شیلات، مجله منابع طبیعی ایران
دوره ۷۶، شماره ۱، بهار ۱۴۰۲
صفحات ۳۸-۲۷

مقایسهٔ عملکرد مدل‌های خطی تعمیم‌یافته (GLM) و جنگل تصادفی (RF) در پیش‌بینی توزیع صید ماهی سفید (*Rutilus frisii*)

فاتح معزی[۱]، هادی پورباقر[۲]*، سهیل ایگدری[۲]، جهانگیر فقهی[۳]

۱. دانش آموخته دکتری بوم‌شناسی آبزیان، دانشکده منابع طبیعی، دانشگاه تهران، کرج، ایران
۲. دانشیار گروه شیلات، دانشکده منابع طبیعی، دانشگاه تهران، کرج، ایران
۳. استاد گروه مهندسی جنگلداری و اقتصاد جنگل، دانشکده منابع طبیعی، دانشگاه تهران، کرج، ایران

تاریخ ارسال: ۱۴۰۱/۰۳/۱۸ تاریخ پذیرش: ۱۴۰۱/۰۶/۲۴

چکیده

هدف از انجام مطالعهٔ حاضر، ارزیابی عملکرد مدل‌های خطی تعمیم‌یافته (GLM) و جنگل تصادفی (RF) در پیش‌بینی توزیع صید ماهی سفید دریای خزر (*Rutilus frisii*) بود. بدین‌منظور، داده‌های صید در واحد تلاش (CPUE) ماهی سفید به‌عنوان متغیر اصلی و داده‌های سنجش از دور ۵ متغیر محیطی شامل دمای روزانهٔ سطح آب (SST)، غلظت کلروفیل-a (CHL)، ضخامت اپتیک ریزگردها (ASL)، محتوای کربن آلی ذره‌ای (POC) و کربن غیرآلی ذره‌ای (PIC) به‌عنوان متغیرهای پیش‌بین مورد استفاده قرار گرفت. جهت سنجش عملکرد توصیفی و پیش‌بینی مدل‌ها از شاخص‌های ضریب تبیین (R^2)، میانگین خطای مطلق (MAE) و ریشهٔ میانگین مربعات خطا (RMSE) استفاده گردید. در بهترین مدل GLM برازش‌یافته تنها دو پارامتر Log(PIC) و POC معنی‌دار بودند، در حالی‌که در مدل RF تمامی متغیرها بکار گرفته شدند. مدل RF از توان تو صیفی بیشتری نسبت به مدل GLM برخوردار بود (GLM: R^2=۰/۰۵۳؛ RF:R^2=۰/۴۷). همچنین، دقت بیشتری برای مدل RF ((kg/hour.seine) RMSE=۱۳۲۶/۱؛ MAE=۹۷۲/۴) در مقایسه با GLM ((kg/hour.seine) RMSE=۱۴۶۵/۶؛ MAE=۱۳۲۸/۷) وجود داشت. پارامترهای ASL و CHL به‌ترتیب دارای بیشترین (۳۳/۳۱٪) و کمترین (۲۸/۸۷٪) سهم اهمیت نسبی در مدل RF بودند. براساس مجموعه نتایج به‌دست آمده، مدل جنگل تصادفی (RF) به‌عنوان یک مدل کارآمد جهت مدل‌سازی توزیع صید ماهیان پیشنهاد می‌گردد.

واژگان کلیدی: مدل خطی تعمیم‌یافته، مدل جنگل تصادفی، متغیرهای زیستگاهی، ماهی سفید، دریای خزر

* نویسنده مسئول: هادی پورباقر
Email: poorbagher@ut.ac.ir

Journal of Fisheries
Vol. 76, No. 1, Spring 2023
pp. 27-38

Comparing the performance of generalized linear model (GLM) and random forest (RF) models in predicting catch distribution of Caspian Kutum (*Rutilus frisii*)

Fateh Moezzi[1], Hadi Poorbagher[2]*, Soheil Eagderi[2], Jahangir Feghhi[3]

1. Ph.D. Graduate, Department of Fisheries, Faculty of Natural Resources, University of Tehran, Karaj, Iran
2. Associate Professor, Department of Fisheries, Faculty of Natural Resources, University of Tehran, Karaj, Iran
3. Professor, Department of Forestry and Forest Economics, Faculty of Natural Resources, University of Tehran, Karaj, Iran

Received: 8-Jun-2022 **Accepted: 15-Sep-2022**

Abstract

The present study aimed to assess the performance of generalized linear model (GLM) and random forest (RF) model in predicting Caspian Kutum (*Rutilus frisii*) catch distribution. Caspian Kutum catch per unit of effort (CPUE) data was used as the response variable. Remotely-sensed data of five environmental parameters were used as model predictors as well, including daily sea surface temperature (SST), chlorophyll-a concentration (CHL), aerosol optical thickness (ASL), particulate organic carbon (POC) and particulate inorganic carbon (PIC) concentrations. The coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) scores were used as measures of model performance and accuracy. The best fitted GLM had only Log(PIC) and POC as significant parameters, while the RF model contained all predictors. RF showed higher explaining potential compared to GLM (RF: R^2=0.47; GLM: R^2=0.053). Also, higher accuracy was observed using RF (MAE=972.4; RMSE=1326.1 (kg/hour.seine)) than GLM (MAE=1328.7; RMSE=1465.6 (kg/hour.seine)). ASL (33.31%) and CHL (28.87%) were parameters with the highest and lowest relative influence in the RF model. Based on the results, random forest modelling is suggested as a practical technique for predicting fish catch distribution.

Keywords: Generalized linear model, Random forest, Habitat parameters, Caspian Kutum, Caspian Sea

∗ Corresponding author: Hadi Poorbagher Email: poorbagher@ut.ac.ir

۱. مقدمه

در سال‌های گذشته، بررسی ارتباط حضور گونه‌ها و فراوانی آن‌ها با شرایط محیطی با بکارگیری روش‌های مدل‌سازی که عمدتاً از آن‌ها به‌عنوان مدل‌های توزیع گونه‌ای[1] (SDMs) نام برده می‌شود، بسیار مورد توجه قرار گرفته است. این مدل‌ها ابزارهایی توانمند در بررسی پیامدهای تغییرات شرایط محیطی بر حضور و فراوانی گونه‌ها به شمار می‌روند (Guisan *et al*., 2013). مدل‌های SDM در ارزیابی مطلوبیت زیستگاهی و پیش‌بینی توزیع فراوانی گونه‌ای نیز قابل استفاده می‌باشند (Gormley *et al*., 2011; Zhang *et al*., 2020). این روش‌ها نقش قابل توجهی در برنامه‌های حفاظتی و مدیریتی گونه‌ها و زیستگاه‌های آن‌ها دارند (Robinson *et al*., 2021).

طیف گسترده‌ای از مدل‌های کلاسیک آماری تا روش‌های پیشرفته یادگیری ماشین در زمینهٔ ارزیابی روابط موجودات با شرایط محیطی محل زیست آن‌ها توسعه یافته است. انتخاب روش مدل‌سازی مناسب تا حد زیادی نتیجه‌گیری و استدلال‌های حاصل از ارزیابی‌های توزیع گونه‌ای را تحت تأثیر قرار می‌دهد (Bučas *et al*., 2013). تعیین بهترین روش با بالاترین توان پیش‌بینی همواره چالش‌برانگیز بوده است؛ زیرا عملکرد پیش‌بینی مدل‌ها با توجه به عوامل متعدد از جمله تفاوت در رویکرد آماری یا الگوریتم‌های مورداستفاده، انتخاب متغیرهای ورودی متفاوت و ویژگی‌های آن‌ها، فرضیات و پیچیدگی مدل‌ها و ویژگی‌های گونه‌های مورد بررسی متغیر است (Ward *et al*., 2015; Shabani *et al*., 2016; Zhang *et al*., 2021). به‌همین دلیل پیشنهاد شده تا مجموعه‌ای از روش‌های مدل‌سازی با الگوریتم‌های مختلف جهت دست‌یابی به مدل بهینه، مورد ارزیابی قرار گیرند (Robinson et al., 2017).

مدل‌های خطی تعمیم‌یافته[2] (GLMs) یکی از معمول‌ترین روش‌های مورد استفاده در بررسی‌های توزیع گونه‌ای محسوب می‌شوند. مدل‌های GLM، با رویکرد بسط یافته‌ای از مدل‌های خطی، مدل‌هایی منعطف و با کاربرد گسترده در مدل‌سازی توزیع گونه‌ای هستند که از طریق محاسبهٔ حضور گونه‌ها به صورت توابعی پارامتریک از متغیرهای محیطی عمل می‌کنند (McCullagh and Nelder, 2019). این روش در مطالعات مختلفی در زمینهٔ ارزیابی ذخایر شیلاتی دریایی مورد استفاده قرار گرفته است (Hart, 2012; Campbell, 2015; Sguotti *et al*., 2016).

روش جنگل تصادفی[3] (RF) یک روش یادگیری ماشین با مبنای درختان طبقه‌بندی یا رگرسیونی است که در مدل‌سازی توزیع گونه‌ای از طریق ایجاد درخت‌های تصمیم‌گیری متعدد و ترکیب خروجی‌های آن‌ها مورد استفاده قرار می‌گیرند (Robinson *et al*., 2021). توان طبقه‌بندی و پیش‌بینی بالای این روش در مطالعات متعدد (Cutler *et al*., 2017) گزارش شده است که تا حد زیادی ناشی از کارآیی آن در شناخت روندهای داده، نداشتن پیش‌فرض‌های اختصاصی در ارتباط با توزیع داده‌ها و قدرت آن در پرداختن به اثرات متقابل بین متغیرها می‌باشد (Olaya-Marín *et al*., 2013). این روش در مطالعات متعددی در ارتباط با اکوسیستم‌های آبی و مباحث شیلاتی از جمله پیش‌بینی تنوع گونه‌ای و برآورد مطلوبیت زیستگاهی مورد استفاده قرار گرفته است (Li *et al*., 2015; Luan *et al*., 2018).

ماهی سفید، *Rutilus frisii* (Nordmann, 1840) از خانواده Leuciscidae و جنس *Rutilus* یکی از گونه‌های با ارزش تجاری دریای خزر است (Abdolhay *et al*., 2012; Eagderi *et al*. 2022). مناطق پراکنش این گونه در دریای خزر عمدتاً سواحل جنوبی آن را شامل می‌شود (Valipour *et al*., 2011). این ماهی توسط تعاونی‌های صیادی پره در امتداد سواحل استان‌های گیلان، مازندران و گلستان صید می‌شود. بخش عمدهٔ صید ماهیان

[1] Species Distribution Models: SDMs
[2] Generalized linear model: GLM
[3] Random Forest: RF

استخوانی خزر (بیش از ۷۰٪) در طول دو دهۀ گذشته متعلق به این گونه بوده است (Ghasemi *et al*., 2014). مطالعات زیادی در رابطه با ویژگی‌های زیستی این گونه انجام شده، اما ارزیابی مناسبی از پراکنش این گونه و ارتباط آن با شرایط محیطی و زیستگاهی تاکنون صورت نگرفته است.

در مطالعۀ حاضر، داده‌های صید ماهی سفید (*R. frisii*) در صیدگاه‌های پرۀ استان مازندران جهت ارزیابی عملکرد مدل‌های GLM و RF در تعیین تأثیر عوامل محیطی بر پراکنش صید ماهی با استفاده از داده‌های سنجش از دور پارامترهای محیطی مورد استفاده قرار گرفت. اهداف این مطالعه ۱) بررسی عملکرد توصیفی و پیش‌بینی مدل‌های GLM و RF در برآورد صید ماهی سفید بر اساس داده‌های محیطی سنجش از دور؛ ۲) تعیین میزان اهمیت پارامترهای تأثیرگذار بر توزیع صید بر اساس مدل با عملکرد بهتر و ۳) یافتن الگوهای تغییرات صید در برابر تغییرات پارامترهای محیطی بود.

۲. مواد و روش‌ها

۱.۲. داده‌های صید ماهی

داده‌های صید ماهی سفید (*R. frisii*) مربوط به صید پره در تعداد ۴۸ صیدگاه واقع در سواحل استان مازندران در سال‌های ۱۳۸۱ تا ۱۳۹۱ از سازمان شیلات ایران به دست آمد. این داده‌ها شامل میزان صید ماهی (kg)، تعداد ساعات تورکشی (hour) و تعداد تورهای پره مورد استفاده (seine) در هر تورکشی بود. به‌منظور استانداردسازی داده‌ها، میزان صید به‌ازای واحد تلاش صیادی[1] (CPUE) در فصول صید در هر صیدگاه با استفاده از رابطۀ (۱) محاسبه گردید:

$$CPUE\ (kg/hour \cdot seine) = \frac{\text{میزان صید } (kg)}{\text{تعداد تور } (seine) \times \text{زمان تورکشی} (hour)} \quad \text{(رابطۀ ۱)}$$

۲.۲. شاخص های محیطی

پنج متغیر محیطی به‌عنوان عوامل بالقوۀ مؤثر بر پراکنش ماهیان به‌عنوان ورودی مدل‌ها مورد استفاده قرار گرفتند که عبارت بودند از: دمای روزانۀ سطح آب[2] (SST)، غلظت کلروفیل-a[3] (CHL)، ضخامت اپتیک ریزگردها[4] (ASL)، محتوای کربن آلی ذره‌ای[5] (POC) و کربن غیرآلی ذره‌ای[6] (PIC). این متغیرها در مطالعات مختلف به‌عنوان عوامل اثرگذار بر شرایط فیزیکو شیمیایی و فرآیندهای اکولوژیک محیط آبی یا شاخص‌های مرتبط با آن‌ها گزارش شده‌اند (Giannoulaki *et al*., 2013; Mahowald *et al*., 2018; Zhang *et a*l., 2019; Hopkins *et al*., 2019; Hua *et al*., 2020). داده‌های سنجش از دور متغیرهای مورد اشاره از پایگاه داده پروژه MODIS (NASA Goddard Space Flight Center, Ocean Ecology Laboratory, 2021) به‌دست آمد. استخراج مقادیر این متغیرها در نقاط صیدگاهی از داده‌های سنجش از دور با استفاده از بستۀ raster (نسخۀ 3.4-10) (Hijmans, 2021) در محیط نرم‌افزار R (R Project Team, 2021) انجام شد.

۳.۲. برازش و ارزیابی مدل

از مجموعۀ ۱۰ سالۀ داده‌ها، داده‌های ۸ سال اول جهت برازش مدل و داده‌های ۲ سال پایانی جهت ارزیابی عملکرد پیش‌بینی مدل‌ها مورد استفاده قرار گرفت. جهت بررسی سطوح همبستگی داده‌های متغیرهای محیطی از آزمون همبستگی اسپیرمن استفاده شد که با تابع cor از

[1] Catch per Unit of Effort: CPUE
[2] Daily Sea Surface Temperature: SST
[3] Near Surface Chlorophyll-a Concentration: CHL
[4] Aerosol Optical Thickness: ASL
[5] Particulate Organic Carbon: POC
[6] Particulate Inorganic Carbon: PIC

بستهٔ stats (نسخهٔ 3.6.2) (R Project Team, 2021) صورت گرفت. همچنین، جهت بررسی میزان سطوح هم‌خطی در متغیر ها م قادیر فاکتور انحراف وار یانس[1] (VIF) آن‌ها با ا ستفاده از تابع vif از بستهٔ usdm (نسخهٔ 1.1-18) (Naimi et al., 2014) محاسبه گردید.

برازش مدل های GLM و RF در محیط نرم‌افزار R انجام شد. تابع glm از بستهٔ stats (نسخهٔ 3.6.2) (R Project Team, 2021) جهت برازش مدل GLM استفاده گردید. با توجه به آنالیز توزیع آماری متغیرهای محیطی، برای متغیر PIC از تبدیل لگاریتمی استفاده شد. یک رویکرد انتخاب گام به گام (stepwise-selection) جهت انتخاب بهترین مدل GLM و تعیین بهترین ترکیب متغیرهای محیطی بر مبنای مقدار شاخص آکایکه (AIC) بکار گرفته شد. برازش مدل RF با استفاده از تابع randomForest از بستهٔ randomForest (نسخهٔ 4.7-1.1) (Liaw and Wiener, 2002) انجام گرفت. میزان اهمیت نسبی متغیرها در مدل با استفاده از تابع varImp از بستهٔ caret (نسخهٔ 6.0-92) (Kuhn, 2022) به‌دست آمد. جهت ر سم نمودارهای وابستگی نسبی برای متغیرهای محیطی در مدل RF از تابع partial از بستهٔ pdp (نسخهٔ 0.8.0) (Greenwell, 2017) استفاده گردید.

ضریب تبیین (R^2) و میانگین خطای مطلق[2] (MAE) به‌عنوان شاخص‌های جهت ارزیابی عملکرد مدل‌ها در توصیف واریانس صید استفاده شد. مقادیر MAE با استفاده از رابطهٔ (۲) محاسبه گردید:

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{N} \qquad \text{(رابطهٔ ۲)}$$

که در آن: $\hat{y}_i$: مقادیر برآورد شده مدل، y_i: مقادیر واقعی پا سخ و N: تعداد م شاهدات ا ست. همچنین، توان پیش‌بینی مدل‌ها بر مبنای مقدار ریشه میانگین مربعات خطا[3] (RMSE) ارزیابی شد که مقادیر آن از رابطهٔ (۳) به‌دست آمد:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \qquad \text{(رابطهٔ ۳)}$$

۳. نتایج

۱.۳. شاخص های محیطی

مقادیر میانگین و دامنهٔ تغییرات (حداقل و حداکثر) متغیر های محیطی مورد استفاده در برازش مدل ها در جدول ۱ ارائه شده است. جدول ۲، نشان‌دهندهٔ مقادیر VIF و همچنین ضرایب همبستگی اسپیرمن برای متغیرهای محیطی است. بیشترین مقدار VIF به‌دست آ مده برای متغیر ها معادل ۳/۱ مربوط به کربن غیرآلی ذره‌ای بود.

جدول ۱- مقادیر میانگین، حداقل و حداکثر پارامترهای محیطی در صیدگاه‌ها.

متغیر محیطی	میانگین	حداقل	حداکثر
دمای روزانهٔ سطح آب (°C) (SST)	۱۵/۴۰	۱۳/۷۹	۱۶/۹۸
غلظت کلروفیل-a (mg/m^3) (CHL)	۴/۹۸۸	۲/۷۷۴	۹/۲۷۵
ضخامت اپتیک ریزگردها (-) (ASL)	۰/۰۸۰۵	۰/۰۵۸۸	۰/۱۱۳۸
کربن آلی ذره‌ای (mg/m^3) (POC)	۶۱۴/۴	۳۸۸/۲	۹/۱۹۰۵
کربن غیرآلی ذره‌ای (mol/m^3) (PIC)	۰/۰۰۷۵۵	۰/۰۰۱۰۴	۰/۰۵۱۵۸

[1] Variance inflation factor: VIF
[2] Mean Absolute Error: MAE
[3] Root Mean Square Error: RMSE

جدول ۲- مقادیر VIF (variance inflation factor) و نتایج آزمون همبستگی اسپیرمن بین متغیرهای محیطی.

متغیر محیطی	VIF	دمای روزانه سطح آب	غلظت کلروفیل-a	ضخامت اپتیک ریزگردها	کربن آلی ذره‌ای	کربن غیرآلی ذره‌ای
دمای روزانه سطح آب	۱/۲۹	**۱/۰۰**	P <۰/۰۰۱	P =۰/۰۰۴	P <۰/۰۰۱	P =۰/۵۳۳
غلظت کلروفیل-a	۱/۲۷	۰/۲۸۱	**۱/۰۰**	P =۰/۰۱۷	P <۰/۰۰۱	P <۰/۰۰۱
ضخامت اپتیک ریزگردها	۱/۹۰	-۰/۱۲۹	۰/۱۰۸	**۱/۰۰**	P <۰/۰۰۱	P <۰/۰۰۱
کربن آلی ذره‌ای	۱/۹۹	۰/۲۳۱	۰/۴۵۰	۰/۴۱۷	**۱/۰۰**	P <۰/۰۰۱
کربن غیرآلی ذره‌ای	۳/۱۰	۰/۰۲۸	-۰/۱۸۱	-۰/۶۴۵	-۰/۶۶۷	**۱/۰۰**

۲.۳. عملکرد مدل‌های GLM و RF

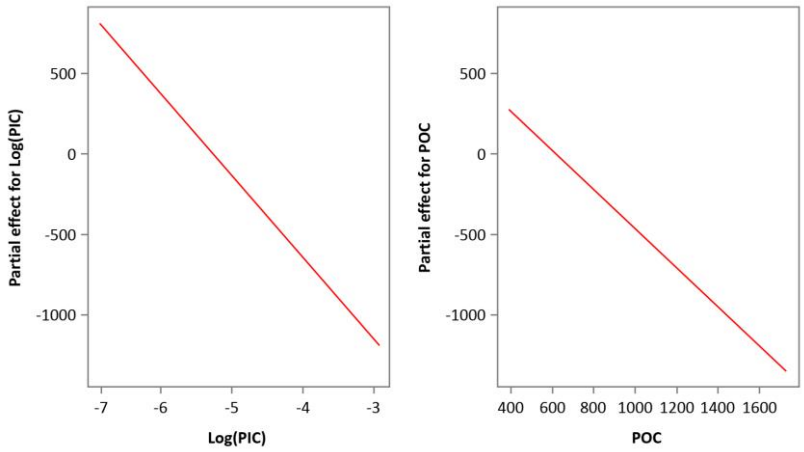
خلاصهٔ مدل‌های GLM و RF برازش‌یافته و همچنین شاخص‌های ارزیابی آن‌ها در جدول ۳ ارائه شده است. در بهترین مدل GLM برازش‌یافته با کمترین مقدار شاخص AIC (۶۸۵۱/۰) تنها دو پارامتر log(PIC) و POC معنی‌دار بودند. بهترین مدل RF تمامی متغیرهای محیطی را در ساختار مدل استفاده کرد. مقدار R^2 به‌دست آمده برای RF به مراتب بیشتر از GLM بود. همچنین، مدل RF برازش یافته با داده‌های ۸ سال اول مقدار MAE کمتری را نسبت به مدل GLM نشان داد. مقدار RMSE محاسبه شده برای مدل RF بر اساس پیش‌بینی داده‌های اعتبارسنجی، کمتر از GLM بود. با توجه به مجموعه شاخص‌های ارزیابی مورد استفاده، مدل RF عملکرد بهتری را در مقایسه با GLM نشان داد.

جدول ۳- خلاصه مدل‌های GLM و RF و مقادیر R^2 و MAE (واسنجی) و RMSE (اعتبارسنجی).
متغیرهای زیرخط‌دار نشان‌دهندهٔ متغیرهای معنی‌دار در مدل GLM هستند.

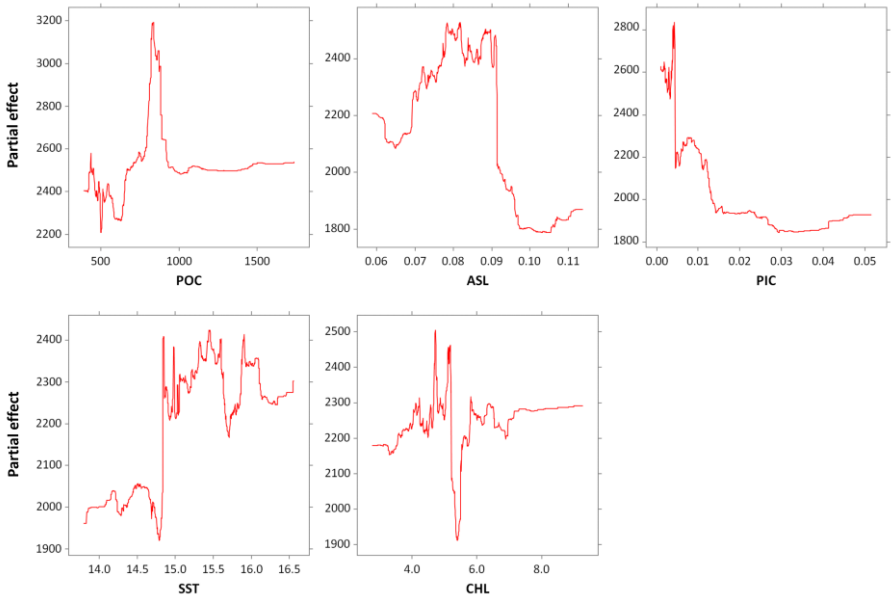
مدل	اجزای مدل	R^2	MAE	RMSE
GLM	CPUE ~ log(PIC) + POC + ASL	۰/۰۵۳	۱۳۲۸/۷	۱۴۶۵/۶
RF	CPUE ~ CHL + SST + PIC + POC + ASL	۰/۴۷	۹۷۲/۴	۱۳۲۶/۱

شکل‌های ۱ و ۲ میزان و روند نسبی تأثیر متغیرهای محیطی بر CPUE را در مدل‌های GLM و RF نشان می‌دهند. مدل GLM روندهای کاهشی کلی را برای اثر نسبی دو پارامتر log(PIC) و POC نشان داد. در مدل RF، روند تغییرات میزان تأثیر نسبی پارامترها الگوهایی غیرخطی را نمایان ساخت. مقادیر POC در بازهٔ نسبی ۸۰۰ تا ۱۰۰۰ (mg/m^3) با بالاترین سطوح CPUE متناظر بودند. برای ASL نیز بیشترین میزان اثرگذاری در توضیح CPUE در بازهٔ ۰/۰۷۵ تا ۰/۰۹ به‌دست آمد. عامل PIC یک روند کاهشی مشخص را از نظر میزان تأثیر در توضیح CPUE نشان داد و بیشترین میزان صید با پیک مشخص در کمترین مقادیر PIC مشاهده گردید. بازهٔ دمایی ۱۵ تا ۱۶ (°C) نیز با بیشترین سطوح صید مرتبط بود. برای CHL، یک روند کلی افزایش تدریجی مشاهده شد هر چند در سطوح میانی مقادیر این پارامتر، نوسانات قابل‌توجهی وجود داشت.

جدول ۴، میزان اهمیت نسبی متغیرهای محیطی در مدل RF را نشان می‌دهد. براساس نتایج، متغیرهای کربن آلی ذره‌ای و غلظت کلروفیل-a به ترتیب بیشترین و کمترین سطوح اهمیت را در مدل RF نشان دادند.



شکل ۱- نمودارهای مربوط به اثرات نسبی متغیرهای معنی‌دار مدل GLM ((Log(PIC) و POC).



شکل ۲- نمودارهای وابستگی نسبی (partial dependence plots) متغیرهای مدل RF.

جدول ۴- مقادیر اهمیت (٪) متغیرهای محیطی در مدل RF.

متغیر	دمای روزانهٔ سطح آب	غلظت کلروفیل-a	ضخامت اپتیک ریزگردها	کربن آلی ذره‌ای	کربن غیرآلی ذره‌ای
میزان اهمیت (٪)	۲۹/۰۱	۲۸/۸۷	۳۳/۳۱	۳۳/۸۲	۳۲/۰۵

۴. بحث و نتیجه گیری نهایی

در مطالعهٔ حاضر، داده‌های سنجش از دور مربوط به پنج پارامتر محیطی تأثیرگذار بر شرایط فیزیکوشیمیایی و اکولوژیک محیط آبی به‌عنوان متغیرهای پیش‌بینی‌کننده در مدل‌های GLM و RF مورد استفاده قرار گرفت. یکی از مباحث مهم در رابطه با ورودی‌های مدل، بررسی سطوح همبستگی و هم‌خطی بین متغیرها ست به‌طوری که نباید پارامترهای ورودی دارای همبستگی یا همخطی

زیاد را به‌صورت همزمان در ساختار یک مدل استفاده کرد. در این مطالعه، سطوح همبستگی قابل‌توجهی بین متغیرهای مورد استفاده وجود نداشت. مقدار ضریب همبستگی ۰/۷ به‌عنوان حد آستانه جهت انتخاب یا حذف متغیر های ورودی مدل ها پیشنهاد شده است (Schickele *et al*., 2020) به‌طوری که اگر بین دو متغیر مقدار ضریب همبستگی r>۰/۷ باشد، باید یکی از آن‌ها از مجموعه متغیرهای ورودی مدل حذف شود. همچنین شاخص VIF به‌عنوان معیار سنجش هم‌خطی بین متغیرها مورد محاسبه قرار گرفت و مقادیر آن تقریباً کمتر از ۳ بود. مقدار VIF کمتر از ۱۰ به‌عنوان بازه قابل‌قبول این شاخص از نظر عدم وجود هم‌خطی چندگانهٔ دارای اثر نامطلوب بر پیش‌بینی مدل‌ها در نظر گرفته شده است (Bui *et al*., 2011). بدین‌ترتیب، مجموعه متغیر های ورودی در مدل‌های مورد بررسی در این مطالعه فاقد سطوح نامطلوب همبستگی و هم‌خطی بودند.

در مطالعهٔ حاضر، مقایسهٔ عملکرد پیش‌بینی مدل‌های RF و GLM نشان داد که مدل RF از قدرت پیش‌بینی بیشتری برای توزیع صید ماهی سفید با استفاده از مجموعه داده‌های محیطی برخوردار بوده است. عملکرد مدل‌های GLM و RF در پیش‌بینی توزیع گونه‌ای ماهیان در مطالعات متعدد دیگری مورد بررسی قرار گرفته (Bučas *et al*., 2013; Kwon *et al*., 2015; Li *et al*., 2011a,b; Li *et al*., 2017; Lin *et al*., 2015; Robinson *et al*., 2021; Smoliński and Radtke, 2017) که مشابه نتایج به‌دست آمده در مطالعهٔ حاضر، در بیشتر آن‌ها عملکرد مناسب‌تر مدل RF گزارش شده است. در مطالعهٔ انجام شده Smoliński and Radtke (۲۰۱۷) دقت پیش‌بینی بالاتری برای روش RF در میان روش‌های یادگیری ماشین و در مقایسه با روش‌های رگرسیونی از جمله GLM در ارزیابی و پیش‌بینی شاخص‌های تنوع جوامع ماهیان به‌دست آمده است. همچنین در مطالعهٔ Bučas و همکاران (۲۰۱۳) عملکرد مدل RF در مقایسه با سایر روش‌های مورد استفاده از جمله GAM و MAXENT مناسب‌تر بوده است. Kwon و همکاران (۲۰۱۵) نیز توان پیش‌بینی بیشتری را برای مدل RF در مقایسه با GLM در پیش‌بینی توزیع ماهیان آب شیرین در مواجهه با تغییرات اقلیمی گزارش کرده‌اند.

براساس نتایج، تنها دو متغیر log(PIC) و POC در بهترین مدل GLM برازش‌یافته معنی‌دار بودند درحالی‌که مدل RF تمامی متغیرهای پیش‌بین را در خود داشت، هر چند متغیرهای مدل GLM در میان متغیرهای با بیشترین سهم اهمیت نسبی در مدل RF قرار داشتند. مدل های GLM پیش‌فرض وجود یک رابطهٔ خطی بین متغیر پاسخ و متغیرهای پیش‌بینی‌کننده را دارا می‌باشند و از این نظر وجود روابط غیرخطی توان این مدل‌ها را در بکارگیری اطلاعات فاکتورهای پیش‌بین در جهت توضیح تغییرات پاسخ کاهش می‌دهد. نمودارهای مربوط به روند تغییرات متغیر پاسخ (CPUE) برای متغیر های مدل ها (شکل‌های ۱ و ۲) نیز به‌وضوح این واقعیت را نشان دادند به‌طوری که روندهای خطی کلی مدل GLM با دقت بسیار بیشتر در مدل RF در بازهٔ تغییرات متغیرهای POC و PIC پردازش شده‌اند که وجود الگوهای غیرخطی را تأیید می‌کنند. روابط غیرخطی بین داده های تراکم ماهیان و فاکتورهای محیطی در مطالعات مختلف گزارش شده است (Walsh and Kleiber, 2001; Denis *et al*., 2002). در نتایج به‌دست آمده، با توجه به مقادیر بسیار کم R2 برای مدل GLM، روشن است که روابط خطی بسیار ضعیفی بین متغیرهای معنی‌دار مدل و CPUE وجود داشته است. این در حالی است که مقدار R^2 مدل RF (۰/۴۷) توان قابل‌ملاحظه این روش را در مقایسه با مدل GLM در توضیح تغییرات CPUE در ارتباط با پارامترهای محیطی مورد استفاده نشان می‌دهد. همچنین، مقدار کمتر MAE متعلق به مدل RF نسبت به GLM نشان‌دهندهٔ سطح دقت بیشتر (خطای کمتر) این مدل بوده است. مدل RF از حساسیت کمتری نسبت به بکارگیری متغیرهای با اهمیت کمتر برخوردار است (Okun and Priisalu, 2007) و این مدل حتی با استفاده از متغیرهای دارای واریانس زیاد می‌تواند پیش‌بینی‌های مناسبی را محقق سازد (Díaz-Uriarte and de Andrés, 2006). پیش‌بینی مدل‌های برازش‌یافته روی داده‌های دو سال پایانی نیز با

توجه به کمتر بودن مقدار RMSE مدل RF، بیانگر قدرت بیشتر پیش‌بینی و دقت بالاتر (Lin *et al*., 2015) این مدل در مقایسه با GLM برای داده‌های جدید است. دقت و عملکرد مناسب‌تر روش RF به‌طور خاص برای مجموعه‌های داده دارای تغییرات زیاد می‌تواند ناشی از الگوریتم‌های مورد استفاده در ساختار این مدل باشد. این روش، یک روش طبقه‌بندی است که در آن درخت‌های تصمیم متعدد با استفاده از تکرارهای bootstrap یک مجموعه داده برازش می‌یابند و در نهایت متوسط برداشت‌های تمامی درختان به‌عنوان خروجی مدل ارائه می‌شود (Breiman, 2001). چنین الگوریتمی منجر به قدرت توصیف بالاتر مدل و کاهش خطای پیش‌بینی آن می‌گردد.

متغیرهای ورودی مورد استفاده در برازش مدل‌های GLM و RF در این مطالعه، فاکتورهای بالقوه تأثیرگذار بر شرایط محیطی سیستم آبی و مرتبط با حضور ماهیان بودند. هر دو مدل GLM و RF، روندهای کاهشی را برای تأثیر PIC بر CPUE نشان دادند. PIC یک فاکتور مرتبط با میزان تولید در محیط‌های آبی است (Hopkins *et al*., 2019) و سطوح بالای آن در آب بیانگر تولید بیشتر می‌باشد، اما در عین حال، غلظت‌های زیاد این پارامتر با کاهش نفوذ نور در آب‌ها مرتبط است (Mitchell *et al*., 2017). بر این اساس، می‌توان وقوع بیشترین CPUEs در کمترین مقادیر PIC را به تأثیرگذاری این عامل بر شرایط روشنایی تودهٔ آبی نسبت داد. فاکتور POC به‌عنوان عامل مرتبط با محتوای کربن آلی ذره‌ای و شاخصی دیگر از میزان تولید آب (Zhang *et al*., 2019)، دو روند تقریباً متفاوت را در مدل‌ها نشان داد. این عامل در محیط‌های آبی تا حد زیادی با افزایش تولید بنتیک مرتبط است و از آنجایی که ماهی سفید دریای خزر عمدتاً از موجودات بنتیک تغذیه می‌کند (Naderi Jolodar *et al*., 2013)، می‌توان گفت مدل RF به‌شکل مناسب‌تری تأثیر آن را در پراکنش ماهی در ساختار مدل بکار گرفته و وجود پیک مشخص در نمودار POC تأییدکننده این موضوع است. ASL یکی از پارامترهای با سهم اهمیت بالا در مدل RF بود و مدل یک بازهٔ مشخص از مقادیر این پارامتر را به‌عنوان بازهٔ دارای بیشترین تأثیر نشان داد. ورود ریزگردها به منابع آبی یکی از عوامل اصلی واردکنندهٔ مواد مغذی بوده و سطوح تولید را به‌طور خاص از طریق انتقال بار مواد مغذی به بستر در این اکوسیستم‌ها افزایش می‌دهد (Boyd *et al*., 2017; Yang *et al*., 2019). با این حال، کاهش تأثیر این پارامتر بر میزان CPUE در مقادیر بالای آن می‌تواند ناشی از اثرات دیگر پارامترها بوده باشد. دمای آب نیز به‌عنوان یک فاکتور اصلی اثرگذار بر حضور و فراوانی ماهیان، بقا، وضعیت فیزیولوژیک و فرآیندهای حیاتی آن‌ها تحت تأثیر قرار داده و تنظیم می‌کند (Chen *et al*., 2012; Kempf *et al*., 2013; Olsen, 2019). براساس نتایج این مطالعه، میزان صید ماهی سفید در نقاط با دمای آب بیشتر، بالاتر بوده است. غلظت کلروفیل-a، متغیر با کمترین سهم اهمیت در مدل RF بود. نمودار سهم تأثیر نسبی این پارامتر یک روند افزایشی جزیی با نوسانات شدید را نشان داد. هرچند این فاکتور بیانگر میزان تولید در آب‌های پلاژیک محیط‌های آبی است (Johanson *et al*., 2013)، اما در مطالعهٔ حاضر سهم کمتری در توضیح نوسانات صید داشته است.

۵. نتیجه‌گیری نهایی

نتایج به‌دست آمده در این مطالعه نشان داد که مدل جنگل تصادفی (RF) از عملکرد بسیار مناسب‌تری نسبت به مدل خطی تعمیم‌یافته (GLM) در پردازش روابط بین مقادیر صید در واحد تلاش صیادی (CPUE) ماهی سفید (به‌عنوان شاخص فراوانی ماهی) و پارامترهای محیطی مورد مطالعه و پیش‌بینی آن داشته است. این مدل با بکارگیری تعداد بیشتری از پارامترهای محیطی، قدرت توصیفی و دقت بیشتری را در شناخت روابط ماهی و شرایط محیطی نسبت به GLM نشان داد و پیش‌بینی‌های حاصل از آن با خطای کمتری همراه بود. بر این اساس، این روش به‌عنوان یک روش با کارآیی مناسب در مدل‌سازی صید ماهیان پیشنهاد می‌شود.

References ۶. منابع

Abdolhay, H.A., Daud, S.K., Rezvani, S., Pourkazemi, M., Siraj, S.S., Laloei, F., Javanmard, A., Hassanzadeh Saber, M., 2012. Population genetic structure of Mahi Sefid (*Rutilus frisii kutum*) in the South Caspian Sea: Implications for fishery management. *Iranian Journal of Animal Biosystematics* 8(1), 15-26.

Boyd, P.W., Ellwood, M.J., Tagliabue, A., Twining, B.S., 2017. Biotic and abiotic retention, recycling and remineralization of metals in the ocean. *Nature Geoscience* 10(3), 167-173.

Breiman, L., 2001. Random forests. *Machine Learning* 45, 5-32.

Bučas, M., Bergström, U., Downie, A.L., Sundblad, G., Gullström, M., Von Numers, M., Šiaulys, A., Lindegarth, M., 2013. Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. *ICES Journal of Marine Science* 70(6), 1233-1243.

Bui, D.T., Lofman, O., Revhaug, I., Dick, O., 2011. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*, 59(3), 1413-1444.

Campbell, R.A., 2015. Constructing stock abundance indices from catch and effort data: Some nuts and bolts. *Fisheries Research* 161, 109-130.

Chen, X., Cao, J., Chen, Y., Liu, B., Tian, S., 2012. Effect of the Kuroshio on the spatial distribution of the red flying squid *Ommastrephes bartramii* in the Northwest Pacific Ocean. *Bulletin of Marine Science* 88(1), 63-71.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. *Ecology* 88 (11), 2783–2792.

Denis, V., Lejeune, J., Robin, J.P., 2002. Spatio-temporal analysis of commercial trawler data using General Additive models: patterns of Loliginid squid abundance in the north-east Atlantic. *ICES Journal of Marine Science* 59(3), 633-648.

Díaz-Uriarte, R., de Andrés, S.A., 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), 1-13.

Eagderi, S., Mouludi-Saleh, A., Esmaeli, H.R., Sayyadzadeh, G., Nasri, M., 2022. Freshwater lamprey and fishes of Iran; a revised and updated annotated checklist-2022. *Turkish Journal of Zoology* 46(6), 500-522.

Ghasemi, M., Zamani, H., Hosseini, S.M., Karsidani, S.H., Bergmann, S.M., 2014. Caspian White Fish (*Rutilus frisii kutum*) as a host for spring viraemia of carp virus. *Veterinary Microbiology* 170(3-4), 408-413.

Giannoulaki, M., Iglesias, M., Tugores, M.P., Bonanno, A., Patti, B., de Felice, A., Leonori, I., Bigot, J.L., Tičina, V., Pyrounaki, M.M., Tsagarakis, K., 2013. Characterizing the potential habitat of European anchovy *Engraulis encrasicolus* in the Mediterranean Sea, at different life stages. *Fisheries Oceanography* 22(2), 69-89.

Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S., Scroggie, M.P., Woodford, L., 2011. Using presence-only and presence–absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology* 48(1), 25-34.

Greenwell, B.M., 2017. pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421-436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., 2013. Predicting species distributions for conservation decisions. *Ecology Letters* 16(12), 1424-1435.

Hart, R.A., 2012. Stock assessment of brown shrimp (*Farfantepenaeus aztecus*) in the US Gulf of Mexico for 2011.

Hijmans, R.J., 2021. Raster: Geographic Data Analysis and Modeling. R package version 3.5-11. https://CRAN.R-project.org/package=raster

Hopkins, J., Henson, S.A., Poulton, A.J., Balch, W.M., 2019. Regional characteristics of the temporal variability in the global particulate inorganic carbon inventory. *Global Biogeochemical Cycles* 33(11), 1328-1338.

Hua, C., Li, F., Zhu, Q., Zhu, G., Meng, L., 2020. Habitat suitability of Pacific saury (*Cololabis saira*) based on a yield-density model and weighted analysis. *Fisheries Research* 221, 105408.

Johanson, A. F., Jenkins, S. R., Hiddink, J. G., Hinz, H., 2013. Linking temperate demersal fish species to habitat: scales, patterns and future directions. *Fish and Fisheries* 14(3), 256-280.

Kempf, A., Stelzenmüller, V., Akimova, A., Floeter, J., 2013. Spatial assessment of predator–prey relationships in the North Sea: the influence of abiotic habitat properties on the spatial overlap between 0-group cod and grey gurnard. *Fisheries Oceanography* 22(3), 174-192.

Kuhn, M., 2022. caret: Classification and Regression Training. R package version 6.0-92. https://CRAN.R-project.org/package=care

Kwon, Y.S., Bae, M.J., Hwang, S.J., Kim, S.H., Park, Y.S., 2015. Predicting potential impacts of climate change on freshwater fish in Korea. *Ecological Informatics* 29, 156-165.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011a. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software* 26(12), 1647-1659.

Li, J., Heap, A.D., Potter, A., Huang, Z., Daniell, J.J., 2011b. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. *Continental Shelf Research* 31(13), 1365-1376.

Li, M., Zhang, C., Xu, B., Xue, Y., Ren, Y., 2017. Evaluating the approaches of habitat suitability modelling for whitespotted conger (*Conger myriaster*). *Fisheries Research*, 195, 230-237.

Li, Z., Ye, Z., Wan, R., Zhang, C., 2015. Model selection between traditional and popular methods for standardizing catch rates of target species: a case study of Japanese Spanish mackerel in the gillnet fishery. *Fisheries Research* 161, 312-319.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2(3), 18-22.

Lin, Y.P., Lin, W.C., Wu, W.Y., 2015. Uncertainty in various habitat suitability models and its impact on habitat suitability estimates for fish. *Water* 7(8), 4088-4107.

Luan, J., Zhang, C., Xu, B., Xue, Y., Ren, Y., 2018. Modelling the spatial distribution of three Portunidae crabs in Haizhou Bay, China. *PloS One* 13(11), p.e0207457.

Mahowald, N.M., Hamilton, D.S., Mackey, K.R., Moore, J.K., Baker, A.R., Scanza, R.A., Zhang, Y., 2018. Aerosol trace metal leaching and impacts on marine microorganisms. *Nature Communications* 9(1), 1-15.

McCullagh, P., Nelder, J.A., 2019. Generalized linear models. Routledge.

Mitchell, C., Hu, C., Bowler, B., Drapeau, D., Balch, W.M., 2017. Estimating particulate inorganic carbon concentrations of the global ocean from ocean color measurements using a reflectance difference approach. *Journal of Geophysical Research: Oceans* 122(11), 8707-8720.

Naderi Jolodar, M., Salarvand, G., Abdoli, A., Fazli, H., Eshaqi Nimvary, M., 2013. The feeding strategy of the Caspian Sea Kutum (*Rutilus frissi kutum* Kamenski, 1901). *Journal of Applied Ichthyological Reseach* 1(3), 63-79. (In Persian)

Naimi, B., Hamm, Na., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., 2014. Where is positional uncertainty a problem for species distribution modelling. *Ecography* 37, 191-203.

Okun, O. and Priisalu, H., 2007, June. Random forest for gene expression based cancer classification: overlooked issues. In Iberian conference on pattern recognition and image analysis (pp. 483-490). Springer, Berlin, Heidelberg.

Olaya-Marín, E.J., Martínez-Capel, F., Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. *Knowledge and Management of Aquatic Ecosystems* (409), 07.

Olsen, Z., 2019. Quantifying nursery habitat function: variation in habitat suitability linked to mortality and growth for juvenile Black Drum in a hypersaline estuary. *Marine and Coastal Fisheries* 11(1), 86-96.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robinson, C.L., Proudfoot, B., Rooper, C.N., Bertram, D.F., 2021. Comparison of spatial distribution models to predict subtidal burying habitat of the forage fish *Ammodytes personatus* in the Strait of Georgia, British Columbia, Canada. *Aquatic Conservation: Marine and Freshwater Ecosystems* 31(10), 2855-2869.

Robinson, N.M., Nelson, W.A., Costello, M.J., Sutherland, J.E., Lundquist, C.J., 2017. A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science* 4, 421.

Schickele, A., Leroy, B., Beaugrand, G., Goberville, E., Hattab, T., Francour, P., Raybaud, V., 2020. Modelling European small pelagic fish distribution: Methodological insights. *Ecological Modelling* 416, 108902.

Sguotti, C., Lynam, C.P., García-Carreras, B., Ellis, J.R., Engelhard, G.H., 2016. Distribution of skates and sharks in the North Sea: 112 years of change. *Global Change Biology* 22(8), 2729-2743.

Shabani, F., Kumar, L., Ahmadi, M., 2016. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecology and Evolution* 6(16), 5973-5986.

Smoliński, S., Radtke, K., 2017. Spatial prediction of demersal fish diversity in the Baltic Sea: comparison of machine learning and regression-based techniques. *ICES Journal of Marine Science* 74(1), 102-111.

Valipour, A., Kanipour, A., Khadivi Nia Moghaddam, M., Valinassab, T., 2011. Kutum: jewel of the Caspian Sea, vol 1. Iranian Fisheries Research Organization, Tehran. (In Persian)

Walsh, W.A., Kleiber, P., 2001. Generalized additive model and regression tree analyses of blue shark (*Prionace glauca*) catch rates by the Hawaii-based commercial longline fishery. *Fisheries Research* 53(2), 115-131.

Ward, E.J., Jannot, J.E., Lee, Y.W., Ono, K., Shelton, A.O., Thorson, J.T., 2015. Using spatiotemporal species distribution models to identify temporally evolving hotspots of species co-occurrence. *Ecological Applications* 25(8), 2198-2209.

Yang, T., Chen, Y., Zhou, S., Li, H., 2019. Impacts of aerosol copper on marine phytoplankton: A review. *Atmosphere* 10(7), 414.

Zhang, M., Wu, Y., Qi, L., Xu, M., Yang, C., Wang, X., 2019. Impact of the migration behavior of mesoplagic fishes on the compositions of dissolved and particulate organic carbon on the northern slope of the South China Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* 167, 46-54.

Zhang, Y., Xu, B., Ji, Y., Zhang, C., Ren, Y., Xue, Y., 2021. Comparison of habitat models in quantifying the spatio-temporal distribution of small yellow croaker (*Larimichthys polyactis*) in Haizhou Bay, China. *Estuarine, Coastal and Shelf Science* 261, 107512.

Zhang, Z., Mammola, S., Xian, W., Zhang, H., 2020. Modelling the potential impacts of climate change on the distribution of ichthyoplankton in the Yangtze Estuary, China. *Diversity and Distributions* 26(1), 126-137.