Ecological Informatics 48 (2018) 1–11

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Predicting the future is hard and other lessons from a population time series data science competition

Humphries G.R.W.[a,*], Che-Castaldo C.[b], Bull P.J.[c], Lipstein G.[c], Ravia A.[d,e], Carrión B.[f], Bolton T.[g], Ganguly A.[h], Lynch H.J.[b,i]

[a] Black Bawks Data Science Ltd., 24 Abertarff Place, Fort Augustus PH32 4DR, UK
[b] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA
[c] DrivenData Inc., Denver, CO, USA
[d] Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel
[e] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
[f] PRDW Consulting Port and Coastal Engineers, Santiago, Chile
[g] Department of Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom
[h] Unaffiliated, Sector III, Saltlake, Kolkata, India.
[i] Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY, USA.

ARTICLE INFO

Keywords:
Penguins
Antarctica
MAPPPD
Forecasting
Uncertainty

ABSTRACT

Population forecasting, in which past dynamics are used to make predictions of future state, has many real-world applications. While time series of animal abundance are often modeled in ways that aim to capture the underlying biological processes involved, doing so is neither necessary nor sufficient for making good predictions. Here we report on a data science competition focused on modelling time series of Antarctic penguin abundance. We describe the best performing submitted models and compare them to a Bayesian model previously developed by domain experts and build an ensemble model that outperforms the individual component models in prediction accuracy. The top performing models varied tremendously in model complexity, ranging from very simple forward extrapolations of average growth rate to ensembles of models integrating recently developed machine learning techniques. Despite the short time frame for the competition, four of the submitted models outperformed the model previously created by the team of domain experts. We discuss the structure of the best performing models and components therein that might be useful for other ecological applications, the benefit of creating ensembles of models for ecological prediction, and the costs and benefits of including detailed domain expertise in ecological modelling. Additionally, we discuss the benefits of data science competitions, among which are increased visibility for challenging science questions, the generation of new techniques not yet adopted within the ecological community, and the ability to generate ensemble model forecasts that directly address model uncertainty.

1. Introduction

Time series lie at the heart of population biology and are increasingly used in conservation applications such as adaptive management of marine stocks and population viability analysis. Using traditional linear modelling approaches, abundance can be modeled as a function of covariates such as time (capturing trends in abundance) or environmental conditions thought to control survivorship or reproduction (Geissler and Noon, 1981). While population models are often driven by the causal mechanisms underlying population change, this is neither necessary nor always sufficient for precise forecasts of future abundance (Shmueli, 2010). In fact, forecast uncertainty for covariates can contribute to forecast uncertainty for abundance, an important consideration when building population models from which future abundance or population viability assessments will be derived (Dietze, 2017).

Phenomenological approaches have many advantages, not the least of which is a reliance on the data in hand rather than a priori knowledge of the underlying ecological or biological system, the need for which narrows the scope of inquiry only to domain experts intimately familiar with the system. Autoregressive-(integrated)-moving-average (AR[*I*]MA) models have been used for ecological time series for some

[*] Corresponding author.
E-mail address: grwhumphries@blackbawks.net (G.R.W. Humphries).

https://doi.org/10.1016/j.ecoinf.2018.07.004
Received 5 April 2018; Received in revised form 10 July 2018; Accepted 11 July 2018
Available online 12 July 2018

time (Goldman et al., 1989; Ives et al., 2010), whereas machine learning algorithms like random forests (Breiman, 2001; Cutler et al., 2007; Prasad et al., 2006) or generalized boosted regression models (Elith et al., 2008; Friedman, 2002) are more recent additions to the population ecologists' toolbox. So-called 'deep-learning' methods (e.g., long short-term memory neural nets) have shown great promise in various 'big data' applications (Hochreiter and Schmidhuber, 1997) but are rarely applicable to animal population time series. Understanding which of these tools are useful and practical for population time series is a growing challenge.

While several excellent public databases exist (e.g., Global Population Dynamics Database) with population time series that ecologists might use to explore different time series methods (Ward et al., 2014), there are relatively few opportunities to test and compare the forecasting accuracy of time series models. This is partly because model development often comes after data collection has ceased. It is also, in part, a reflection on the discipline of population ecology, which lags other disciplines such as terrestrial ecosystem, biogeochemical, or climate modelling in terms of formalized procedures for model comparison and validation. To address this shortfall in the context of Antarctic ecology, we developed the Mapping Application for Penguin Populations and Projected Dynamics (MAPPPD; Humphries et al., 2017), an open-access database of all known data on the breeding abundance and distribution of Antarctic penguins (Adélie [*Pygoscelis adeliae* Hombron & Jacquinot], gentoo [*P. papua* Forster], chinstrap [*P. Antarctica* Forster] and emperor [*Aptenodytes forsterii*]). While far smaller in taxonomic scope than the Global Population Dynamics Database, MAPPPD was assembled specifically to assist in conservation planning and management of Antarctic resources and, being geographically complete for all known penguin breeding locations south of 60 °S, is one of the most spatially and temporally comprehensive sets of population census data for any taxonomic group.

One of the species included in MAPPPD is the Adélie penguin which, having a circumpolar distribution and widely considered one of Antarctica's 'canaries-in-the-coal mine', is by far the most well studied of the Antarctic penguins (Ainley, 2002; Ainley et al., 2010; Boersma, 2008). For at least the last 40 years, researchers have debated why Adélie penguin populations have been changing and, to a lesser degree, what drives interannual fluctuations in abundance at the breeding colony (Che-Castaldo et al., 2017 and references therein). Accordingly, several major efforts to model the population dynamics of Adélie penguins have been developed (Ainley et al., 2010; Che-Castaldo et al., 2017; Fraser et al., 1992; Jenouvrier et al., 2009; Lynch et al., 2012; Lynch and Larue, 2014; Trivelpiece et al., 2011; Wilson et al., 2001). Reflecting the logistical challenges of collecting census data in the Antarctic, population models are often parameterized using data from a small number of neighboring populations with a focus on understanding past events rather than predicting future abundance. Unfortunately, Adélie penguin dynamics are spatially heterogeneous because the Antarctic is comprised of many disparate biogeographic zones and penguins breeding in each region face different bottlenecks on their survival and reproduction. Adélie penguin populations are increasing in the Ross Sea region and in Eastern Antarctica (Che-Castaldo et al., 2017; Larue et al., 2013; Lyver et al., 2014) even as they decline in parts of the Antarctic Peninsula (Cimino et al., 2016; Lynch et al., 2012). These disparate trends are further complicated by significant inter-annual fluctuations in abundance not easily tied to environmental drivers (Che-Castaldo et al., 2017).

While much attention has been paid to the Adélie penguin, even less is known about the detailed population dynamics of gentoo or chinstrap penguins, despite the relative ease of access afforded by their concentration on the Antarctic Peninsula. Gentoo penguin abundance has surged in recent decades, while chinstrap penguin abundance has notably declined at most sites (Lynch et al., 2012; Trivelpiece et al., 2011). As with the Adélie penguin, these long-term trends have emerged over decades amidst significant year-to-year variability in abundance at each site, which challenges short-term forecasts even where trends are unambiguous. Until MAPPPD, there was no dynamic central database for the abundance of Antarctic penguins, which has precluded the development of models to understand how the individual results emerging from different research groups fit together into an integrated understanding of penguin population biology.

Despite years spent modelling penguin population dynamics, we (GH, CC, HL) were frustrated by the relatively poor performance of models built using our own biological knowledge of the system. Decades of analysis by the community had produced a suite of models with reasonable explanatory power, but there has been no way to compare different population models in a quantitative setting or explore the extent to which inferred causal drivers in one region could be transferred to another. Moreover, significant unexplained process error made it difficult to forecast abundance for any given breeding population (Che-Castaldo et al., 2017) and, as such, there was no way to link a dynamical model for abundance with long-term predictions of suitability under climate change (Cimino et al., 2016). This led us to question our own methods and whether there were other, perhaps newer, approaches available that we should be using. To address these concerns and generate a larger suite of models that could be used to understand the impact of model uncertainty, we leveraged the opportunities afforded by a data science competition.

Data science competitions are an up-and-coming trend among data scientists. These competitions offer prizes to teams who can solve a problem, the results of which are typically assessed by predictions on held-out subsets of the data. While their popularity has vastly increased alongside the growth of data science as a discipline, competitions for time series modelling are not a new idea. In this regard, Weigand and Gershenfeld's (1994) classic description of a modelling competition from 1991 remains surprisingly relevant; while computing has changed radically over the last 25 years (datasets no longer have to be distributed by floppy disk, for one), we are still facing many of the very same difficulties building good predictive time series models as we always have (e.g., non-stationarity and non-linearity, missing data) despite a surging interest in 'data science' and prediction algorithms based on machine learning. Online platforms such as Driven Data (http://www.drivendata.org), Kaggle (http://www.kaggle.com), and Top Coder (http://www.topcoder.com) engage data scientists and other quantitative experts from around the world in building models for a combination of prize money, recognition, practice, and real-world impact. The challenges enable participants to try out thousands of models for a given problem using whatever backgrounds, skills, and approaches they see fit, with the solutions that perform best (by predicting held-out data) rising to the top of the leaderboard. This approach represents a parallelization of effort and vastly improves the efficiency of the modelling process, as many models are developed and tested simultaneously (analogous to multiple cores of a computer working on the same problem). Examples of such competitions can be found throughout scientific literature (e.g., Ben Taieb and Hyndman, 2014; Glaeser et al., 2016; Narayanan et al., 2011). Such competitions offer the ability to explore new, and potentially better, ways to approach modelling problems (Bull et al., 2016; Carpenter, 2011).

Antarctic penguin population ecology is a good test case for harnessing the power of data science competitions and community modelling in ecology; while the data are spatially comprehensive (including all 660 known penguin colonies in the Antarctic), their shortcomings (e.g., only 30–40 years in length, lots of missing data) are stereotypical for time series of animal populations. Given the ongoing debate regarding key drivers of penguin population dynamics in Antarctica and the urgent need to understand how management of fisheries may impact penguin populations, the need for better time series models is clear.

The goal of this paper is to describe a data science competition focused on population time series modelling of Antarctic penguin abundance, identify and describe which techniques performed well and should be considered for other ecological applications, and demonstrate the application of ensemble-based approaches for ecological prediction. We will discuss how relatively "domain agnostic" (i.e. general) statistical approaches might be used for rapid assessment even as information required for more biologically-motivated mechanistic models is being collected, and how data science competitions can vastly expand our understanding of model uncertainty in ecological and conservation contexts.

2. Materials and methods

2.1. Data science competition

Our competition was hosted by DrivenData (www.drivendata.org), a platform that specializes in data science challenges with positive social impact. DrivenData was started in early 2014 as a graduate school project, then incubated out of the Harvard Innovation Lab. From a couple hundred users mostly in the university network, the challenge community has since grown to > 15,000 data enthusiasts from > 140 countries. The overall number of prize-based competitions and the number of people participating in them has seen similar growth. With demand for advanced data skills outpacing the supply, these challenges have become a powerful mechanism for engaging the cognitive surplus of the global data science community.

2.2. MAPPPD

The raw count data used by competitors (Table 1), as provided by MAPPPD, come from several sources including published literature (peer-reviewed articles and reports), and contributed data (i.e. unpublished databases).

The front end of MAPPPD (www.penguinmap.com) allows customized map- or text-based spatial queries to all publicly-available abundance data, their associated citations, and population estimates derived from our initial population model (Che-Castaldo et al., 2017; Humphries et al., 2017). MAPPPD data are housed in a Postgres database written using structured query language (SQL). While the database is under continual development, the initial database is described in depth in Humphries et al. (2017). MAPPPD currently hosts 1405 records for Adélie penguins, 907 records for chinstrap penguins and 933 records for gentoo penguins and cover the entire continental range for all three species (Fig. 1). Emperor penguins were not included due to data scarcity. The largest contributor to the MAPPPD database is the Antarctic Site Inventory and related papers (e.g., Casanovas et al., 2015; Lynch et al., 2013), which comprise 41.1% of the data, with 12.3% coming from the Landcare Research dataset (http://www.landcareresearch.co.nz/resources/data/adelie-census-data). All other contributions make up the remaining 46.6% of the data. We restricted the data science competition to the period 1982–2013, which covers the majority of data available, and withheld data from the 2014–2016 seasons for model validation. Data from the 2014–2016 seasons were also redacted from the online database. At the time of the competition, there were no data for gentoo penguins available for the 2016 season, and so models were assessed using nest counts from 2014 to 2016 for chinstrap and Adélie penguins, and 2014–2015 for gentoo penguins.

Table 1
Field names in database table containing data for counts used in data science competition.

Field name	Description
preprocessed_id	integer ID of each count for table referencing
site	four letter ID of the site
common_name	common name of species counted
citekey	citation ID for the specific count
day	day the count was performed ('none') if unknown
month	month the count was performed ('none') if unknown
season_day	integer value of the day of the season starting from June 1
season	season the count was performed. Counts performed in January, February or March are assigned to season value of the year – 1. For example, January 2015 counts were part of the 2014 season.
year	the calendar year that a count was performed
presence	presence (1) or absence (0) of species counted
count	the total count for the specific record
accuracy	accuracy flag for the count with 1 being the highest accuracy and 5 being the lowest accuracy.
count_type	the type of count (chicks, adults or nests)
vantage	the platform / vantage from which the count was taken (e.g., ground, aerial, satellite)
note	Comments associated with each count

Quality flags for each abundance count (i.e. 'accuracy' field) were codified according to the scale used by Croxall and Kirkwood (1979) and subsequently by most other reports of penguin abundance. Quality flags of '1' are associated with the highest accuracy ground counts (± 5%), while quality flags of '5' are considered good only to an order of magnitude. For the purposes of this competition, the quality flags were assigned the following confidence intervals: 1 (± 5%), 2 (± 10%), 3 (± 25%), 4 (± 50%) and 5 (± 90%). This mapping matches what is generally accepted in the Antarctic penguin literature.

An important characteristic of this dataset, typical of many ecological time series, is the large fraction of years for which no counts are available. The patchiness of these data made the problem both challenging and technically interesting as the success of the predictive model hinged on both methods for data imputation and time series modelling. Of the 660 sites in MAPPPD, only 14 contained full count data spanning the length of the database (1982–2016).

2.3. Model assessment

Mean absolute percentage error (MAPE) is one of the most popular measures of forecast accuracy (Bowerman et al., 2004; Hanke and Reitsch, 1995), but it does not account for differences in observation error across the dataset (Tofallis, 2015). We therefore used an adjusted MAPE (AMAPE; Eq. (1))

$$AMAPE = \frac{1}{N}\sum_{n=1}^{N}\frac{\left|\frac{\hat{y}_n - y_n}{max(1, y_n)}\right|}{e_n} \tag{1}$$

which scaled absolute percent error ($\hat{y}_n$ is the predicted count, y_n is the actual count) to the observation error (e_n)). For example, a prediction that was 50% off an observed count with 50% observation error would be weighted the same as a count that was 25% off an observed count with 25% observation error and ten times as much as one with an observation error of only 5%. If the count y_n is 0, a value of 1 is used in the denominator of the fractional error to avoid dividing by 0.

For the competition, the AMAPE values were calculated using public and private hold out subsets, which represented a 50/50 split of the 2014–2016 data. The public subset was used to give competitors an idea of how well they were doing with each submission, and the private subset was used to determine the winners of the competition. For this paper, we have re-calculated AMAPE values using the entire 2014–2016 data to get a more realistic representation of model success.

2.4. Models

The competition yielded models from 97 competitors over the 62 days. Assuming each competitor spent 2 h per day working on model development, this amounts to approximately 500 person-days working on the problem. In our model comparison, we consider six models: the four top models (out of a total of 567) submitted to the data science competition (models AG, TB, AR and BC), one model developed previously for the MAPPPD project (and published in Che-Castaldo et al., 2017; model CC), and an ensemble model of these five models (model

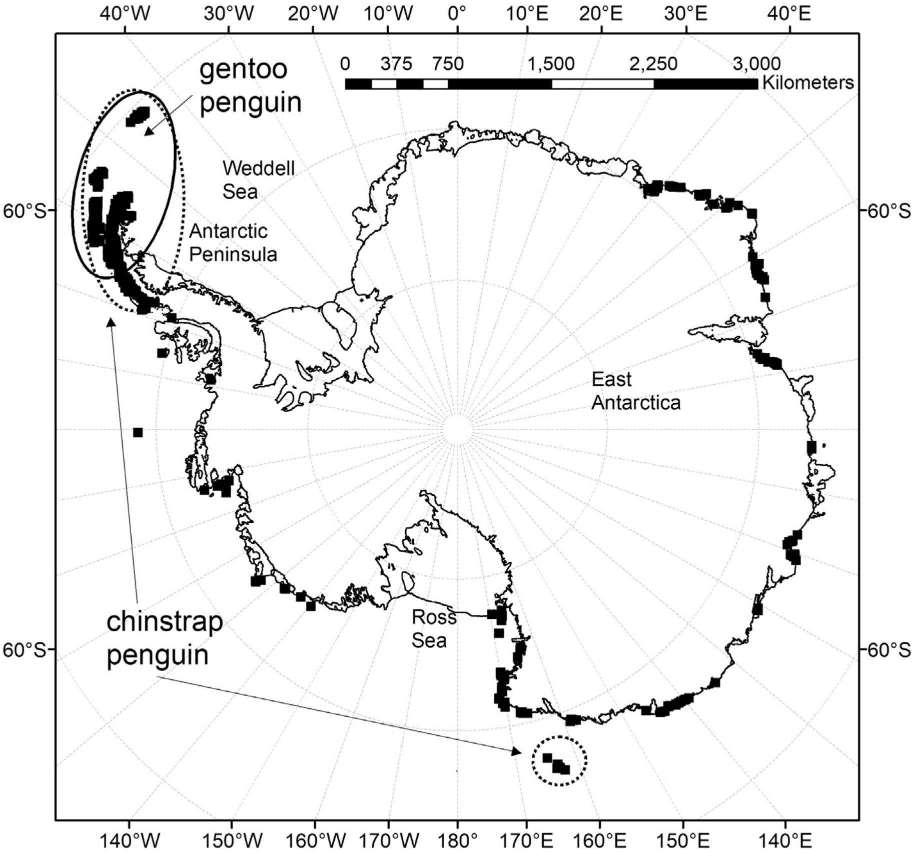

Fig. 1. Distribution of all sites in MAPPPD for Adélie, chinstrap and gentoo penguins (black squares). The regions where chinstrap and gentoo penguins can be found nesting sympatrically with each other, on their own, or with Adélies are identified on the map by a solid (gentoo penguins) or dashed (chinstrap penguins) oval.

EN). Model CC was developed by scientists over several years with extensive expert biological knowledge of the ecosystem, while models AG, TB, AR and BC were developed by individuals in the data science community during the 62 days between the launch of the competition and the deadline for submissions.

2.5. Model AG

Model AG was split into two parts: (1) data imputation, where missing data in the time series were imputed, and (2) model building and forecasting, where the imputed time series were modeled and then forecasted to years 2014, 2015, and 2016. Model AG used a combination of five imputation methods: Linear extrapolation, last observation carried forward, next observation carried backwards, replace by zero, and Stineman extrapolation (Stineman, 1980). The Stineman imputation algorithm (Stineman, 1980) works by altering the shape of the imputation curve depending on whether trends are monotonic or not, making it a commonly used imputation technique. The five imputation techniques were applied to all data to create five sets of data to be used for prediction.

Five modelling algorithms were also implemented: in R, the ARIMA model (McKenzie, 1984), the error, trend, seasonality (ETS) model (Hyndman and Khandakar, 2007), and the Prophet (Taylor and Letham, 2017; Box 1) models were used. In Python, XGBoost (Chen and Guestrin, 2016) and random forest (Breiman, 2001) algorithms were used. In all cases, covariates used for prediction were previous counts. For each time series (i.e. each colony), Model AG would run each algorithm using a variable window of time prior to each year used for prediction (i.e. variable amounts of training data). The window length was chosen to optimize the quality of the 2010 to 2013 predictions for each algorithm and the combination of algorithm and window length that yielded the most accurate 2010–2013 prediction was used to predict 2014–2016. This provided the flexibility for the model to select whatever combination of imputation technique, modelling algorithm, or window length optimized the predictions.

2.6. Model TB

Model TB used an ‘analyst-in-the-loop’ approach, whereby predictive models were tailored to each site-specific time series. TB explored the datasets and, finding that gentoo penguin populations were generally increasing, assumed a 1% per annum growth rate for all site with only a single data point. For other sites (those with Adélies and chinstraps) that had only one data point, the competitor used the most recent count as the prediction for upcoming years, as would be ideal for a time series undergoing a random walk (Ward et al., 2014). Time series with 10–13 counts were fit by an auto-regressive model (Akaike, 1969). Because the auto-regressive model was unable to improve predictive performance for time series with > 13 counts, sites with > 13 counts were fit using either linear regression or exponential linear regression as determined by visual assessment of the existing data (linear vs. exponential change over time).

2.7. Model AR

Model AR was similar to Model TB in that it integrated several techniques depending on the time series, adapted from the Facebook Prophet approach (Taylor and Letham, 2017). For time series with two or fewer counts, the most recent count and subsequent predicted values were multiplied by a constant value of 1.083 (determined by trial and error around a value of 1, which would indicate no growth). For time series with more than two counts, a multi-tiered approach was taken

Box 1

In the sphere of social media, it is important to predict the behaviour of users at particular times of a year. For this application Facebook developed an algorithm (Prophet) that can be tuned in terms of its complexity, whose uses extend beyond social media into other time series applications. The Prophet algorithm is open access and can be accessed from both the Python and R programming languages (https://facebook.github.io/prophet) and is described in detail by Taylor and Letham (2017). In brief, the algorithm works by way of an “analyst in the loop” approach, where the user can adjust parameters such as timing of regular events (e.g., holidays in business models), growth rate (e.g. linear or logistic), or the number of regressors (e.g., covariates) for an additive regressive model with four components: a piecewise linear or logistic growth curve that detects changepoints, a yearly component modeled using a Fourier series, a weekly component using dummy variables, and a user supplied list of important dates. The algorithm itself is written using ‘STAN’, a language commonly used by Bayesian modelers, and also includes the ability to run predictions through a Markov Chain Monte Carlo simulation. The use of the Prophet algorithm by two of the winners of our competition suggests further exploration of Prophet for ecological time series modeling and prediction is warranted, particularly in cases where rapid predictions might be useful while more mechanistically-motivated predictions are being developed.

where the predictions were derived by a weighted average of the following three factors: the most recent prediction, weighted by 0.6, a short-term linear trend and a long-term linear trend. The short-term linear trend was computed using the last six observations and was weighted by 0.35. The long-term linear trend, computed using all observations, had a weight of 0.05. These weights were determined by trial and error and reflect the hypothesis that the near-term prediction should be influenced most heavily by the recent past; comparing predicted values against a randomly held back test set, with those values selected being those that led to the best predictions.

2.8. Model BC

Model BC, the most basic of all the implementations, used basic exploratory analysis to estimate growth rate over time. First, BC assumed a standard growth rate for all species at 1.0, multiplying the latest count by this value and tested this against a held-out subset from 2010 to 2013. Values around 1.0 were tested using trial and error until the best AMAPE score was found. The approximate growth rates as determined by the BC implementation were 1.075 for chinstrap and gentoo penguins, and 0.9 for Adélie penguins. The latest counts and subsequent predicted counts were multiplied by these constant growth rates.

2.9. Model CC

Model CC was developed within a Bayesian framework and uses the data on abundance of the three species of penguin from the 1982 season to the present. In Che-Castaldo et al. (2017), the model is described for Adélie penguins, but has been refit (without changes to model structure) to the gentoo and chinstrap data for the purposes of this paper. Model CC is parameterized by sea-ice data extracted from satellite, which is used as a proxy for a number of different biological processes thought to be important for penguins, such as krill recruitment (a dominant component of penguin diet) and colony access for new recruits. In addition, Model CC estimates a random effect for each year, which incorporates additional (spatially global but temporally varying) random variation not otherwise captured by the sea ice covariates included in the model. It is important to note that Bayesian models provide information on prediction uncertainty but to accommodate a direct comparison to the other models we used the median of the posterior predictive distributions for all site, year, and visit combinations withheld for validation.

2.10. EN models

Ensemble models are becoming widely accepted and used across a number of fields because they balance out uncertainty across multiple models (Wichard and Ogorzalek, 2004). We constructed an ensemble prediction from the point estimates provided by the five models described above. There are several methods by which to create an ensemble model estimate ranging from basic model averaging to complex methods in which models are combined during the ‘tuning’ phase of a machine learning algorithm (Casanova and Ahrens, 2009; Dietterich, 2000). In this case, following the lead of other studies integrating models of different origins (Weigel et al., 2008) we tested four different methods of model averaging (Table 2).

The first method is a simple model average, where we take the mean of the five model predictions (model EN-UW). The second method (model EN-WE) is a weighted model average where the weights are defined by the inverse of AMAPE, analogous to how mean squared error is used for weighting in other work (Casanova and Ahrens, 2009).

$$w_i \propto \frac{1}{AMAPE_i} \tag{2}$$

The third method (model EN-WN) is an extension of EN-WE with normalized weights

$$w_i \propto \frac{\frac{1}{AMAPE_i} - min\left(\frac{1}{AMAPE_i}\right)}{max\left(\frac{1}{AMAPE_i}\right) - min\left(\frac{1}{AMAPE_i}\right)} \tag{3}$$

Normalizing the weights in this fashion puts more weight on the best model and less on the worst model, which may be desirable when the inverse of the AMAPE scores are very similar and the unnormalized model weights would be nearly uniform across models in the set. The fourth method (model EN-LM), inspired by Krishnamurti et al. (1999), uses a weighting derived from linear regression, wherein each model is weighted according to the deviation of the slope of predicted values to observed values from 1.0 (larger weights for models with a slope closer to 1.0). Note that, unlike the other weighting schemes, models with a fixed bias for all abundances would not be penalized under this weighting scheme. It is also important to note that this method minimizes the root mean squared error and is not, therefore, optimized for AMAPE. Once the best of the four ensemble model schemes was

Table 2
Model weightings under each model weighting scheme. Note that due to rounding, not all rows will add to 1.0 exactly.

Model	AG	TB	AR	BC	CC
EN-UW	0.20	0.20	0.20	0.20	0.20
EN-WE	0.21	0.20	0.20[a]	0.19	0.19[a]
EN-WN	0.35	0.33	0.29	0.03	0.00
EN-LM	0.00	0.02	0.08	0.35	0.54

[a] Denotes values that were lower than other values by a marginal amount in the same row that seem like a tie.

selected, we reweighted (i.e., re-tuned) this scheme (model EN-WNrt) by first subsetting the submissions by species and calculating species-specific AMAPE scores. These AMAPE scores were used to calculate new weights for the ensemble model. We next created the new species-specific ensemble model and then re-calculated the AMAPE score. We then re-tuned using the same process except by year instead of species. This allowed us to test if ensemble models always outperformed other models.

The four schemes for weighting point estimates each provide a measure of the average prediction across the different models developed. However, it is not a priori clear how to calculate a confidence interval on the ensemble prediction, particularly because four of the models yielded only point estimates. The total uncertainty in the ensemble model prediction should reflect both the uncertainty between models as well as the uncertainty within models, however the latter is known only for the Bayesian model (Model CC) whose posterior prediction inherently captured prediction uncertainty. Here we calculate the standard error of the five-point estimates and use ± 1.96SE (which measures inter-model uncertainty) as a lower bound on total ensemble estimate uncertainty.

3. Results

From the original submitted models, model AG (AMAPE = 4.57) was ranked as the best model (lowest AMAPE score), followed by models TB (AMAPE = 4.59), AR (AMAPE = 4.61) and BC (AMAPE = 4.80), respectively. The previously developed Bayesian model (Model CC) was ranked 5th overall (AMAPE = 4.82). These results differ slightly from the official results at the end of the competition (https://www.drivendata.org/competitions/47/penguins/leaderboard) because we re-calculated AMAPE using the entire hold-out data subset from 2014 to 2016 versus only the private subset used by DrivenData. It is notable that two of the winning models (AG and AR) exploited the recently-developed Prophet model, which accommodates cyclic dynamics that arise naturally in many ecological time series. Prophet (Box 1) employs an ‘analyst in the loop’ approach in which users can choose automated functionality or, if they wish, alter aspects of the algorithm to better suit their goals. Model AG used the Prophet algorithm with automated functionality, while model AR used the Prophet algorithm for inspiration in designing their approach.

Breaking down model performance by species, we see that model CC, which had been developed initially for Adélie penguins, saw its highest ranking for gentoo penguins (2nd place from the original submitted models) (Table 3), though an extreme imbalance in the size of the three validation datasets (8 counts for Adélie penguins, 57 for chinstrap penguins, and 62 for gentoo penguins) suggests caution in our interpretation of differences in fit by species. Caution in interpretation of the AMAPE values is also warranted here due to its sensitivity to a small number of very large colonies with large leverage that arise from the log-normal distribution of colony sizes and the higher error term associated with them which heavily penalizes differences. For these largest colonies, the CC model tended to better predict abundance compared to the other submitted models (AG, AR, TB, BC), which tended to overpredict (Fig. 2).

Table 3
Ranked AMAPE values for each of the models calculated on seasons 2014–2016 that were held back from the modelling process broken down for the whole dataset and by species, and a model retuned using the EN-WN method (Eq. (3)) with weights for the retuned models in parentheses.

Model	Whole dataset	Adélie	chinstrap	gentoo
EN-WN	4.05[a]	3.06	4.51[a]	4.02
EN-WNrt	–	2.90[a]	4.56	3.73[a]
AG	4.57	3.38 (0.29)	5.18 (0.28)	4.17 (0.25)
TB	4.59	2.85 (0.44)	5.35 (0.11)	4.11 (0.32)
AR	4.61	4.85 (0.05)	4.97 (0.48)	4.25 (0.15)
BC	4.80	3.73 (0.22)	5.38 (0.13)	4.40 (0.00)
CC	4.82	5.36 (0.00)	5.50 (0.00)	4.14 (0.28)

[a] Best model as per AMAPE.

As expected, all the ensemble models (i.e., EN-WN and EN-WNrt for the whole dataset and the three penguin species) provided better predictions than any of the individual models considered (Table 3). The best implementation of the ensemble model for the whole dataset was the normalized weighted mean (AMAPE = 4.05; model EN-WN; Eq. (2)), followed by the weighted mean (AMAPE = 4.12; model EN-WE), the unweighted mean (AMAPE = 4.13; model EN-UW), and the weighted linear model implementation (AMAPE = 4.16; model EN-LM). Model EN-WN scored an AMAPE of 4.05 for the whole dataset — nearly 13% better than the best submitted competition model (model AG). While model EN-WN's performance did not decline over the period of forecasting, the AMAPE scores for all individual component models steadily increase from 2014 (mean = 4.00) to 2016 (mean = 5.77), demonstrating the divergence of model predictions from actual values over time (Fig. 3). Surprisingly, the best overall submitted model (model AG) was not uniformly the best model; while it underperformed in 2014 or 2015, it was significantly better than the next best submitted model in 2016. In other words, it was the overall best model because the rate of decline in predictive performance was slower than for the other submitted models (Table 3). Among the ensemble models we created from the individual component models, the EN-WNrt ensemble model, which used species-specific model weights, had the best fit (lowest AMAPE score) for both Adélie and gentoo penguins, whereas the best model for chinstrap penguins was model EN-WN.

Finally, we note that mean AMAPE values of submitted models from the top competitors decreased steadily over time up to the end of the competition, as models were continually improved through tuning of model parameters (Fig. 4). This suggests that a longer competition may have resulted in models with even better forecasting ability.

4. Discussion

While our focal application was specific to Antarctic ecology, time series forecasting is a common application across most disciplines and, as such, is ripe for further development. However, while looking for a host for our competition, it was surprising to see how few data science competitions involved time series datasets, especially since other biological applications, particularly those involving imagery data or visual computing, have seen tremendous benefits from this kind of interdisciplinary crowd sourcing. For example, image recognition tools for monitoring right whale *Eubalaena glacialis* populations have been developed through the Kaggle platform (http://www.kaggle.com; Kabani and El-Sakka, 2016). DrivenData also hosted two similar competitions: one to identify fish from images on board fishing vessels (https://www.drivendata.org/competitions/48/identify-fish-challenge), and another to identify animals on camera traps by species (https://www.drivendata.org/competitions/49/deep-learning-camera-trap-animals). The code and algorithms that come from competitions like these can be made open access and can therefore be tailored to new problems, providing benefits for domain scientists far beyond those directly involved in the competition itself and society in general (Bull et al., 2016). Our competition was only open for 62 days and yet produced nearly 600 individual models from 97 competitors, and the top models provided better AMAPE scores over the withheld data than a detailed and biologically-motivated hierarchical Bayesian model constructed over years by domain experts.

4.1. Lessons learned about ecological modelling

As ecologists, we tend to approach time series modelling from the perspective of those environmental covariates we think may be driving the dynamics of a system. As such, one of the surprising lessons of this

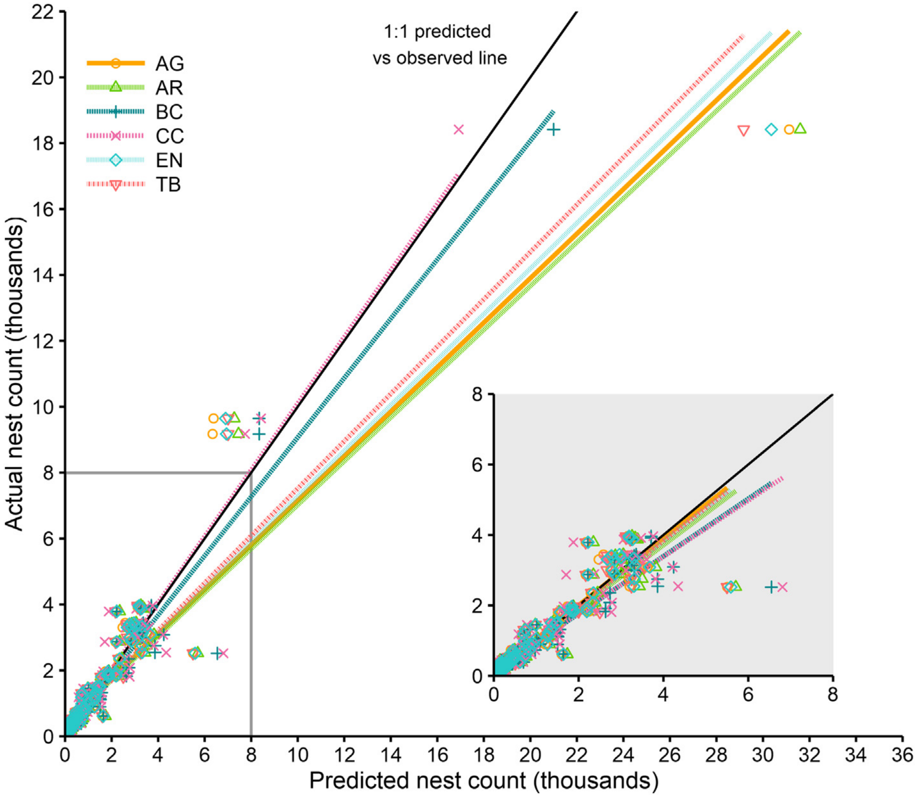

Fig. 2. Predicted versus observed regression lines compared for the six models against the 1:1 slope ('perfect' predictions) for all data (inset shows the 95th percentile of data). Note that the regression lines used for illustration purposes here are based on the root mean squared error rather than AMAPE.

competition was that the environmental covariates we have and frequently use for the Antarctic (i.e. those things that can be measured by satellites and are therefore available over a large spatial extent) do not improve forecast accuracy. Only two of the top competitors (BC and TB) attempted to include environmental covariates (sea ice extent and sea surface temperature), and neither was able to improve predictive performance by including them. This is surprising because sea ice extent has been linked to the distribution and abundance of krill (Loeb et al., 1997), a primary food source for the *Pygoscelis* spp. penguins, as well as access to foraging areas (Wilson et al., 2001) and the role of both sea ice extent and sea surface temperature have been demonstrated in other studies focused on narrower portions of their range (Fraser et al., 1992; Hinke et al., 2007; Loeb et al., 1997; Lynch et al., 2012; Ribic et al., 1998). However, the fact that including these covariates did not improve predictive performance suggests that the modelling techniques were inadequate for examining the relationships properly, the lags associated with sea ice are so long as to obfuscate the relationship, or perhaps that any influence of sea ice or sea surface temperature is either swamped by other (unmeasured) drivers or buffered by compensatory dynamics (Youngflesh et al., 2017). As such, while it is tempting to assume that dynamics over short time scales are reasonable proxies for the dynamics we might expect to play out over long time scales (Forcada et al., 2006) or vice versa, this may not be the case. That environmental covariates did not add to predictive performance may also simply reflect the difficulty of measuring them. We therefore cannot eliminate the possibility that strongly predictive covariates do exist and that better tools for measuring these environmental characteristics would yield a better understanding of penguin population dynamics and better short-term forecasts (Che-Castaldo et al., 2017).

On a more positive note, the results of our competition demonstrate that 'domain agnostic' time series forecasting approaches using relatively tractable and well-studied algorithms can yield short term predictions that are as good or better than those derived using more biologically-driven models. While the winning model was in fact quite complex, forecasts of similar quality were produced by exceptionally simple models. In noisy systems, simple and complex models alike yield forecasts of similar accuracy; as such, the investment in more complex forecasts may not be worth the effort. At the very least, a simple model with reasonable predictive accuracy should be considered the appropriate null model against which to compare more complex or biologically-motivated models.

Despite the inclusion of information on measurement uncertainty (i.e. observation error), and the fact that data varied wildly in the accuracy associated with each count (from < 5% to ~90%), none of the top competitors integrated measurement uncertainty into their models and yet all of those models outperformed the CC model, which included an explicit model for observation error. This could mean that the CC model does not handle uncertainty in a way that improves overall performance, or that uncertainty is not important when it comes to making predictions in this system because stochasticity in the dynamics (i.e. process error) is larger than observation error (Che-Castaldo et al., 2017). Revisiting the models in the future with a focus on uncertainty would help determine its importance on prediction. Although some machine learning methods can provide or incorporate uncertainty estimates (Durga and Solomatine, 2006), they are not commonly applied. This could be viewed as an advantage of the Bayesian approach, where a level of certainty can be given to predicted values. However, we note as well that our choice of AMAPE as the selection criteria for models may not fairly reflect the predictive power of model CC because it was created and optimized prior to the competition using different criteria for estimating error. For example, using an R^2 to quantify predicted versus observed (Fig. 2), we might expect model CC, with its slope so close to 1, to have performed the best. Our experience highlights that model selection criteria, often selected out of tradition or convenience, are of utmost importance in any ecological modelling exercise and should be carefully selected to suit the goals of the modelling process.

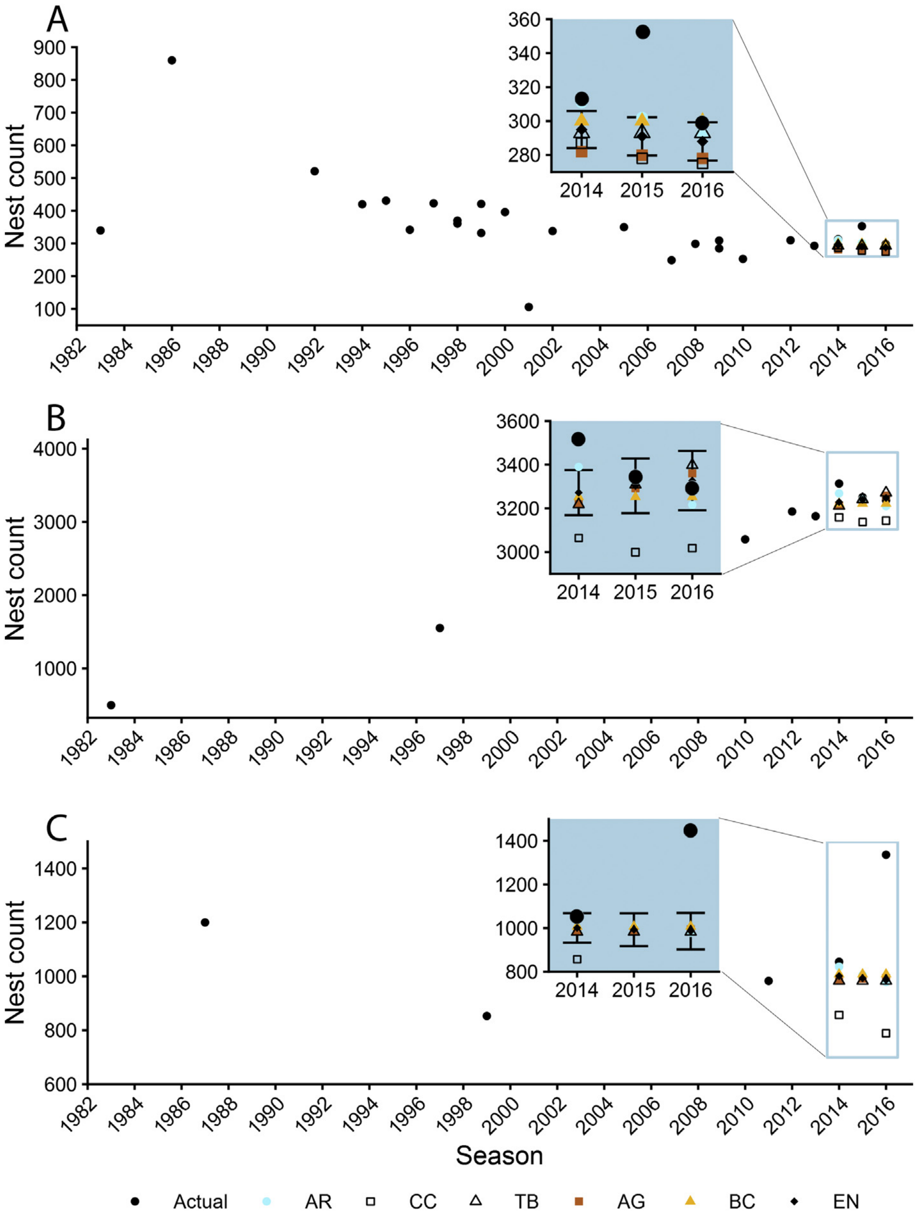

Fig. 3. Entire time series of nest counts available for Orne Islands (A), Pinguino Island (B) and Fort Point (C) from MAPPPD for chinstrap penguins compared to predictions from all models for 2014–2016. Inserts are enlarged regions highlighted in light blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

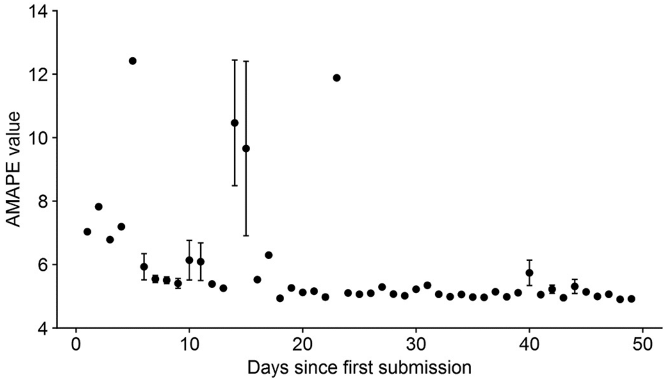

Fig. 4. Mean AMAPE of submitted models AR, AG, TB and BC starting from the first day that either of these models were submitted to the competition. Standard error bars are plotted on the figure for days when multiple models were submitted.

Arguably, creating thousands of models could leave us vulnerable to ‘over-fitting’, in some algorithms may predict well by luck alone, and inferences drawn from such an exercise could lack generalizability. While this is undoubtedly a concern an ensemble model average buffers us from the possibly arbitrary selection of one single “best” model among several that may be nearly equivalent. Here, we have focused on a relatively small set of “top” models, but a more robust approach might include many more models, performing sensitivity tests to determine the combination of models that produce the best overall prediction and using shifts in model weights as an indication of changing system dynamics (Runge et al., 2016).

4.2. Short-term versus long-term predictions

Despite our effort to generate many candidate models by way of the data science competition, and develop ensemble model forecasts that might (and, in fact, do) outperform any of the individual models developed, all the models considered (including the ensemble models) yielded predictions that diverged from the actual values by the 2016 season. Although we did not quantify a forecast proficiency threshold (see Petchey et al., 2015), our findings suggest that the practical forecast horizon for penguin population dynamics remains stubbornly short (e.g., less than a generation of approximately 5–8 years). Consequently, we still have no way to link time series models for abundance (parameterized using data on past abundance) to long-term projections of extinction risk or range shift. We have no way to predict abundance accurately even a few years in advance (see Che-Castaldo et al., 2017), which compromises our ability to use forecasted values to inform management decisions on sustainable krill catches or tourism activities at penguin colonies. While the current state of the art still falls short of the ideal in terms of data-management feedback, this framework for generating a suite of models of differing structure and using them in an ensemble model forecast paves the way forward.

Finally, it is worth noting that while the data being modeled in this competition included time series of abundance only, age- or stage-structured models provide an additional mechanism to link specific demographic parameters (e.g., breeding success, age- or stage-specific survival) to environmental drivers, which may yield more accurate forecasts of total abundance than the models considered here. While mark-recapture data are available for only a very small number of populations (e.g., Ballerini et al., 2009; Clarke et al., 2003; Dugger et al., 2010; Hinke et al., 2017; Jenouvrier et al., 2006; Lescroël et al., 2009), it remains an open research question whether the parametrization of age-structured models, either in isolation or when combined with (unstructured) dynamical models for total abundance, may yield better population forecasts.

4.3. Future of competitions in ecology

Although data science competitions are not new, their increasing popularity in industry and science deserves attention from the ecological community. Over the last several decades, ecologists have made major strides in collecting, curating, and organizing 'big data' datasets (NEON, LTER, etc.) to address long-standing questions in population and community ecology. While our dataset was approximately an order of magnitude smaller in size than those typically facing data scientists, continued monitoring will improve the length of many ecological datasets. At the same time, ecologists need to look carefully through the "modern data scientist's toolbox" to find those approaches that will be applicable to the kinds of smaller datasets we have now. The Prophet model is one such approach, previously unknown to several of us, that we think is worth consideration and may be adapted for use in other ecological applications.

Beyond the scientific benefits, data science competitions yield other benefits as well. Academic scientists are always looking for ways to raise public awareness, and data science competitions yield a concrete way to generate excitement over a scientific challenge and to engage non-academics in a modelling challenge (including both researchers working in other disciplines, working in industry, as well as non-scientists interested in the application). Additionally, the return on investment for a funder is enormous; for a nominal sum of money (on the scale of a major research project), you can garner many orders of magnitude more person-hours than would otherwise be possible and get independent models that are free from a priori assumptions. Because all models involved use the same training set and are judged on the basis of the same test set, model competitions provide an opportunity to directly compare models on the basis of their prediction accuracy.

To help other ecologists interested in future data science competitions, we have several suggestions based on our own experience with the process.

- We had difficulty finding a host for our competition because the data set was considered too small to be of interest for the "data science community", which raised concerns among some of us (HL, CC, GH) that the data science community had become so enamored of "big data" challenges that equally important and arguably more difficult "mesoscale data" challenges were left underserved. In DrivenData, we were able to find an organization willing to host a competition for a dataset of modest size. Ecologists looking to host future competitions should not underestimate the time required to find a willing host for their competition.
- The amount of data used to train and validate models can have profound effect on the outcome of the competition. In our case, an (unavoidable) imbalance between the different species in the test set meant that the overall winning model was disproportionately influenced by its fit to the species that happened to be best represented in the test set. A more even balance among the species would have been preferred.
- The data used to test the outcomes should be representative of the question being asked. For example, if a geographic prediction is to be made, then the prediction should be to an independent geographic subset, and if making predictions to the future, the test set should be an independent subset of the latest values (like in our competition). If possible once the competition is completed, the models could be continuously evaluated as long as new data are being collected.
- A metric for determining the winners should be based on sound theory and its sensitivity should be tested. The choice of metric will have an impact on the choice of best model and must be considered carefully, particularly in lieu of assumptions regarding the distribution of the data itself (e.g., Gaussian, Bernoulli, etc.). While our metric was deliberately simple, more nuanced metrics could be designed to account for additional criteria, such as declining predictive accuracy over time.
- If the end goal is to create an ensemble model, then uncertainty estimates should be part of the output given by competitors, and code should be standardized to a single language as a best practice (e.g., Python).
- If possible, data to be used for model testing should not be placed online before the competition, although if data are already public there are strategies to obfuscate data that can be discussed with data competition hosts. Models fit with knowledge of the withheld data could be made arbitrarily good (with respect to predicting withheld data) and even data that is retracted from the web prior to competition may be available to competitors through an old "image" of the internet. Our preference would have been to withhold more data, but this risked the integrity of the competition given that much of the older data had already been published. Determining the proportion of data to withhold is one of the most challenging and important elements of designing a fair competition that yields models that are likely to also perform well in future years.
- The length of time to run the competition should be considered with respect to the possibility of achieving better results as time progresses, while balancing the needs of the project (i.e. timing of funding and deadlines). We ran our competition for 2 months, but depending on the size of the dataset and complexity of the problem, longer might be recommended.

5. Conclusions

Our motivation for this competition was driven in part by our own frustration that "good" models for penguin population dynamics were elusive, and that few efforts have been made to benchmark models against each other. The data science competition framework offers the opportunity to compare one domain-knowledge inspired population model to domain-agnostic methods commonly used in data science. For our dataset at least, simple models performed comparably to complex models for prediction over the short term, and covariates strongly supported by prior knowledge of the system did not improve prediction accuracy. The domain-knowledge inspired model was competitive with the very best models submitted by the competition, which suggests that at least some of the commonly used tools in statistical ecology (in this case, hierarchical Bayesian time series modelling) are reasonable. Not surprisingly, techniques developed specifically for prediction (e.g., machine learning methods) scored highest. Although limited inference on mechanisms can be made by machine learning based methods, more traditional ecological modelling techniques are more appropriate for understanding cause and effect from hypothesis testing. Data science competitions provide one avenue for jumpstarting development of better predictive models, encourage community-level aggregation of 'clean' datasets, and directly facilitate public engagement. For all these reasons, we look forward to seeing future data science competitions for ecological research.

Acknowledgements

HJL, CC-C, GH gratefully acknowledge funding provided by US National Aeronautics and Space Administration [Award No. NNX14AC32G], without which there would have been no prize money for this competition, and especially Woody Turner and Cindy Schmidt for their help with that process. HJL would also like to acknowledge the National Science Foundation Office of Polar Programs Award No. NSF/OPP-1255058. We would like to thank C. Youngflesh for comments on drafts our manuscript. We would also like to greatly thank all the participants in the data science competition, and to the entire DrivenData team who managed the process.

Data accessibility

All data are available on the MAPPPD website (www.penguinmap.com), and details can be found in Humphries et al. (2017). DOI: https://doi.org/10.1017/S0032247417000055

References

Ainley, D.G., 2002. The Adelie Penguin: Bellweather of Climate Change. Columbia University Press, New York.

Ainley, D., Russell, J., Jenouvrier, S., Woehler, E., Lyver, P.O., Fraser, W.R., Kooyman, G.L., 2010. Antarctic penguin response to habitat change as Earth's troposphere reaches 2°C above preindustrial levels. Ecol. Monogr. 80, 49–66.

Akaike, H., 1969. Fitting autoregressive models for prediction. Ann. Inst. Stat. Math. *21*, 243–247.

Ballerini, T., Tavechia, G., Olmastroni, S., Pezzo, F., Focardi, S., 2009. Nonlinear effects of winter sea ice on the survival probabilities of Adélie penguins. Oecologia 161, 253–265.

Ben Taieb, S., Hyndman, R.J., 2014. A gradient boosting approach to the Kaggle load forecasting competition. Int. J. Forecast. 30, 382–394.

Boersma, P.D., 2008. Penguins as marine sentinels. Bioscience 58, 597–607.

Bowerman, B., O'Connell, R., Koehler, A., 2004. Forecasting, Time Series and Regression: An Applied Approach. Thomson Brooks, Belmont, CA.

Breiman, L., 2001. Random forests. In: Machine Learning. Vol. 45. pp. 5–32.

Bull, P., Slavitt, I., Lipstein, G., 2016. Harnessing the power of the crowd to increase capacity for data science in the social sector. In: 2016 International Conference on Machine Learning. Workshop on #Data4Good: Machine Learning in Social Good Applications. International Machine Learning Society, New York, pp. 31–35.

Carpenter, J., 2011. May the best analyst win. Science 331, 698–699.

Casanova, S., Ahrens, B., 2009. On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. Mon. Weather Rev. 137, 3811–3822.

Casanovas, P.V., Naveen, R., Forrest, S., Poncet, J., Lynch, H.J., 2015. A comprehensive coastal seabird survey maps out the front lines of ecological change on the Western Antarctic Peninsula. Polar Biol. 38, 927–940.

Che-Castaldo, C., Jenouvrier, S., Youngflesh, C., Shoemaker, K.T., Humphries, G.R.W., McDowall, P., Landrum, L., Holland, M.M., Yun, L., Rubao, J., Lynch, H.J., 2017. Pan-Antarctic analysis aggregating spatial estimates of Adélie penguin abundance reveals robust dynamics despite stochastic noise. Nat. Commun. 8, 832.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. ACM Press, pp. 785–794.

Cimino, M.A., Lynch, H.J., Saba, V.S., Oliver, M.J., 2016. Projected asymmetric response of Adélie penguins to Antarctic climate change. Sci. Rep. 6, 28785.

Clarke, J., Emmerson, L.M., Townsend, A., Kerry, K.R., 2003. Demographic characteristics of the Adelie penguin population on Bechervaise Island after 12 years of study. CCAMLR Sci. 10, 53–74.

Croxall, J.P., Kirkwood, E.D., 1979. The Distribution of Penguins on the Antarctic Peninsula and Islands of the Scotia Sea. British Antarctic Survey, Cambridge.

Cutler, D., Edwards, T., Beard, K., Cutler, A., Hess, K., Gibson, J., Lawler, J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.

Dietterich, T.G., 2000. Ensemble methods in machine learning. Mult. Class. Syst. 1857, 1–15.

Dietze, M.C., 2017. Ecological Forecasting. Princeton University Press.

Dugger, K.M., Ainley, D.G., Lyver, P.O., Barton, K., Ballard, G., 2010. Survival differences and the effect of environmental instability on breeding dispersal in an Adelie penguin meta-population. Proc. Natl. Acad. Sci. U. S. A. 107, 12375–12380.

Durga, L.S., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19, 225–235.

Elith, J., Leathwick, J., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

Forcada, J., Trathan, P.N., Reid, K., Murphy, E.J., Croxall, J.P., 2006. Contrasting population changes in sympatric penguin species in association with climate warming. Glob. Chang. Biol. 12, 411–423.

Fraser, W.R., Trivelpiece, W.Z., Ainley, D.G., Trivelpiece, S.G., 1992. Increases in Antarctic penguin populations: reduced competition with whales or a loss of sea ice due to environmental warming? Polar Biol. 11, 525–531.

Friedman, J., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Geissler, P.H. & Noon, B.R. (1981) Estimates of avian population trends from the North American Breeding Bird Survey. *Estimating Numbers of Terrestrial Birds* (eds C.J. Ralph & J.M. Scott), pp. 42–51. Stud. Avian Biol. 6.

Glaeser, E.L., Hillis, A., Kominers, S.D., Luca, M., 2016. Crowdsourcing City government: using tournaments to improve inspection accuracy. Am. Econ. Rev. 106, 114–118.

Goldman, C.R., Jassby, A., Powell, T., 1989. Interannual fluctuations in primary production: meteorological forcing at two subalpine lakes. Limnol. Oceanogr. 34, 310–323.

Hanke, J.E., Reitsch, A.G., 1995. Business Forecasting. Prentice-Hall, Englewood Cliffs, NJ.

Hinke, J.T., Salwicka, K., Trivelpiece, S.G., Watters, G.M., Trivelpiece, W.Z., 2007. Divergent responses of Pygoscelis penguins reveal a common environmental driver. Oecologia 153, 845–855.

Hinke, J.T., Trivelpiece, S.G., Trivelpiece, W.Z., 2017. Variable vital rates and the risk of population declines in Adélie penguins from the Antarctic peninsula region. Ecosphere 8 (1), e01666.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Humphries, G., Che-Castaldo, C., Naveen, R., Schwaller, M., McDowall, P., Schrimpf, M., Lynch, H., 2017. Mapping application for penguin populations and projected dynamics (MAPPPD): data and tools for dynamic management and decision support. Polar Rec. 3, 1–7.

Hyndman, R.J., Khandakar, Y., 2007. Automatic Time Series for Forecasting: The Forecast Package for R. Monash University: Department of Econometrics and Business Statistics.

Ives, A.R., Abbott, K.C., Ziebarth, N.L., 2010. Analysis of ecological time series with ARMA(*p*,*q*) models. Ecology 91, 858–871.

Jenouvrier, S., Barbraud, C., Weimerskirch, H., 2006. Sea ice affects the population dynamics of Adélie penguins in Terre Adélie. Polar Biol. 29, 413–423.

Jenouvrier, S., Caswell, H., Barbraud, C., Holland, M., Stroeve, J., Weimerskirch, H., 2009. Demographic models and IPCC climate projections predict the decline of an emperor penguin population. Proc. Natl. Acad. Sci. 106, 1844–1847.

Kabani, A., El-Sakka, M.R., 2016. North Atlantic right whale localization and recognition using very deep and leaky neural network. Math. Appl. 5, 155–170.

Krishnamurti, T.N., Kishtawal, C.M., Larow, T.E., Bachiochi, D.R., Zhang, Z., Williford, C.E., Gadgil, S., Surendran, S., 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285, 1548–1550.

Larue, M.A., Ainley, D.G., Swanson, M., Dugger, K.M., Lyver, P.O., Barton, K., Ballard, G., 2013. Climate change winners: receding ice fields facilitate colony expansion and altered dynamics in an Adélie penguin metapopulation. PLoS One 8, e60568.

Lescroël, A., Dugger, K.M., Ballard, G., Ainley, D.G., 2009. Effects of individual quality, reproductive success and environmental variability on survival of a long-lived seabird. J. Anim. Ecol. 78, 798–806.

Loeb, V., Siegel, V., Holm-Hansen, O., Hewitt, R., Fraser, W., Trivelpiece, W., Trivelpiece, S., 1997. Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. Nature 387, 897–900.

Lynch, H.J., Larue, M.A., 2014. First global census of the Adélie penguin. Auk 131, 457–466.

Lynch, H.J., Naveen, R., Trathan, P.N., Fagan, W.F., 2012. Spatially integrated assessment reveals widespread changes in penguin populations on the Antarctic peninsula. Ecology 93, 1367–1377.

Lynch, H.J., Naveen, R., Casanovas, P.V., 2013. Antarctic site inventory breeding bird survey data 1994/95-2012/13. Ecology 94, 2653.

Lyver, P.O., Barron, M., Barton, K.J., Ainley, D.G., Pollard, A., Gordon, S., McNeill, S., Ballard, G., Wilson, P.R., 2014. Trends in the breeding population of Adélie penguins in the Ross Sea, 1981–2012: a coincidence of climate and resource extraction effects. PLoS One 9, e91188.

McKenzie, E., 1984. General exponential smoothing and the equivalent Arma process. J. Forecast. 3, 333–344.

Narayanan, A., Shi, E., Rubinstein, B.I.P., 2011. Link prediction by de-anonymization: how we won the Kaggle social network challenge. In: The 2011 International Joint Conference on Neural Networks. IEEE, pp. 1825–1834.

Petchey, O.L., Pontarp, M., Massie, T.M., Kéfi, S., Ozgul, A., Weilenmann, M., Palamara, G.M., Altermatt, F., Matthews, B., Levine, J.M., Childs, D.Z., 2015. The ecological forecast horizon, and examples of its uses and determinants. Ecol. Lett. 18, 597–611.

Prasad, A., Iverson, L., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

Ribic, C., Ainley, D., Spear, L., 1998. Scale-related seabird-environmental relationships in Pacific equatorial waters, with reference to El Niño-Southern Oscillation events. Mar. Ecol. Prog. Ser. 156, 183–203.

Runge, M.C., Stroeve, J.C., Barrett, A.P., McDonald-Madden, E., 2016. Detecting failure of climate predictions. Nat. Clim. Chang. 6, 861–864.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310.

Stineman, R.W., 1980. A consistently well-behaved method of interpolation. Creat. Comput. 6, 54–57.

Taylor, S.J., Letham, B., 2017. Forecasting at scale. PeerJ Preprints *5*, e3190v2.

Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. J. Oper. Res. Soc. 66, 1352–1362.

Trivelpiece, W.Z., Hinke, J.T., Miller, A.K., Reiss, C.S., Trivelpiece, S.G., Watters, G.M., 2011. Variability in krill biomass links harvesting and climate warming to penguin population changes in Antarctica. Proc. Natl. Acad. Sci. 108, 7625–7628.

Ward, E.J., Holmes, E.E., Thorson, J.T., Collen, B., 2014. Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. Oikos 123, 652–661.

Weigand, A.S., Gershenfeld, N.A., 1994. Time Series Prediction. Santa Fe Institute, Santa Fe.

Weigel, A.P., Liniger, M.A., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134, 241–260.

Wichard, J., Ogorzalek, M., 2004. Time series prediction with ensemble models. In: IEEE International Joint Conference on Neural Networks. IEEE, pp. 1625–1630.

Wilson, P.R., Ainley, D.G., Nur, N., Jacobs, S.S., Barton, K.J., Ballard, G., Comiso, J.C., 2001. Adélie penguin population change in the pacific sector of Antarctica: relation to sea-ice extent and the Antarctic circumpolar current. Mar. Ecol. Prog. Ser. 213, 301–309.

Youngflesh, C., Jenouvrier, S., Li, Y., Ji, R., Ainley, D.G., Ballard, G., Barbraud, C., Delord, K., Dugger, K.M., Emmerson, L.M., Fraser, W.R., Hinke, J.T., Lyver, P.O.B., Olmastroni, S., Southwell, C.J., Trivelpiece, S.G., Trivelpiece, W.Z., Lynch, H.J., 2017. Circumpolar analysis of the Adélie penguin reveals the importance of environmental variability in phenological mismatch. Ecology 98, 940–951.