Vol. 645: 125–139, 2020
https://doi.org/10.3354/meps13385

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 9

Predicting species richness and abundance of tropical post-larval fish using machine learning

Henitsoa Jaonalison[1,*], Jean-Dominique Durand[2], Jamal Mahafina[1], Hervé Demarcq[3], Nils Teichert[4], Dominique Ponton[5]

[1]Institut Halieutique et des Sciences Marines, Université de Toliara, BP 141 Rue Dr Rabesandratana, Mahavatse II, 601 Toliara, Madagascar

[2]MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Bat 24 cc 093 Place Eugène Bataillon, 34095 cedex Montpellier, France

[3]MARBEC, IRD, Univ Montpellier, CNRS, Ifremer, CS 30171 Avenue Jean Monnet, 34203 cedex Sète, France

[4]Laboratoire de Biologie des Organismes et Ecosystèmes Aquatiques (BOREA), Muséum National d'Histoire Naturelle, CNRS, IRD, SU, UCN, UA – Station Marine de Dinard – CRESCO, 38 rue du Port Blanc, 35800 Dinard, France

[5]ENTROPIE, IRD - Université de La Réunion – CNRS - Université de la Nouvelle-Calédonie - Ifremer, c/o Institut Halieutique et des Sciences Marines, Université de Toliara, BP 141 Rue Dr Rabesandratana, Mahavatse II, 601 Toliara, Madagascar

ABSTRACT: Post-larval prediction is important, as post-larval supply allows us to understand juvenile fish populations. No previous studies have predicted post-larval fish species richness and abundance combining molecular tools, machine learning, and past-days remotely sensed oceanic conditions (RSOCs) obtained in the days just prior to sampling at different scales. Previous studies aimed at modeling species richness and abundance of marine fishes have mainly used environmental variables recorded locally during sampling and have merely focused on juvenile and adult fishes due to the difficulty of obtaining accurate species richness estimates for post-larvae. The present work predicted post-larval species richness (identified using DNA barcoding) and abundance at 2 coastal sites in SW Madagascar using random forest (RF) models. RFs were fitted using combinations of local variables and RSOCs at a small-scale (8 d prior to fish sampling in a 50 × 120 km^2 area), meso-scale (16 d prior; 100 × 200 km^2), and large-scale (24 d prior; 200 × 300 km^2). RF models combining local and small-scale RSOC variables predicted species richness and abundance best, with accuracy around 70 and 60%, respectively. We observed a small variation of RF model performance in predicting species richness and abundance among all sites, highlighting the consistency of the predictive RF model. Moreover, partial dependence plots showed that high species richness and abundance were predicted for sea surface temperatures <27.0°C and chlorophyll *a* concentrations <0.22 mg m^{-3}. With respect to temporal changes, these thresholds were solely observed from November to December. Our results suggest that, in SW Madagascar, species richness and abundance of post-larval fish may only be predicted prior to the ecological impacts of tropical storms on larval settlement success.

KEY WORDS: Fish post-larvae · DNA barcoding · Surface water masses · Remote sensing · Random Forests · Modeling

1. INTRODUCTION

Several studies have pointed out the critical need to survey pre-settlement fish (hereafter referred to as post-larvae) for marine ecosystem monitoring (Hsieh et al. 2005, Cowen et al. 2007, Koslow et al. 2013). Post-larval surveys are important because variations in the supply of post-larvae will influence the size and composition of juvenile fish populations (Jenkins & King 2006) and thus determine adult fish renewal (Takahashi & Watanabe 2004). Post-larval distribution, richness, and abundance can be affected by bio-

*Corresponding author: jaonasat@gmail.com

physical processes (Jackson et al. 2001, Leathwick et al. 2006, Mavruk et al. 2018) and increasing stress from anthropogenic activities (Jackson et al. 2001). Because post-larval supply onto the nearshore habitat is a function of spawning stock size (Moser & Watson 1990), determining how post-larval species richness and abundances vary through time would help us understand ecosystem modifications (Wernberg et al. 2013) and detect changes in fish communities (Koslow & Wright 2016). The central role of post-larvae in fish population dynamics highlights the urgent need for a better understanding of how species richness and abundance vary in nearshore habitats.

Since the early life stage of fishes is strongly impacted by biophysical drivers (including hydrodynamic conditions and food availability), we should be able to use those drivers to predict post-larval species richness and abundance in nearshore habitats. These predictions may inform the future structure of settled fishes, and thus adult populations, which is important information for resource management (Francis et al. 2011). Therefore, by integrating environmental variables, species richness and abundance of post-larvae are also central for ecological modeling (Nicolas et al. 2010, França et al. 2012). Most previous modeling studies have been based on regression methods for predicting the species richness and abundance of fishes. For example, Francis et al. (2005) and Klemas (2012) used generalized additive models (GAMs) for predicting the abundance of fish, while Vasconcelos et al. (2015) used generalized linear models (GLMs) for predicting the species richness of estuarine fish. However, França & Cabral (2015) reported the outperformance of classification and regression trees (CARTs) for predicting species richness of juvenile estuarine fishes compared to GLM, GAM, and boosted regression trees. CART can be highly sensitive to changes in training and test data sets. Knudby et al. (2010) identified the high performance of random forest (RF) models in predicting the richness of fish compared to the modeling techniques mentioned above, including CART. The modeling studies mentioned above have always focused on juvenile and adult estuarine fish because an accurate identification of post-larvae to species level is often difficult. The few previous modeling studies of post-larvae included only a single species (Jenkins et al. 1999, Koehl et al. 2007) or specimens identified to the family level (Burgess et al. 2007).

All of the aforementioned studies modeling species richness and abundance were based on variables recorded during sampling, either *in situ* or using remotely sensed oceanic conditions (RSOCs). However, environmental variables recorded before the sampling period using RSOCs could be important for modeling species richness and abundance because these conditions can influence post-larval spatial distribution and survival and the structure of larval fish supply into coastal habitat. To our knowledge, RSOCs during the few days preceding sampling have never been used for predicting species richness and abundance of post-larvae, so the usefulness of this approach remains to be demonstrated. Although developing predictive models based on RSOCs remains challenging, this approach has been successfully used to predict the occurrence of jellyfish blooms (Albajes-Eizagirre et al. 2011).

Based on accurate estimates of species richness using DNA barcoding, this study aimed to (1) identify the best RSOC scales for accurately predicting variations in post-larval species richness and abundance; (2) define the main RSOC variables affecting species richness and abundance; and (3) discover how model performance varied between 2 contrasting coastal sites. We hypothesized that (1) small-scale RSOC variables would predict the variations in species richness and abundance of post-larval tropical fishes better than local, meso-, and large-scale variables, and (2) the important variables should be similar for both species richness and abundance at both sites. To achieve these goals and test these hypotheses, we used the RF machine learning technique to model species richness and abundance of post-larval tropical fishes sampled at 2 contrasting coastal sites in southwestern Madagascar. RF models were based on locally recorded information (local variables) and RSOC variables that were recorded at different spatial and temporal scales before each sampling period.

2. MATERIALS AND METHODS

2.1. Study site

This study was carried out in southwestern Madagascar (Fig. 1) because of (1) the presence of a barrier reef and contrasting sites in terms of water mass characteristics; (2) the availability of previous data related to these sites; and (3) the presence of small-scale fisheries activities which are in need of effective management. Due to logistical constraints, the study was only conducted at 2 coral reefs, located approximately 50 km apart. The first site, off Anakao (Anakao reef; ANA), was a flat coral reef surrounding Nosy Ve Island, situated 10 km south of the permanent Onilahy River. This site is influenced by the plume of the Onilahy River when northerly winds

blow. The influence of this river flow explains the variability in the water mass characteristics between December and February (sea surface temperature [SST]: 23.5–29°C; salinity: 32–35), i.e. during the warm and rainy season (Jaonalison et al. 2016). The second site was located in the north of the great barrier reef of Toliara (GRT) (Fig. 1). This reef stretches over 19 km and represents approximately 33 km^2 of structurally diverse, shallow reef areas where coral diversity has declined since the 1960s (Bruggemann et al. 2012). The GRT site was situated 4.5 km south of the non-permanent Fiherenana River, and 25 km north of Onilahy River. Characteristics of water masses at GRT also vary, with SSTs ranging from 24–28°C and salinity between 33 and 36 (Jaonalison et al. 2016). Nevertheless, the characteristics of the water masses at GRT can be considered less variable than at ANA because diurnal tidal currents induce a regular, intense mixing in the northern part of the great barrier reef (R. Arfi unpubl. data).

2.2. Post-larval sampling

This work was based on a monthly sampling during 3 austral warm seasons (November–April) in the 2014–2015, 2016–2017, and 2017–2018. Sampling was performed during 3 consecutive nights of the new

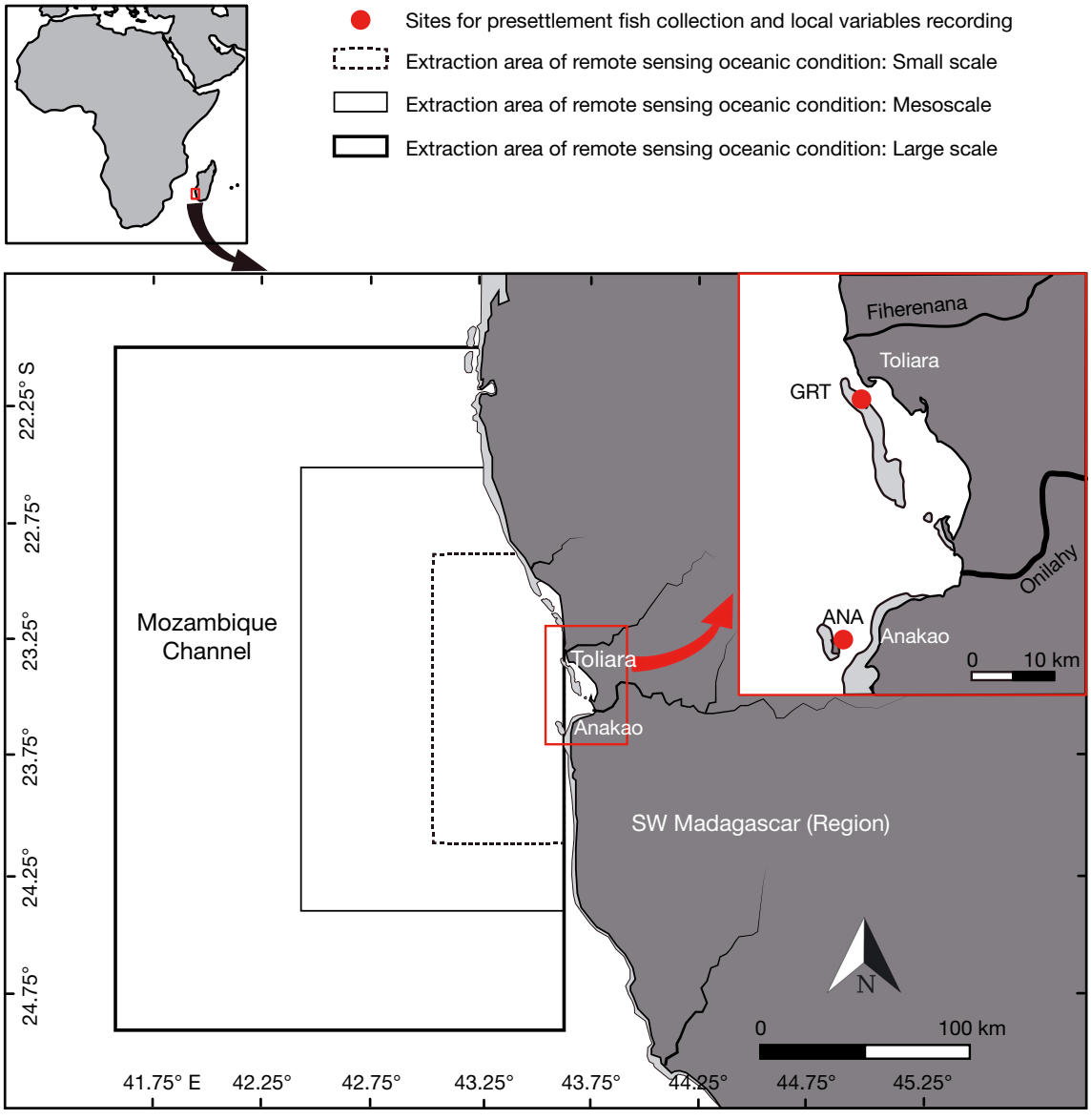

Fig. 1. Sampling site locations for post-larvae (red circles) in southwestern Madagascar. ANA: Anakao reef; GRT: great barrier reef of Toliara

moon period, using Système Lumineux Electronique d'Echantillonnage des Post-larves (SLEEP) light traps, as in Collet et al. (2018). The new moon period of the warm season was chosen because it corresponds to the post-larval supply peak (Robertson et al. 1988). Although light traps are selective and their efficiency is influenced by hydrodynamic conditions and water mass turbidity (Marchetti et al. 2004, Lindquist & Shaw 2005), they were used because they can catch fish post-larvae before they settle onto benthic habitats (Wilson 2001, Anderson et al. 2002). At each site, 3 light traps were set up at dusk and retrieved the following morning. The collected post-larvae were live-transported to the laboratory, where they were sorted to morphospecies.

2.3. Identification process: high definition photo and DNA barcoding

One specimen of each morphospecies was randomly selected and photographed with a Nikon model D90 camera equipped with a Sigma 105 mm macro lens. The camera was connected directly to a computer equipped with Adobe Lightroom® software, which was used for managing the photos and all information related to each specimen. A tissue fragment of each photographed specimen was preserved in 90% ethanol and stored at −20°C until DNA extraction. Because morphological identification of post-larval fish to species level remains challenging (Frantine-Silva et al. 2015), genetic analysis has been used to effectively identify fish larvae (Ko et al. 2013). DNA extraction and sequencing were performed at the Mediterranean Center for the Environment and Biodiversity (CEMEB) at the University of Montpellier, France, following the methods of Collet et al. (2018). DNA sequences of the cytochrome oxidase I (COI) gene were manually adjusted after visual inspection using 'Chromas v.2.6.4' (http://technelysium.com.au/wp/chromas/). The adjusted sequences were then edited and aligned with 'Clustak W' using MEGA 7.0 (Kumar et al. 2016). COI sequences, with the image of the corresponding specimen, were then uploaded into the Barcode of Life Data System (BOLD) database (public data set DS-PHDJAO). In BOLD, each sequence was automatically assigned a Barcode Index Number (BIN; Ratnasingham & Hebert 2013). To assign a species name to each BIN, we first identified the specimen as 'Genus+species' if (1) the BIN corresponded to only one species in BOLD and (2) the species was only observed in this BIN. Secondly, if the BIN corresponded to more than one species of the same genus, or if the species' name corresponded to multiple BINs in BOLD, the specimen was identified as 'Genus+BIN' (e.g. *Lethrinus* [BOLD:AAB0511]). Third, if the BIN corresponded to species from different genera, but belonging to the same family, the specimen was identified as 'Family+BIN' (e.g. Gobiidae [BOLD:ACV9382]). Note that the identifications Genus+BIN and Family+BIN signify identification to species level because each BIN corresponds to an operational taxonomic unit, and thus to a putative species (Ratnasingham & Hebert 2013). When DNA barcoding failed, identification remained at the morphospecies level (e.g. Congridae_gen sp_1HJ).

The total abundance and species richness in each sample did not take into account small pelagic fish species (e.g. Clupeiformes and Atheriniformes), which can be abundant in some light-trap catches. However, their abundances were considered in the models among the local variables because large numbers of small pelagic fishes may reduce the efficiency of the light traps for catching reef fish post-larvae.

2.4. Environmental variables

Several local variables were recorded at each site when the light traps were set: SST (using a thermometer), water transparency (using a Secchi disk), and wind speed and direction (using an anemometer and a compass). These variables were selected because they can influence either the efficiency of light traps (Hickford & Schiel 1999) or the species richness and abundance of the post-larvae (Harris et al. 2001, Chen et al. 2018). Due to technical problems, sea surface salinity was recorded for the 2017–2018 sampling season only and was thus not retained for analyses. Due to logistical constraints, light-trap setting time was not fixed but ranged from 16:45–20:45 h. The difference between the time of sunset and light-trap setting and between the time of sunrise and light-trap collection were thus calculated and used as local variables to reflect the potential effect of high tide that always occurred around sunset.

RSOCs were extracted from 3 different spatial and temporal scales (Fig. 1) based on James et al. (2002), who suggested that fish larvae could be transported 21–43 km over 6 d. According to a similar proportion, we hypothesized that larvae could be influenced by environmental factors (1) up to 50 km from the coast over an 8 d period before the sampling night at the small scale (RSOC-SS); (2) up to 100 km and 16 d at the meso scale (RSOC-MS); and (3) up to 200 km from the coast and 24 d at the large scale (RSOC-LS). The RSOC-SS, -MS and -LS included the composite MUR

SST product (JPL 2015), AQUARIUS sea surface salinity (Meissner & Wentz 2016) (both datasets available through https://oceandata.sci.gsfc.nasa.gov/Aquarius/), and the level-3 daily data set of chlorophyll *a* (chl *a*) concentration from MODIS (NASA Ocean Biology Processing Group 2017) (https://oceancolor.gsfc.nasa.gov/). The WindSat data (www.remss.com) are produced by Remote Sensing Systems and sponsored by the NASA Earth Science MEaSUREs DISCOVER Project and the NASA Earth Science Physical Oceanography Program (Wentz et al. 2013). Finally, the OSCAR surface current velocity (ESR 2009) was obtained from www.esr.org/research/oscar/oscar-surface-currents/. All data sets and variable names are summarized in Table 1. The spatial and temporal resolution of these RSOC variables, as well as their corresponding number of pixels (for each variable, one pixel denotes one piece of information), are detailed in Table 1. For each variable, information from these pixels for each day was averaged over 8, 16, and 24 d.

2.5. Data sets used for RF modeling

For each sampling site, species richness and abundance were modeled separately with local variables, RSOC-SS, RSOC-MS, and RSOC-LS (Fig. 2a, Step 1). This first step (1) selected the important local variables and (2) compared the goodness-of-fit of those models using local variables against those using the 3 scales of RSOCs. Because local variables can play an essential role in the prediction of species richness and abundance, the most important variables were identified following a RF-recursive feature elimination (RF-RFE) algorithm, and then added to the RSOC-SS, RSOC-MS, and RSOC-LS variables (Fig. 2a, Step 2). This second modeling step allowed comparison between the goodness-of-fit of models without (Step 1) and with (Step 2) the most important local variables.

2.6. RF modeling

The RF algorithm was chosen because this machine learning technique can be used with data presenting non-constant variance distributions, or unbalanced data. RF modeling can also easily deal with missing values (Potts & Elith 2006) and allow for nonlinear relationships between predictors (Darst et al. 2018). The number of predictors randomly sampled at each node ('mtry') and the number of trees ('ntree') are the main parameters of the RF algorithm (Liaw & Wiener 2002). Because the number of predictors (p) for each model was always ≤14, mtry = $\sqrt{p}$ was used by default for all the models. The ntree parameter was visually selected by plotting the minimum mean square error (MSE) according to the number of trees. For all the models, 500

Table 1. Response and explanatory variables included in random forest modeling, and spatial resolution of remote sensing oceanic condition (RSOC) variables and the number of pixels corresponding each RSOC variable over small- (SS), meso- (MS), and large-scales (LS) extracted at 8, 16, and 24 d preceding sampling, respectively. (–) not applicable

Variable types	Variable codes	Variable description	No. of pixel scale^{-1} SS	MS	LS	Spatial resolution (km)
Response variables	Species richness	No. of species sample^{-1}	–	–	–	–
	Abundance	No. of ind. sample^{-1}	–	–	–	–
Local variables recorded during sampling night (LO)	SST_LO (°C)	Sea surface temperature (°C)	–	–	–	–
	Wind_ U_LO	Cross-shelf wind velocity (m s^{-1})	–	–	–	–
	Wind_V_LO	Alongshore wind velocity (m s^{-1})	–	–	–	–
	Water_turb_LO	Water turbidity (m)	–	–	–	–
	Dif_set_LO	Difference between sunset and light-trap setting time (h)	–	–	–	–
	Dif_col_LO	Difference between the sunrise and light-trap collection time (h)	–	–	–	–
	S.Pelagic_LO	Pelagic fish (ind.)	–	–	–	–
RSOC	SST_SS	Sea surface temperature (°C)	12998	70208	284216	4
	SSS_SS	Sea surface salinity	0	127	1159	70
	Chl_*a*_SS	Chlorophyl *a* concentration (mg m^{-3})	6780	39406	148978	4
	Current _U_SS	Cross-shelf current velocity (m s^{-1})	48	318	1402	33
	Current _V_SS	Alongshore current velocity (m s^{-1})	48	318	1402	33
	Wind _U_SS	Cross-shelf wind velocity (m s^{-1})	34	473	6923	25
	Wind _V_SS	Alongshore wind velocity (m s^{-1})	34	473	6923	25

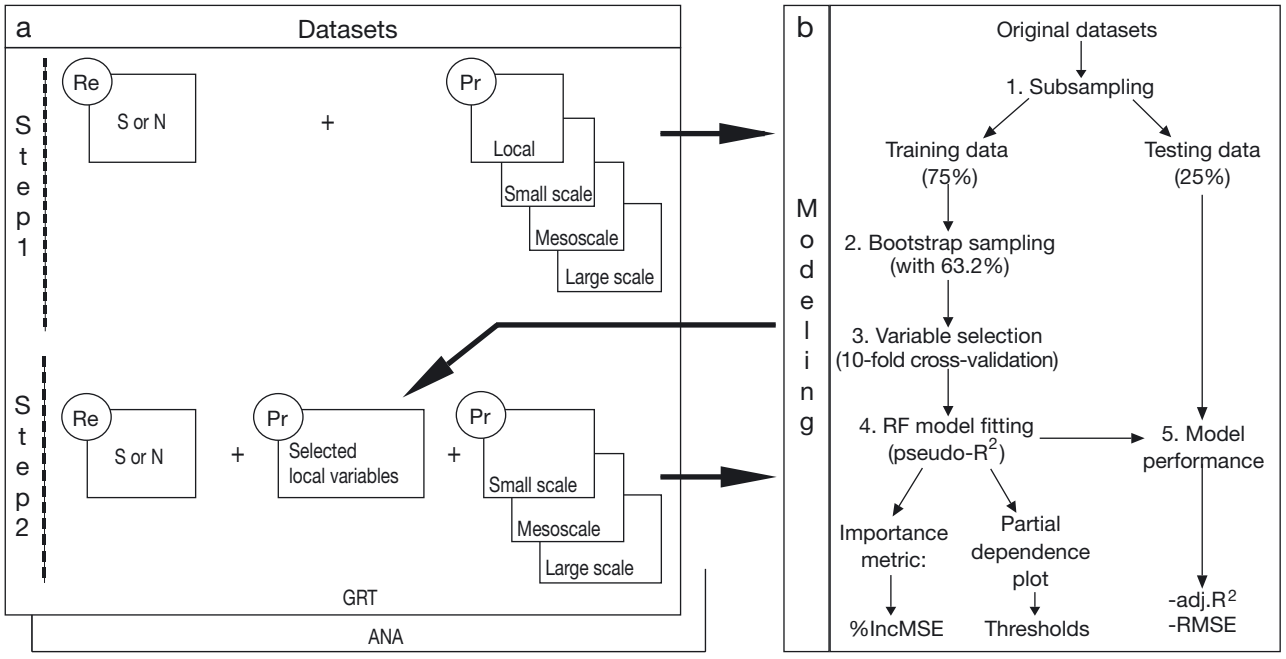

Fig. 2. Schematic representation of the modeling process: (a) structure of data sets for each site (GRT: great barrier reef of Toliara; ANA: Anakao reef), where Re are the response variables (S: species richness; N: abundance) and Pr are predictor variables (local: local variables; small-scale, meso-scale, and large-scale: remotely sensed oceanic conditions [RSOCs] at different scales). See Section 2.6 for the detailed descriptions of the steps and all the outputs. (b) Random forest (RF) modeling process, including the percentage of explained deviance (pseudo-R^2), importance measure based on decreasing importance of mean square error (%IncMSE), threshold values obtained from partial dependence plot of each predictor (thresholds), adjusted determination coefficient (adj.R^2), and root mean square error (RMSE)

trees were retained because this value provided the minimum error estimates and, after this threshold, the error remained stationary. The original data set was split into a training data set containing 75% of the original data and a testing data set containing 25% (Fig. 2b). RF models were, however, implemented using bootstrap samples (with replacement) that were generated from the training data. These bootstrap samples contained an average of 63.2% of the training data, following Han & Kamber (2006). RFs model goodness-of-fit based on a percentage of explained deviance (pseudo-R^2) obtained from models trained with bootstrap samples. The best-fitted models are those with the highest pseudo-R^2.

Multicollinearity and redundancy can influence RF model goodness-of-fit and its interpretability (Murphy et al. 2010), so variable selection is an important step before training RF models. We used the RF-RFE algorithm (Fig. 2b) that corresponds to the 'wrapper' selection method (Guyon & Elisseeff 2003, Genuer et al. 2010). According to Darst et al. (2018), this algorithm mitigates the impact of correlated predictors on a RF model by selecting the group of predictor variables that corresponds to the lowest root MSE of prediction (RMSE). Variable selection with the RF-RFE algorithm was applied with 10-fold cross-validation to avoid over-fitting (Reunanen 2003). The RMSE was then plotted to determine the best predictor set for fitting the RF model.

For RF models based on regression trees, the mean decrease accuracy (%IncMSE) is the most widely used, and more reliable, metric for measuring the relative importance of variables (Genuer et al. 2010). In this study, %IncMSE was used to rank the relative contribution of predictor variables to variations in species richness and abundance. Partial dependence plots were then used to describe the marginal effects of predictors on response variables (Friedman 2001) and to identify the threshold values corresponding to the first abrupt shift along predictor gradients (Cutler et al. 2007).

Testing model performance in predicting response variables was the last step of the modeling processes. Two metrics were used for assessing the predictive model performance. The first was the adjusted determination coefficient (adj.R^2), which measures the error between predicted and observed values in independent testing data (Fig. 2b). The adj.R^2 values vary between 0 and 1, with values close to 1 indicating high prediction performance and values close to 0 denot-

ing low prediction performance. The second metric was RMSE. RMSE is used to compare the predictive performance of models when their adj.R^2 values are equal. The lowest RMSE value denotes the best predictive model.

Data analyses and modeling were all performed with R v.3.5.1 (R Core Team 2018), using 'caret' package (v.6.0-84; Kuhn 2019) for variable selection with RF-RFE and training data bootstrapping and the 'randomForest' package (v.4.6-14; Breiman et al. 2018) for RF model fitting.

3. RESULTS

3.1. Temporal change of variables inside and outside of the bay

The characteristics of the water masses inside the bay differed between sites (Fig. 3a). Water masses were more turbid at ANA than at GRT (ranging from 1–9 and 3–15 m, respectively). Sea surface salinities varied widely at ANA (ranging from 31.5–35.7), while little variation (34.3–35.9) was recorded at GRT. No clear difference in SST was detected between the sites. For the RSOC variables outside of the bay (Fig. 3b), high SSTs and chl *a* concentrations occurred from January–April. Cross-shelf winds were mainly westerly (i.e. positive values), whereas alongshore winds were northerly (i.e. negative values). Similarly, alongshore currents were also mainly eastward, while alongshore currents were southward.

3.2. Species richness and abundance of post-larvae

A total of 277 light-trap samples, 141 at GRT and 136 at ANA, were obtained over the 3 austral warm seasons. The difference in sampling effort among sites was associated with logistic problems. A total of 238 species (114 to species level, 76 to Genus+BIN level, 15 to Family+BIN level, and 33 morphospecies) were caught, 190 at GRT and 165 at ANA, with 116 observed at both sites (Table S1 in the Supplement at www.int-res.com/articles/suppl/m645p125_supp.pdf). Maximum species richness and abundance was 29 species and 673 ind. trap^{-1} night^{-1} at GRT, and 17 species and 202 ind. trap^{-1} night^{-1} at ANA. The highest values of species richness and abundance were observed from November–December for both sites (Fig. 4). At both sites, catches were dominated by Apogonidae, Lethrinidae, Pomacentridae, Lutjanidae, Siganidae, Chaetodontidae, and Acanthuridae.

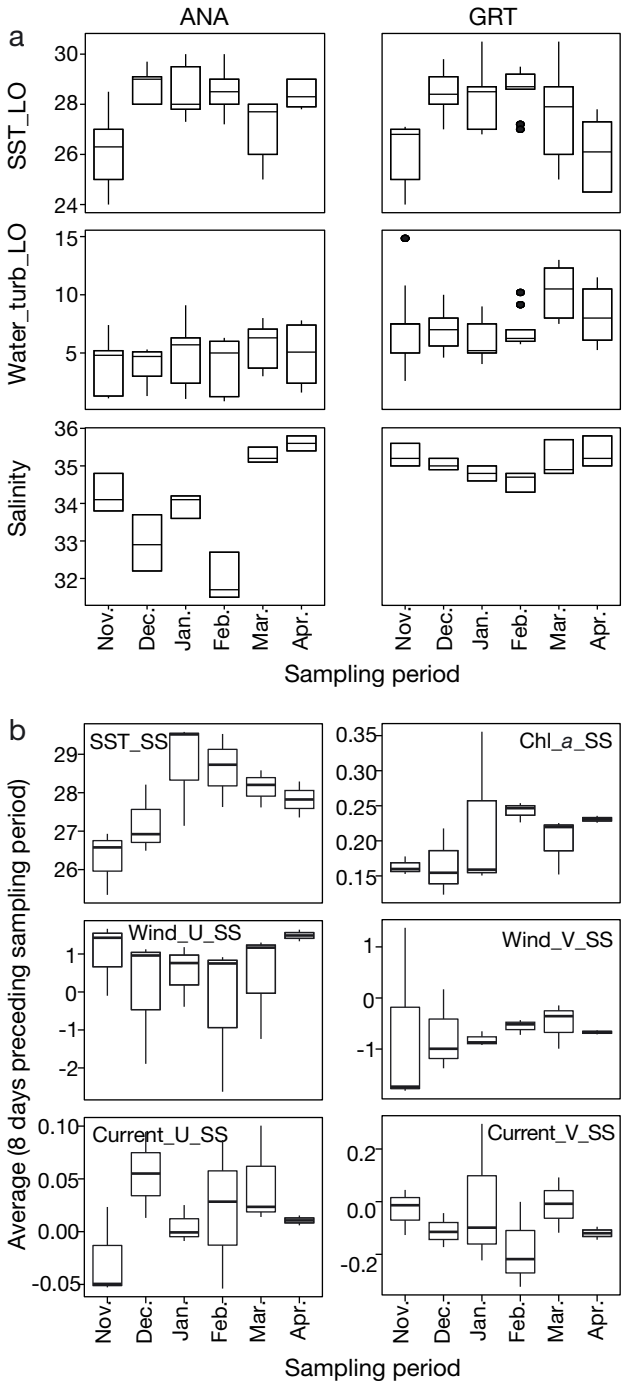

Fig. 3. Temporal variability in the characteristics of water masses from November to April (a) observed locally at the great barrier reef of Toliara (GRT) and Anakao reef (ANA), and (b) obtained through small-scale remotely sensed oceanic conditions (RSOC-SS). LO characteristics included sea surface temperature (SST_LO, in °C), water turbidity (Water_turb_LO, in m), and salinity. RSOC_SS included sea surface temperature (SST_SS, in °C), chlorophyll *a* (chl_*a*_SS, in mg m^{-3}), cross-shelf (Wind_U_SS, in m s^{-1}) and alongshore (Wind_V_SS, in m s^{-1}) wind speed, and cross-shelf (Current_U_SS, in m s^{-1}) and alongshore (Current_V_SS, in m s^{-1}) current speed

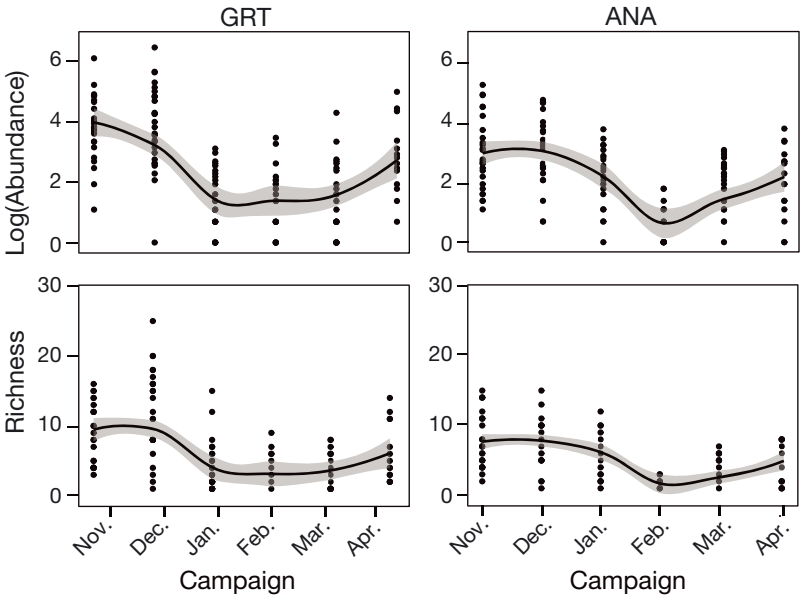

Fig. 4. Temporal change in post-larvae fish abundance and richness at the great barrier reef of Toliara (GRT) and Anakao reef (ANA) from November to April. Each black dot corresponds to 1 sample station^{-1} night^{-1}. Black smooth curves: 'loess' smoothing to foresee trends; gray shading: confidence interval around each smoothed curve. Refer to Fig. 1 for the locations of the GRT and ANA sites

3.3. Goodness-of-fit of models without and with local variables

Deviance explained by models (pseudo-R^2) trained separately with RSOCs and local variables ranged from 28–39% at GRT and 22–35% at ANA for species richness. The pseudo-R^2 for abundance ranged from 14–28% (GRT) and 33–50% (ANA) (Table 2). For both sites, models trained with local variables explained more variations in species richness and abundance than RSOC-based models, except for species richness in GRT, where variability was mainly explained by RSOC variables (Table 2). However, models combining RSOCs and local variables explained a larger part of the response variables than models trained separately. For species richness, deviance explained by models combining RSOC and local variables ranged from 51–53% for GRT and from 38–40% for ANA. Abundance models combining RSOCs and local variables explained 29–30 and 52–54% of model deviance for GRT and ANA, respectively (Table 2).

3.4. Performance of RF models for predicting species richness and abundance

The adj.R^2 and RMSE remained quite stable from small-scale to large-scale variables at ANA compared to GRT (Table 3). For species richness, the adj.R^2 remained similar from small- to large-scale variables (adj.R^2 = 0.66). RMSE values varied between 2.08 and 2.11 for ANA, whereas the adj.R^2 (between 0.68 and 0.61) and RMSE values (between 3.04 and 3.34) were more variable for GRT (Table 3). The adj.R^2 and RMSE from small- to large-scale variables of the abundance models varied from 0.58–0.57 and 17.65–18.30, respectively, for ANA while they ranged from 0.61–0.53 and 46.45–50.88 for GRT (Table 3). The lowest RMSE values for small-scale models suggest that small-scale variables more accurately predict species richness (2.08 for

Table 2. Percentage of explained deviance of random forest models (pseudo-R^2) fitted independently with local variables (LO; including sea surface temperature, water turbidity, cross-shelf and alongshore wind speed, trap setting and collection time, and pelagic fish) and remote sensing oceanic conditions (RSOCs; including sea surface temperature, chl *a*, cross-shelf and alongshore wind speed, and cross-shelf and alongshore current speed) at a small scale (RSOC-SS), meso scale (RSOC-MS), and large scale (RSOC-LS), as well as the combination of both (i.e. local variables with each RSOC at each scale) for explaining species richness and abundance. See Fig. 2 for a description of the modeling steps; ANA: Anakao reef; GRT: great barrier reef of Toliara

	Models: without local variables (Independent)				Models: with local variables (Combination)			
	Richness		Abundance		Richness		Abundance	
	GRT	ANA	GRT	ANA	GRT	ANA	GRT	ANA
LO	28	35	28	50				
RSOC-SS	39	23	16	33	51	40	30	52
RSOC-MS	39	23	14	34	51	38	30	54
RSOC-LS	39	22	15	35	53	40	29	54

Table 3. Random forest (RF) model performance, combining local variables (LO; including sea surface temperature, water turbidity, cross-shelf and alongshore wind speed, trap setting and collection time, and pelagic fish) and remote sensing oceanic conditions (RSOCs; including sea surface temperature, chl *a*, cross-shelf and alongshore wind speed, and cross-shelf and alongshore current speed) at a small-scale (SS), meso-scale (MS), and large-scale (LS). RF model performances are measured by the adjusted determination coefficient (adj.R^2) and the prediction root mean squared error (RMSE) for the great barrier reef of Toliara (GRT) and Anakao reef (ANA) using independent testing data (25%). See Fig. 2 for a description of the modeling steps

	Richness models				Abundance models			
	GRT		ANA		GRT		ANA	
	adj.R^2	RMSE	adj.R^2	RMSE	adj.R^2	RMSE	adj.R^2	RMSE
RSOC-SS+LO	0.68	3.04	0.66	2.08	0.61	46.45	0.58	17.65
RSOC-MS+LO	0.64	3.25	0.66	2.09	0.54	48.82	0.55	18.09
RSOC-LS+LO	0.61	3.34	0.66	2.11	0.53	50.88	0.57	18.30

ANA and 3.04 for GRT) and abundance (17.65 for ANA and 46.45 for GRT) of post-larvae fish than those that incorporate meso- and large-scale variables (Table 3). The adj.R^2 of the models with small-scale variables also suggested a small degree of variation in the models' performance among sites, with 0.66–0.68 for species richness and 0.58–0.61 for abundance. Finally, the model performances, visualized in Fig. 5, indicate the relevance of species richness and abundance predictions for the testing data set.

3.5. Importance and contribution of variables

Although the mechanisms driving species richness and abundance in post-larval samples differed among sites (Table 4), 5 common important variables were identified: SST (SST_SS), chl *a* concentration (chl_*a*_SS), alongshore wind speed (Wind_V_SS), water turbidity (Water_turb_LO), and trap-setting time (Dif_set_LO) (Table 4). However, at GRT, species richness was explained by the local alongshore wind speed (Wind_V_LO) instead of Dif_set_LO (Table 4). The partial dependence plots indicated that high northerly Wind_V_SS (i.e. negative values; >1.5 m s^{-1}), low SST (<27°C), and low chl *a* (<0.22 mg m^{-3}) were generally related to high species richness and abundance values at the 2 sites (Fig. 6). Surprisingly, southerly Wind_V_SS (positive values) and low chl *a* (<0.15 mg m^{-3}) predicted high abundance values only at GRT (Fig. 6). Regarding local variables, partial dependence plots indicated that turbid water (i.e. low value of Water_turb_LO) and trap-setting time 0.5 h before sunset (negative value) were associated with high species richness and abundance (Fig. 6). Moreover, high abundance values would be expected when the traps were set 1.5 h after sunset (positive value) at GRT (Fig. 6). Current speed along the 2 directions was only important for predicting species richness and abundance at ANA. However, the importance of current speeds always ranked lower than SST, chl *a*, and wind speed. Finally, high species richness and abundance was predicted for specific current conditions: eastward (positive value of Current_U_SS) and low westward current speed (negative value of Current_U_SS), and northward (positive value of Current_V_SS) and low southward current speed (negative value of Current_V_SS).

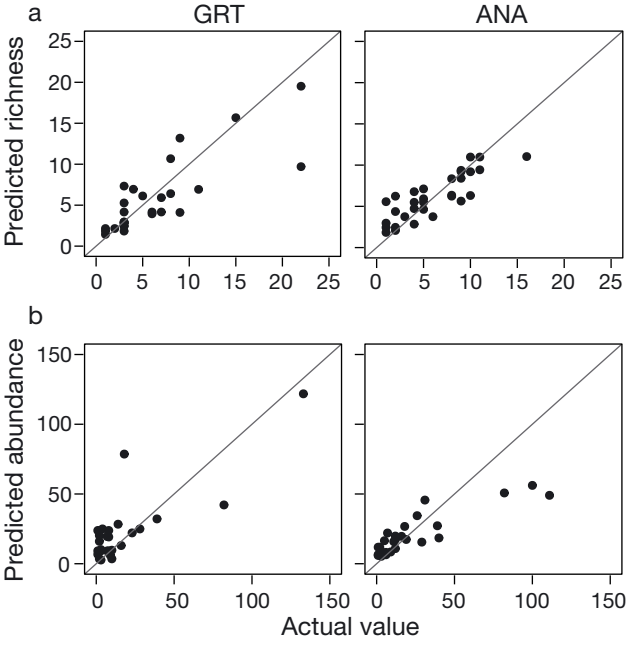

Fig. 5. (a) Species richness and (b) abundance of post-larvae predicted from the random forest models for the independent validation data set (testing data) at the great barrier reef of Toliara (GRT) and Anakao reef (ANA). Diagonal lines: the 1:1 line. Black dots: observations

Table 4. Importance of local (LO) and remote sensing oceanic condition (RSOC) variables based on mean square error (%IncMSE) for explaining the variation in richness and abundance of post-larvae at the great barrier reef of Toliara (GRT) and Anakao reef (ANA). Rank: the rank of each variable based on the importance score (%IncMSE), where 1 corresponds to the highest importance score for a given response variable. LO variables included trap-setting time (Dif_set_LO, in hours), water turbidity (Water_turb_LO, in m), and alongshore wind speed (Wind_V_LO, in m s^{-1}). RSOCs included sea surface temperature (SST_SS, in °C), Chlorophyll *a* (Chl_*a*_SS, in mg m^{-3}), cross-shelf (Wind_U_SS, in m s^{-1}) and alongshore (Wind_V_SS, in m s^{-1}) wind speed, and cross-shelf (Current_U_SS, in m s^{-1}) and alongshore (Current_V_SS, in m s^{-1}) current speeds. (–) variables not retained for model fitting

	Richness				Abundance			
	GRT		ANA		GRT		ANA	
LO & RSOC	%IncMSE	Rank	%IncMSE	Rank	%IncMSE	Rank	%IncMSE	Rank
Wind_V_SS	19.24	4	17.39	2	16.94	1	12.60	1
Chl_*a*_SS	13.20	5	19.38	1	11.15	4	8.19	7
SST_SS	23.59	2	13.50	5	9.55	5	10.37	3
Water_turb_LO	25.24	1	15.34	4	16.01	2	9.10	5
Dif_set_LO	–	–	15.78	3	12.07	3	10.08	4
Wind_V_LO	22.36	3	–	–	–	–	–	–
Current_V_SS	–	–	10.84	7	–	–	8.54	6
Current_U_SS	–	–	10.18	8	–	–	7.91	8
Wind_U_SS	–	–	11.07	6	–	–	11.17	2

4. DISCUSSION

To our knowledge, this is the first study to predict species richness and abundance of post-larvae fish based on (1) precise species identification using DNA barcoding, (2) a machine learning technique (here, RF models), and (3) RSOCs extracted over several days preceding sampling and at different scales. Our findings revealed that RF models combining local and RSOC-SS variables more accurately predicted species richness and abundance of post-larval fish. Only a few previous studies have used RSOCs to model marine communities. For example, Albajes-Eizagirre et al. (2011) extracted RSOC variables in the days just prior to sampling to determine factors affecting the occurrence of jellyfish blooms in the Catalan coast of the northwestern Mediterranean Sea. The RSOC variables obtained in the days just prior to sampling used by these authors included water mass (SST and salinity) and chl *a* as a proxy for food availability. Using a computational intelligence algorithm, the authors found a strong relationship between the minimum value of salinity and the appearance of jellyfish blooms. Avendaño-Ibarra et al. (2013) used RSOCs extracted during sampling in a redundancy analysis to describe larval fish abundance variance in the Gulf of California. They demonstrated the influence of SST, salinity, and chl *a* on the variation in abundance of larvae in plankton nets. Complementing these previous studies, the present research using RF models was able to detect thresholds in local and RSOC-SS variables for explaining the variation in species richness and abundance. The technique we adopted, combining molecular tools, machine learning, and RSOCs obtained in the days just prior to sampling and at different scales, can be spatially and temporally transferable for addressing questions in other regions or habitats.

4.1. Importance of local variables for predicted species richness and abundances

The high importance of local variables highlights that post-larvae are sensitive to the conditions when the light traps are set. Indeed, high species richness and abundance values were observed when traps were set 30 min or more before sunset (see Dif_set-LO in Fig. 6). To our knowledge, no previous studies have investigated the effect of trap-setting time on species richness and abundance in light traps. These high values of species richness and abundance observed before sunset may be due to the interaction between trap-setting time and tidal currents. Indeed, post-larvae can be transported by onshore tidal currents, with maximum transport occurring before high tide (Sponaugle & Cowen 1996). In the present study, sampling took place during the new moon period, when high tide always occurs before sunset. However, the high species richness and abundance observed in the light traps could also be associated with larval fish activities because they migrate vertically throughout the water surface around sunset to feed on zooplankton (McLaren & Avendaño 1995).

High species richness and abundance were also detected when the water turbidity was high (see Fig. 6). This finding was unexpected. We anticipated that post-larval species richness and abundance would be

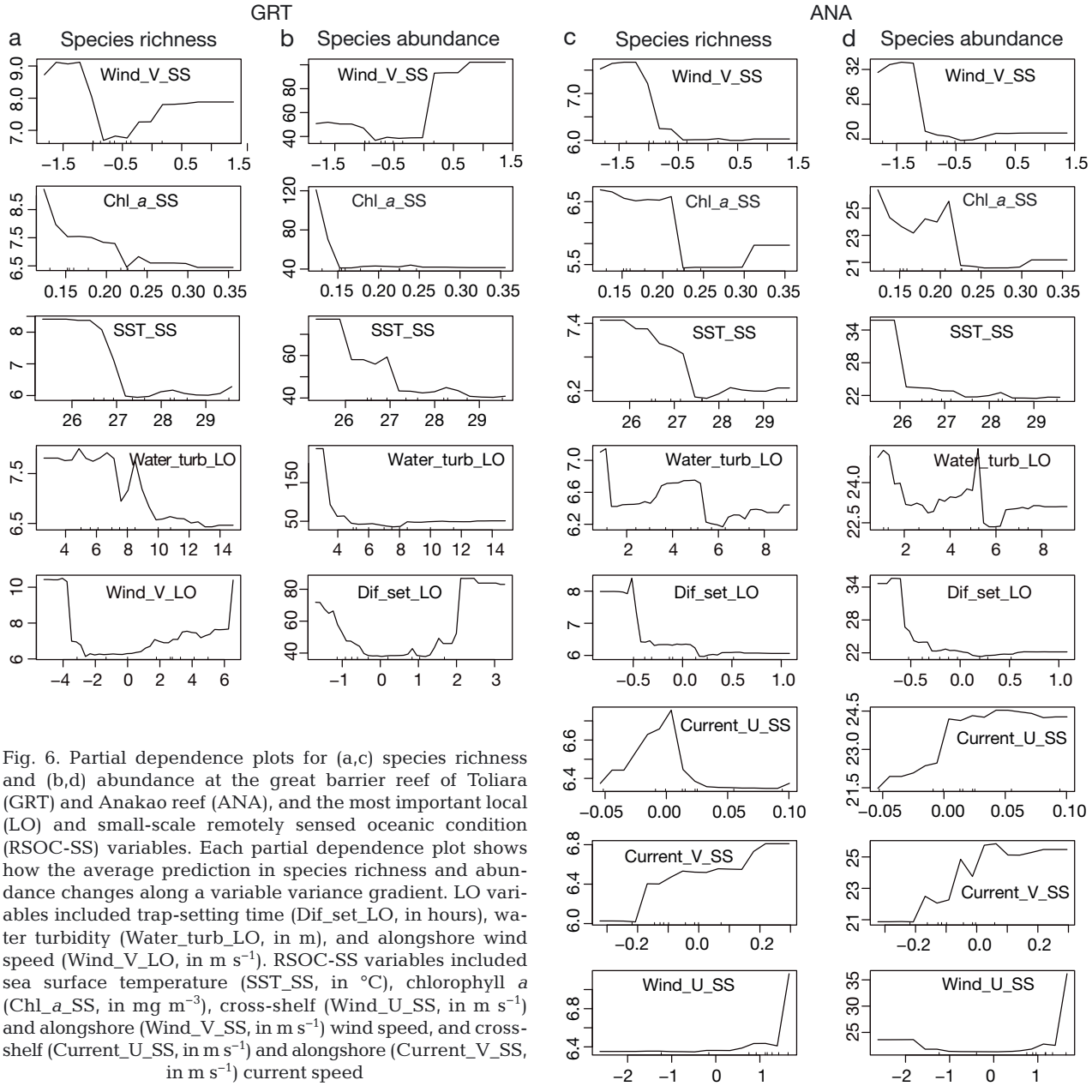

Fig. 6. Partial dependence plots for (a,c) species richness and (b,d) abundance at the great barrier reef of Toliara (GRT) and Anakao reef (ANA), and the most important local (LO) and small-scale remotely sensed oceanic condition (RSOC-SS) variables. Each partial dependence plot shows how the average prediction in species richness and abundance changes along a variable variance gradient. LO variables included trap-setting time (Dif_set_LO, in hours), water turbidity (Water_turb_LO, in m), and alongshore wind speed (Wind_V_LO, in m s^{-1}). RSOC-SS variables included sea surface temperature (SST_SS, in °C), chlorophyll *a* (Chl_*a*_SS, in mg m^{-3}), cross-shelf (Wind_U_SS, in m s^{-1}) and alongshore (Wind_V_SS, in m s^{-1}) wind speed, and cross-shelf (Current_U_SS, in m s^{-1}) and alongshore (Current_V_SS, in m s^{-1}) current speed

low when water turbidity was high, as high water turbidity usually reduces the efficiency of the light traps (Hickford & Schiel 1999, Lindquist & Shaw 2005) because the phototactic responses of post-larvae decrease with low light intensities (Stearns et al. 1994). The high species richness and abundance observed during high water turbidity in the current study may be linked to strong winds that increase water turbidity (Cho 2007) and generate surface currents (Kingsford & Finn 1997) in coastal waters. A strong wind effect during sampling (Wind_V_LO) was observed for species richness in GRT (Fig. 6a). In such environmental conditions, light traps set in more turbid waters may be able to capture high species richness and abundance by including non-phototactic species that are driven by wind-generated currents.

4.2. Combination of local variables with RSOCs: useful for predicting species richness and abundance

RF models combining RSOCs and local variables more accurately predicted species richness and abundance. This may be related to the spatial distribution of post-larvae, whereby there is higher spe-

cies richness and abundance in near-shore waters than in oceanic waters (Hsieh et al. 2010). Coastal oceanic conditions can therefore be an important driver for larval dispersal. For example, circular coastal currents such as tidal eddies may retain post-larval fish due to their recirculating properties, thereby reducing post-larval potential dispersion (Kingsford & Finn 1997, Burgess et al. 2007).

RF model performance in predicting species richness and abundance differed between sites. As indicated by the adj.R^2, RF models better predicted species richness and abundance at GRT than at ANA. Such differences could be explained by the contrasting surface water characteristics. Anakao reef is located about 10 km from the mouth of the Onilahy River, which is influenced by permanent river discharge that averages 145 m^3 s^{-1} and reaches up to 1500 m^3 s^{-1} during flood events (R. Arfi et al. unpubl. data). Sentinel-2A images (European Spatial Agency) have identified that the turbidity of the Onilahy River plume can easily reach Anakao reef (T. Jallera unpubl. data). This explains the high water turbidity and the low sea surface salinity recorded at ANA (Fig. 3) compared to GRT (Fig. 3). GRT is only occasionally under the influence of flow from the Fiherenana River (R. Arfi et al. unpubl. data). Moreover, the massive and dominant cross-reef tidal inflow can renew up to 80% of the water at GRT (Chevalier et al. 2014). Nevertheless, the restricted variation in the RF models' performance between sites reflects the consistency of RF models in predicting species richness and abundance despite the spatial variability in surface water characteristics.

4.3. Variables contributing to predicted species richness and abundance

RF models predicted high species richness and abundance when the average SST remained under 27.0°C. High values of abundance were also predicted when chl *a* remained under 0.22 mg m^{-3}. These threshold values were observed between November and December for the 3 sampling seasons (Fig. 3b). This period corresponds to the peak in species richness and abundance, indicating reproductive activity of most reef fish species (Reynalte-Tataje et al. 2012). Interestingly, chl *a* values above this threshold (observed between January and April) coincided with low species richness and abundance (Fig. 4). This was unexpected because high chl *a* concentrations should lead to high post-larval species richness (Leathwick et al. 2006) and abundance (Falfán-Vázquez et al. 2008). This paradoxical finding may be related to tropical storms that always occur between January and April in Madagascar; tropical storms enhance chl *a* concentrations (Lin et al. 2003), but they are also known to weaken fish reproductive success by transporting larvae away from their settlement areas (Morsink 2018).

Despite the presence of prevailing southerly alongshore winds in southwestern Madagascar, the strong northerly alongshore winds were the best predictor of species richness and abundance. The northerly alongshore winds determine the warm and wet seasons in the region (R. Arfi et al. unpubl. data) and play a role in the intensification of the southward Southwestern Madagascar Alongshore Coastal Currents (SMACC; Ramanantsoa et al. 2018). Indeed, winds are major driving forces for ocean currents (Cowen & Sale 2002), which partly explains the tight correlation between these 2 factors. According to Murphy et al. (2010), the presence of strong relationships between variables reduces the RF models' goodness-of-fit. This explains why the southward alongshore currents were not retained in RF models for GRT. This strong relationship between winds and currents is also associated with the better statistical fit of the alongshore wind (Wind_V_SS) for GRT compared to ANA (Table 4). Thus, the lesser importance of the alongshore wind at ANA may explain why southward alongshore currents (Current_V_SS) were retained in the RF models despite their strong relationship, when the importance of alongshore currents was ranked very low. Moreover, strong currents (>0.2 m s^{-1}) are associated with low values in species richness and abundance (Fig. 6), while the SMACC speed can reach 0.3 m s^{-1} during the austral warm season (Ramanantsoa et al. 2018). These findings support the fact that wind-driven surface currents constitute one of the major factors explaining the transport and concentration of fish larvae (Schlaefer et al. 2018).

4.4. Possible contribution of other factors

The comparison of predicted and observed species richness and abundance highlights that additional factors should be considered to improve RF predictive model performance. River flow, which contributed to the difference between the present study sites, may be an important candidate predictor. Models involving freshwater, sediment, and organic material driven by the river should be developed for model performance improvement. In addition, larval fish behavior may be also important because it may influ-

ence species richness and abundance. Active larval fish behavior, particularly shoreward movements which enable them to swim against water flow, helps with their retention near coastal habitats (Tricklebank et al. 1992, Tzeng & Wang 1993, Hickford & Schiel 1999, Leis 2007). Larval fish species are also known to migrate vertically (Paris & Cowen 2004, Irisson et al. 2009, Huebert et al. 2011, Berenshtein et al. 2018). These vertical migrations, as observed in Pomacentridae (Paris & Cowen 2004) and flatfish (Bailey et al. 2005), help the larvae avoid strong surface currents (Shanks 2009). The most dominant fish families in the present study (Siganidae, Lutjanidae, Pomacentridae, Lethrinidae, Chaetodontidae, and Acanthuridae) are able to swim faster than 0.3 m s^{-1}, reaching speeds up to 0.65 m s^{-1} (Fisher 2005). This suggests an ability of the post-larvae fish to maintain their position by swimming against currents such as the SMACC in southwest Madagascar (Ramanantsoa et al. 2018).

In conclusion, the present work demonstrates the usefulness of DNA barcoding techniques for inventorying the biodiversity of reef fish in their early life stages. This molecular tool permitted post-larval fish identification, enabling precise and accurate calculations of species richness, which are essential for RF modeling. RF models performed well in predicting species richness and abundance. The weak spatial variability in RF performances suggests that these models were also consistent. RF models provided information on how alongshore wind speed, SST, and chl *a* concentration best explained the post-larval supply in terms of species richness and abundance. RF models, assessing the differences between predicted and observed species richness and abundances, should be considered when calculating future changes in tropical fish post-larval supply. However, the thresholds detected for each variable may change over time, so they would need to be validated over a longer time series. Such long-term monitoring would take into account (1) the effect on El Niño-Southern Oscillation on the characteristics of surface water, hydrodynamic conditions, and food availability (Hoareau et al. 2012); and (2) the contribution of tropical storms, which can transport fish larvae over large distances (Reid et al. 2016). The validation of post-larval supply predictions through long-term monitoring will be useful for fisheries management. Indeed, post-larval supply is a part of spawning stock size (Moser & Watson 1990) and can be used as a biological index for predicting juvenile fish recruitment (Milicich et al. 1992, Stige et al. 2013). Accordingly, post-larval supply success in coastal habitats is expected to shape juvenile and adult fish communities.

Acknowledgements. We thank J. J. Marcellin, D. Fiandria, R. Tsipy, Tovondrainy, and Noelson for help with field collections and the staff of the GenSeq technical facilities of the LabEx 'Centre Méditerranéen de l'Environnement et de la Biodiversité' of the Université de Montpellier for sequencing collected samples. We also thank Camille DeSisto and Dr. Christopher Golden for help in reviewing the language accuracy. This work was financially supported by the Critical Ecosystem Partnership Fund (CEPF/IH.SM-MG 66 341), project POE 2.10 POCT FED – FEDER 'Biodiversité de l'Océan Indien', the French National Research Institute for Sustainable Development (JEAI-ACOM project), and Institut Halieutique et des Sciences Marines (materials support). This a contribution from 'Laboratoire Mixte International MIKAROKA', Institut Halieutique et des Sciences Marines - Centre National de Recherche Océanographique – IRD - University of La Réunion – CNRS - Ifremer.

LITERATURE CITED

Albajes-Eizagirre A, Romero L, Soria-Frisch A, Vanhellemont Q (2011) Jellyfish prediction of occurrence from remote sensing data and a non-linear pattern recognition approach. Proc SPIE 8174:18

Anderson TW, Bartels CT, Hixon MA, Bartels E, Carr MH, Shenker JM (2002) Current velocity and catch efficiency in sampling settlement-stage larvae of coral-reef fishes. Fish Bull 100:404–413

Avendaño-Ibarra R, Godínez-Domínguez E, Aceves-Medina G, González-Rodríguez E, Trasviña A (2013) Fish larvae response to biophysical changes in the Gulf of California, Mexico (winter–summer). J Mar Biol 2013:1–17

Bailey KM, Nakata H, Van der Veer HW (2005) The planktonic stages of flatfishes: physical and biological interactions in transport processes. In: Gibson RN (ed) Flatfishes. Blackwell Science, Oxford, p 94–119

Berenshtein I, Paris CB, Gildor H, Fredj E, Amitai Y, Lapidot O, Kiflawi M (2018) Auto-correlated directional swimming can enhance settlement success and connectivity in fish larvae. J Theor Biol 439:76–85

Breiman L, Cutler A, Liaw A, Wiener M (2018) Breiman and Cutler's random forests for classification and regression, version 4.6-14. www.stat.berkeley.edu/~breiman/RandomForests/ (accessed 8 May 2018)

Bruggemann JH, Rodier M, Guillaume MMM, Andréfouët S and others (2012) Wicked social ecological problems forcing unprecedented change on the latitudinal margins of coral reefs: the case of southwest Madagascar. Ecol Soc 17:47

Burgess SC, Kingsford MJ, Black KP (2007) Influence of tidal eddies and wind on the distribution of presettlement fishes around One Tree Island, Great Barrier Reef. Mar Ecol Prog Ser 341:233–242

Chen LC, Lan KW, Chang Y, Chen WY (2018) Summer assemblages and biodiversity of larval fish associated with hydrography in the northern South China Sea. Mar Coast Fish 10:467–480

Chevalier C, Devenon JL, Rougier G, Blanchot J (2014) Hydrodynamics of the Toliara reef lagoon (Madagascar): example of a lagoon influenced by waves and tides. J Coast Res 31:1403–1417

Cho HJ (2007) Effects of prevailing winds on turbidity of a shallow estuary. Int J Environ Res Public Health 4:185–192

Collet A, Durand JD, Desmarais E, Cerqueira F, Cantinelli T, Valade P, Ponton D (2018) DNA barcoding post-larvae

can improve the knowledge about fish biodiversity: an example from La Reunion, SW Indian Ocean. Mitochondrial DNA A DNA Mapp Seq Anal 29:905–918

Cowen RK, Sale PF (2002) Larval dispersal and retention and consequences for population connectivity. In: Sale PF (ed) Coral reef fishes: dynamics and diversity in a complex ecosystem. Academic Press, San Diego, CA, p 149–170

Cowen RK, Gawarkiewicz G, Pineda J, Thorrold SR, Werner FE (2007) Population connectivity in marine systems: an overview. Oceanography (Wash DC) 20:14–21

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

Darst BF, Malecki KC, Engelman CD (2018) Using recursive feature elimination in random forests to account for correlated variables in high dimensional data. BMC Genet 19:65–70

ESR (Earth & Space Research) (2009) OSCAR third degree resolution ocean surface currents. NASA Physical Oceanography Distributed Active Archive Center, Pasadena, CA

Falfán-Vázquez E, Ordóñez-López U, Órnelas-Roa M (2008) Spatial variation of snapper and grouper larvae in Yucatan Shelf. Hidrobiologica 18:69–76

Fisher R (2005) Swimming speeds of larval coral reef fishes: impacts on self-recruitment and dispersal. Mar Ecol Prog Ser 285:223–232

França S, Cabral HN (2015) Predicting fish species richness in estuaries: Which modeling technique to use? Environ Model Softw 66:17–26

França S, Vasconcelos RP, Fonseca VF, Tanner SE, Reis-Santos P, Costa MJ, Cabral HN (2012) Predicting fish community properties within estuaries: influence of habitat type and other environmental features. Estuar Coast Shelf Sci 107:22–31

Francis MP, Morrison MA, Leathwick J, Walsh C, Middleton C (2005) Predictive models of small fish presence and abundance in northern New Zealand harbours. Estuar Coast Shelf Sci 64:419–435

Francis MP, Morrison MA, Leathwick J, Walsh C (2011) Predicting patterns of richness, occurrence and abundance of small fish in New Zealand estuaries. Mar Freshw Res 62:1327–1341

Frantine-Silva W, Sofia SH, Orsi ML, Almeida FS (2015) DNA barcoding of freshwater ichthyoplankton in the Neotropics as a tool for ecological monitoring. Mol Ecol Resour 15:1226–1237

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31:2225–2236

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. Morgan Kaufmann, San Francisco, CA

Harris SA, Cyrus DP, Beckley LE (2001) Horizontal trends in larval fish diversity and abundance along an ocean–estuarine gradient on the Northern KwaZulu-Natal Coast, South Africa. Estuar Coast Shelf Sci 53:221–235

Hickford MJH, Schiel DR (1999) Evaluation of the performance of light traps for sampling fish larvae in inshore temperate waters. Mar Ecol Prog Ser 186:293–302

Hoareau TB, Boissin E, Berrebi P (2012) Evolutionary history of a widespread Indo-Pacific goby: the role of Pleistocene sea-level changes on demographic contraction/expansion dynamics. Mol Phylogenet Evol 62:566–572

Hsieh CH, Reiss C, Watson W, Allen MJ and others (2005) A comparison of long-term trends and variability in populations of larvae of exploited and unexploited fishes in the Southern California region: a community approach. Prog Oceanogr 67:160–185

Hsieh HY, Lo WT, Liu DC, Su WC (2010) Influence of hydrographic features on larval fish distribution during the south-westerly monsoon in the waters of Taiwan, western North Pacific Ocean. J Fish Biol 76:2521–2539

Huebert KB, Cowen RK, Sponaugle S (2011) Vertical migrations of reef fish larvae in the Straits of Florida and effects on larval transport. Limnol Oceanogr 56:1653–1666

Irisson JO, Guigand C, Paris CB (2009) Detection and quantification of marine larvae orientation in the pelagic environment. Limnol Oceanogr Methods 7:664–672

Jackson JB, Kirby MX, Berger WH, Bjorndal KA and others (2001) Historical overfishing and the recent collapse of coastal ecosystems. Science 293:629–637

James MK, Armsworth PR, Mason LB, Bode L (2002) The structure of reef fish metapopulations: modelling larval dispersal and retention patterns. Proc R Soc B 269: 2079–2086

Jaonalison H, Mahafina J, Ponton D (2016) Fish post-larvae assemblages at two contrasted coral reef habitats in southwest Madagascar. Reg Stud Mar Sci 6:62–74

Jenkins GP, King D (2006) Variation in larval growth can predict the recruitment of a temperate, seagrass-associated fish. Oecologia 147:641–649

Jenkins GP, Black KP, Keough MJ (1999) The role of passive transport and the influence of vertical migration on the presettlement distribution of a temperate, demersal fish: numerical model predictions compared with field sampling. Mar Ecol Prog Ser 184:259–271

JPL (Jet Propulsion Laboratory) (2015) GHRSST Level 4 MUR global foundation sea surface temperature analysis. NASA Physical Oceanography Distributed Active Archive Center, Pasadena, CA

Kingsford M, Finn M (1997) The influence of phase of the moon and physical processes on the input of presettlement fishes to coral reefs. J Fish Biol 51:176–205

Klemas V (2012) Remote sensing of environmental indicators of potential fish aggregation: an overview. Baltica 25: 99–112

Knudby A, Brenning A, LeDrew E (2010) New approaches to modeling fish–habitat relationships. Ecol Modell 221: 503–511

Ko HL, Wang YT, Chiu TS, Lee MA and others (2013) Evaluating the accuracy of morphological identification of larval fishes by applying DNA barcoding. PLOS ONE 8:e53451

Koehl MAR, Strother JA, Reidenbach MA, Koseff JR, Hadfield MG (2007) Individual-based model of larval transport to coral reefs in turbulent, wave-driven flow: behavioral responses to dissolved settlement inducer. Mar Ecol Prog Ser 335:1–18

Koslow JA, Wright M (2016) Ichthyoplankton sampling design to monitor marine fish populations and communities. Mar Policy 68:55–64

Koslow JA, Goericke R, Watson W (2013) Fish assemblages in the Southern California Current: relationships with climate, 1951–2008. Fish Oceanogr 22:207–219

Kuhn M (2019) Classification and regression training, version 6.0-84. https://github.com/topepo/caret/ (accessed 8 May 2018)

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar Ecol Prog Ser 321:267–281

Leis J (2007) Behaviour as input for modeling dispersal of fish larvae: behaviour, biogeography, hydrodynamics, ontogeny, physiology and phylogeny meet hydrography. Mar Ecol Prog Ser 347:185–193

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Lin I, Liu WT, Wu CC, Wong GTF and others (2003) New evidence for enhanced ocean primary production triggered by tropical cyclone. Geophys Res Lett 30:1718

Lindquist DC, Shaw RF (2005) Effects of current speed and turbidity on stationary light-trap catches of larval and juvenile fishes. Fish Bull 103:438–444

Marchetti MP, Esteban E, Limm M, Kurth R (2004) Evaluating aspects of larval light trap bias and specificity in the northern Sacramento river system: Do size and color matter? Am Fish Soc Symp 39:269–279

Mavruk S, Bengil F, Yüksek A, Özyurt CE, Kiyağa VB, Avşar D (2018) Intra-annual patterns of coastal larval fish assemblages along environmental gradients in the northeastern Mediterranean. Fish Oceanogr 27:232–245

McLaren IA, Avendaño P (1995) Prey field and diet of larval cod on Western Bank, Scotian Shelf. Can J Fish Aquat Sci 52:448–463

Meissner T, Wentz FJ (2016) Remote sensing systems SMAP level 3 sea surface salinity standard mapped image 8-day running mean V2.0 validated dataset. NASA Physical Oceanography Distributed Active Archive Center, Pasadena, CA

Milicich M, Meekan MG, Doherty PJ (1992) Larval supply: a good predictor of recruitment of three species of reef fish (Pomacentridae). Mar Ecol Prog Ser 86:153–166

Morsink K (2018) Hurricanes, typhoons, and cyclones. http://ocean.si.edu/planet-ocean/waves-storms-tsunamis/hurricanes-typhoons-and-cyclones (accessed 28 May 2019)

Moser HG, Watson W (1990) Distribution and abundance of early life history stages of the California halibut, *Paralichthys californicus* and comparisons with the fantail sole, *Xystreurys liolepis*. Fish Bull 174:31–84

Murphy MA, Evans JS, Storfer A (2010) Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. Ecology 91:252–261

NASA Ocean Biology Processing Group (2017) MODIS-Aqua level 3 mapped chlorophyll data version R2018.0. NASA Ocean Biology Distributed Active Archive Center, Greenbelt, MD

Nicolas D, Lobry J, Lepage M, Sautour B and others (2010) Fish under influence: a macroecological analysis of relations between fish species richness and environmental gradients among European tidal estuaries. Estuar Coast Shelf Sci 86:137–147

Paris CB, Cowen RK (2004) Direct evidence of a biophysical retention mechanism for coral reef fish larvae. Limnol Oceanogr 49:1964–1979

Potts JM, Elith J (2006) Comparing species abundance models. Ecol Modell 199:153–163

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Ramanantsoa JD, Penven P, Krug M, Gula J, Rouault M (2018) Uncovering a new current: the southwest Madagascar coastal current. Geophys Res Lett 45:1930–1938

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLOS ONE 8:e66213

Reid K, Crochelet E, Bloomer P, Hoareau TB (2016) Investigating the origin of vagrant dusky groupers, *Epinephelus marginatus* (Lowe, 1834), in coastal waters of Réunion Island. Mol Phylogenet Evol 103:98–103

Reunanen J (2003) Overfitting in making comparisons between variable selection methods. J Mach Learn Res 3: 1371–1382

Reynalte-Tataje DA, Zaniboni-Filho E, Bialetzki A, Agostinho AA (2012) Temporal variability of fish larvae assemblages: influence of natural and anthropogenic disturbances. Neotrop Ichthyol 10:837–846

Robertson DR, Green DG, Victor BC (1988) Temporal coupling of production and recruitment of larvae of a Caribbean reef fish. Ecology 69:370–381

Schlaefer JA, Wolanski E, Lambrechts J, Kingsford MJ (2018) Wind conditions on the Great Barrier Reef influenced the recruitment of snapper (*Lutjanus carponotatus*). Front Mar Sci 5:193

Shanks AL (2009) Pelagic larval duration and dispersal distance revisited. Biol Bull (Woods Hole) 216:373–385

Sponaugle S, Cowen RK (1996) Nearshore patterns of coral reef fish larval supply to Barbados, West Indies. Mar Ecol Prog Ser 133:13–28

Stearns DE, Holt GJ, Forward RB Jr, Pickering PL (1994) Ontogeny of phototactic behavior in red drum larvae (Sciaenidae: *Sciaenops ocellatus*). Mar Ecol Prog Ser 104:1–11

Stige LC, Hunsicker ME, Bailey KM, Yaragina NA, Hunt GL Jr (2013) Predicting fish recruitment from juvenile abundance and environmental indices. Mar Ecol Prog Ser 480:245–261

Takahashi M, Watanabe Y (2004) Growth rate-dependent recruitment of Japanese anchovy *Engraulis japonicus* in the Kuroshio–Oyashio transitional waters. Mar Ecol Prog Ser 266:227–238

Tricklebank KA, Jacoby CA, Montgomery JC (1992) Composition, distribution and abundance of neustonic ichthyoplankton off northeastern New Zealand. Estuar Coast Shelf Sci 34:263–275

Tzeng WN, Wang YT (1993) Hydrography and distribution dynamics of larval and juvenile fishes in the coastal waters of the Tanshui River estuary, Taiwan, with reference to estuarine larval transport. Mar Biol 116:205–217

Vasconcelos RP, Henriques S, França S, Pasquaud S, Cardoso I, Laborde M, Cabral HN (2015) Global patterns and predictors of fish species richness in estuaries. J Anim Ecol 84:1331–1341

Wentz FJ, Ricciardulli L, Gentemann C, Meissner T, Hilburn KA, Scott J (2013) Remote sensing systems Coriolis Windsat [Daily]. Environmental Suite on 0.25 deg grid, version 7.0.1. Remote Sensing Systems, Santa Rosa, CA

Wernberg T, Smale DA, Tuya F, Thomsen MS and others (2013) An extreme climatic event alters marine ecosystem structure in a global biodiversity hotspot. Nat Clim Chang 3:78–82

Wilson D (2001) Patterns of replenishment of coral-reef fishes in the nearshore waters of the San Blas Archipelago, Caribbean Panama. Mar Biol 139:735–753

Editorial responsibility: Myron Peck, Hamburg, Germany

Submitted: July 16, 2019; Accepted: June 8, 2020
Proofs received from author(s): July 3, 2020