Vol. 316: 269–283, 2006 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 3

OPEN ACCESS

Modeling habitat distribution from organism occurrences and environmental data: case study using anemonefishes and their sea anemone hosts

J. M. Guinotte[1,2,5,*], J. D. Bartley[1], A. Iqbal[1], D. G. Fautin[1,3], R. W. Buddemeier[1,4]

[1]Kansas Geological Survey, 1930 Constant Avenue, Lawrence, Kansas 66047, USA
[2]School of Tropical Environmental Studies and Geography, James Cook University, Townsville, Queensland 4810, Australia
[3]Department of Ecology and Evolutionary Biology, and Natural History Museum and Biodiversity Research Center, and
[4]Department of Geography, University of Kansas, Lawrence, Kansas 66045, USA

[5]*Present address:* Marine Conservation Biology Institute, 2122 112th Ave NE, Suite B-300, Bellevue, Washington 98004, USA

ABSTRACT: We demonstrate the KGSMapper (Kansas Geological Survey Mapper), a straightforward, web-based biogeographic tool that uses environmental conditions of places where members of a taxon are known to occur to find other places containing suitable habitat for them. Using occurrence data for anemonefishes or their host sea anemones, and data for environmental parameters, we generated maps of suitable habitat for the organisms. The fact that the fishes are obligate symbionts of the anemones allowed us to validate the KGSMapper output: we were able to compare the inferred occurrence of the organism to that of the actual occurrence of its symbiont. Characterizing suitable habitat for these organisms in the Indo-West Pacific, the region where they naturally occur, can be used to guide conservation efforts, field work, etc.; defining suitable habitat for them in the Atlantic and eastern Pacific is relevant to identifying areas vulnerable to biological invasions. We advocate distinguishing between these 2 sorts of model output, terming the former maps of realized habitat and the latter maps of potential habitat. Creation of a niche model requires adding biotic data to the environmental data used for habitat maps: we included data on fish occurrences to infer anemone distribution and vice versa. Altering the selection of environmental variables allowed us to investigate which variables may exert the most influence on organism distribution. Adding variables does not necessarily improve precision of the model output. KGSMapper output distinguishes areas that fall within 1 standard deviation (SD) of the mean environmental variable values for places where members of the taxon occur, within 2 SD, and within the entire range of values; eliminating outliers or data known to be imprecise or inaccurate improved output precision mainly in the 2 SD range and beyond. Thus, KGSMapper is robust in the face of questionable data, offering the user a way to recognize and clean such data. It also functions well with sparse datasets. These features make it useful for biogeographic meta-analyses with the diverse, distributed datasets that are typical for marine organisms lacking direct commercial value.

KEY WORDS: Biogeography · Clownfish · Ecological niche · Range · GIS

INTRODUCTION

Biogeographic information, whether about taxa, guilds, or groups of associated organisms, is fundamental to human use and understanding of the environment. Electronic resources are rapidly enhancing the volume and diversity of information that can be brought to bear on problems such as the identification and protection of biodiversity, actual or potential invasive species, and diagnosis and prediction of the effects of climate change (e.g. Soberón & Peterson 2004, and references cited therein). As distributed biogeographical and environmental datasets become more available and better integrated, the need for

*Email: john@mcbi.org

simple but flexible tools to exploit them will grow, and the outputs will be extended to more uses. It is vital to understand the nature of the data and the uses to which tools and their outputs can appropriately be put. In this proof of concept study, we explore some characteristics of mapping tools and their output.

The most fundamental biogeographic data concern organism distribution. One convention for depicting distribution is plotting known occurrences as dots (points) on a map. With rare exceptions, these dots are not intended to represent the entire distribution of the taxon in question. A range map is commonly derived from such a dot map, the outermost bounds of a polygon which represents the taxon's distribution connecting the most peripheral dots of the taxon's known occurrence or the points at which organism density falls below a particular threshold (e.g. MacArthur 1972). Such a polygon commonly overestimates the taxon's range. In the marine realm, the range of a shallow-water species occurring throughout the tropical Pacific would cover the entire tropical Pacific Ocean, including the deep water between islands (as in e.g. Fautin & Allen 1992). A more ecologically realistic approach is to correlate actual occurrences with physical, chemical, or biological data (e.g. MacArthur 1972), so, for example, a shallow marine species would be depicted as occurring only around land masses or on banks and shoals.

Thus, more than a collection of geographic coordinates, a range is a manifestation of characteristics of the habitat (biotic and abiotic) that limit or support the organism of interest. A range is inherently a large-scale concept based on observed occurrences; however, range analysis does not necessarily predict organism presence at any specific point. We illustrate some alternative approaches to modeling and understanding habitat distributions for marine organisms by analyzing data from 3 databases with the KGSMapper (Kansas Geological Survey Mapper), an application for interactive analysis of georeferenced occurrence records of marine organisms with gridded environmental data. It is one of a class of electronic tools that, by making it progressively easier to develop correlative analyses from occurrence and environmental data, are rapidly supplanting traditional approaches to interpretive mapping, which tend to be tedious and difficult to replicate. We discuss some issues in evaluating these sorts of analyses. Computer tools and databases cannot substitute completely for knowledge and judgment, however, and the tool we discuss provides ways in which the investigator can interact with and modify the datasets used in order to explore or test hypotheses and tune the nature of the output to the question of interest, rather than simply generating a 'hard-wired' occurrence prediction.

Applications such as WhyWhere (http://biodi.sdsc.edu/ww_home.html) and GARP (www.lifemapper.org/desktopgarp/, http://biodi.sdsc.edu/Doc/GARP/Manual/manual.html), which offer computationally sophisticated approaches to associating environmental and occurrence data (e.g. genetic algorithms), provide the user limited control over datasets and particularly data processing. Tools such as BIOCLIM (http://cres.anu.edu.au/outputs/anuclim/doc/bioclim.html) are confined to or work best in terrestrial habitats. No single approach will be optimal for all questions, or for the needs of all potential users (Fielding & Bell 1997; compare assessments of GARP by Beauvais et al. 2004, Drake & Bossenbroek 2004); in making a choice, consideration must be given to types, scale, quality, and quantity of data available, questions to be addressed, and verifiability of the product (e.g. Fielding & Bell 1997, Manel et al. 2001, Beauvais et al. 2004, Drake & Bossenbroek 2004).

We investigated the issues listed below by generating probabilistic maps of potential habitat occurrence, depicting large-scale areas suitable for survival of these organisms, not organism presence–absence inferences. We used the KGSMapper to analyze the occurrence of habitat suitable for anemonefishes (which may be referred to as clownfishes) and their host sea anemones. The fact that the fishes are obligate symbionts of the anemones (although individual anemones may be found without anemonefish) make this an ideal test case for validating model output: we did not have to go to the field to determine if the organism occurs where we inferred it would, but could compare the inferred occurrence of suitable habitat for the organism to that of the actual occurrence of its symbiont. It is also ideal as a test case in being typical of datasets available for non-fisheries marine species. We discuss model outputs, often termed range, habitat, and niche predictions. Such outputs are commonly used within the natural range of a taxon to guide field work, conservation efforts, etc., and outside the natural range to identify areas vulnerable to biological invasion.

1. Sampling issues. Datasets for a diversity of environmental parameters may be available. The outcome of occurrence predictions or range inferences will be affected by which variables are selected, and how. True niche models (e.g. Peterson 2001, Raxworthy et al. 2003, Soberón & Peterson 2004) must include parameters of the biotic environment beyond strictly habitat characteristics.

2. Data quality. Models must be robust in the presence of questionable or erroneous data points. Particularly for meta-analyses, which use datasets from a variety of sources, the data are likely to vary in accuracy, precision, and resolution, making it unlikely that data quality will uniformly meet the desired standards of

any individual user or application. Therefore, tools are needed for evaluating and/or cleaning datasets when there is a basis for doing so, and the criteria for these actions must be clear.

3. Data quantity. The effect of the number of data on inferences is vital to recognize (e.g. Stockwell & Peterson 2002). A common use of modeling is to infer the biogeographic range of a taxon for which the documented occurrence records almost certainly fall far short of encompassing the actual range. This situation is extremely common for marine invertebrates, particularly for analyses at the species level, but is by no means restricted to them (e.g. Beauvais et al. 2004). Models can provide insight into the areas in which data will be most economical or efficient to sample in order to verify the true extent of the range.

4. Validating or testing results. Assessing predictions or inferences is a desideratum (Fielding & Bell 1997) in this, as in any hypothesis-testing. The end-members on the predictive continuum are a broad-brush approach that minimizes errors of omission and a focused approach that minimizes errors of commission (Fielding & Bell 1997, Anderson et al. 2003). In dealing with the continuum of quality and/or extent inherent in habitat assessment at large spatial scales, omission and commission are not binary no–yes choices, as is typically the case in dealing with presence–absence of organisms; different tests and criteria are called for.

5. Identifying controlling factors. Drake & Bossenbroek (2004, p. 939–940) appealed to scientists to 'develop methods to identify the factors that causally determine species range, and not simply make predictions based on correlations.' Characteristics of a taxon's range or physiology may suggest that particular environmental parameters control its occurrence.

DATA AND METHODS

Data sources and organisms. The organism distribution data for both taxa are georeferenced point occurrences; the third dataset includes gridded coverages of environmental parameters. Having come from 3 proximate providers, all of which compiled data from multiple ultimate sources, our data are unlikely to be homogeneous in quality and scale.

Anemonefishes, which are widespread in the tropical and subtropical Indo–West Pacific but are absent from the eastern Pacific and the Atlantic, occur in nature only with sea anemones of 10 species belonging to 5 genera in 3 families; the fish population is limited by the number of suitable hosts (Fautin & Allen 1992). Anemonefishes belong to 2 genera (*Amphiprion*, with 25 species, and *Premnas*, with 1) in a single subfamily, and vary in host specificity, some associating with only 1 species of host, but most occurring with multiple hosts (Fautin & Allen 1992). Because all host anemones possess photosymbionts, they and their fish symbionts occur only in shallow water (Dunn 1981), typically in waters less than 100 m deep. The distribution of these animals, therefore, is constrained both environmentally and biologically.

In a first approximation, the 10 species of anemones are ecologically similar, and the 26 species of fishes are likewise similar; this allows us to use as our units of taxonomic analysis all host anemones and all anemonefishes. We extracted occurrence data for the anemones from the online resource 'Biogeoinformatics of Hexacorals' (www.kgs.ku.edu/Hexacoral; hereafter referred to as 'Hexacoral'). In the biological database of Hexacoral, which was assembled from the published literature, all names used to refer to a single species are linked, and names that have been applied to more than 1 species are distinguished. Anemonefish occurrence data were downloaded from FishBase (www.fishbase.org), which has been assembled from published records, museum catalogs, and other sources.

The environmental data, also served from Hexacoral, were assembled from public-domain datasets (sources identified in the metadata associated with each dataset) that are global in coverage. Data were gridded in a register at 0.5° resolution (~55 km per side at the equator), which is a typical resolution for global environmental datasets. Datasets with native resolutions other than 0.5° were sampled or aggregated to conform to the grid; for a variable with a native resolution finer than 0.5° (such as the 2′ ETOPO2 bathymetry), within-cell variability and extremes were calculated. Most values are annual or monthly averages. Of the >200 datasets in Hexacoral, 13 especially relevant to anemonefishes and their hosts are currently available for use with KGSMapper; future versions will make the other datasets accessible. In addition to limitations imposed by the size of grid cells, a significant caveat is that the marine datasets used to generate many of the variables typically fail to represent much of the temporal and spatial variability in nearshore environments.

Tools and analytical procedures. KGSMapper is an interactive web-based mapping tool that permits a user to create maps of inferred distribution in a straightforward manner. The basic calculations can be done in a spreadsheet, although much of the power of KGSMapper derives from its ability to display and manipulate the data in a Geographical Information System (GIS) environment. Its flexibility allows a user to select approaches relevant to the goals of the study and to apply expert judgment in editing datasets. It currently uses a tightly integrated environmental data-

base and front end (Oracle 9i RDBMS with Cold Fusion) with, on the server side, ArcIMS web-mapping software (www.esri.com/software/arcgis/arcims/index.html). Occurrence records are plotted in real time on a map through an XML-coded data structure based on the Ocean Biogeographic Information System (OBIS) schema, an extension of Darwin Core 2 (http://iobis.org/obis/obis.xsd). KGSMapper and its associated environmental data are freely available; it is operational through the Hexacoral website (above), the OBIS website (www.iobis.org), and those of some OBIS partners (e.g. CephBase: www.cephbase.org; FishBase: above).

In our analyses, locality records are the 0.5° grid cells containing organism occurrences. Thus, the number of occurrences may not equal the number of locality records in the dataset. Cells with 1 or multiple occurrences are indistinguishable in our analyses — a single occurrence serves to qualify a cell and its environmental variable values as habitat. Conversely, for an occurrence falling on a cell boundary between 2 or among 4 cells, all cells are included in the analysis.

The version of KGSMapper used for this study (http://hercules.kgs.ku.edu/website/specimen_mapper) currently interacts only with the data discussed here. Table 1 summarizes the features of the KGSMapper. Fig. 1 shows the KGSMapper web page; its functions and features are described below, and in the figure caption. KGSMapper plots organism occurrences and provides summary values of 52 environmental variables for all cells in which there is at least 1 occurrence record. Our tests were constrained by the variables available from the main database, and by inherent resolution limitations of working at global scales with primarily marine parameters. These are practical matters — neither is constrained in theory.

Inferences of where suitable habitat occurs for members of the taxon are based on the environment of places where they are known to occur. The user selects the variables by checking the relevant boxes under 'Use to Find Similar Areas' (Fig. 1, Panel f). When the user selects 'Update Map,' KGSMapper builds and executes a query to find the 0.5° cells having all values within 1 standard deviation (SD) of the means of the environmental variables at the occurrence locations, those within 2 SD, and those within the total value range for all selected variables. The results, displayed as an interactive map (Fig. 1), are also available as tabulated statistics (by clicking a link in Fig. 1, Panel c). For 0.5° cells to be classed as within 1 SD, depicted as dull red on the map, all the selected variables must be within 1 SD of the mean of the values of the same variable in cells containing occurrence records. Orange signifies cells in which the value for all selected variables falls within 2 SD of the mean of the values for the selected variable(s), but at least 1 falls beyond 1 SD, and yellow signifies cells beyond 2 SD to the full range of the values known ('outliers'). This probabilistic approach is appropriate in dealing with habitat, which is a continuum from favorable to marginal. It also allows a user to focus attention where habitat or data are optimal, by recalculating a map that eliminates those original cells that have values in the outlier region or beyond 1 SD. This is done by selecting, respectively, the 'Remove All Cells Outside 2 Std. Deviation ranges from cart' or 'Remove All Cells Outside 1 Std. Deviation ranges from cart' options that appear at the bottom of the statistics pop-up page.

Table 1. Features of the KGSMapper tool used for analyses reported here. Last 3 points refer to features still under development

Features
1. Dynamic mapping of selected occurrences
2. Selectable map background
3. Data point identification with link to source database
4. Short list of selectable environmental variables
5. Viewable environmental variable metadata
6. Viewable distribution histogram of individual environmental variables for selected occurrences
7. Correlation matrix of environmental variables for selected occurrences
8. Pairwise scatterplots of environmental variables for selected occurrences
9. Map zoom controls region of analysis, occurrences selected
10. Environmental data table reflects selected locations
11. Range localities classified within 1 SD, 2 SD, and total range of selected environmental variables
12. Downloadable file of occurrences
13. Downloadable shape file of inferred areas
14. Downloadable table of relationships among occurrences, sample cells, and inferred areas
15. Downloadable table of cell IDs for all areas in analysis
16. Eliminating individual records from working dataset
17. Limiting maximum and/or minimum values for environmental variables
18. Comparing or combining 2 datasets
19. Ability to use a random 50% of locations, tested with others
20. User can save and return to a modified dataset
21. User can upload an independent dataset for analysis

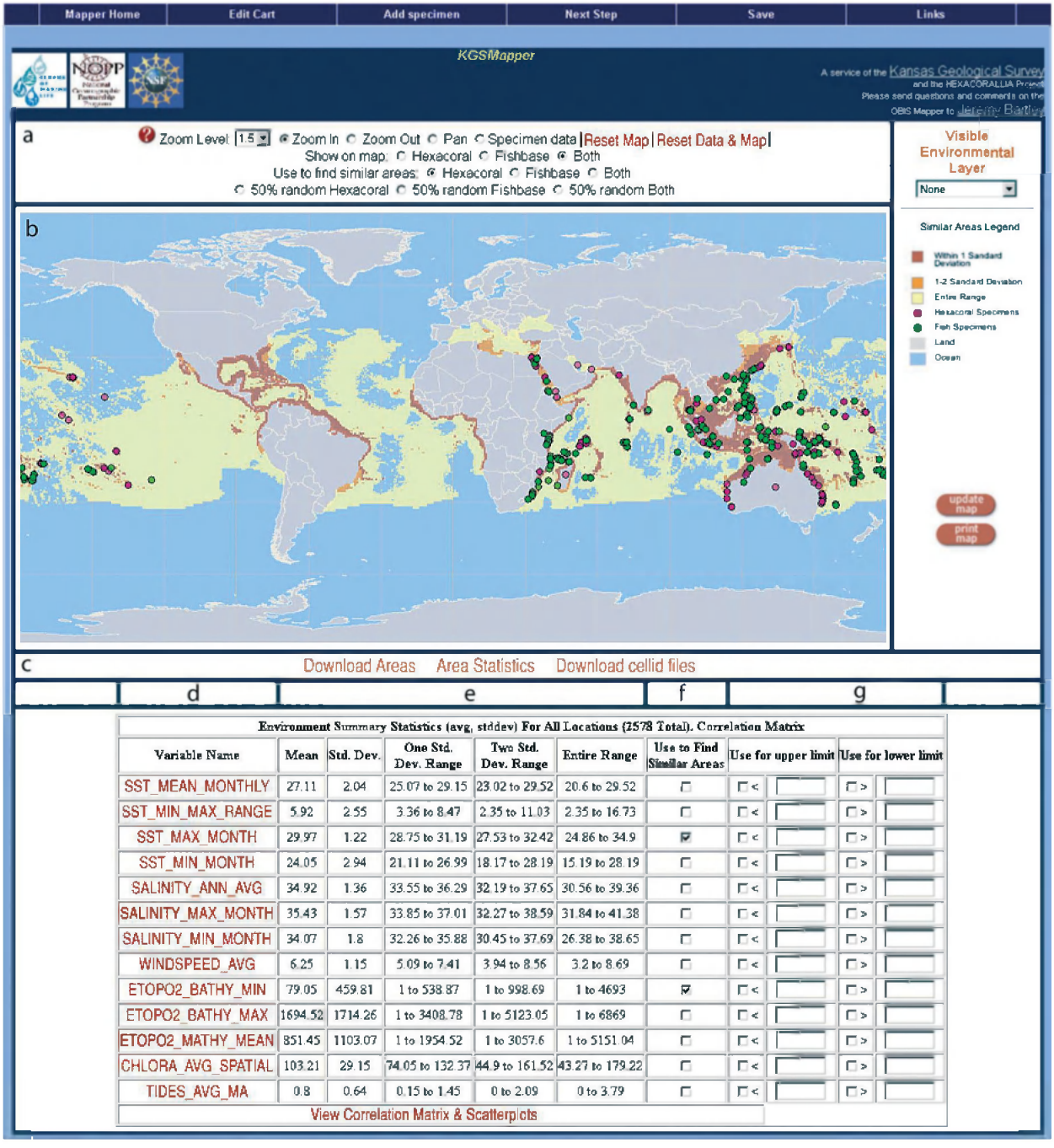

Fig. 1. KGSMapper page: the inferred range displayed is based on anemone distributions, maximum monthly sea-surface temperature (SST), and minimum depth value for the grid cells containing anemone occurrences (checked in boxes below map). (a) Zoom and pan controls on top line select region and scale. Clicking a point with 'Specimen data' activated produces a pop-up window containing species name(s) and coordinates, values for environmental variables in each cell in the selected area, summary of environmental statistics for all cells containing an occurrence record, and the option of removing the point from the analysis. Second line selects sample points displayed. Third line selects sample set of cells used. Fourth line randomly selects ~50% of one or both datasets to make a range inference to be tested with the remaining cells. (b) Map shows both datasets with localities distinguished by color of points (purple: sea anemones; green: anemonefishes) and inferred distribution of suitable habitat based on the selected environmental variables (below). Cells in areas colored dull red have values for all variables used for the inference within 1 SD of their means in the record-containing cells, orange is for cells between 1 and 2 SD, and yellow is for the rest of the total range. (c) Links below map provide a download of shapefiles for the areas, a table of statistics of occurrences in both datasets relative to the cells in each range class, or a set of tables of the grid cell identifiers for the cells in each SD category by record contents. (d) Link from the variable name brings up a histogram showing distribution of values and statistics for variable values from the selected locations. Environmental parameters are SST (monthly mean, maximum, minimum, and range), salinity (annual averages, and monthly minimum and maximum), average windspeed, depth (based on ETOPO2 bathymetry: minimum, maximum, and mean), average chlorophyll *a* concentration, and average tidal amplitude. (e) Statistics for each variable reflect the dataset selected by the map display (Panel b). (f) Check boxes for selecting variables with which to 'update map' and infer ranges. (g) Check boxes for entering minimum and/or maximum values to restrict the locations selected. Bottom line: link displays a correlation matrix (Table 2) showing linear regression coefficients for each pair of environmental variables, based on values selected in the map display. Values for the correlation coefficients in the matrix cells are linked dynamically to scatterplots of the selected values of each pair of variables. Quality of Figs. 1 to 4 corresponds to that of the images on the computer screen

The menu bar at the top of the page (Fig. 1) provides links to other parts of the site and 2 editing functions. 'Add Specimen' permits a user to augment the occurrence dataset; the 'Edit Cart' link allows a user to eliminate entries from the list of occurrences. The user can also review and edit individual location records with 'Specimen Data' (Fig. 1, Area a). The link 'Next Step' takes the user to the menu of all 200+ environmental datasets in Hexacoral—these currently do not otherwise interact with KGSMapper, but a later version will allow a user to select from all datasets. The 2 right-hand columns (Fig. 1, Area g) allow the user to select upper and lower limits for environmental data, eliminating cells with values outside a specified interval from the analysis. Statistical analyses of both the variables and the inferred ranges can be viewed and downloaded, as can lists of cell IDs and ESRI shapefiles (Fig. 1, Area c). The KGSMapper, which can show 2 groups of taxa concurrently, provides the option to choose which taxa will be displayed (fish, anemones, or both) and/or used as the basis for the range inference (Fig. 1, Area a). In addition, the user can withhold a random selection of ~50% of the records for either dataset or for both datasets, infer a range with the remaining half, and test the product using the withheld records (Fig. 1, Area a).

Because organism occurrences are points (which define the 0.5° cells of analysis), not coverages, inferring the distribution of the habitat suitable for 1 taxon based on distribution records for another differs from inferences using environmental data. Only qualitative matches are possible using maps. A quantitative assessment can be made by determining the number of cells inferred to contain suitable habitat for 1 taxon, based on occurrence records for that taxon, then determining the proportion of known occurrences for the other taxon falling within those cells.

Analyses. We considered the effects of various aspects of the data on model outcome, addressing the issues we raised in the 'Introduction'.

Selection and effects of environmental variables (Issues 1 and 5, see 'Introduction'). We investigated which variables can explain occurrence of the subjects and, if a selection is to be made among them, the basis for choice. We tested 5 variables individually and combined into 4 groups (below). Some of these are known to affect occurrence of anemonefishes and their sea anemone hosts (sea-surface temperature [SST], depth, salinity); others (tidal amplitude and productivity, for which chlorophyll *a* concentration is a proxy) were tested to determine if they might have an effect. We also examined alternative parameters (maximum, minimum, and mean) of some variables (results not reported); minimum SST was chosen because the restriction of the animals to the tropics makes it likely that minimum temperature limits their distribution more than mean or maximum. The correlation matrix (Table 2) assesses the degree to which environmental variables covary within the region selected; this tool permits the investigator to explore the effects of spatial auto-correlation and covariance between variables, in order to help guide variable selection for the question being addressed. The strongest correlations among variables used in this study are within the variants of SST, salinity, and depth; only 1 from each category was used. For example, as might be expected, maximum and mean SST are highly correlated (but minimum SST is less so).

The following groups were selected to determine the effect on output of a number of variables: (1) minimum SST and minimum depth, (2) as Group 1 plus minimum salinity, (3) as Group 2 plus average chlorophyll *a* concentration, and (4) as Group 3 plus tidal amplitude.

Table 2. Correlation matrix of variables for the datapoints associated with fishes and anemones as it appears on screen. 1A: SST_mean_monthly; 1B: SST_min_max_range; 1C: SST_max_month; 1D: SST_min_month; 2A: Salinity_ann_avg; 2B: Salinity_max_month; 2C: Salinity_min_month; 3: Windspeed_avg; 4A: ETOPO2_bathy_min; 4B: ETOPO2_bathy_max; 4C: ETOPO2_bathy_mean; 5: CHLORA_avg_spatial (CHLORA: chlorophyll *a* concentration); 6: Tides_AVG.MA (Tides, Average Maximum Amplitude)

	1A	1B	1C	1D	2A	2B	2C	3	4A	4B	4C	5	6
1A	1	−0.7148	0.791	0.9466	−0.448	−0.3571	−0.2782	−0.703	−0.1105	0.0228	−0.0329	0.0153	−0.0916
1B	−0.7148	1	−0.1629	−0.8921	0.4939	0.486	0.1148	0.56	−0.0222	−0.2255	−0.1698	0.2258	0.1446
1C	0.791	−0.1629	1	0.5912	−0.2069	−0.0804	−0.3012	−0.4896	−0.177	−0.1759	−0.1977	0.2008	−0.0183
1D	0.9466	−0.8921	0.5912	1	−0.4985	−0.4341	−0.2317	−0.6907	−0.0629	0.1038	0.0482	−0.0905	−0.1266
2A	−0.448	0.4939	−0.2069	−0.4985	1	0.9494	0.4293	0.3026	0.0345	0.0938	0.0872	−0.222	−0.1283
2B	−0.3571	0.486	−0.0804	−0.4341	0.9494	1	0.1772	0.1744	−0.0178	−0.0235	−0.0157	−0.0775	−0.1416
2C	−0.2782	0.1148	−0.3012	−0.2317	0.4293	0.1772	1	0.3792	0.1102	0.2677	0.2382	−0.346	0.0042
3	−0.703	0.56	−0.4896	−0.6907	0.3026	0.1744	0.3792	1	0.1233	0.1369	0.1718	−0.2946	−0.0884
4A	−0.1105	−0.0222	−0.177	−0.0629	0.0345	−0.0178	0.1102	0.1233	1	0.5028	0.7361	−0.2887	−0.1417
4B	0.0228	−0.2255	−0.1759	0.1038	0.0938	−0.0235	0.2677	0.1369	0.5028	1	0.9033	−0.6949	−0.2789
4C	−0.0329	−0.1698	−0.1977	0.0482	0.0872	−0.0157	0.2382	0.1718	0.7361	0.9033	1	−0.6195	−0.2404
5	0.0153	0.2258	0.2008	−0.0905	−0.222	−0.0775	−0.346	−0.2946	−0.2887	−0.6949	−0.6195	1	0.3395
6	−0.0916	0.1446	−0.0183	−0.1266	−0.1283	−0.1416	0.0042	−0.0884	−0.1417	−0.2789	−0.2404	0.3395	1

Uncertain data quality (Issue 2). Both organism datasets contain locations known to be inaccurate (of course, we cannot know if there are additional inaccurate locations). Inaccurate locations can sometimes be identified by their associated depth; because the anemonefishes are constrained to live within the photic zone (operationally to ~100 m) by the photosymbionts of the host anemones, depths greater than 100 m strongly suggest an erroneous location. We inferred potential habitats of both fish and anemones, with and without eliminating cells, at minimum depths of >100 m.

Validating or testing range inferences (Issue 4), including making inferences about the effects of data quantity (Issue 3). We compared the outcomes of inferring habitat of each taxon based on records of another, and inferring the habitat of each taxon based on ~50% of the records for a taxon selected randomly by the KGSMapper tool. We also demonstrated the effects of eliminating from the initial dataset points in cells with values for environmental parameters >1 SD and >2 SD.

In this case study dealing with the continuum of quality and/or extent inherent in habitat assessment, KGSMapper output ranks probability of matching habitat characteristics rather than a dichotomous occurrence or not of organisms; for this reason and because assessment of known absences at the scales used (global extents and ~2500 km^2 grid cells) are impractical, output cannot be evaluated by confusion matrix measures (Fielding & Bell 1997, Manel et al. 2001). We evaluated output by what we term 'effectiveness' and 'efficiency,' assessing the distribution of cells among the intervals 0 to 1 SD, >1 to 2 SD, and >2 SD. The assumption, as in most habitat models, is that the distribution of cells inferred to contain suitable habitat will reflect that of occurrence-containing cells. For each interval i, the number of cells containing an occurrence is a_i, and the number of cells within the range is n_i. a_T is the total number of cells containing an occurrence record over n_T (the total of cells over all n_i). 'Effectiveness' is the ratio a_i/a_T—for each interval, the fraction of occurrences contained within the cells of that interval; a high value indicates inclusiveness or relative lack of false negatives. 'Efficiency' is the fraction of total occurrences per area (number of cells) inferred; we use the ratio $(a_i/a_T)/n_i$. This represents the density of positive occurrences; increasing values indicate a decrease in false positives. Effectiveness and efficiency, which are related but not identical to the confusion matrix measures of predictive power, sensitivity and prevalence, function within a run of the model; effectiveness minimizes errors of omission, and efficiency minimizes errors of commission. The data selection and editing tools permit the ratio of efficiency to effectiveness to be adjusted according to the questions and data of interest; like the output itself, evaluation of the results will necessarily be application specific.

RESULTS

Environmental variables

For each set of environmental variables, we did 3 analyses, 1 for each group of organisms individually and 1 for the 2 together. We illustrate examples of inferring the distribution of suitable habitat for each combination. Of datasets in the KGSMapper, the parameters of chlorophyll *a* concentration (Fig. 2a), minimum salinity (Fig. 2b), and tidal amplitude and wind speed (not shown) did not discriminate suitable habitat at the geographic scale of this analysis. Combinations of 2 or more of these variables provided no more resolution than any single variable analyzed individually. SST discriminated best for the habitat of these organisms latitudinally, with results differing somewhat depending on the parameter used (compare Fig. 2c,d for maximum and minimum monthly SST, respectively). Two approaches were tried to consider depth, which also restricts distribution of these animals: Fig. 2e resulted from using occurrence data alone, whereas Fig. 2f excluded the cells with minimum depths of >100 m. The number of cells inferred to contain suitable habitat (total range) was reduced by >85% as a result of editing for depth (Table 3, Fig. 2e,f). The outlier

Table 3. Inferences of suitable habitat using minimum SST and minimum depth as environmental variables, and occurrence data. Edited inferences (right-hand column for each taxon) used only records in cells in which minimum depth was <100 m. The line '0–2 SD' is the total of the preceding 2 lines. Ctot: total number of cells inferred to contain suitable habitat; Crec: number of record-containing cells; Rec: number of occurrence records; n = 641 for anemones; n = 1937 for fish

	Anemones						Fish					
	Unedited			Minimum depth <100 m			Unedited			Minimum depth <100 m		
	Ctot	Crec	Rec	Ctot	Crec	Rec	Ctot	Crec	Rec	Ctot	Crec	Rec
0–1 SD	6187	261	385	5331	244	385	7661	250	1281	4791	221	1211
1–2 SD	3450	103	207	1853	90	188	9150	119	538	1719	88	492
0–2 SD	9637	364	592	7184	334	573	16811	369	1819	6510	309	1703
>2 SD	49142	63	35	915	6	30	39379	58	107	731	27	20
Total range	58779	427	627	8099	340	603	56190	427	1926	7241	336	1723

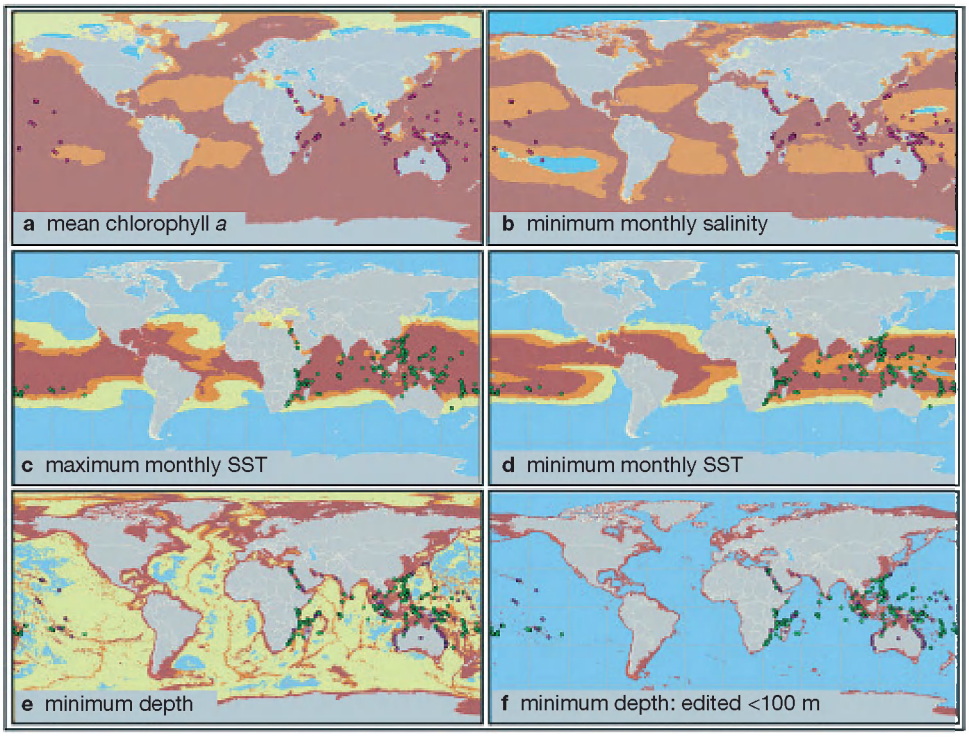

Fig. 2. Suitable habitat inferred on the basis of single variables and organism distributions. Habitat suitable for anemones inferred from anemone occurrences is based on values from the cells containing occurrence records: (a) chlorophyll *a* concentration and (b) minimum monthly salinity. Habitat suitable for fish inferred from fish occurrences: (c) maximum monthly SST and (d) minimum monthly SST. Combined fish and anemone habitat inferences: (e) minimum depth and (f) minimum depth excluding cells with values >100 m. Chlorophyll and salinity do not account for habitat, individually or in combination; similar results were obtained with tidal amplitude and wind speed (not shown). The depth constraints and the latitudinal controls imposed by SST provide a powerful combination (see Figs. 3 & 4)

category (>2 SD) was most heavily affected for anemones; editing reduced the number of 0 to 1 SD cells by 14% and of 0 to 2 SD cells by 25% in the case of anemones. The figures for fishes were 37 and 61%, respectively.

Fig. 3 illustrates the way datasets can be cleaned or edited based on either specific knowledge or statistical evaluation; to allow details to be seen clearly, they show only the part of the world where most species of these animals occur, but the analyses which led to these results made use of global data. Fig. 3a,b shows inferences of anemone and fish habitat, respectively, based on minimum depth and minimum SST, which individually provided reasonable first approximations to defining appropriate habitat (above). Fig. 3c,d shows the improvements in both inferred ranges generated by eliminating cells with a minimum depths of >100 m. Fig. 3e,f has been remapped after elimination of all cells >2 SD in Fig. 2a,b. Fig. 3g,h shows the effects of removing all cells >1 SD from the datasets used in Fig. 3a,b. This rigorous cleaning shrinks the geographic range noticeably, but the 0 to 1 and 0 to 2 SD intervals remain relatively similar throughout.

Occurrence data quantity and quality

After removal of 2 anemone localities in the Mediterranean Sea that were clearly due to misidentification of specimens, misapplication of a name, or misstatement of provenance, the datasets contained 641 anemone and 1937 fish records. They included some suspect data points and some of low precision; we retained all to provide a realistic test of habitat inference using the sort of data likely to be available for analysis of non-fisheries species.

Four anemone and 9 fish records fell on land outside a coastal cell; because marine variables are not associated with inland cells, these points were ignored in the analyses. Points on land in a coastal cell were analyzed using the marine variables associated with that cell. Some records on land do not reflect errors: the anemone dataset (for which a precision value is assigned to each georeferenced point) contains low-precision records assigned by a convention that plots the locality in the center of a country or region given as the only location information in the original publication. This results in points on land

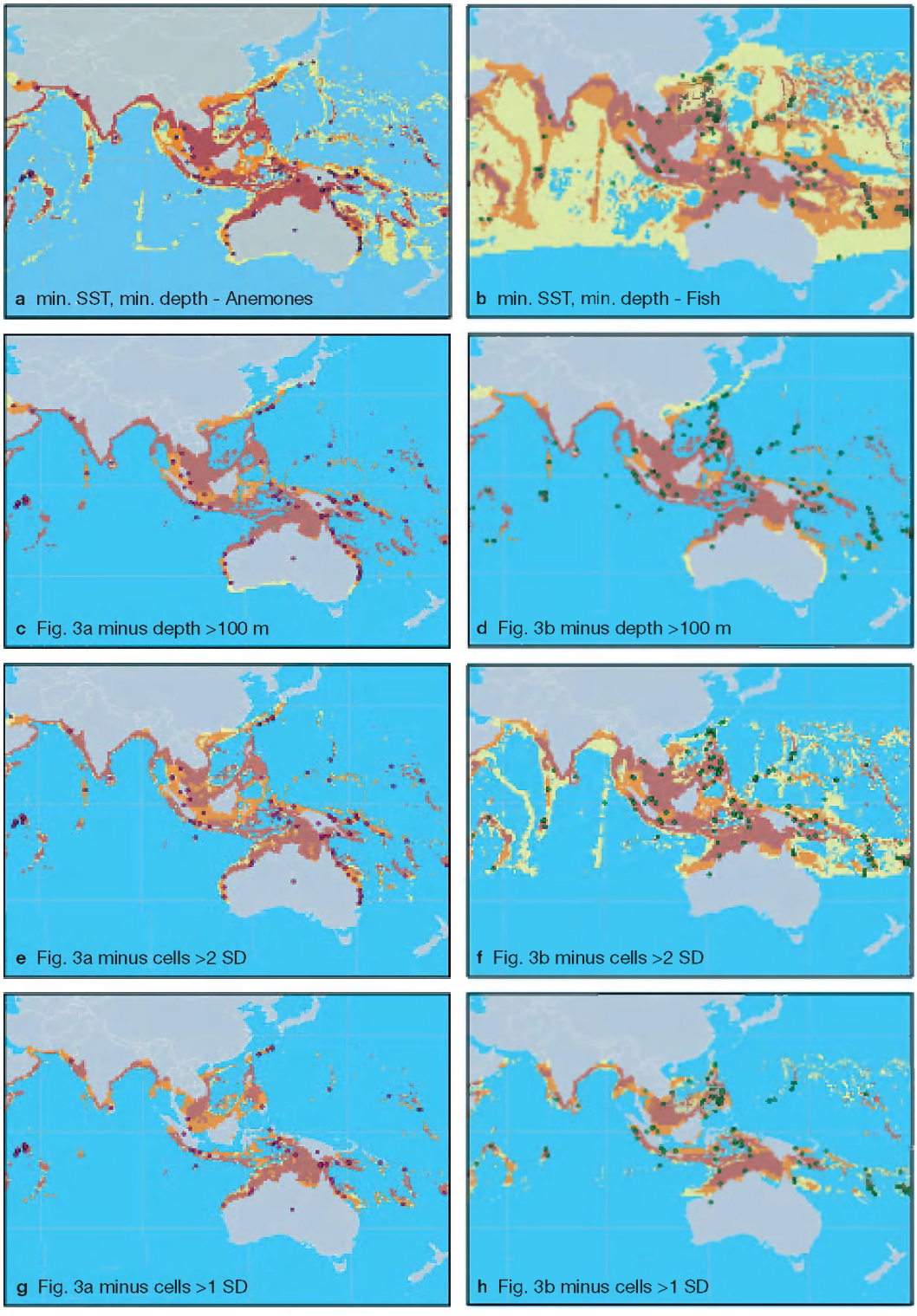

Fig. 3. Dataset clean-up and editing features displaying zoomed views of the Australasian region after inferring ranges based on the global dataset. (a) Habitat suitable for host anemones based on anemone occurrences and (b) anemonefish habitat based on fish occurrence records using unedited datasets, with minimum monthly SST and minimum depth. (c), (d) as (a) and (b), respectively, but with datasets edited to eliminate cells having minimum depths of >100 m (see Fig. 1g). (e), (f) as (a) and (b), respectively, but recalculated after eliminating cells in the >2 SD category in the initial analysis. (g), (h) as (a) and (b), respectively, but recalculated after eliminating cells >1 SD in the original analysis. The datasets can also be edited point by point, if desired (Fig. 1, Area a)

Table 4. Data using 50% of anemone-containing cells and minimum SST and minimum depth to infer habitat suitable for the remaining anemones. After editing to exclude cells with depths >100 m, 136 cells were used in this analysis

Trial	No./% cells used for inference	No./% remaining records inferred
1	57/41.91	76/96.20
2	63/46.32	73/100
3	74/54.41	61/98.39
4	64/47.06	69/95.83
5	65/47.79	71/100
6	69/50.74	67/100
7	63/46.32	72/98.63
8	67/49.26	68/98.55
9	62/45.59	71/95.95
10	65/47.79	71/100
11	57/41.91	74/93.67
12	66/48.53	70/100
13	65/47.79	71/100
14	68/50.00	68/100
15	73/53.68	63/100
16	71/52.21	65/100
17	75/55.15	61/100
18	62/45.59	74/100
19	66/48.53	69/98.57
20	71/52.21	62/95.38
Average	66.15/48.64	68.80/98.56
SD	5.01/3.68	4.44/2.01

(e.g. the centroid of Australia for localities given as 'Australia'), and over water far deeper than that in which these anemones live (e.g. the center point of Fiji for localities given as 'Fiji').

Editing to eliminate occurrence-containing cells with minimum depths of >100 m reduced anemone records by ~4% (24) and fish records by ~11% (203), but, because one 0.5° cell may contain >1 occurrence record of a fish or anemone, the number of record-containing cells was reduced by ~20 and ~21%, respectively (Table 3).

Cross-comparison and validation

Areas of suitable habitat for anemones, as inferred using 50% of anemone-containing cells, included between 93.7 and 100% of the remaining known occurrences (Table 4)—as well as many places where the anemones are not recorded as living. Clearly, the best test of our model output would be to seek the animals in places where suitable habitat is inferred to exist and the animals are not known to occur. That being impractical, we ran an analysis using KGSMapper, appropriate environmental parameters, and the native distribution (from FishBase) of anemonefishes.

On a map, known fish occurrences fell largely within areas of inferred habitat suitable for anemones and vice versa. In a quantitative assessment, using minimum SST and minimum depth (see Table 5), areas inferred by anemone occurrences included virtually all places fish are known to occur, a result somewhat improved by editing both datasets for depth. Fish occurrences were less effective in identifying areas suitable for anemones, and editing had little effect (Fig. 3). Thus, at the scale of this analysis, suitable habitat is inferred not to occur where it does not occur (at high latitude and at depth).

To explore the effects of number of environmental variables on inferred ranges, we used the 4 groups of environmental variables listed in 'Data and methods.' Fig. 4a,b shows the number of cells within each interval (the former for raw data, the latter for data edited to exclude cells with minimum depths of >100 m), Fig. 4c, d shows effectiveness, and Fig. 4e,f shows efficiency. As single variables were added, effectiveness of the output in the 0 to 1 SD interval declined. However, efficiency increased because the inferred number of cells (n_{0-1}) decreased more rapidly than the number of occurrence-containing cells (a_{0-1}). We found the same pattern within groups of related variables—inferences using maximum or minimum SST plus minimum depth and maximum or minimum SST plus the 4 variables used to generate Fig. 4d indicate that the use of maximum SST is more effective than minimum SST, which is somewhat more efficient than maximum SST.

DISCUSSION

Environmental variables

Individually, the variables of minimum salinity, chlorophyll *a* concentration, tidal amplitude, and wind speed do not identify the occurrence of habitat suitable for anemonefishes and sea anemones (Fig. 2a,b): much

Table 5. Using minimum SST and minimum depth plus occurrence of a symbiotic partner to infer occurrence of habitat suitable for another symbiotic partner, as evaluated by the percentage of target organism occurrences in the various categories of inferred habitat cells (anemones were used to infer fish habitat and vice versa). Unedited inferences used all data; edited inferences eliminated records in cells in which minimum depth was >100 m

Category	Fish habitat inferred from anemones (%)		Anemone habitat inferred from fish (%)	
	Unedited	Edited	Unedited	Edited
0–1 SD	69.4	74.0	53.4	52.4
1–2 SD	24.4	25.9	27.6	28.0
>2 SD	5.9	0.1	14.2	14.6
Total range	99.7	100	94.2	95.0

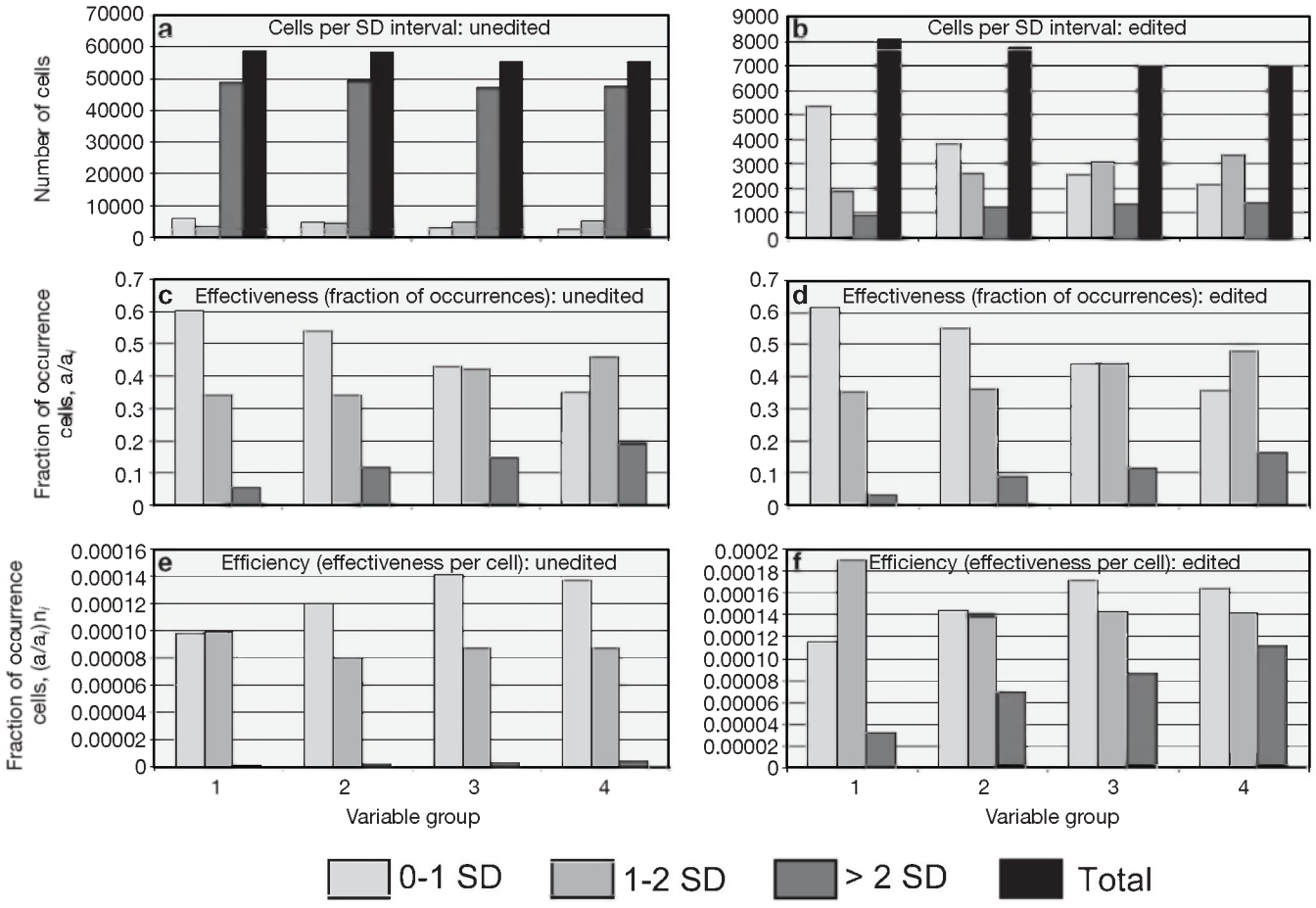

Fig. 4. Distribution of cells, effectiveness, and efficiency of habitat inference for sea anemones as functions of kind and number of variables. (a,c,e) Use values from all data. (b,d,f) Use data edited to exclude cells having minimum depths of >100 m. Numbers on abscissa are variable groups listed in the 'Data and methods' section. (a) and (b) show number of cells; (c) and (d) show effectiveness; (e) and (f) show efficiency

of the ocean has values equal to those of waters in which these animals occur. Although low salinity is negatively associated with anemone occurrence (most sea anemones, including the species that host anemonefishes, are stenohaline; Shick 1991), the resolution of our datasets both temporally (monthly averages) and spatially (0.5° cells based mainly on oceanic measurements) is too coarse to capture its effect. Similar arguments can be made for chlorophyll *a* and for the energy- and exchange-related tide and wind variables. Further, tidal amplitude is unlikely to exert systematic control because it is the relative, rather than absolute, position to low tide that affects anemone survival.

Even highly correlated parameters (Table 2) may not have the same effect. For single variables, maximum and minimum SST (Fig. 2c,d, respectively) infer somewhat different distributions of suitable habitat overall, and in the intervals 0 to 1, 1 to 2, and >2 SD. This is also true in combination with other parameters.

Adding parameters sequentially to minimum temperature and depth (Fig. 4) did not provide increasingly good inferences, from which we conclude that more variables are not necessarily better (cf. Stockwell & Peterson 2002). Quality of the variables, as judged by relevance to occurrence of the taxon in question (Fielding & Bell 1997), seems more important than the number of variables. Quality can be improved by basing the output on values that do not include the outliers (>2 SD) or >1 SD (Fig. 3).

Even with the limited number of environmental variables available in the KGSMapper, choosing variables expected to be relevant to the distribution of any taxon requires some expert judgment, as does determining which relevant variables to use for a given purpose. For example, although maximum SST is quantitatively more effective than minimum SST, it identifies a larger range overall and overextends the northern extent of the fish distribution (Fig. 1, map result). KGSMapper can help to reveal which parameters are most closely

correlated with occurrence, and thus may be important in controlling, or describing, distribution.

Occurrence data quality and quantity

The linkage of taxonomic synonyms allows Hexacoral to map occurrences for the species rather than for the name; this also helps to increase the number of records for a species. Thus, rather than synonymous names being viewed as a problem (Soberón & Peterson 2004), if handled appropriately, they can serve to enhance data quantity and taxonomic quality.

The 2 Mediterranean records we removed illustrate the need for expert judgment in selecting both occurrence and environmental data. Machine algorithms that cleanse datasets by purging records from areas well beyond known occurrences risk removing information on range extensions or invasions. An expert may be able to differentiate among potential sources of error by considering date, similar species, taxonomic history of a name, etc., to make suspect records useful, and thereby improve data quality and quantity.

Using only cells with minimum depth values <100 m resulted in a more precisely defined range (Fig. 2f) than merely selecting minimum depth as a variable (Fig. 2e), presumably because some actual occurrences fall in cells with minimum depths of >100 m, due to either error or convention (such as using the center point of the Fiji Islands for all localities given only as 'Fiji' in the anemone dataset).

Such editing for a feature relevant to organism distribution provides a crude assay of data quality. Editing produced a less dramatic change for anemones than for fishes; compare Fig. 3a,c,e,g with Fig. 3b,d,f,h, respectively. This finding is concordant with what is known of the data sources: the anemone records were assembled as a single project (by D. G. Fautin) and have been extensively checked, whereas the fish records are from multiple sources with unknown and diverse authentication procedures. Thus, KGSMapper deals with suspected, inferred, or known erroneous data to provide a justifiable way to limit consideration to reasonable habitat possibilities. By doing so, it inferentially takes absences into account.

It is commonly thought that more environmental variables will improve the sensitivity or precision of a prediction. Fielding & Bell (1997) call this into question in their discussion of the issues of inappropriate variables, the 'costs' of misclassification, and the contexts in which predictive models are evaluated. Fig. 4 illustrates how choice and number of environmental variables affect output in our study system. As we added single variables to the analysis, the number of cells identified as containing suitable habitat declined by ~10%, but it would be a mistake to interpret this as increasing precision; the fraction within 1 SD declined by ~40% in both edited and unedited analyses. KGSMapper statistics are calculated in a univariate manner; as variables are added, the probability declines that any cell will contain values within 1 SD for all of them. Thus, adding a variable that would be expected, based on biology and analysis, to have little control over organism occurrence can eliminate cells that contain suitable habitat—a high price to pay for minimal return in terms of genuinely improved results.

Others have also found that quality of prediction is not necessarily improved by quantity of data. 'Accuracy' of 4 modeling methods, including GARP, used by Stockwell & Peterson (2002) did not increase beyond about 20 data points, 10 producing 59 to 64% 'accuracy' (90% of potential achievement rate using with their methods). Beauvais et al. (2004) achieved 'validation success' rates of 40.0 to 88.2%, the lowest with a dataset of 18 records, another dataset of 20 records had a rate of 80.0%. The effects of geographic scale and habitat heterogeneity on quality of model output have not been addressed formally, but, based on what is now known, this is an issue which should be addressed. The methods of Stockwell & Peterson (2002, p 11) modeled 'widespread species … less accurately'; Raxworthy et al. (2003) achieved a similar result using GARP. Attention must be paid to this subject for marine species, many of which have larger geographic ranges than is typical of terrestrial species for which predictive algorithms were developed (the animals we studied range through about 180° of longitude and 50° of latitude) and occur in 3 dimensions. In one of the few published modeling studies for the distribution of marine species (fish living in the central western Atlantic), Wiley et al. (2003, p 124) also found that, using GARP, results for widespread species were 'weak.'

In addition to large numbers of points, a desideratum for this sort of analysis is independence of data (Fielding & Bell 1997). However, many of the anemone records we used came from a small number of areas and/or investigators; we have found that records for other poorly studied marine organisms may not be truly independent.

Validating or testing results

Use of training data for assessing quality of model output is a common practice (e.g. Anderson et al. 2003). Such data may constitute a portion of known occurrences (e.g. 50% in Peterson et al. 2002, 75% in Beauvais et al. 2004) or areas of occurrence (e.g.

states in Peterson 2001). KGSMapper has a tool that randomly selects ~50% of reported occurrences and uses the associated locality records (grid cells) to infer the remainder of the localities and their associated occurrences (Fig. 1, Area a). If grid cells are the basis for analysis when using a gridded environmental database and a one-to-one relationship between cells and occurrence records does not exist, the use of occurrence records will not be reliable. A random sample of (e.g. 50%) of the locality cells may contain far more than the stated proportion of the sample occurrences (up to 75–80% in 50% of the tests we conducted with anemone data). This greatly increases the apparent quality of the results and is misleading if that level of performance is ascribed to 50% of the occurrences.

A drawback involved in withholding some records as training data is that 'the algorithm cannot take advantage of all known locality records' (Anderson et al. 2003, p 213). The symbiosis allowed us both to use all data and to implement the desideratum of incorporating interspecific information into the model (Fielding & Bell 1997); we used records of 1 organism to infer areas of suitable habitat for another. We ascribe the asymmetry in our results (Table 5) to that in the relationship—although an anemonefish never occurs without an anemone in nature, individual anemones may occur without fish in some areas. Thus, anemone data will somewhat overestimate suitable habitat for fish. This result is consistent with the potential problem in modeling pointed out by Fielding & Bell (1997) of undersaturation of habitat. Accordingly, saturated symbiotic systems such as this case should be particularly favorable as tests of habitat models.

As an indirect assessment of KGSMapper, we used environmental variables from Hexacoral with occurrence data from FishBase for the tropical Indo-Pacific lionfish *Pterois volitans*. The inferred distribution of suitable habitat resembles that of anemonefishes, and includes the coast of the southeastern United States, where it has recently established viable populations (e.g. Semmens et al. 2004).

The addition of environmental variables that do not, and are not expected to, have any real explanatory power has the effect of increasing the apparent efficiency of the range inference. This is an artifact of constraining the basis on which cells are selected, whether or not that constraint has anything to do with organism occurrence. For a group of organisms that has been extensively sampled over most of its range, this will have little effect other than to distort the apparent quality of the range inference. However, for sparsely sampled organisms, such as most marine organisms, inclusion of gratuitous variables could significantly alter the inferred range.

Although we can readily envision application of KGSMapper to dichotomous problems, the analyses presented here cannot be usefully evaluated by confusion matrix methods (Manel et al. 2001), because of the unavailability of useful absence data at the scale of interest. A half-degree grid cell can be as large as 3000 km^2 in area; the organisms of interest range from a few cm^2 to about 1 m^2 in area, and habitat patches may be <100 m^2. The grid cell is best treated as a mosaic of potential habitats, ranging from favorable to stressful to impossible. To provide some assessment of the quality and characteristics of the inferences, we use efficiency and effectiveness, which allow a user to tune the results for a particular purpose based on the relative importance or cost (Fielding & Bell 1997) assigned to errors of commission and omission. For example, a user planning an expedition to sample particular taxa or to devise a scheme for protected areas would probably want to emphasize efficiency (i.e. maximize the probability of finding organisms per unit area covered), while a study concerned with invasion potential, marginal habitats, or range limits would need the most effective (complete) inventory of potential habitat. Moreover, such analysis allows a user to allocate effort where it will most enhance a product of prediction—adding occurrences would improve the product more than adding environmental features. Similarly, Graybeal (1998) found that adding taxa improved resolution of phylogenetic trees more than adding characters.

Maps and model outputs

An occurrence (or dot) map plots localities where members of the taxon have been documented (for example, Fig. 1, Area b, without the inferred areas of occurrence); subdividing occurrences temporally allows comparing distributions through time. An inference about where members of the taxon may occur beyond the known occurrences constitutes a range map. This, too, may be temporally defined, showing, for example, where organisms formerly occurred, but do not occur currently. It may consist of discontinuous patches, as for the anemonefish and their host anemones around land masses. When drawn up the 'old-fashioned way,' a range map is a simple abstraction of occurrences, an inference of where members of the taxon may occur within the same geographical region. A map generated electronically by a tool such as KGSMapper, by correlating environmental parameters with known distributions, is essentially a habitat map, plotting places compatible with the life of the organism of interest.

A habitat map may contain areas of 2 types, and we advocate that these be distinguished from one

another. Areas on a habitat map that fall within the broad ambit of the taxon constitute, as defined above, a range map. Such maps are useful for planning, e.g. field research and conservation strategies in that, by depicting realized habitat, they provide reasonable precise inferences about where members of a taxon may actually live. Some habitat maps include areas that fall well outside the known distribution of the taxon, as illustrated in Fig. 1: anemonefishes and their hosts only occur naturally in the Indo–West Pacific, but ostensibly suitable habitat for them occurs in some areas of the Atlantic (especially the Caribbean) and the eastern Pacific. Such a map depicts potential habitat, which is ideal for identifying places vulnerable to invasion. Because the word 'prediction' literally refers to the future, it is appropriately used for areas outside the natural range—that is for areas subject to invasion. Within the general geographical area in which members of a taxon are known to occur, where direct evidence of their occurrence may currently be absent, a model actually infers—rather than predicts—appropriate habitat.

Some model outputs are said to be niche maps; whereas a habitat is defined on the basis of abiotic parameters, a niche also includes biotic parameters (e.g. Peterson 2001, Anderson et al. 2003). Including explicit biotic information in automated tools such as KGSMapper is difficult, because such information is rarely in the form of coverages. The 1 biotic parameter common in oceanographic data is chlorophyll *a* concentration, but this lacks discriminatory value for the occurrence of most organisms such as those we studied (Fig. 2a). We found that, although appropriate habitat for anemonefishes exists outside the Indo–West Pacific, when we included a vital component of the animal's biotic environment, a host anemone, those areas were no longer identified as habitable. We therefore advocate that such relevant biotic factors be explicitly incorporated into models if they are to be considered niche models. In this case we used symbionts, some pairs of which are mutualistic, precisely because this provided a clearly relevant biotic factor with which to test model output. The relevant biotic factors in other analyses may be less obvious.

Thus, anemonefishes are less likely than lionfish to establish viable populations in the coastal southern United States: although abiotic attributes of the habitat, such as temperature and depth, appear suitable for anemonefish existence, anemones that naturally host anemonefishes do not occur there (Fautin & Allen 1992). One way to infer absence is to eliminate deep water cells (cells in which minimum depth is >100 m). A second way to infer absence is to eliminate all fish habitat cells outside the Indo–West Pacific. This is justifiable based on the absence of an obligate symbiont. By contrast, the potential for Hawaii to be invaded by anemonefish is real, because 1 species of host anemone occurs in Hawaiian waters (Fautin & Allen 1992). On the other hand, for species of these anemones that can live in nature without fish symbionts (most of them), we infer that the suitable habitat outside the Indo–West Pacific is vulnerable to invasion. Once individuals of a host anemone are present in a non-native place, they might be follwed by the species of fish able to live with this particular species of host anemone.

Modeling tools

It is difficult and/or impractical to control quality when using merged, distributed datasets. Therefore, analytical and predictive tools must have features that ensure robust output in the presence of questionable data and that offer the user ways to modify the datasets and to assess the results—by improving data quality, by testing hypotheses derived from them, or both. We have shown that the number and distribution of outlier points is an indicator of both the quality of occurrence data and the relevance of the environmental variables selected. Thus, an output that segregates results into categories of diminishing accuracy allows a user to select appropriate subsets of the output. User decisions can be based on the level of data confidence and the purposes for which the output is to be used. With KGSMapper, for example, we found the 1 SD range to be a robust initial estimate of range, even in noisy datasets.

Beyond this passive evaluative approach, KGSMapper has data-editing features that are broadly useful in assessment and research. A user can edit occurrence data: (1) point by point on the map or data list, (2) by geographic area (using the zoom control), (3) by taxon, and/or (4) by editing environmental variables. Future versions of the KGSMapper will have more versatile means of selecting geographic extent and will include explicit absence as well as presence data. In addition to selecting geographic extent, the ability to edit variables provides a means of exercising expert judgment by cleaning the datasets of points that do not conform to relevant environmental controls and of refining the geographic limits of potential ranges; these are ways to incorporate knowledge of absence. An important means of improving the precision of the habitat inferences is provided by allowing a user to remove records that fall beyond a predetermined statistical limit. The user can then recalculate the model with the remaining cells. KGSMapper allows application of expert judgment at both input and output ends

of the process; algorithms such as GARP apply it only at the output end (e.g. Anderson et al. 2003, Drake & Bossenbroek 2004).

KGSMapper outputs go beyond simple map visualizations, providing statistical analyses of the individual variables, of the relationships among the variables, and of the occurrence–environment relationships. In addition to allowing analyses in a manipulative GIS environment, KGSMapper has options that permit dynamic data assessment, which enables the user to identify covarying parameters, variables to be edited, and specific ranges of values to be included or excluded.

Models of organism occurrence may contain 2 types of errors: predicting the organism will occur where it does not (false positive, commission, or overprediction) and not predicting the organism to occur where it does (false negative, omission, or underprediction) (e.g. Fielding & Bell 1997, Anderson et al. 2003). Unlike many algorithms, the objective of KGSMapper is to infer the locations of habitat suitable for occurrence of organisms, not organism occurrence itself. Finding the organisms in the habitat clearly demonstrates it is suitable; not finding them, termed by Anderson et al. (2003) 'apparent commission error,' is due to well-known contingencies in occurrence. To regard prediction of habitat in a place that has not been searched as a false positive is to imply perfect knowledge of organism occurrence. Selecting areas for fieldwork is a potential use of the output of such modeling, particularly for poorly sampled taxa; overestimation of habitat occurrence is therefore neither unexpected nor necessarily undesirable. Moreover, a model that identifies all, but only, the places of known occurrence would be tautologous.

Acknowledgements. Financial support was provided by US National Science Foundation Grants OCE 00-03970 (through the National Oceanographic Partnership Program), DEB 99-78106 (in the PEET program: Partnerships for Enhancing Expertise in Taxonomy), and DEB 95-21819 (PEET). Collaborations with LOICZ (Land–Ocean Interactions in the Coastal Zone, an IGBP project) and FishBase (www.fishbase.org), an element of OBIS (the Ocean Biogeographic Information System), were central to this effort. Additional gratitude is due to the US Fish and Wildlife Service; the US National Biological Information Infrastructure (NBII), especially T. D. Beard and M. Fornwall; K. Look and K. Nelson of the Kansas Geological Survey; A. Ardelean, and G. P. Beauvais, University of Wyoming, USA.

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

LITERATURE CITED

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol Model 162:211–232

Beauvais GP, Keinath D, Thurston R (2004) Predictive range maps for 5 species of management concern in southwestern Wyoming. Report prepared for Advanced Resources International by the Wyoming Natural Diversity Database, University of Wyoming, Laramie. Available from http://uwadmnweb.uwyo.edu/WYNDD/Reports/pdf_beauvais/sw_wyo_predictive_rm_04.pdf

Drake JM, Bossenbroek JM (2004) The potential distribution of zebra mussels in the United States. Bioscience 54: 931–941

Dunn DF (1981) The clownfish sea anemones: Stichodactylidae (Coelenterata: Actiniaria) and other sea anemones symbiotic with pomacentrid fishes. Trans Am Phil Soc 71(1):1–115

Fautin DG, Allen GR (1992) Field guide to anemonefishes and their host sea anemones. Western Australian Museum, Perth. Available from www.nhm.ku.edu/inverts/ebooks/intro.html

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetics problem? Syst Biol 47:9–17

MacArthur RH (1972) Geographical ecology: patterns in the distribution of species. Harper & Row, New York

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Peterson AT (2001) Predicting species' geographic distributions based on ecological niche modeling. Condor 103: 599–605

Peterson AT, Ball LG, Cohoon KP (2002) Predicting distributions of Mexican birds using ecological niche modeling methods. Ibis (online) 144:E27–E32

Raxworthy CJ, Martinez-Meyer E, Horning N, Nussbaum RA, Schneider GE, Ortega-Huerta MA, Peterson AT (2003) Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837–841

Semmens BX, Buhle ER, Salomon AK, Pattengill-Semmens CV (2004) A hotspot of non-native marine fishes: evidence for the aquarium trade as an invasion pathway. Mar Ecol Prog Ser 266:239–244

Shick JM (1991) A functional biology of sea anemones. Chapman & Hall, London

Soberón J, Peterson AT (2004) Biodiversity informatics: managing and applying primary biodiversity data. Phil Trans R Soc Lond Ser B Biol Sci 359:689–698

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148: 1–13

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16:120–127

Submitted: December 27, 2004; Accepted: December 12, 2005
Proofs received from author(s): June 14, 2006