RESEARCH ARTICLE

Toward a Monte Carlo approach to selecting climate variables in MaxEnt

John L. Schnase*, Mark L. Carroll, Roger L. Gill, Glenn S. Tamkin, Jian Li, Savannah L. Strong, Thomas P. Maxwell, Mary E. Aronne, Caleb S. Spradlin

Office of Computational and Information Sciences and Technology, NASA Goddard Space Flight Center, Greenbelt, Maryland, United States of America

* john.l.schnase@nasa.gov

Abstract

MaxEnt is an important aid in understanding the influence of climate change on species distributions. There is growing interest in using IPCC-class global climate model outputs as environmental predictors in this work. These models provide realistic, global representations of the climate system, projections for hundreds of variables (including Essential Climate Variables), and combine observations from an array of satellite, airborne, and *in-situ* sensors. Unfortunately, direct use of this important class of data in MaxEnt modeling has been limited by the large size of climate model output collections and the fact that MaxEnt can only operate on a relatively small set of predictors stored in a computer's main memory. In this study, we demonstrate the feasibility of a Monte Carlo method that overcomes this limitation by finding a useful subset of predictors in a larger, externally-stored collection of environmental variables in a reasonable amount of time. Our proposed solution takes an ensemble approach wherein many MaxEnt runs, each drawing on a small random subset of variables, converges on a global estimate of the top contributing subset of variables in the larger collection. In preliminary tests, the Monte Carlo approach selected a consistent set of top six variables within 540 runs, with the four most contributory variables of the top six accounting for approximately 93% of overall permutation importance in a final model. These results suggest that a Monte Carlo approach could offer a viable means of screening environmental predictors prior to final model construction that is amenable to parallelization and scalable to very large data sets. This points to the possibility of near-real-time multiprocessor implementations that could enable broader and more exploratory use of global climate model outputs in environmental niche modeling and aid in the discovery of viable predictors.

OPEN ACCESS

Citation: Schnase JL, Carroll ML, Gill RL, Tamkin GS, Li J, Strong SL, et al. (2021) Toward a Monte Carlo approach to selecting climate variables in MaxEnt. PLoS ONE 16(3): e0237208. https://doi.org/10.1371/journal.pone.0237208

Editor: Yangyang Xu, Texas A&M University, UNITED STATES

Received: July 30, 2020

Accepted: February 15, 2021

Published: March 3, 2021

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

MaxEnt is one of the most popular software packages in use today by the ecological research community [1–3]. Based on a machine learning approach to maximum entropy modeling, MaxEnt allows researchers to construct ecological niche models (ENMs) that estimate the habitat suitability of a species using occurrence data and a set of environmental variables [1, 2, 4–6]. An abundant literature points to MaxEnt's effectiveness across a wide range of applications in fields as diverse as biogeography and phylogeny [7], conservation biology and epidemiology [8, 9], invasion biology [10–12], and archaeology [13]. Its merits compared to alternative approaches have been the subject of numerous statistical and methodological analyses, many of which have led to software improvements and refinements to the way MaxEnt is used [14–23]. In this paper, we contribute to this ongoing dialog by describing our efforts to overcome a specific technical limitation of the MaxEnt software that makes the tool difficult to use with large predictor data sets.

In recent years, MaxEnt has become a particularly important aid in understanding the influence of climate change on species distributions [24–29]. The need for reliable climate projections in this work is leading to greater use of global climate model (GCM) outputs as predictors [25]. While creating important new opportunities for research, this trend is also creating a "Big Data" challenge for the MaxEnt community [16]. The largest and most sophisticated GCMs—sometimes referred to as "IPCC-class" models because of the critical role they play in the work of the Intergovernmental Panel on Climate Change (IPCC)—produce petabyte-scale data sets comprising hundreds of variables, a volume that vastly exceeds what is generally used in bioclimatic modeling today [30–32]. Moreover, the direct outputs of these systems are being transformed into derived climate data products on an unprecedented scale [26, 33]. As a result, model tuning and variable selection, which are crucial aspects of any species distribution modeling effort, are becoming more complicated issues [22, 23, 34, 35].

Part of the problem lies in the fact that MaxEnt, like many machine learning systems, acts on its inputs as a piece: predictors and observations must be memory-resident for the program to work [36]. This results in run-times and space requirements that scale linearly with the size of a model's inputs. In most cases, these scaling properties pose few difficulties. But when the number of predictors under consideration becomes large, compute times can become impractically long, models can become overly complex, and efforts to understand any particular variable's contribution to model formation, either as an aspect of model analysis or as a way of selecting subsets of variables for further model refinement, can become challenging [17, 34, 37–39]. Clearly, an effective way of dealing with large, externally-stored environmental data sets that preserves the many advantages of MaxEnt while overcoming its current limitations would benefit the MaxEnt community.

In this study, we investigated the potential of an out-of-core Monte Carlo method to help accomplish such an outcome. Monte Carlo optimizations are a common way of finding approximate answers to problems that are solvable in principle but lack a practical means of solution [40]. Out-of-core (or "external memory") algorithms process data sets that are too large to fit a computer's main memory [41, 42]. Our objective was to find a useful subset of predictors in a larger collection of environmental variables in a reasonable amount of time. Our proposed solution takes an ensemble approach wherein many MaxEnt runs, each drawing on a small random subset of variables stored in the filesystem, converges on a global estimate of the top contributing subset of variables in the larger collection.

Preliminary results suggest that the method reliably selects a suitable subset of the original predictors that could be explored in more detailed ways and further refined prior to final model construction. Since each model run in the Monte Carlo screening process is independent and uses a set number of variables, the method is totally parallelizable, independent of the intrinsic scaling properties of MaxEnt, and amenable to implementation as an external memory algorithm. If proved to be effective, such an approach could contribute to the ecological modeling process when there is a need to preselect a small set of predictors in a pool comprising a potentially very large number of predictors. This could lead to greater use of climate model outputs by the ecological research community and aid the search for viable predictors when variable selection through ecological reasoning is not apparent.

Materials and methods

We used Cassin's Sparrow as a target species in our development efforts. Cassin's Sparrow (*Peucaea cassinii* Woodhouse, 1852) is an elusive resident of arid shrub grasslands of Middle America and the Southwestern United States [43]. Desert-adapted birds, such as Cassin's Sparrow, appear to be especially vulnerable to climate change [44, 45]. While the current work does not address a Cassin's Sparrow science question *per se*, we chose Cassin's Sparrow as an example of a species whose study could benefit from the technical advances described here. Occurrence data was obtained from the Global Biodiversity Information Facility (GBIF) for the year 2016 [46]. After removing replicates, a total of 1865 records were acquired. To reduce sampling bias and avoid double counting the same individual, we only kept non-overlapping observations within a 16 km buffer, which resulted in a total of 609 observations [47–51].

For predictors, we used Worldclim Version 2.1's standard (19) Bioclimatic (bioclim) environmental variables at a resolution of 5.0 arc-minutes throughout (Table 1) [52, 53]. These predictor layers were clipped to the coverage area of our observational data, reprojected, and formatted for use with MaxEnt using the Geospatial Data Abstraction Library Version 3.0 (GDAL) software package [54] following the guidelines of Hijmans et al. [55]. We used Variance Inflation Factor analysis to identify collinearities in the predictor data set [56] (S1 Table); however, we did not attempt to minimize collinearity by removing variables, because the current study focuses on stochastic down-selection from a full variable set as a preliminary screening step, which presumably would be followed by refinements such as this prior to final model construction.

In addition to GDAL, our computing environment comprised MaxEnt Version 3.4.1 [57], R Version 4.0.1 [58], the ENMeval Version 0.3.0 R package [59], RStudio Version 1.2.5033 [60], and ENMTools Version 1.4.4 [61] running on a 2.8 GHz Intel Quad-Core i7 MacBook Pro with 16 GB of memory.

One of the most common uses for ecological niche models is to identify important variables [62, 63]. In this study, we used MaxEnt in two different ways to find the six most influential

Table 1. Worldclim bioclimatic variables.

bio01	Annual Mean Temperature
bio02	Mean Diurnal Range (Mean of monthly (max temp—min temp))
bio03	Isothermality (BIO2/BIO7) (×100)
bio04	Temperature Seasonality (standard deviation ×100)
bio05	Max Temperature of Warmest Month
bio06	Min Temperature of Coldest Month
bio07	Temperature Annual Range (BIO5-BIO6)
bio08	Mean Temperature of Wettest Quarter
bio09	Mean Temperature of Driest Quarter
bio10	Mean Temperature of Warmest Quarter
bio11	Mean Temperature of Coldest Quarter
bio12	Annual Precipitation
bio13	Precipitation of Wettest Month
bio14	Precipitation of Driest Month
bio15	Precipitation Seasonality (Coefficient of Variation)
bio16	Precipitation of Wettest Quarter
bio17	Precipitation of Driest Quarter
bio18	Precipitation of Warmest Quarter
bio19	Precipitation of Coldest Quarter

https://doi.org/10.1371/journal.pone.0237208.t001 range- and niche-defining bioclim variables for Cassin's Sparrow. The choice of "top six" for this evaluation was based on our experience that six or fewer predictors generally predominate in such models.

First, we developed a baseline model using the stand-alone MaxEnt program operated through its graphical user interface (GUI). MaxEnt users can apply various combinations of five mathematical transformations ('feature classes' or FCs) to predictor variables to enable more complex fits to the observational data. The available feature types for continuous variables are linear (L), quadratic (Q), hinge (H), product (P), and threshold (T) [4]. Users can also adjust a regularization multiplier (RM) to maximize predictive accuracy and offset the overfitting that FC adjustments can introduce. By default, MaxEnt uses the LQHP feature classes and a regularization multiplier of 1.0 when there are more than 80 training samples, which was the case here [57]. We confirmed the appropriateness of these settings for our data by performing a comprehensive ENMeval scan of all five FC classes across RMs ranging from 0.5 to 4.0 in half-step intervals [38, 59] (S2 Table). We applied MaxEnt's default FC and RM settings (i.e. the "Auto features" setting) with 10 replicate cross-validation and jackknife evaluation of variable importance. Ten thousand background points were selected from across the study area following the recommendations of Phillips et al. [64] and Fourcade et al. [48].

MaxEnt provides three algorithm-specific indicators of variable importance: percent contribution, permutation importance, and change-in-gain based on jackknife analysis of individual variables [3, 62, 65]. No single measure is sufficient to identify which variables are best for producing a final model [23]; however, for screening purposes and to simplify comparisons in this initial evaluation, we used permutation importance as our sole indicator of variable importance. We determined the average permutation importance for each variable in three replicated runs. The top six predictors in the three-run ensemble constituted our preselected variable set. These were then used to develop a final MaxEnt baseline model.

We then developed an alternative method to identify the top six variables using a random selection of variables to produce sets of predictors for repeated MaxEnt runs. We implemented our Monte Carlo approach as an R script that invokes MaxEnt through ENMeval, which provides convenient control over model settings, built-in evaluation metrics, and improved performance [38, 59]. To reduce variability and isolate outcomes as much as possible to the effects of the sampling process, we again used MaxEnt's default feature class setting of LQHP and a regularization multiplier setting of 1.0 as fixed parameters in all the Monte Carlo runs. We defined ensemble, in this case, to mean a collection of 100 sprints, where each sprint consisted of ten runs. A tally table was used to maintain a count of the number of times a variable was used in a run along with a cumulative sum of the variable's permutation importance. The tally table thus provided the information needed to determine the average permutation importance of a predictor at any point along the way.

To process a sprint, we initialized each of its ten model runs with a random subset of environmental variables read from the filesystem. Random integers drawn from a uniform distribution ranging 1–19 corresponding to the 19 bioclim predictors were used to make the selection. At the conclusion of each run, the tally table was updated appropriately. At the conclusion of each sprint, we computed a MaxEnt model using the six predictors in the original starting set having the highest average permutation importance values at that point. This process was repeated 100 times to produce a complete ensemble. This resulted in an evolving progression of models that converged on a stable assemblage of top six predictors over the course of an ensemble. We assessed the algorithm's performance in two ensembles. In the first, we chose two random variables for each sprint run; in the second, six random variables were used for each run. This resulted in an overall total of 2000 MaxEnt runs.

Given our focus on variable screening as an initial step in the modeling process, we did not perform a comprehensive analysis on any of the six-variable final models. We did, however, look at several attributes of these models to gain a general understanding of how the Monte Carlo algorithm was performing. The predictive distribution maps produced by the models were judged for reasonableness based on first-hand knowledge of the species, its habitat preferences, and known range [50]. We further compared these predictions to observational records from Cornell Lab's eBird citizen-scientist database [66]. We used the area under the operating curve (AUC) [67] as an indication of a model's classification accuracy (higher values indicating greater accuracy) and the Akaike information criterion corrected for small sample size (AICc) [68] as a measure of relative explanatory power (lower values indicating less information loss). Model similarity was compared with Warren's I-statistic [69] and Schoener's D statistic [70] (higher values in both indicating greater similarity) using ENMTools. Single-processor run times were recorded to aid our understanding of algorithm performance and help identify opportunities for multiprocessor parallelization. Input data and the R script used in the study are provided as S1 File.

Results

On the basis of permutation importance, 13 of the 19 original bioclim variables were among the top ten most contributory predictors across all three replicated runs of the MaxEnt baseline: bio02, bio03, bio05, bio06, bio08–bio12, bio14, bio15, bio17, and bio18 (Table 2). Of those, bio02, bio05, and bio14 appeared in only one run each at 10th place. Bio18 showed strong dominance throughout. When performance was averaged across all three runs, the top six contributory variables in the ensemble collectively accounted for 65% of overall permutation importance (ensemble average). In descending order of importance, the top six predictors included bio18, bio03, bio10, bio15, bio11, and bio06. When these six top-contributing variables were used in a final MaxEnt run, the model's four most contributory variables (bio18, bio03, bio10, and bio15) accounted for approximately 86% of overall permutation importance, and its predicted habitat suitability distribution corresponded well with what is known about the natural history of the species and observational records for Cassin's Sparrow for the year 2016 (Fig 1) [66].

A distinct pattern of progression toward a stable subset of key variables was observed in the Monte Carlo ensembles (Figs 2 and 3). In both cases, the top three contributory variables among the top six were selected early in the sprint runs, and AICc values fluctuated within a narrow range around an average that changed little over the course of the selection process. Greater variability in the composition of the top six subset was seen in Ensemble #1 where two random variables at a time were selected for each sprint run (Table 2 and Fig 2). In Ensemble #2, where six random variables at a time were selected for the MaxEnt runs, the top six variables were identified by the 25th sprint and had settled into their final rank order by sprint 54 (Fig 3). Ensemble #2 appeared to produce the best overall results and shared four variables in common with the top six selected by the MaxEnt baseline (bio18, bio03, bio11, and bio06) (Table 2). Ensemble #2's final model had the lowest overall AICc, and its four most contributory variables accounted for approximately 93% of overall permutation importance, the highest attained overall.

Ensemble #1 had only one variable in common with the top six selected by both the baseline run and Ensemble #2. What accounts for this difference is not immediately apparent; however, we speculate that the random pair-wise comparisons occurring in Ensemble #1 may alter the relative global influence of the collinearities known to exist in the bioclim variables [73–75]. The average number of times a variable was sampled appeared to have a marginal, positive

Table 2. Results of MaxEnt baseline and Monte Carlo selection trials.

	Bioclim Environmental Variables Permutation Importance																					Permutation Importance		Total Run	Total Run Time		Avg Random
MODELS	bio01	bio02	bio03	bio04	bio05	bio06	bio07	bio08	bio09	bio10	bio11	bio12	bio13	bio14	bio15	bio16	bio17	bio18	bio19	AUC	AICc	% Top 3	% Top 4	Count	Mins	Hrs	Samples / Variable
Maxent Baseline																											
Run #1	1.0	4.3	5.6	2.8	0.6	0.9	0.9	5.2	4.2	8.4	4.9	4.6	1.0	2.7	8.3	0.0	2.2	30.6	1.7					1	151	2.5	–
Run #2	1.7	2.9	6.4	1.9	3.5	9.9	0.5	3.5	3.9	5.4	7.7	4.8	1.0	3.5	8.3	0.8	0.5	30.1	3.6					1	120	2.0	–
Run #3	0.8	3.4	6.1	1.3	1.9	10.4	1.0	2.5	4.5	7.6	8.3	4.6	2.3	3.4	7.1	0.1	3.4	28.9	2.5					1	100	1.7	–
Ensemble avg	1.2	3.5	6.0	2.0	2.0	7.1	0.8	3.7	4.2	7.1	7.0	4.7	1.4	3.2	7.9	0.3	2.0	29.9	2.6								
Final model			14.1			6.6				13.8	7.3				10.9			47.4		0.818	12,222	75.3	86.2	1	18.0	0.3	–
Monte Carlo Selection																											
Ensemble #1	*(two random variables per sprint run)*																										
Sprint 025					17.6			10.8	9.3				22.2			14.3		25.9		0.801	12,229	65.7	80.0	250	181	3.0	26
Sprint 050					15.1			11.7	11.2				21.2			15.2		25.5		0.802	12,231	61.9	77.0	500	355	5.9	53
Sprint 100				24.1	18.2			6.2					7.6			4.1		39.7		0.806	12,252	82.0	89.6	1000	710	11.8	105
Ensemble #2	*(six random variables per sprint run)*																										
Sprint 025			25.4			3.6					15.6		7.7			3.8		43.9		0.801	12,376	84.9	92.6	250	438	7.3	79
Sprint 050			26.6			4.9					13.5		3.8			3.7		47.7		0.807	12,168	87.8	92.7	500	856	14.3	158
Sprint 100			26.5			5.5					18.9		4.1			2.6		42.4		0.805	12,152	87.8	93.3	1000	1793	29.9	316

https://doi.org/10.1371/journal.pone.0237208.t002

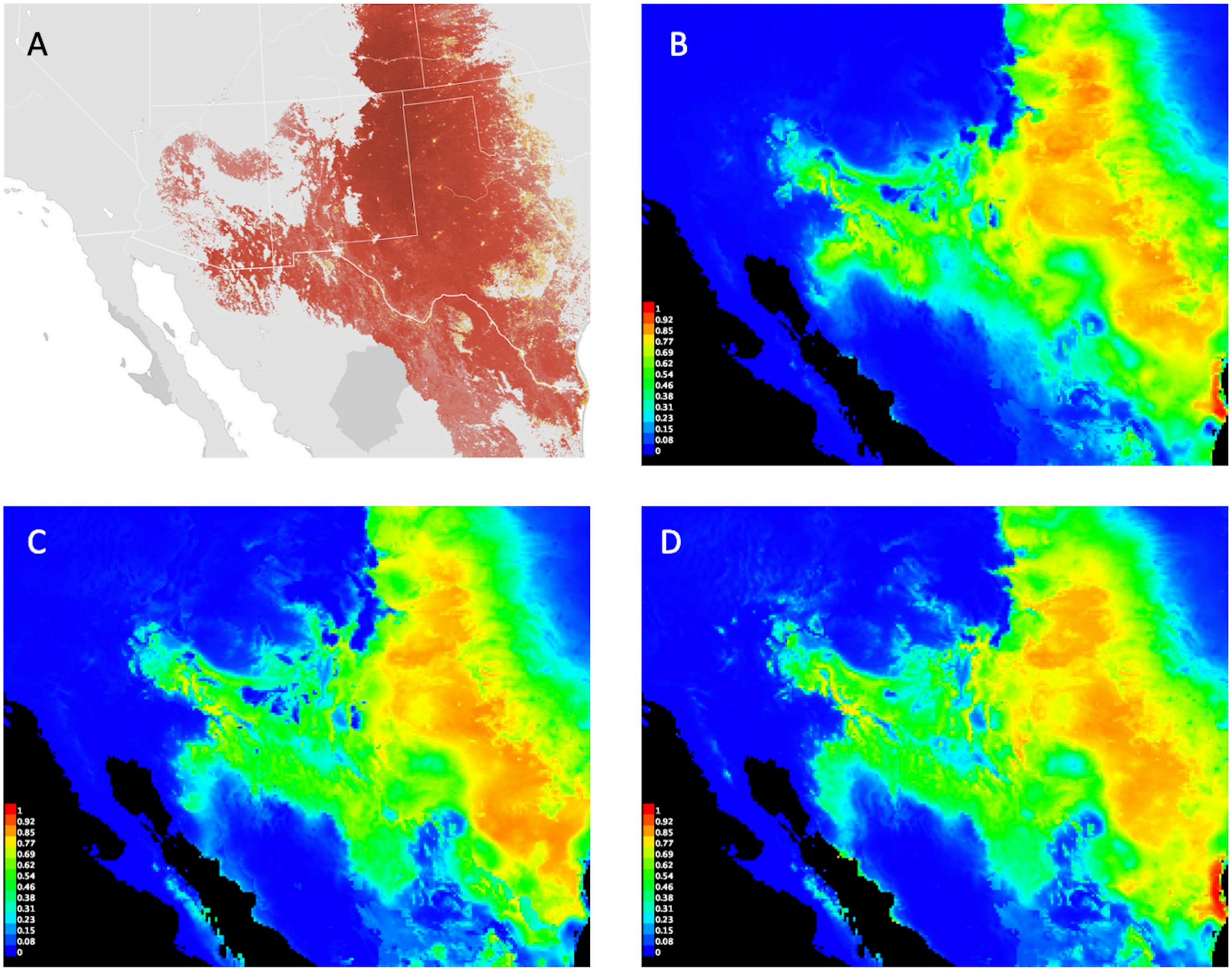


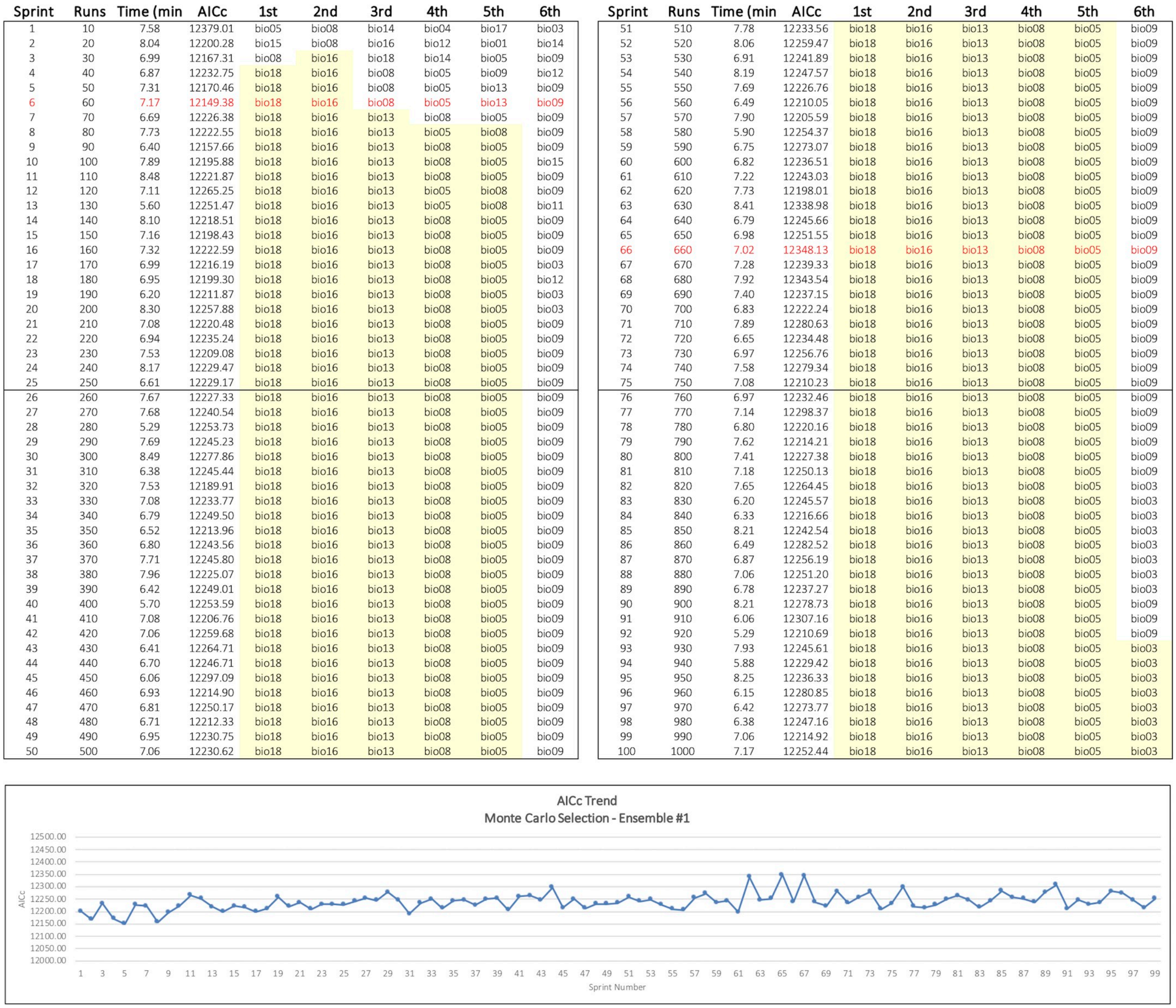
Fig 1. Cassin's Sparrow distribution maps. Cassin's Sparrow range map (A) compared to the species' predicted habitat suitability distributions obtained from the MaxEnt baseline (B), Monte Carlo Ensemble #1 (C), and Monte Carlo Ensemble #2 (D). Image (A) provided by eBird (www.ebird.org), created 28 July 2020, and reprinted from [71] under a CC BY license, with permission from the Cornell Lab of Ornithology. Images (B)–(D) created by the authors show MaxEnt logistic output, which can be interpreted as an estimated probability of presence between 0 and 1 with warmer colors indicating better predicted conditions [72].

https://doi.org/10.1371/journal.pone.0237208.g001

influence on resulting model quality once an adequate minimum was attained. Ensemble #2 results suggest that at least 80 uniformly distributed samples per starting-set variable are needed to identify a reasonable top six set of variables; the best overall model resulted from over 300 samples per variable (Table 2).

Discussion

The most striking outcome of the study is the Monte Carlo method's ability to select a set of top-contributing predictors by randomly sampling a collection of variables that is comparable to the top-contributing predictors identified by MaxEnt when it operates on the collection as a whole (Table 2). The top six selected predictors in the MaxEnt baseline and the Monte Carlo ensembles produced predicted habitat suitability distributions that are nearly indistinguishable



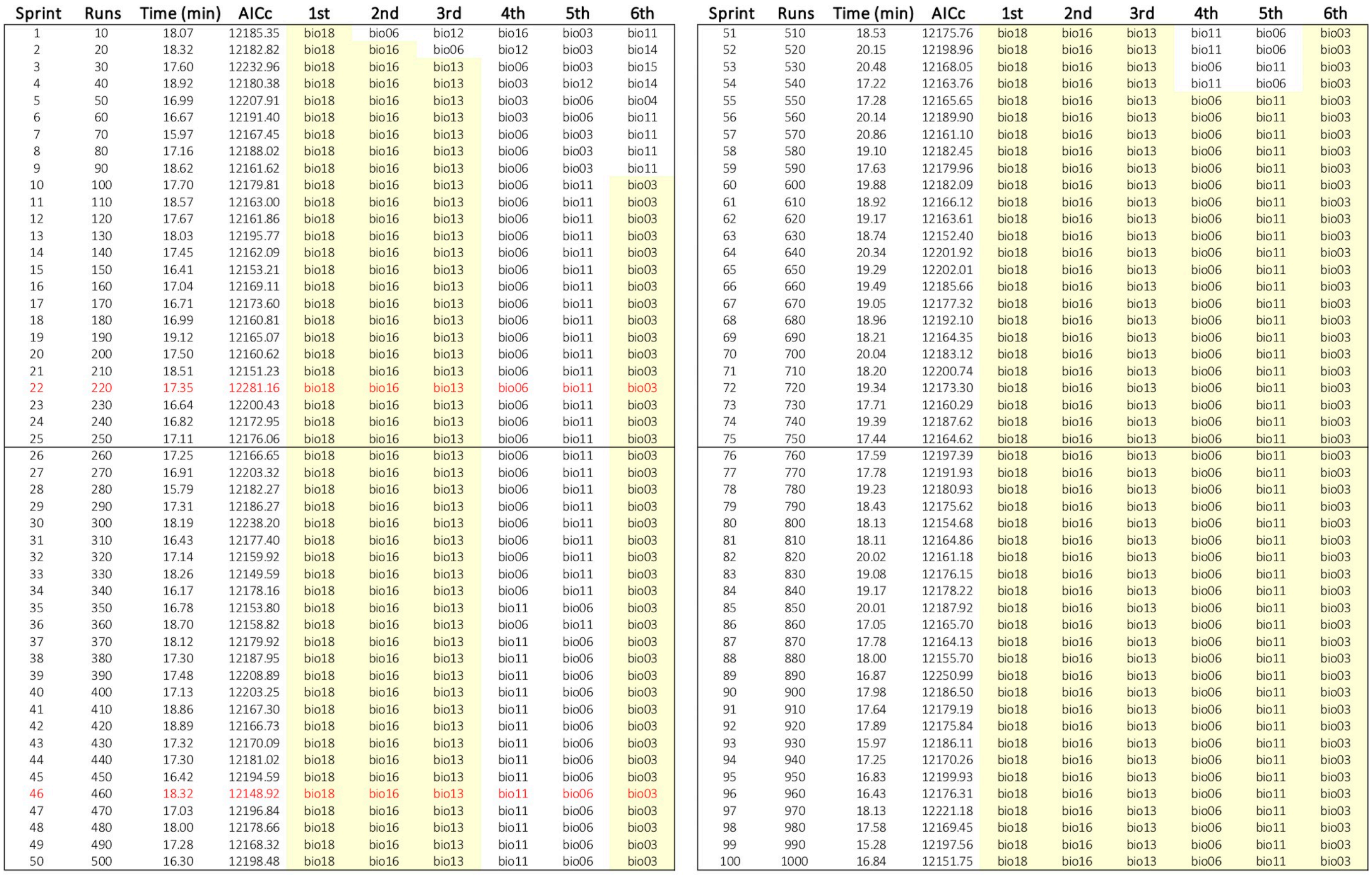
Sprint	Runs	Time (min	AICc	1st	2nd	3rd	4th	5th	6th
1	10	7.58	12379.01	bio05	bio08	bio14	bio04	bio17	bio03
2	20	8.04	12200.28	bio15	bio08	bio16	bio12	bio01	bio14
3	30	6.99	12167.31	bio08	bio16	bio18	bio14	bio05	bio09
4	40	6.87	12232.75	bio18	bio16	bio08	bio05	bio09	bio12
5	50	7.31	12170.46	bio18	bio16	bio08	bio05	bio13	bio09
6	60	7.17	12149.38	bio18	bio16	bio08	bio05	bio13	bio09
7	70	6.69	12226.38	bio18	bio16	bio13	bio08	bio05	bio09
8	80	7.73	12222.55	bio18	bio16	bio13	bio05	bio08	bio09
9	90	6.40	12157.66	bio18	bio16	bio13	bio08	bio05	bio09
10	100	7.89	12195.88	bio18	bio16	bio13	bio08	bio05	bio15
11	110	8.48	12221.87	bio18	bio16	bio13	bio08	bio05	bio09
12	120	7.11	12265.25	bio18	bio16	bio13	bio05	bio08	bio09
13	130	5.60	12251.47	bio18	bio16	bio13	bio05	bio08	bio11
14	140	8.10	12218.51	bio18	bio16	bio13	bio08	bio05	bio09
15	150	7.16	12198.43	bio18	bio16	bio13	bio08	bio05	bio09
16	160	7.32	12222.59	bio18	bio16	bio13	bio08	bio05	bio09
17	170	6.99	12216.19	bio18	bio16	bio13	bio08	bio05	bio03
18	180	6.95	12199.30	bio18	bio16	bio13	bio08	bio05	bio12
19	190	6.20	12211.87	bio18	bio16	bio13	bio08	bio05	bio03
20	200	8.30	12257.88	bio18	bio16	bio13	bio08	bio05	bio03
21	210	7.08	12220.48	bio18	bio16	bio13	bio08	bio05	bio09
22	220	6.94	12235.24	bio18	bio16	bio13	bio08	bio05	bio09
23	230	7.53	12209.08	bio18	bio16	bio13	bio08	bio05	bio09
24	240	8.17	12229.47	bio18	bio16	bio13	bio08	bio05	bio09
25	250	6.61	12229.17	bio18	bio16	bio13	bio08	bio05	bio09
26	260	7.67	12227.33	bio18	bio16	bio13	bio08	bio05	bio09
27	270	7.68	12240.54	bio18	bio16	bio13	bio08	bio05	bio09
28	280	5.29	12253.73	bio18	bio16	bio13	bio08	bio05	bio09
29	290	7.69	12245.23	bio18	bio16	bio13	bio08	bio05	bio09
30	300	8.49	12277.86	bio18	bio16	bio13	bio08	bio05	bio09
31	310	6.38	12245.44	bio18	bio16	bio13	bio08	bio05	bio09
32	320	7.53	12189.91	bio18	bio16	bio13	bio08	bio05	bio09
33	330	7.08	12233.77	bio18	bio16	bio13	bio08	bio05	bio09
34	340	6.79	12249.50	bio18	bio16	bio13	bio08	bio05	bio09
35	350	6.52	12213.96	bio18	bio16	bio13	bio08	bio05	bio09
36	360	6.80	12243.56	bio18	bio16	bio13	bio08	bio05	bio09
37	370	7.71	12245.80	bio18	bio16	bio13	bio08	bio05	bio09
38	380	7.96	12225.07	bio18	bio16	bio13	bio08	bio05	bio09
39	390	6.42	12249.01	bio18	bio16	bio13	bio08	bio05	bio09
40	400	5.70	12253.59	bio18	bio16	bio13	bio08	bio05	bio09
41	410	7.08	12206.76	bio18	bio16	bio13	bio08	bio05	bio09
42	420	7.06	12259.68	bio18	bio16	bio13	bio08	bio05	bio09
43	430	6.41	12264.71	bio18	bio16	bio13	bio08	bio05	bio09
44	440	6.70	12246.71	bio18	bio16	bio13	bio08	bio05	bio09
45	450	6.06	12297.09	bio18	bio16	bio13	bio08	bio05	bio09
46	460	6.93	12214.90	bio18	bio16	bio13	bio08	bio05	bio09
47	470	6.81	12250.17	bio18	bio16	bio13	bio08	bio05	bio09
48	480	6.71	12212.33	bio18	bio16	bio13	bio08	bio05	bio09
49	490	6.95	12230.75	bio18	bio16	bio13	bio08	bio05	bio09
50	500	7.06	12230.62	bio18	bio16	bio13	bio08	bio05	bio09
51	510	7.78	12233.56	bio18	bio16	bio13	bio08	bio05	bio09
52	520	8.06	12259.47	bio18	bio16	bio13	bio08	bio05	bio09
53	530	6.91	12241.89	bio18	bio16	bio13	bio08	bio05	bio09
54	540	8.19	12247.57	bio18	bio16	bio13	bio08	bio05	bio09
55	550	7.69	12226.76	bio18	bio16	bio13	bio08	bio05	bio09
56	560	6.49	12210.05	bio18	bio16	bio13	bio08	bio05	bio09
57	570	7.90	12205.59	bio18	bio16	bio13	bio08	bio05	bio09
58	580	5.90	12254.37	bio18	bio16	bio13	bio08	bio05	bio09
59	590	6.75	12273.07	bio18	bio16	bio13	bio08	bio05	bio09
60	600	6.82	12236.51	bio18	bio16	bio13	bio08	bio05	bio09
61	610	7.22	12243.03	bio18	bio16	bio13	bio08	bio05	bio09
62	620	7.73	12198.01	bio18	bio16	bio13	bio08	bio05	bio09
63	630	8.41	12338.98	bio18	bio16	bio13	bio08	bio05	bio09
64	640	6.79	12245.66	bio18	bio16	bio13	bio08	bio05	bio09
65	650	6.98	12251.55	bio18	bio16	bio13	bio08	bio05	bio09
66	660	7.02	12348.13	bio18	bio16	bio13	bio08	bio05	bio09
67	670	7.28	12239.33	bio18	bio16	bio13	bio08	bio05	bio09
68	680	7.92	12343.54	bio18	bio16	bio13	bio08	bio05	bio09
69	690	7.40	12237.15	bio18	bio16	bio13	bio08	bio05	bio09
70	700	6.83	12222.24	bio18	bio16	bio13	bio08	bio05	bio09
71	710	7.89	12280.63	bio18	bio16	bio13	bio08	bio05	bio09
72	720	6.65	12234.48	bio18	bio16	bio13	bio08	bio05	bio09
73	730	6.97	12256.76	bio18	bio16	bio13	bio08	bio05	bio09
74	740	7.58	12279.34	bio18	bio16	bio13	bio08	bio05	bio09
75	750	7.08	12210.23	bio18	bio16	bio13	bio08	bio05	bio09
76	760	6.97	12232.46	bio18	bio16	bio13	bio08	bio05	bio09
77	770	7.14	12298.37	bio18	bio16	bio13	bio08	bio05	bio09
78	780	6.80	12220.16	bio18	bio16	bio13	bio08	bio05	bio09
79	790	7.62	12214.21	bio18	bio16	bio13	bio08	bio05	bio09
80	800	7.41	12227.38	bio18	bio16	bio13	bio08	bio05	bio09
81	810	7.18	12250.13	bio18	bio16	bio13	bio08	bio05	bio09
82	820	7.65	12264.45	bio18	bio16	bio13	bio08	bio05	bio03
83	830	6.20	12245.57	bio18	bio16	bio13	bio08	bio05	bio03
84	840	6.33	12216.66	bio18	bio16	bio13	bio08	bio05	bio03
85	850	8.21	12242.54	bio18	bio16	bio13	bio08	bio05	bio03
86	860	6.49	12282.52	bio18	bio16	bio13	bio08	bio05	bio03
87	870	6.87	12256.19	bio18	bio16	bio13	bio08	bio05	bio03
88	880	7.06	12251.20	bio18	bio16	bio13	bio08	bio05	bio03
89	890	6.78	12237.27	bio18	bio16	bio13	bio08	bio05	bio03
90	900	8.21	12278.73	bio18	bio16	bio13	bio08	bio05	bio09
91	910	6.06	12307.16	bio18	bio16	bio13	bio08	bio05	bio09
92	920	5.29	12210.69	bio18	bio16	bio13	bio08	bio05	bio09
93	930	7.93	12245.61	bio18	bio16	bio13	bio08	bio05	bio03
94	940	5.88	12229.42	bio18	bio16	bio13	bio08	bio05	bio03
95	950	8.25	12236.33	bio18	bio16	bio13	bio08	bio05	bio03
96	960	6.15	12280.85	bio18	bio16	bio13	bio08	bio05	bio03
97	970	6.42	12273.77	bio18	bio16	bio13	bio08	bio05	bio03
98	980	6.38	12247.16	bio18	bio16	bio13	bio08	bio05	bio03
99	990	7.06	12214.92	bio18	bio16	bio13	bio08	bio05	bio03
100	1000	7.17	12252.44	bio18	bio16	bio13	bio08	bio05	bio03

Fig 2. Monte Carlo Ensemble #1 results. Two random variables at a time were chosen for each MaxEnt sprint run. The sprint log on top shows the progressive selection of a stable set of top six variables in yellow. The graph on the bottom shows the narrow range of fluctuating AICc values over the course of the ensemble runs. Maximum and minimum AICc values are shown in red.

https://doi.org/10.1371/journal.pone.0237208.g002

from one another (Fig 1). The two approaches each identified four variables that collectively contributed more than 80% to the formulation of their respective models. And across the board, models based on selected predictors showed a high degree of similarity in Schoener's D and the I-statistic (Table 3). This gives us confidence that the Monte Carlo method will be able to preselect viable predictors when applied to a larger variable pool.

We note that among the top six variables resulting from all the MaxEnt runs, only two (bio06 and bio16) present collinearity issues with respect to the other selected variables: bio06/bio11, bio16/bio13, and bio16/bio18 are potentially problematic pairs. (Tables 2 and S1). However, there were no collinearity issues among the top four variables in any of the runs, and the top four selected by the Monte Carlo method contributed significantly to their models, with permutation importance ranging from 77% to 93% (Table 2). When used as an initial



Sprint	Runs	Time (min)	AICc	1st	2nd	3rd	4th	5th	6th
1	10	18.07	12185.35	bio18	bio06	bio12	bio16	bio03	bio11
2	20	18.32	12182.82	bio18	bio16	bio06	bio12	bio03	bio14
3	30	17.60	12232.96	bio18	bio16	bio13	bio06	bio03	bio15
4	40	18.92	12180.38	bio18	bio16	bio13	bio03	bio12	bio14
5	50	16.99	12207.91	bio18	bio16	bio13	bio03	bio06	bio04
6	60	16.67	12191.40	bio18	bio16	bio13	bio03	bio06	bio11
7	70	15.97	12167.45	bio18	bio16	bio13	bio06	bio03	bio11
8	80	17.16	12188.02	bio18	bio16	bio13	bio06	bio03	bio11
9	90	18.62	12161.62	bio18	bio16	bio13	bio06	bio03	bio11
10	100	17.70	12179.81	bio18	bio16	bio13	bio06	bio11	bio03
11	110	18.57	12163.00	bio18	bio16	bio13	bio06	bio11	bio03
12	120	17.67	12161.86	bio18	bio16	bio13	bio06	bio11	bio03
13	130	18.03	12195.77	bio18	bio16	bio13	bio06	bio11	bio03
14	140	17.45	12162.09	bio18	bio16	bio13	bio06	bio11	bio03
15	150	16.41	12153.21	bio18	bio16	bio13	bio06	bio11	bio03
16	160	17.04	12169.11	bio18	bio16	bio13	bio06	bio11	bio03
17	170	16.71	12173.60	bio18	bio16	bio13	bio06	bio11	bio03
18	180	16.99	12160.81	bio18	bio16	bio13	bio06	bio11	bio03
19	190	19.12	12165.07	bio18	bio16	bio13	bio06	bio11	bio03
20	200	17.50	12160.62	bio18	bio16	bio13	bio06	bio11	bio03
21	210	18.51	12151.23	bio18	bio16	bio13	bio06	bio11	bio03
22	220	17.35	12281.16	bio18	bio16	bio13	bio06	bio11	bio03
23	230	16.64	12200.43	bio18	bio16	bio13	bio06	bio11	bio03
24	240	16.82	12172.95	bio18	bio16	bio13	bio06	bio11	bio03
25	250	17.11	12176.06	bio18	bio16	bio13	bio06	bio11	bio03
26	260	17.25	12166.65	bio18	bio16	bio13	bio06	bio11	bio03
27	270	16.91	12203.32	bio18	bio16	bio13	bio06	bio11	bio03
28	280	15.79	12182.27	bio18	bio16	bio13	bio06	bio11	bio03
29	290	17.31	12186.27	bio18	bio16	bio13	bio06	bio11	bio03
30	300	18.19	12238.20	bio18	bio16	bio13	bio06	bio11	bio03
31	310	16.43	12177.40	bio18	bio16	bio13	bio06	bio11	bio03
32	320	17.14	12159.92	bio18	bio16	bio13	bio06	bio11	bio03
33	330	18.26	12149.59	bio18	bio16	bio13	bio06	bio11	bio03
34	340	16.17	12178.16	bio18	bio16	bio13	bio06	bio11	bio03
35	350	16.78	12153.80	bio18	bio16	bio13	bio11	bio06	bio03
36	360	18.70	12158.82	bio18	bio16	bio13	bio06	bio11	bio03
37	370	18.12	12179.92	bio18	bio16	bio13	bio11	bio06	bio03
38	380	17.30	12187.95	bio18	bio16	bio13	bio11	bio06	bio03
39	390	17.48	12208.89	bio18	bio16	bio13	bio11	bio06	bio03
40	400	17.13	12203.25	bio18	bio16	bio13	bio11	bio06	bio03
41	410	18.86	12167.30	bio18	bio16	bio13	bio11	bio06	bio03
42	420	18.89	12166.73	bio18	bio16	bio13	bio11	bio06	bio03
43	430	17.32	12170.09	bio18	bio16	bio13	bio11	bio06	bio03
44	440	17.30	12181.02	bio18	bio16	bio13	bio11	bio06	bio03
45	450	16.42	12194.59	bio18	bio16	bio13	bio11	bio06	bio03
46	460	18.32	12148.92	bio18	bio16	bio13	bio11	bio06	bio03
47	470	17.03	12196.84	bio18	bio16	bio13	bio11	bio06	bio03
48	480	18.00	12178.66	bio18	bio16	bio13	bio11	bio06	bio03
49	490	17.28	12168.32	bio18	bio16	bio13	bio11	bio06	bio03
50	500	16.30	12198.48	bio18	bio16	bio13	bio11	bio06	bio03
51	510	18.53	12175.76	bio18	bio16	bio13	bio11	bio06	bio03
52	520	20.15	12198.96	bio18	bio16	bio13	bio11	bio06	bio03
53	530	20.48	12168.05	bio18	bio16	bio13	bio06	bio11	bio03
54	540	17.22	12163.76	bio18	bio16	bio13	bio11	bio06	bio03
55	550	17.28	12165.65	bio18	bio16	bio13	bio06	bio11	bio03
56	560	20.14	12189.90	bio18	bio16	bio13	bio06	bio11	bio03
57	570	20.86	12161.10	bio18	bio16	bio13	bio06	bio11	bio03
58	580	19.10	12182.45	bio18	bio16	bio13	bio06	bio11	bio03
59	590	17.63	12179.96	bio18	bio16	bio13	bio06	bio11	bio03
60	600	19.88	12182.09	bio18	bio16	bio13	bio06	bio11	bio03
61	610	18.92	12166.12	bio18	bio16	bio13	bio06	bio11	bio03
62	620	19.17	12163.61	bio18	bio16	bio13	bio06	bio11	bio03
63	630	18.74	12152.40	bio18	bio16	bio13	bio06	bio11	bio03
64	640	20.34	12201.92	bio18	bio16	bio13	bio06	bio11	bio03
65	650	19.29	12202.01	bio18	bio16	bio13	bio06	bio11	bio03
66	660	19.49	12185.66	bio18	bio16	bio13	bio06	bio11	bio03
67	670	19.05	12177.32	bio18	bio16	bio13	bio06	bio11	bio03
68	680	18.96	12192.10	bio18	bio16	bio13	bio06	bio11	bio03
69	690	18.21	12164.35	bio18	bio16	bio13	bio06	bio11	bio03
70	700	20.04	12183.12	bio18	bio16	bio13	bio06	bio11	bio03
71	710	18.20	12200.74	bio18	bio16	bio13	bio06	bio11	bio03
72	720	19.34	12173.30	bio18	bio16	bio13	bio06	bio11	bio03
73	730	17.71	12160.29	bio18	bio16	bio13	bio06	bio11	bio03
74	740	19.39	12187.62	bio18	bio16	bio13	bio06	bio11	bio03
75	750	17.44	12164.62	bio18	bio16	bio13	bio06	bio11	bio03
76	760	17.59	12197.39	bio18	bio16	bio13	bio06	bio11	bio03
77	770	17.78	12191.93	bio18	bio16	bio13	bio06	bio11	bio03
78	780	19.23	12180.93	bio18	bio16	bio13	bio06	bio11	bio03
79	790	18.43	12175.62	bio18	bio16	bio13	bio06	bio11	bio03
80	800	18.13	12154.68	bio18	bio16	bio13	bio06	bio11	bio03
81	810	18.11	12164.86	bio18	bio16	bio13	bio06	bio11	bio03
82	820	20.02	12161.18	bio18	bio16	bio13	bio06	bio11	bio03
83	830	19.08	12176.15	bio18	bio16	bio13	bio06	bio11	bio03
84	840	19.17	12178.22	bio18	bio16	bio13	bio06	bio11	bio03
85	850	20.01	12187.92	bio18	bio16	bio13	bio06	bio11	bio03
86	860	17.05	12165.70	bio18	bio16	bio13	bio06	bio11	bio03
87	870	17.78	12164.13	bio18	bio16	bio13	bio06	bio11	bio03
88	880	18.00	12155.70	bio18	bio16	bio13	bio06	bio11	bio03
89	890	16.87	12250.99	bio18	bio16	bio13	bio06	bio11	bio03
90	900	17.98	12186.50	bio18	bio16	bio13	bio06	bio11	bio03
91	910	17.64	12179.19	bio18	bio16	bio13	bio06	bio11	bio03
92	920	17.89	12175.84	bio18	bio16	bio13	bio06	bio11	bio03
93	930	15.97	12186.11	bio18	bio16	bio13	bio06	bio11	bio03
94	940	17.25	12170.26	bio18	bio16	bio13	bio06	bio11	bio03
95	950	16.83	12199.93	bio18	bio16	bio13	bio06	bio11	bio03
96	960	16.43	12176.31	bio18	bio16	bio13	bio06	bio11	bio03
97	970	18.13	12221.18	bio18	bio16	bio13	bio06	bio11	bio03
98	980	17.58	12169.45	bio18	bio16	bio13	bio06	bio11	bio03
99	990	15.28	12197.56	bio18	bio16	bio13	bio06	bio11	bio03
100	1000	16.84	12151.75	bio18	bio16	bio13	bio06	bio11	bio03

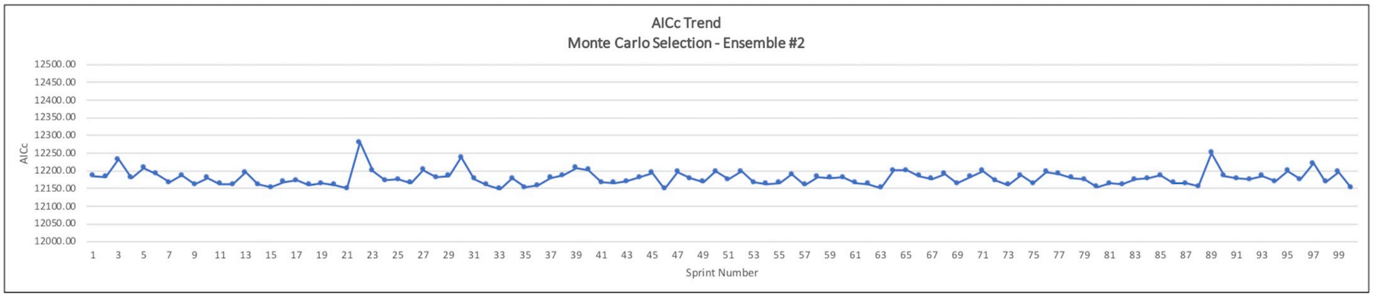


Fig 3. Monte Carlo Ensemble #2 results. Six random variables at a time were chosen for each MaxEnt sprint run. The sprint log on top shows the progressive selection of a stable set of top six variables in yellow. The graph on the bottom shows the narrow range of fluctuating AICc values over the course of the ensemble runs. Maximum and minimum AICc values are shown in red.

https://doi.org/10.1371/journal.pone.0237208.g003

screening step, it would be crucial, at this point, for modelers to perform other quality control steps prior to final model construction.

While perfecting an ENM for Cassin's Sparrow was not a goal in this study, we also note that the selected variables have biological relevance. In particular, the three temperature-derived variables, bio03 (isothermality), bio06 (minimum temperature of coldest month), and bio11 (mean temperature of coldest quarter), and the precipitation-derived bio18 (precipitation of warmest quarter) have been identified as important influences on the distribution of arid-adapted birds in general and Cassin's Sparrow in particular [50, 76–79].

Table 3. Model similarity metrics.

Schoener's D Statistic										
MODELS ↓ →	Maxent-Run1	Maxent-Run2	Maxent-Run3	Maxent-Final	MC-E1-025	MC-E1-050	MC-E1-100	MC-E2-025	MC-E2-050	MC-E2-100
Maxent-Run1	1	0.9712	0.9738	0.9355	0.8629	0.8630	0.9044	0.8882	0.8903	0.8906
Maxent-Run2	x	1	0.9793	0.9397	0.8639	0.8648	0.9078	0.8915	0.8949	0.8947
Maxent-Run3	x	x	1	0.9354	0.8579	0.8586	0.9039	0.8871	0.8903	0.8902
Maxent-Final	x	x	x	1	0.8667	0.8673	0.9214	0.9104	0.9138	0.9142
MC-E1-025	x	x	x	x	1	0.9880	0.9006	0.8810	0.8801	0.8785
MC-E1-050	x	x	x	x	x	1	0.9026	0.8813	0.8804	0.8786
MC-E1-100	x	x	x	x	x	x	1	0.9393	0.9389	0.9365
MC-E2-025	x	x	x	x	x	x	x	1	0.9815	0.9810
MC-E2-050	x	x	x	x	x	x	x	x	1	0.9844
MC-E2-100	x	x	x	x	x	x	x	x	x	1
Warren's I Statistic										
MODELS ↓ →	Maxent-Run1	Maxent-Run2	Maxent-Run3	Maxent-Final	MC-E1-025	MC-E1-050	MC-E1-100	MC-E2-050	MC-E2-100	MC-E2-100
Maxent-Run1	1	0.9991	0.9994	0.9948	0.9760	0.9758	0.9874	0.9828	0.9833	0.9834
Maxent-Run2	x	1	0.9995	0.9949	0.9760	0.9760	0.9878	0.9831	0.9838	0.9839
Maxent-Run3	x	x	1	0.9945	0.9748	0.9748	0.9869	0.9823	0.9830	0.9831
Maxent-Final	x	x	x	1	0.9755	0.9753	0.9894	0.9871	0.9878	0.9880
MC-E1-025	x	x	x	x	1	0.9998	0.9887	0.9798	0.9797	0.9784
MC-E1-050	x	x	x	x	x	1	0.9889	0.9795	0.9794	0.9781
MC-E1-100	x	x	x	x	x	x	1	0.9910	0.9912	0.9906
MC-E2-025	x	x	x	x	x	x	x	1	0.9996	0.9994
MC-E2-050	x	x	x	x	x	x	x	x	1	0.9996
MC-E2-100	x	x	x	x	x	x	x	x	x	1

https://doi.org/10.1371/journal.pone.0237208.t003

The most significant drawback identified in the study was the long run times. MaxEnt's linear scaling behavior can be challenging in a single-processor environment. In the baseline runs, producing a single model through MaxEnt's GUI using our selected settings involved writing many files to disk and took from 18 minutes (with six variables) to over two hours (with all 19 variables). MaxEnt in the R environment outputs memory-resident objects, which results in faster run times. Still, with its repeated invocations of MaxEnt, Ensemble #2 took nearly 30 hours to complete (Table 2). This too is a result of a linear scaling property; however, the Monte Carlo method's scaling behavior is not determined by the MaxEnt program, since each of the MaxEnt runs in the Monte Carlo method operates on a set number of predictors. The method's linear scaling property is determined, instead, by the need to adequately sample the starting set of environmental variables in order to obtain a good result.

This is an important distinction. It means that each of the Monte Carlo MaxEnt runs is entirely independent from all other runs in the ensemble. This high level of subtask independence is sometimes referred to as an "embarrassingly parallel" workload. It makes practical, multiprocessor implementations of the method possible. If 1000 processors were recruited into service—which is becoming increasingly convenient with the proliferation of multiprocessor, high-performance cloud computing—a 1000-run ensemble could conceivably take as long as a single MaxEnt run.

The potential significance of this advantage becomes apparent when one considers the method's use with large collections of environmental data. The Monte Carlo approach described here provides an approximate solution to the problem of finding a useful *k*-size subset of an *n*-size collection of variables. In principle, there are $n! / [k!(n-k)!]$ variable combinations to consider in such an evaluation, a staggering 27,000-plus six-variable subsets with the 19 bioclim variables alone. Algorithms that accomplish variable selection through stepwise removal or are otherwise bound to the linear scaling properties of underlying software components are inherently unable to exhaustively explore this combinatorial space. A Monte Carlo method makes such a search possible by randomly sampling the universe of possible combinations and returning approximate solutions in practical amounts of time, particularly if implemented as a high-performance cloud service.

While these findings are preliminary, they address an important issue facing the modeling community. There is heightened awareness of the significance of dimensionality in understanding environmental spaces and the importance of variable selection in modeling those spaces [23, 34, 80]. This awareness is accompanied by a recognition that logistic difficulties often preclude examining large numbers of variables [62]. This has led to a search for alternative means of variable selection and calls for process automation [22, 23, 62, 78, 81]. A comprehensive review of these approaches is beyond the scope of this paper; however, it is worth nothing that even among the most recent work in this area, many of the solutions put forward —such as manual prescreening for collinear variables, greater use of biological insight in variable selection, broader use of memory-resident machine language-based analysis software, etc. —do not, in general, scale well. They are unlikely to accommodate the petabyte and even larger size data collections on the horizon.

Cobos et al. [22] provide a useful framework for understanding where the results presented here might fit (Fig 4). The work of ecological niche modeling can be thought of as a multi-step process ranging from initial data preparation and cleaning, to model calibration, final model construction, model evaluation, and the assessment of extrapolation risk. Among the tasks

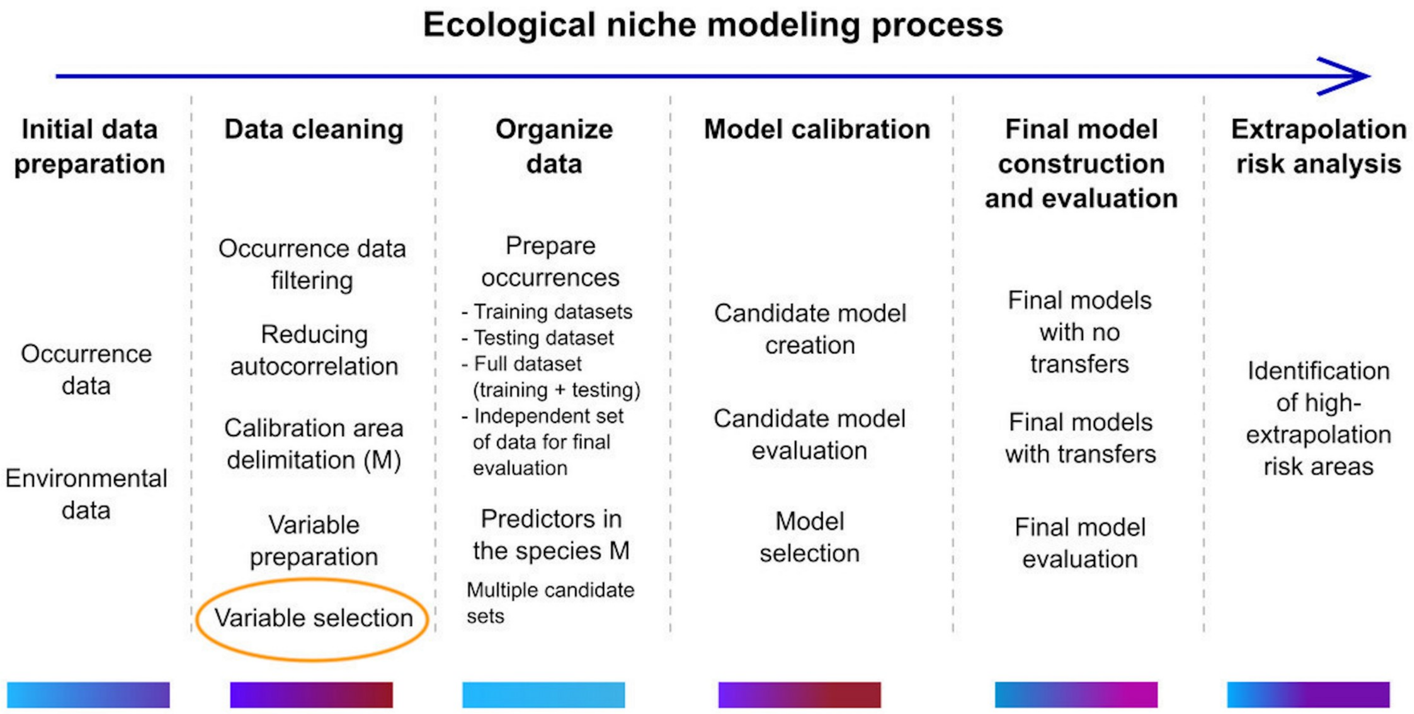

Ecological niche modeling process
Initial data preparation
Occurrence data
Environmental data
Data cleaning
Occurrence data filtering
Reducing autocorrelation
Calibration area delimitation (M)
Variable preparation
Variable selection
Organize data
Prepare occurrences
- Training datasets
- Testing dataset
- Full dataset (training + testing)
- Independent set of data for final evaluation
Predictors in the species M
Multiple candidate sets
Model calibration
Candidate model creation
Candidate model evaluation
Model selection
Final model construction and evaluation
Final models with no transfers
Final models with transfers
Final model evaluation
Extrapolation risk analysis
Identification of high-extrapolation risk areas

Fig 4. Ecological niche modeling process. Schematic description of the ecological niche modeling process. Color bars under each step reflect an approximate range of times that may be needed, ranging from low (blue) to high (red). Use of a Monte Carlo method to prescreen a large collection of predictors could support variable selection in the data cleaning step. Image provided by [22] and adapted for use here under a CC-BY license.

https://doi.org/10.1371/journal.pone.0237208.g004

associated with data cleaning, the selection of viable predictors is crucial, time-consuming, and the place where a means for rapid, automatic, preselection, however coarse, could improve the overall workflow, especially if it enabled exploration of a large universe of predictors.

The use of IPCC-class climate model outputs in efforts to assess the impacts of climate change on biodiversity and other ecosystem processes is growing. Exploring the potential of these massive data sets, expanded use of ensemble modeling, and the actual work of fitting models for the thousands of species scientists wish to study will require hundreds to thousands of projections [16, 25]. An improved capacity to use large environmental data sets in MaxEnt modeling would greatly benefit this work. We are encouraged to think that innovative use of Monte Carlo techniques might provide a helpful means of meeting this challenge.

Conclusions

This small-scale, proof-of-concept study leaves many practical and theoretical questions unanswered. Preliminary results, however, suggest that a Monte Carlo approach might be an effective way to screen environmental predictors prior to final model construction that could be parallelized and scaled to large data sets, including externally-stored collections. This points to the possibility of near-real-time multiprocessor implementations that would enable broader and more exploratory use of global climate model outputs in environmental niche modeling and aid in the discovery of new predictors.

Next steps will focus on implementing a parallel, high-performance version of this capability in NASA's science cloud, evaluating the method's behavior using products generated by the Goddard Earth Observing System, Version 5 (GEOS-5) climate modeling system, extending stochasticity to feature class and regularization multiplier selection, and making various improvements to the algorithm, such as developing automatic stopping rules and developing better measures of variable importance. We also look forward to evaluating the method's effectiveness in addressing research questions relating to climate change influences on Cassin's Sparrow distribution.

Supporting information

S1 Table. Bioclim correlation analysis. Table shows values of Pearson correlation coefficient (r), Pearson coefficient of determination (r^2), and Variance Inflation Factor (VIF) for the Worldclim Bioclimatic variables for the study area [56]. Values of $r > 0.8$, $r2 > 0.8$, and $VIF > 10.0$ are highlighted and indicate highly correlated variables.
(PDF)

S2 Table. ENMeval feature class and regularization multiplier scan. Table shows results from a comprehensive ENMeval scan of the 19 Bioclim variables over the study area [38, 59]. MaxEnt's default feature class and regularization multiplier settings (LQHP, 1.0) resulted in the lowest AICc value and best overall model in the scan.
(PDF)

S1 File. Study data and script. Compressed file folder containing the input data and R script used in the study.
(ZIP)

Acknowledgments

The authors wish to thank our academic editor and a four anonymous reviewers for helpful comments on earlier drafts of the manuscript. We appreciate input from Zeng Yiwen that helped shape our early thinking about the work described here. Steven Phillips provided helpful guidance on the use of MaxEnt maps.

Author Contributions

Conceptualization: John L. Schnase, Mark L. Carroll.

Formal analysis: John L. Schnase, Mark L. Carroll.

Investigation: John L. Schnase.

Methodology: John L. Schnase, Mark L. Carroll.

Resources: Caleb S. Spradlin.

Software: Roger L. Gill, Glenn S. Tamkin, Jian Li, Thomas P. Maxwell.

Validation: Savannah L. Strong.

Visualization: Mary E. Aronne.

Writing – original draft: John L. Schnase.

Writing – review & editing: Mark L. Carroll.

References

1. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Diversity and distributions. 2011; 17: 43–57.
2. Phillips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME. Opening the black box: An open-source release of Maxent. Ecography. 2017; 40: 887–893.
3. Phillips SJ. A Brief Tutorial on Maxent. AT&T Research. 2005; 190: 231–259.
4. Phillips SJ, Anderson RP, Schapire RE. Maximum Entropy Modeling of Species Geographic Distributions. Ecological Modelling. 2006; 190: 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026
5. Kalinski CE. Building Better Species Distribution Models with Machine Learning: Assessing the Role of Covariate Scale and Tuning in Maxent Models. 2019; 129.
6. Merow C, Smith MJ, Silander JA. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography. 2013; 36: 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x
7. Schmidt-Lebuhn AN, Knerr NJ, Miller JT, Mishler BD. Phylogenetic diversity and endemism of Australian daisies (Asteraceae). Journal of Biogeography. 2015; 42: 1114–1122. https://doi.org/10.1111/jbi.12488
8. Warren DL, Wright AN, Seifert SN, Shaffer HB. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. Diversity and distributions. 2014; 20: 334–343.
9. Cardoso-Leite R, Vilarinho AC, Novaes MC, Tonetto AF, Vilardi GC, Guillermo-Ferreira R. Recent and future environmental suitability to dengue fever in Brazil using species distribution model. Transactions of The Royal Society of Tropical Medicine and Hygiene. 2014; 108: 99–104. https://doi.org/10.1093/trstmh/trt115 PMID: 24463584
10. Morisette JT, Jarnevich CS, Ullah A, Cai W, Pedelty JA, Gentle JE, et al. A Tamarisk Habitat Suitability Map for the Continental United States. Frontiers in Ecology and the Environment. 2006; 4: 11–17. https://doi.org/10.1890/1540-9295(2006)004
11. Stohlgren TJ, Schnase JL. Risk Analysis for Biological Hazards: What We Need to Know about Invasive Species. Risk Analysis. 2006; 26: 163–73. https://doi.org/10.1111/j.1539-6924.2006.00707.x PMID: 16492190
12. Beauchamp VB, Koontz SM, Suss C, Hawkins C, Kyde KL, Schnase JL. An Introduction to Oplismenus Undulatifolius (Ard.) Roem. & Schult (Wavyleaf Basketgrass), a Recent Invader in Mid-Atlantic Forest Understories 1,2. The Journal of the Torrey Botanical Society. 2013; 140: 391–413. https://doi.org/10.3159/torrey-d-13-00033.1
13. Muttaqin LA, Murti SH, Susilo B. MaxEnt (Maximum Entropy) model for predicting prehistoric cave sites in Karst area of Gunung Sewu, Gunung Kidul, Yogyakarta. In: Wibowo SB, Rimba AB, A. Aziz A, Phinn

S, Sri Sumantyo JT, Widyasamratri H, et al., editors. Sixth Geoinformation Science Symposium. Yogyakarta, Indonesia: SPIE; 2019. p. 3. https://doi.org/10.1117/12.2543522

14. Feng X, Park DS, Walker C, Peterson AT, Merow C, Papeş M. A checklist for maximizing reproducibility of ecological niche models. Nature Ecology & Evolution. 2019; 3: 1382–1395. https://doi.org/10.1038/s41559-019-0972-5 PMID: 31548646

15. Morales NS, Fernández IC, Baca-González V. MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. PeerJ. 2017; 5: e3093. https://doi.org/10.7717/peerj.3093 PMID: 28316894

16. Araújo M, New M. Ensemble forecasting of species distributions. Trends in Ecology & Evolution. 2007; 22: 42–47. https://doi.org/10.1016/j.tree.2006.09.010 PMID: 17011070

17. Zeng Y, Low BW, Yeo DCJ. Novel methods to select environmental variables in MaxEnt: A case study using invasive crayfish. Ecological Modelling. 2016; 341: 5–13. https://doi.org/10.1016/j.ecolmodel.2016.09.019

18. Araújo MB, Anderson RP, Barbosa AM, Beale CM, Dormann CF, Early R, et al. Standards for distribution models in biodiversity assessments. Science Advances. 2019; 5: eaat4858. https://doi.org/10.1126/sciadv.aat4858 PMID: 30746437

19. Qiao H, Soberón J, Peterson AT. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Kriticos D, editor. Methods in Ecology and Evolution. 2015; 6: 1126–1136. https://doi.org/10.1111/2041-210X.12397

20. Ashraf U, Peterson AT, Chaudhry MN, Ashraf I, Saqib Z, Rashid Ahmad S, et al. Ecological niche model comparison under different climate scenarios: a case study of Olea spp. in Asia. Ecosphere. 2017; 8: e01825. https://doi.org/10.1002/ecs2.1825

21. Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecological modelling. 2000; 135: 147–186.

22. Cobos ME, Peterson AT, Barve N, Osorio-Olvera L. kuenm: an R package for detailed development of ecological niche models using Maxent. PeerJ. 2019; 7: e6281. https://doi.org/10.7717/peerj.6281 PMID: 30755826

23. Cobos ME, Peterson AT, Osorio-Olvera L, Jiménez-García D. An exhaustive analysis of heuristic methods for variable selection in ecological niche modeling and species distribution modeling. Ecological Informatics. 2019; 53: 100983. https://doi.org/10.1016/j.ecoinf.2019.100983

24. Li Y, Li M, Li C, Liu Z. Optimized Maxent Model Predictions of Climate Change Impacts on the Suitable Distribution of Cunninghamia lanceolata in China. Forests. 2020; 11: 302. https://doi.org/10.3390/f11030302

25. Cavanagh RD, Murphy EJ, Bracegirdle TJ, Turner J, Knowland CA, Corney SP, et al. A Synergistic Approach for Evaluating Climate Model Output for Ecological Applications. Frontiers in Marine Science. 2017; 4: 308. https://doi.org/10.3389/fmars.2017.00308

26. Harris RMB, Grose MR, Lee G, Bindoff NL, Porfirio LL, Fox-Hughes P. Climate projections for ecologists: Climate projections for ecologists. Wiley Interdisciplinary Reviews: Climate Change. 2014; 5: 621–637. https://doi.org/10.1002/wcc.291

27. Stock CA, Alexander MA, Bond NA, Brander KM, Cheung WW, Curchitser EN, et al. On the use of IPCC-class models to assess the impact of climate on living marine resources. Progress in Oceanography. 2011; 88: 1–27.

28. Bojinski S, Verstraete M, Peterson TC, Richter C, Simmons A, Zemp M. The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. Bulletin of the American Meteorological Society. 2014; 95: 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1

29. Braunisch V, Coppes J, Arlettaz R, Suchant R, Schmid H, Bollmann K. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. Ecography. 2013; 36: 971–983. https://doi.org/10.1111/j.1600-0587.2013.00138.x

30. Schnase JL. Climate Analytics as a Service. Cloud Computing in Ocean and Atmospheric Sciences. 2016. pp. 187–219. https://doi.org/10.1016/b978-0-12-803192-6.00011–6

31. Edwards PN. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge, MA: MIT Press; 2010.

32. IPCC—Intergovernmental Panel on Climate Change. 2020 [cited 14 Mar 2020]. Available: https://www.ipcc.ch/

33. Responding to the Challenge of Climate and Environmental Change: NASA's Plan for a Climate-Centric Architecture for Earth Observations and Applications from Space. National Aeronautics and Space Administration; 2010. Available: https://gmao.gsfc.nasa.gov/reanalysis/MERRA/.

34. Araújo MB, Guisan A. Five (or so) challenges for species distribution modelling. Journal of Biogeography. 2006; 33: 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

35. Peterson AT, Cobos ME, Jiménez-García D. Major challenges for correlational ecological niche model projections to future climate conditions: Climate change, ecological niche models, and uncertainty. Annals of the New York Academy of Sciences. 2018; 1429: 66–77. https://doi.org/10.1111/nyas.13873 PMID: 29923606

36. Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. International Journal of Information Management. 2019; 48: 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

37. Galante PJ, Alade B, Muscarella R, Jansa SA, Goodman SM, Anderson RP. The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. Ecography. 2018; 41: 726–736.

38. Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, et al. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods in Ecology and Evolution. 2014; 5: 1198–1205. https://doi.org/10.1111/2041-210X.12261

39. Radosavljevic A, Anderson RP. Making better MaxEnt models of species distributions: complexity, overfitting and evaluation. Araújo M, editor. Journal of Biogeography. 2014; 41: 629–643. https://doi.org/10.1111/jbi.12227

40. Kroese DP, Brereton T, Taimre T, Botev ZI. Why the Monte Carlo method is so important today: Why the MCM is so important today. Wiley Interdisciplinary Reviews: Computational Statistics. 2014; 6: 386–392. https://doi.org/10.1002/wics.1314

41. Ito Y, Imai H, Duc TL, Negishi Y, Kawachiya K, Matsumiya R, et al. Profiling based Out-of-core Hybrid Method for Large Neural Networks. arXiv:190705013 [cs]. 2019 [cited 11 Jan 2021]. Available: http://arxiv.org/abs/1907.05013

42. Chen T, Xu B, Zhang C, Guestrin C. Training Deep Nets with Sublinear Memory Cost. arXiv:160406174 [cs]. 2016 [cited 11 Jan 2021]. Available: http://arxiv.org/abs/1604.06174

43. Dunning, Jr. JB, Bowers, Jr. RK, Suter SJ, Bock CE. Cassin's Sparrow (Peucaea cassinii), Version 1.0. In: Birds of the World (P. G. Rodewald, Editor) [Internet]. 2020 [cited 22 May 2020]. Available: https://doi.org/10.2173/bow.casspa.01

44. Iknayan KJ, Beissinger SR. Collapse of a desert bird community over the past century driven by climate change. Proc Natl Acad Sci USA. 2018; 115: 8597. https://doi.org/10.1073/pnas.1805123115 PMID: 30082401

45. Radchuk V, Reed T, Teplitsky C, van de Pol M, Charmantier A, Hassall C, et al. Adaptive responses of animals to climate change are most likely insufficient. Nature Communications. 2019; 10: 3109. https://doi.org/10.1038/s41467-019-10924-4 PMID: 31337752

46. GBIF.org (21 February 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.0s8yak.

47. Jiménez-Valverde A. Threshold-dependence as a desirable attribute for discrimination assessment: implications for the evaluation of species distribution models. Biodiversity and Conservation. 2014; 23: 369–385. https://doi.org/10.1007/s10531-013-0606-1

48. Fourcade Y, Engler JO, Rödder D, Secondi J. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. Valentine JF, editor. PLoS ONE. 2014; 9: e97122. https://doi.org/10.1371/journal.pone.0097122 PMID: 24818607

49. Boria RA, Olson LE, Goodman SM, Anderson RP. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecological Modelling. 2014; 275: 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

50. Schnase JL, Grant WE, Maxwell TC, Leggett JJ. Time and energy budgets of Cassin's sparrow (Aimophila cassinii) during the breeding season: evaluation through modelling. Ecological Modelling. 1991; 55: 285–319.

51. Schnase JL, Maxwell TC. Use of song patterns to identify individual male Cassin's Sparrows. Journal of Field Ornithology. 1989; 60: 12–19.

52. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology. 2017; 37: 4302–4315. https://doi.org/10.1002/joc.5086

53. Worldclim bioclimatic variables. 2020 [cited 22 May 2020]. Available: https://worldclim.org/data/worldclim21.html

54. GDAL/OGR Geospatial Data Abstraction Software Library. Open Source Geospatial Foundation; 2020. Available: https://gdal.org/

55. Hijmans RJ, Phillips S, Elith J, Leathwick J. dismo: Species Distribution Modeling. 2017. Available: https://CRAN.R-project.org/package=dismo

56. Pradhan P. Strengthening MaxEnt modelling through screening of redundant explanatory bioclimatic variables with variance inflation factor analysis. Researcher. 2016; 8: 29–34.

57. Maxent Version 3.4.1 Download Site. [cited 22 May 2020]. Available: https://biodiversityinformatics.amnh.org/open_source/maxent/

58. R: The R Project for Statistical Computing. [cited 22 May 2020]. Available: https://www.r-project.org/

59. Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Anderson MU, et al. ENMeval: Automated Runs and Evaluations of Ecological Niche Models. 2018. Available: https://CRAN.R-project.org/package=ENMeval

60. RStudio | Open source & professional software for data science teams. [cited 27 May 2020]. Available: https://rstudio.com/

61. Warren DL, Glor RE, Turelli M. ENMTools: a toolbox for comparative studies of environmental niche models. Ecography. 2010 [cited 27 Mar 2020]. https://doi.org/10.1111/j.1600-0587.2009.06142.x

62. Smith AB, Santos MJ. Testing the ability of species distribution models to infer variable importance. Ecography. 2020; 43: 1801–1813. https://doi.org/10.1111/ecog.05317

63. Bradie J, Leung B. A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. Journal of Biogeography. 2017; 44: 1344–1361. https://doi.org/10.1111/jbi.12894

64. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological applications. 2009; 19: 181–197. https://doi.org/10.1890/07-2153.1 PMID: 19323182

65. Phillips SJ, Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography. 2008; 31: 161–175.

66. (Peucaea cassinii)—Species Map—eBird. 2020 [cited 31 May 2020]. Available: https://ebird.org/map/casspa

67. Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation. 1997; 24: 38–49. https://doi.org/10.1017/S0376892997000088

68. Akaike H.. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705

69. Warren DL, Glor RE, Turelli M. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution. 2008; 62: 2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x PMID: 18752605

70. Schoener TW. The Anolis Lizards of Bimini: Resource Partitioning in a Complex Fauna. Ecology. 1968; 49: 704–726. https://doi.org/10.2307/1935534

71. Fink D, Auer T, Johnston A, Strimas-Mackey M, Robinson O, Ligocki S, et al. Cassin's Sparrow—Abundance map—eBird Status and Trends. In: eBird Status and Trends, Data Version: 2018; Released: 2020 [Internet]. 2020 [cited 5 Oct 2020]. Available: https://ebird.org/ebird/science/status-and-trends/casspa/abundance-map

72. Phillips SJ, Research T. A Brief Tutorial on Maxent. 2017; 39.

73. Tang Y, Winkler JA, Viña A, Liu J, Zhang Y, Zhang X, et al. Pearson pairwise correlation matrix between the bioclimatic variables. 2018. https://doi.org/10.1371/journal.pone.0189496.g003

74. O'Donnell MS, Ignizio D a. Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States. Reston, VA: US Geological Survey; 2012 p. 10. Report No.: 691. Available: https://pubs.usgs.gov/ds/691/

75. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013; 36: 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

76. Ruth JM. Cassin's Sparrow (Aimophila cassinii) status assessment and conservation plan. Denver, CO; 2000. Report No.: BTP-R6002-2000. Available: http://pubs.er.usgs.gov/publication/2002055

77. Salas EAL, Seamster VA, Boykin KG, Harings NM, dixon K w., Department of Fish, Wildlife and Conservation Ecology, New Mexico State University, Las Cruces, New Mexico 88003, USA. Modeling the impacts of climate change on Species of Concern (birds) in South Central U.S. based on bioclimatic variables. AIMS Environmental Science. 2017;4: 358–385. https://doi.org/10.3934/environsci.2017.2.358

78. Barbet-Massin M, Jetz W. A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. Heikkinen R, editor. Diversity and Distributions. 2014; 20: 1285–1295. https://doi.org/10.1111/ddi.12229

79. Beaumont LJ, Pitman AJ, Poulsen M, Hughes L. Where will species go? Incorporating new advances in climate modelling into projections of species distributions. Global Change Biology. 2007; 13: 1368–1385. https://doi.org/10.1111/j.1365-2486.2007.01357.x

80. Peterson AT, Nakazawa Y. Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. Global Ecology and Biogeography. 2008;0: 071113201427001-??? https://doi.org/10.1111/j.1466-8238.2007.00347.x

81. Warren DL, Matzke NJ, Iglesias TL. Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications. Journal of Biogeography. 2020; 47: 167–180. https://doi.org/10.1111/jbi.13705