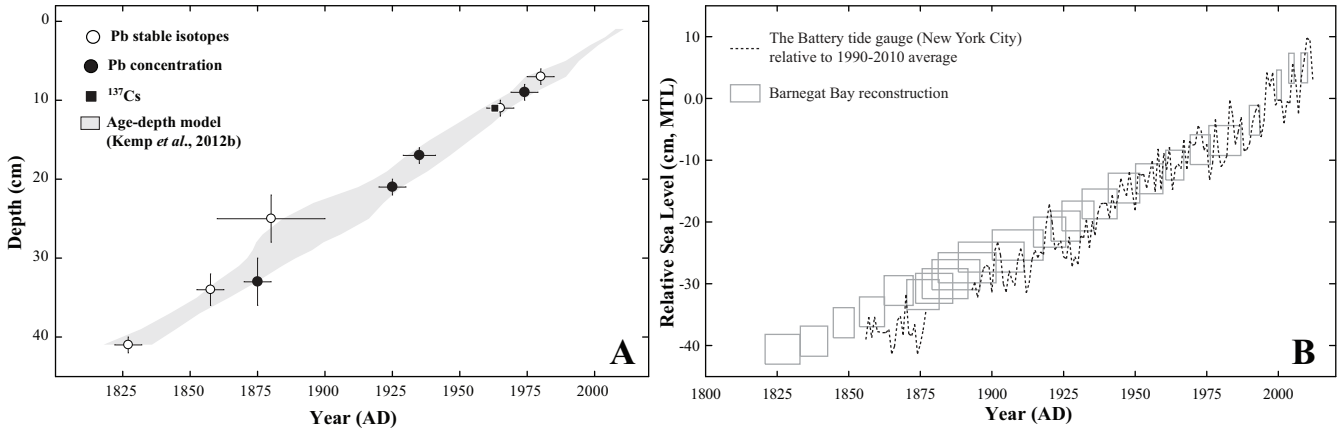

Figure 8. Relative sea level (RSL) reconstruction from core BB1. (A) Age–depth model based upon lead concentrations, stable lead isotopes and ^{137}Cs activity. Reproduced from Kemp *et al.* (2012b). (B) RSL reconstruction with data points represented by boxes that include the 1σ confidence interval from the age–depth model and sample-specific uncertainties of reconstructed elevation from the transfer function. Tide-gauge data from The Battery in New York City are presented relative to the 1990–2010 AD average and are uncorrected for glacio-isostatic subsidence.

range. We followed a stepwise series of tests to evaluate transfer function performance and application. This suite of techniques includes several recently developed approaches that have been under utilized in sea-level research. Analysis of the training set revealed that:

1. Clustered training sets of salt-marsh foraminifera appear robust to the influence of spatial autocorrelation introduced by sampling along transects, as evidenced by the similarity of model performance estimated by LOO and of LOSO cross validation.
2. Uneven sampling of the environmental gradient, which is common in sea-level research, causes over optimistic estimates of model performance. Segment-wise analysis demonstrated that transfer function performance varies across the gradient of elevation and is best in heavily sampled segments.

The transfer function was applied to a dated core of salt-marsh sediment to reconstruct relative sea-level changes in Barnegat Bay, New Jersey, during the last ~200 years. Evaluation of the reconstruction showed that:

3. A diverse, multi-site training was necessary to provide modern analogs for sub-regional variability in assemblages and physiographic conditions.
4. Samples from core BB1 passively projected into NMDS of the training set fell within the scatter of modern samples and indicated that core samples are adequately represented by modern equivalents. The trajectory of core samples agreed with reconstructed values of elevation.
5. Goodness of fit identified core samples that were not well fitted to elevation.
6. The reconstruction failed a test of statistical significance because unbroken sequences of high salt-marsh sediment do not provide sufficient downcore variability for a transfer function trained on modern data to outperform alternative models trained on random environmental data. This is an important caveat for future application of this test in sea-level research.
7. Reconstructed sea level since the mid-19th century agreed with instrumental measurements from The Battery tide gauge in New York City, confirming the utility of the expanded training set dataset and transfer function for reconstructing sea level in the US mid-Atlantic region.

Supporting information

Additional supporting information can be found in the online version of this article:

Fig. S1. Transect sampled at Bass River.
Fig. S2. Transect sampled at Cape May Courthouse.
Fig. S3. Transect sampled at Cedar River.
Fig. S4. Transect sampled at Cold Spring Harbor.
Fig. S5. Transect sampled at Jayne Drive.
Fig. S6. Three transects sampled at Rutgers Marine Station.
Fig. S7. Transect sampled at Sea Breeze.
Fig. S8. Transect sampled at Steelmanville.
Fig. S9. Transect sampled at Stephen's Creek.
Fig. S10. Transect sampled at Tuckahoe.

Table S1. Foraminiferal data and sample elevations, distinguishing samples newly described in this study from those in Kemp *et al.* (2012a).

Acknowledgements. A.C.K. thanks a Yale University Climate and Energy Institute post-doctoral fellowship. Funding for this study was provided by NICRR grant DE-FC02-06ER64298, National Science Foundation award EAR-0951686 and NOAA grant NA11OAR4310101. Norwegian Research Council FriMedBio project palaeoDrivers (213607) helped to support R.J.T. We thank Rutgers University (Roland Hagan) for helping us to access sites and the benchmark at their marine station. Support for fieldwork came in part from the Earthwatch Institute Student Challenge Award Program. We thank Simon Engelhart, Nicole Khan, Carol Wilson and Candace Grand-Pre for helping with sample collection. We thank Ian Shennan and an anonymous reviewer for their comments and suggestions. This is a contribution to PALSEA and IGCP Project 588 'Preparing for Coastal Change'.

Abbreviations. CCA, canonical correspondence analysis; DCA, detrended correspondence analysis; LOO, leave one out; LOSO, leave one site out; MHHW, mean higher high water; MHW, mean high water; MLLW, mean lower low water; MTL, mean tide level; NGS, National Geodetic Survey; NMDS, non-metric multi-dimensional scaling; PAM, Partitioning Around Medoids; PLS, partial least squares; RMSEP, root mean squared error of prediction; RSL, relative sea level; SWLI, standardized water level index; WA, weighted-averaging

References

Anderson PM, Bartlein PJ, Brubaker LB, *et al.* 1989. Modern analogues of Late-Quaternary pollen spectra from the western interior of North America. *Journal of Biogeography* **16**: 573–596.

Birks HJB. 1995. Quantitative palaeoenvironmental reconstructions. In *Statistical Modelling of Quaternary Science Data*, Maddy D, Brew JS (eds). Quaternary Research Association: Cambridge; 161–254.

Callard SL, Gehrels WR, Morrison BV, *et al.* 2011. Suitability of salt-marsh foraminifera as proxy indicators of sea level in Tasmania. *Marine Micropaleontology* **79**: 121–131.

Daddario JJ. 1961. A lagoon deposit profile near Atlantic City New Jersey. *Bulletin of the New Jersey Academy of Science* **6**: 7–14.

de Rijk S. 1995. *Agglutinated Foraminifera as Indicators of Salt Marsh Development in Relation to Late Holocene Sea Level Rise*. Febo: Utrecht.

Dixit S, Cumming B, Birks HJB, *et al.* 1993. Diatom assemblages from Adirondack lakes (New York, USA) and the development of inference models for retrospective environmental assessment. *Journal of Paleolimnology* **8**: 27–47.

Edwards RJ, Wright AJ, van de Plassche O. 2004. Surface distributions of salt-marsh foraminifera from Connecticut, USA: modern analogues for high-resolution sea level studies. *Marine Micropaleontology* **51**: 1–21.

Eleuterius L. 1976. The distribution of *Juncus roemerianus* in the salt marshes of North America. *Chesapeake Science* **17**: 289–292.

Ellison RL, Nichols MM. 1976. Modern and Holocene foraminifera in the Chesapeake Bay region. In *International Symposium on Benthonic Foraminifera of Continental Margins, Part A, Ecology and Biology: Halifax, Nova Scotia, Canada, Maritime Sediments Special Publication*, Schaefer CT, Pelletier BR (eds). Atlantic Geoscience Society Fredericton, Canada. 131–151.

Ellison RL, Nichols MM, Hughes J. 1965. Distribution of Recent Foraminifera in the Rappahannock River Estuary, Special Report in Marine Science and Ocean Engineering. Virginia Institute of Marine Science: Gloucester Point, VA; 1–35.

Fatela F, Taborda R. 2002. Confidence limits of species proportions in microfossil assemblages. *Marine Micropaleontology* **45**: 169–174.

Ferland MA. 1990. Holocene depositional history of the southern New Jersey barrier and back barrier regions. US Army Corps of Engineers.

Figueira BO, Grenfell HR, Hayward B, *et al.* 2012. Comparison of Rose Bengal and CellTracker Green staining for identification of live salt-marsh foraminifera. *Journal of Foraminiferal Research* **42**: 206–215.

Gehrels WR. 1994. Determining relative sea-level change from salt-marsh foraminifera and plant zones on the coast of Maine, U.S.A. *Journal of Coastal Research* **10**: 990–1009.

Gehrels WR, Belknap DF, Black S, *et al.* 2002. Rapid sea-level rise in the Gulf of Maine, USA, since AD 1800. *Holocene* **12**: 383–389.

Gehrels WR, Kirby JR, Prokoph A, *et al.* 2005. Onset of recent rapid sea-level rise in the western Atlantic Ocean. *Quaternary Science Reviews* **24**: 2083–2100.

Goldstein ST, Watkins GT, Kuhn RM. 1995. Microhabitats of salt marsh foraminifera: St Catherine's Island, Georgia, USA. *Marine Micropaleontology* **26**: 17–29.

Hayward BW, Grenfell HR, Scott DB. 1999. Tidal range of marsh foraminifera for determining former sea-level heights in New Zealand. *New Zealand Journal of Geology and Geophysics* **42**: 395–413.

Hill MO, Gauch HG. 1980. Detrended correspondence analysis: an improved ordination technique. *Plant Ecology* **42**: 47–58.

Hippensteel SP, Martin RE, Nikitina D, *et al*. 2000. The formation of Holocene marsh foraminiferal assemblages, middle Atlantic coast, U.S.A.: implications for Holocene sea-level change. *Journal of Foraminiferal Research* **30**: 272–293.

Horton BP. 1999. The distribution of contemporary intertidal foraminifera at Cowpen Marsh, Tees Estuary, UK: implications for studies of Holocene sea-level changes. *Palaeogeography, Palaeoclimatology, Palaeoecology* **149**: 127–149.

Horton BP, Edwards RJ. 2005. The application of local and regional transfer functions to the reconstruction of Holocene sea levels, north Norfolk, England. *Holocene* **15**: 216–228.

Horton BP, Edwards RJ. 2006. Quantifying Holocene sea-level change using intertidal foraminifera: lessons from the British Isles. *Cushman Foundation for Foraminiferal Research Special Publication* **40**: 97.

Jackson ST, Williams JW. 2004. Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences* **32**: 495–537.

Juggins S. 2009. Rioja: an R package for the analysis of quaternary science data, version 0.5-6.

Juggins S, Birks HJB. 2012. Quantitative environmental reconstructions from biological data. In *Data Handling and Numerical Techniques*, Birks HJB, Lotter AF, Juggins S, Smol JP (eds). Springer: Berlin; 431–494.

Kaufman L, Rousseeuw PJ. 1990. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley Interscience: New York.

Kemp AC, Horton BP, Culver SJ. 2009a. Distribution of modern salt-marsh foraminifera in the Albemarle–Pamlico estuarine system of North Carolina, USA: implications for sea-level research. *Marine Micropaleontology* **72**: 222–238.

Kemp AC, Horton BP, Culver SJ, *et al*. 2009b. Timing and magnitude of recent accelerated sea-level rise (North Carolina, United States). *Geology* **37**: 1035–1038.

Kemp AC, Horton BP, Vann DR, *et al*. 2012a. Quantitative vertical zonation of salt-marsh foraminifera for reconstructing former sea level; an example from New Jersey, USA. *Quaternary Science Reviews* **54**: 26–39.

Kemp AC, Sommerfield CK, Vane CH, *et al*. 2012b. Use of lead isotopes for developing chronologies in recent salt-marsh sediments. *Quaternary Geochronology* **12**: 40–49.

Kennish MJ. 2001. Coastal salt marsh systems in the U.S.: a review of anthropogenic impacts. *Journal of Coastal Research* **17**: 731–748.

Murray JW, Bowser SS. 2000. Mortality, protoplasm decay rate, and reliability of staining techniques to recognize 'living' foraminifera: a review. *Journal of Foraminiferal Research* **30**: 66–70.

Oksanen J, Guillaume Blanchet F, Kindt R, *et al*. 2012. Vegan Community Ecology Package. R package version 2.0-5. http://CRAN.R-project.org/package=vegan

Overpeck JT, Webb T, Prentice IC. 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research* **23**: 87–108.

Payne RJ, Telford RJ, Blackford JJ, *et al*. 2012. Testing peatland testate amoeba transfer functions: appropriate methods for clustered training-sets. *Holocene* **22**: 819–825.

Rousseeuw P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster techniques. *Journal of Computational and Applied Mathematics* **20**: 53–65.

Scott DB, Medioli FS. 1978. Vertical zonations of marsh foraminifera as accurate indicators of former sea levels. *Nature* **272**: 528–531.

Scott DB, Medioli FS, Schaefer CT. 2001. *Monitoring Coastal Environments*. Cambridge University Press: Cambridge.

Scott DB, Williamson MA, Duffett TE. 1981. Marsh foraminifera of Prince Edward Island: their recent distribution and application for former sea-level studies. *Maritime Sediments and Atlantic Geology* **17**: 98–129.

Scott DK, Leckie RM. 1990. Foraminiferal zonation of Great Sippewissett salt marsh (Falmouth, Massachuesetts). *Journal of Foraminiferal Research* **20**: 248–266.

Shennan I, Horton B. 2002. Holocene land- and sea-level changes in Great Britain. *Journal of Quaternary Science* **17**: 511–526.

Simpson GL. 2007. Analogue methods in palaeoecology: using the analogue package. *Journal of Statistical Software* **22**: 1–29.

Simpson GL. 2012. Analogue methods. In *Data Handling and Numerical Techniques*, Birks HJB, Lotter AF, Juggins S, Smol JP (eds). Springer: Berlin; 495–522.

Simpson GL, Hall RI. 2012. Human impacts: applications of numerical methods to evaluate surface water acidification and eutrophication. In *Data Handling and Numerical Techniques*, Birks HJB, Lotter AF, Juggins S, Smol JP (eds). Springer: Berlin; 579–614.

Smith DA, Scott DB, Medioli FS. 1984. Marsh foraminifera in the Bay of Fundy: modern distribution and application to sea-level determinations. *Maritime Sediments and Atlantic Geology* **20**: 127–142.

Stuckey IH, Gould LL. 2000. *Coastal Plants from Cape Cod to Cape Canaveral*. University of North Carolina: Chapel Hill.

R Core Development Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, [URL: http://www.R-project.org/]

Telford RJ. 2012. palaeoSig: Significance tests for palaeoenvironmental reconstructions. R package version 1.1-1. http://CRAN.R-project.org/package=palaeoSig

Telford RJ, Birks HJB. 2011a. Effect of uneven sampling along an environmental gradient on transfer-function performance. *Journal of Paleolimnology* **46**: 99–106.

Telford RJ, Birks HJB. 2011b. A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages. *Quaternary Science Reviews* **30**: 1272–1278.

Tiner RW. 1985. *Wetlands of New Jersey. National Wetlands Inventory*. United States Fish and Wildlife Service: Newton Corner, MA.

Törnqvist TE, Gonzalez JL, Newsom LA, *et al*. 2004. Deciphering Holocene sea-level history on the US Gulf Coast: a high-resolution record from the Mississippi Delta. *Geological Society of America Bulletin* **116**: 1026–1039.

Watcham EP, Shennan I, Barlow NLM. 2013. Scale considerations in using diatoms as indicators of sea-level change: lessons from Alaska. *Journal of Quaternary Science* **28**: 165–179.

Wright AJ, Edwards RJ, van de Plassche O. 2011. Reassessing transfer-function performance in sea-level reconstruction based on benthic salt-marsh foraminifera from the Atlantic coast of NE North America. *Marine Micropaleontology* **81**: 43–62.