Received: 17 September 2019 | Revised: 30 December 2019 | Accepted: 7 January 2020

DOI: 10.1111/geb.13070

MACROECOLOGICAL METHODS

Global Ecology and Biogeography A Journal of Macroecology WILEY

Calibration of probability predictions from machine-learning and statistical models

Carsten F. Dormann

Biometry and Environmental System Analysis, University of Freiburg, Freiburg, Germany

Correspondence
Carsten F. Dormann, Biometry and Environmental System Analysis, University of Freiburg, Tennenbacher Str. 4, 79106 Freiburg, Germany.
Email: carsten.dormann@biom.uni-freiburg.de

Editor: Petr Keil

Abstract

Aim: Predictions from statistical models may be uncalibrated, meaning that the predicted values do not have the nominal coverage probability. This is easiest seen with probability predictions in machine-learning classification, including the common species occurrence probabilities. Here, a predicted probability of, say, .7 should indicate that out of 100 cases with these environmental conditions, and hence the same predicted probability, the species should be present in 70 and absent in 30.

Innovation: A simple calibration plot shows that this is not necessarily the case, particularly not for overfitted models or algorithms that use non-likelihood target functions. As a consequence, 'raw' predictions from such a model could easily be off by .2, are unsuitable for averaging across model types, and resulting maps hence be substantially distorted. The solution, a flexible calibration regression, is simple and can be applied whenever deviations are observed.

Main conclusions: 'Raw', uncalibrated probability predictions should be calibrated before interpreting or averaging them in a probabilistic way.

KEYWORDS

calibration plot, machine learning, model averaging, prediction bias, separation, species distribution model

1 | INTRODUCTION

Calibration refers to the comparison of measurements with reference values: electric conductivity against gravimetric soil moisture; chlorophyll concentrations against light absorption at a specific wavelength; a new sensor's pressure readings against an established barometer; wet-lab determined acid-digestible fibre against near-infrared spectroscopy readings; and so forth. Calibration of *model predictions* routinely takes place for example in meteorology, when the predicted probability of rainfall is regressed against actual rainfall incidences under the exact-same conditions; as a result, a 20% probability of precipitation means just that: of 100 days with these atmospheric conditions, it will rain on 20 (Silver, 2012).

Calibration is a statistical step in the employment of virtually any measuring and predicting approach. However, as ecologists we do not suspect uncalibrated output from statistical models. Rather, it would seem obvious that any predicted 'probability' has indeed the reported, so-called 'nominal coverage probability', be it probability of rain or probability of the occurrence of a species under certain environmental conditions. In other words: when a boosted regression tree predicts a value of .35 occurrence 'probability', then we expect the species to occur under these conditions with a frequency of 35 out of 100 cases (either across species, or within a species across different cells with same environmental conditions). For species distribution analyses, this view was rectified by Guillera-Arroita et al. (2015), who clarified what kind of inference can be drawn from different types of data.

Here I want to draw attention to the phenomenon of uncalibrated predictions, in particular for machine-learning models of binary data. The reason for a discrepancy between a machine-learning prediction and the actual frequency stems from overfitting, potentially leading to perfect separation of classes and from the use of non-likelihood loss functions, which lead to a non-probabilistic weighting of prediction misfits. As such, also other classifications and even discrete distributions may cause similar problems. This observation is not new (Pearce & Ferrier, 2000), and indeed corrections have been proposed and re-examined by, among others, Platt (2000), Niculescu-Mizil and Caruana (2005) and Lin, Lin, and Weng (2007). Correct probability predictions are particularly important, as Platt (2000) points out, when they form part of an actual probability-based decision, or when they are averaged with other methods, so that a common measure is required. In the specific field of species distribution models of presence–absence data, Pearce and Ferrier (2000) featured such calibration prominently, yet hardly any study or even standard has picked it up (Araújo et al., 2019; Peterson et al., 2011; Sofaer et al., 2019); for notable exceptions see Franklin (2010); Johnston et al. (2015, 2019); Guisan, Thuiller, and Zimmermann (2017) and Fink et al. (2020).

Diagnosing such non-probabilistic behaviour is simple, and, for all practical purposes, calibration is straightforward. Hence it should be applied to any model type as part of the prediction process, *before* predicting, cross-validating and making effect plots and maps or using predictions in any other probabilistic interpretation.

2 | THE CALIBRATION PLOT AND A DEMONSTRATION OF BIAS

A little analysis demonstrates where the calibration comes into a statistical analysis. The example here is that of an Australian bird species' distribution (as presence–absence data at a scale of 50 × 50 km^2), using climate and land-cover predictors. The example itself is immaterial and merely illustration [R code (R Core Team, 2019) and data can be found in Supporting Information Appendix S1], although in the context of species distribution analysis presence–absence data and predicted occurrence probability are particularly common (some recent examples are Derville, Torres, Iovan, & Garrigue, 2018; Marca et al., 2019; Martínez et al., 2018; Robinson, Ruiz-Gutierrez, & Fink, 2018; Sabatini et al., 2018; Sofaer et al., 2019).

The data were analysed using a traditional binomial generalized linear model (GLM) and some machine-learning approaches: random forest, (simple) neural networks, boosted regression trees and support vector machines. All approaches managed to fit the data well, measured as root mean square error and log-likelihood in five-fold cross-validation [Table 1, for area under the curve (AUC) see Supporting Information Appendix S1].

The calibration plot (see Box 1, also known as a reliability diagram) has been a statistical goodness-of-fit measure for a long time, but seems to have fallen out of fashion (despite recommendations by Harrell, 2001, 2015). It simply regresses observed data against model fits (step I.4 in Box 1) and thus provides information additional to the common (pseudo-)R^2-value (Chalcraft, 2019). For binary data, this requires a binomial GLM and link-scale predictions. The slope of this regression line, the calibration slope β1, should be unity, and the calibration intercept, β0, zero. Figure 1 shows some calibration curves. In this specific illustration, the GLM, the simple neural network and the support vector machine actually truthfully lie on the expected calibration line (with intercept near 0 and slope near 1); the two tree-based approaches display misfit.

TABLE 1 Spatial-block cross-validation RMSE and log-likelihood of the five model types without ('raw') and with ('cal') calibration of their predictions. The models were fitted on the eastern/western half of the data and predicted to the other half (and vice versa). For log-likelihood (ℓ) the sum of the two folds is given, RMSE values represent means. Rank-independent metrics, such as area under the curve (AUC), show no effect of calibration (see Supporting Information Appendix S1)

Model type	$RMSE_{raw}$	$RMSE_{cal}$	ℓ_{raw}	ℓ_{cal}
GLM	0.404	0.408	−5,320	−2,660
RF	0.325	0.363	−1,250	−2,850
ANN	0.435	0.306	−1,690	−3,340
SVM	0.333	0.336	−1,680	−1,940
BRT	0.352	0.367	−2,000	−8,390

Abbreviations: ANN = artificial neural network; BRT = boosted regression tree; GLM = generalized linear model; RF = randomForest; RMSE = root mean square error; SVM = support vector machine.

In this case, the predictions can be calibrated using a flexible calibration regression, such as a structurally constrained generalized additive model (GAM; see step II.1, Box 1). A simple correction based on the binomial GLM estimates (step I.4 in Box 1) is typically insufficient, as lines need not be sigmoidal. Thus, the observed values are regressed against the fitted values using a GAM constrained to be monotonically increasing, as suggested for this purpose in the appendix of Johnston et al. (2015) and detailed in Pya & Wood (2015), yielding the panels on the right of Figure 1.

Platt (2000), after whom the probability-rescaling is sometimes referred to as 'Platt scaling', suggested that this kind of calibration regression will overfit and suggests regularization, cross-validation or replacement of the actual observed values by values moved away from the margin (i.e., 0 or 1). The latter step is motivated by an application of Bayes rule, and the replacement target values are then, for 1s: $t_1 = \frac{N_1+1}{N_0+2}$ and for 0s: $t_0 = \frac{1}{N_0+2}$, with N_0 and N_1 representing the number of 0s and 1s in the training data. Platt sketches a customized calibration procedure, upon which Lin et al. (2007) improve. In their case of support vector machines, the deviation followed a clear sigmoidal pattern (as also seen in Figure 1, e.g., bottom left), and hence Platt suggested to fit a sigmoidal function. For more serpentine curves, one could jackknife the calibration curve to prevent it from overfitting, but the structurally constrained GAM will often curb this problem sufficiently.

Note that any calibration requires the observed data to be unbiased. If, for example, 0s are unreliable and possibly largely attributable to low detection probabilities, calibration cannot restore actual probabilities (Fithian & Hastie, 2013). This applies particularly to any use-availability

Box 1 Calibration steps

After fitting a model to binary data, a calibration plot indicates whether there is a need to calibrate the model's predictions. The following two parts outline the necessary steps to diagnose and treat uncalibrated probabilities. Note that all steps but one (I.4) require response-scale predictions. Code to demonstrate these steps in R is provided in Supporting Information Appendix S1.

I. The calibration plot. Create a calibration plot and compute calibration statistics.

1. Compute model fits (= predicted values for observed data) at the *link* and *response* scale.
2. Plot observed 0/1s on the *y* against response-scale model fits on the *x* axis. Jitter *y*-values to better visualize data density. Add 1:1 line for reference (Figure 1).
3. Add locally weighted scatterplot smoothing or spline smooths to guide the eye.
4. Compute a binomial generalized linear model (GLM) with observed data as function of *link*-scale model fits. (Through the logit link both predictions and fits are at the link-scale in the calibration regression). Ideally, these calibration statistics are an intercept of 0 and a slope of 1, approximately.

This yields the calibration plot, as depicted in the left column of Figure 1, and the calibration statistics.

II. Calibrating predictions. Transform model predictions into calibrated probabilities.

1. Fit a *flexible* binomial model, for example, a structurally constrained monotonously increasing generalized additive model (GAM), to observed data as function of response-scale model fits. This is the calibration regression for later use on model predictions.
2. Compute model predictions, for example, for global change scenarios, effect plots or maps, at the response scale.
3. Use these response-scale model predictions as input in the calibration regression, and compute calibrated predictions at the response scale. Done.

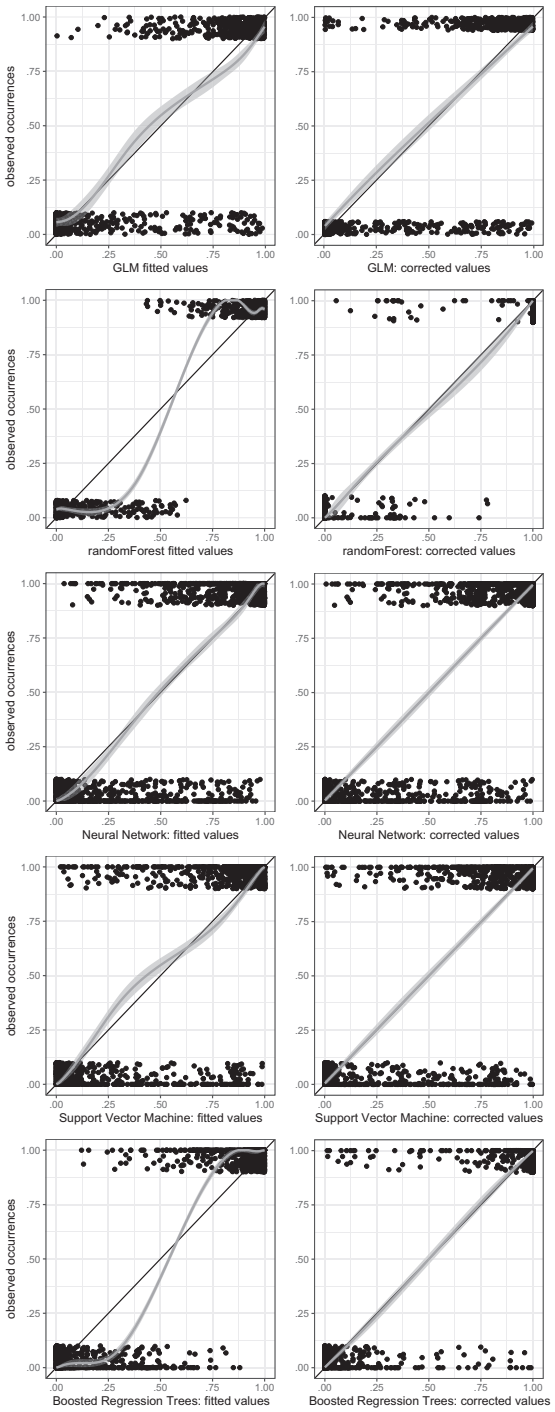

FIGURE 1 Raw prediction-based calibration plots (left) and calibrated prediction plots (right) for five models [generalized linear model (GLM), randomForest, neural network, support vector machine and boosted regression tree] fitted to the same binary data (indicated by dots), with fits on the *x* and observed values on the *y* axis. Diagonal line indicates ideal 1:1 line, grey line a smooth of the predictions. Observations were jittered to better display density of data, which are highly concentrated at 0 and 1 in the case of randomForest. The calibration curve was fitted as a generalized additive model (GAM) constrained to be monotonously increasing (see Box 1 and Supporting Information Appendix S1). Note that, in effect, calibration spreads 0s and 1s more widely across the *x* axis

analysis as employed, for example, in resource-selection or presence-only data, which requires quantification of detection probabilities (Guillera-Arroita et al., 2015; Royle, Kéry, Gautier, & Schmid, 2007).

3 | WHERE DOES THE BIAS COME FROM?

Prediction bias has been reported repeatedly in the ecological literature, often with a clear pattern of overprediction of rare, and underprediction of common events (e.g., Calabrese, Certain, Kraan, & Dormann, 2014; Detto, Visser, Wright, & Pacala, 2019). The causes may be manifold, from

regression dilution (Frost & Thompson, 2000; McInerny & Purves, 2011), to incomplete model structures (Mod, Roux, Guisan, & Luoto, 2015; Pellissier et al., 2012) to overfitting and non-probabilistic estimators, discussed here.

For machine learning, one cause for the discrepancy between model predictions and the 1:1 line may lie in the target function of the modelling approaches. A GLM (and GAM, for that matter) maximizes the log-likelihood, which is founded in probability theory. In contrast, machine-learning algorithms may minimize Gini impurity or maximal risk (minimax), or maximize variance reduction (Hastie, Tibshirani, & Friedman, 2009). While this may make for good class separation, it does not bode well for nominal probabilities. Comparing, for instance, the GLM to a target function that maximizes accuracy (i.e., the proportion of correctly predicted values, using prevalence as the threshold) displays distorted probabilities (Figure 2).

In this illustration, the cause behind the misfit for boosted regression trees must be something different, as it uses the likelihood as target function. The most likely candidate is overfitting, which can lead to perfect separation of the two classes (0s and 1s) and resulting misfits (e.g., Heinze & Schemper, 2002). In neural networks, we can tune the back-propagation by the decay rate, and setting this to very low values increases the chance of overfitting – and deviation from the calibration line. Similarly, the cost-parameter of a support vector machine can be tuned in such a way that deviations are either prominent or absent.

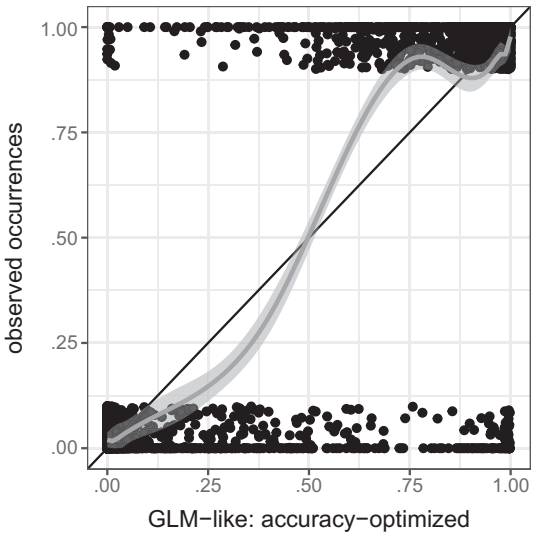

FIGURE 2 Generalized linear model (GLM) fitted using accuracy rather than log-likelihood as optimization target; compare to Figure 1 top left. This is to illustrate that different optimization targets can lead to very different calibration curves

4 | DOES IT MATTER?

But does it matter: will a map of occurrence probabilities look all that different with calibrated predictions to the current practice? That depends to a large extent on the dominant environmental conditions in the depicted region. If for this region predictions are either close to 0 or close to 1, then the map will be nearly indistinguishable. The more shades of grey the prediction has, the more the maps will indeed differ. In this arbitrary case study, prediction maps for randomForest with and without calibration look very similar, apart from south of Perth and Tasmania (Figure 3).

Also the functional relationship between predictors and response is affected by calibration (Figure 4). The already strong 'raw' effects become de-facto thresholds in the calibration. This is not an effect of the calibration, but rather depicts more truthfully the actual predictions of the model itself: the randomForest really fits a threshold, and only because its predictions are uncalibrated probabilities does it not appear so strongly in the raw predictions.

The ultimate aim of calibration is better predictions: if model predictions are biased, that should show under external validation. Splitting the data into an eastern and a western half, using that for training and predicting to the other (spatial block cross-validation: Roberts et al., 2017), I found that prediction error and log-likelihood were both increased and decreased (Table 1). That is to be expected, as also class-separating algorithms should yield very good validation results for these relatively small samples, particularly for rank-based validation measures such as AUC (see Supporting Information Appendix S1).

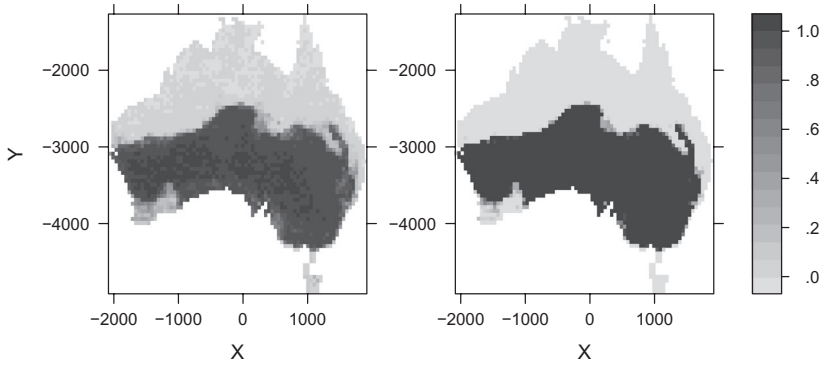

FIGURE 3 Maps of fitted occurrence probability without (left) and with (right) calibration. Note that the randomForest predictions become much crisper through the correction. The calibration plot (Figure 1 left, second from the top) shows that the actual frequency in the observations is much higher than randomForest predictions above .5. Hence, when randomForest predicts $p(Y = 1|X) = .75$ it actually means that the corrected probabilities are close to 1. Units are Universal Transverse Mercator projection coordinates in km

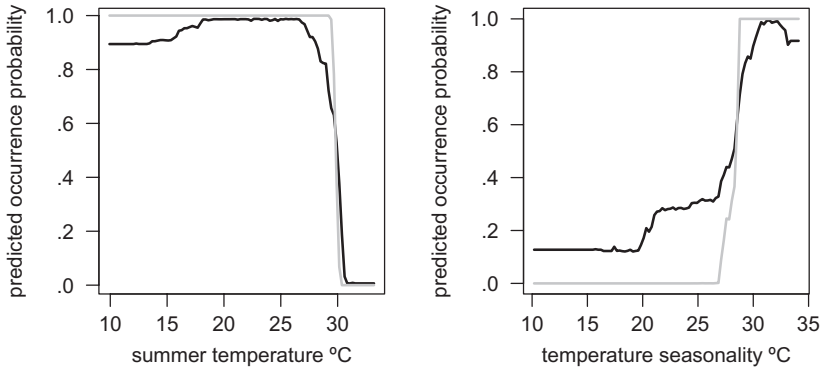

FIGURE 4 Conditional effect plots for the effect of summer temperature (left) and temperature seasonality (right) on occurrence probability. Black lines are uncalibrated predictions of the randomForest model, grey are calibrated. The effect of such a threshold-like function can be seen as much crisper calibrated predictions in Figure 3

While it is beyond dispute that machine-learning predictions may require calibration before being interpretable as probabilities (Platt, 2000), the best way to achieve such calibration is a matter of continuous refinement (e.g., Lin et al., 2007). The pragmatic GAM-based approach presented here is not the final say on the topic. It is important to notice, however, that calibration effects can be stark (Figure 1) and without calibration regression probabilities can easily be misinterpreted.

5 | RECOMMENDATIONS

Overfitting may lead to full separation of categories in the fitted model. As a consequence, uncalibrated probability predictions may well be biased. This is the case for any classification model, including not only data following Bernoulli, binomial and multinomial distributions, but also discrete data such as Poisson and negative binomial. Calibrating predictions is straightforward and approximately restores the actual interpretation as predicted probabilities. Such calibration is necessary when averaging probabilities and interpreting predictions as probabilities, but not when options are merely ranked. Rarely will ecologists or practitioners use such predicted probabilities alone to decide on a management strategy, and often the uncertainty of these predictions may be substantially larger than the calibration misfit. However, as part of a sound craftsmanship, statistical analysts can be expected to do what they know is right: calibrate their predictions.

ACKNOWLEDGMENTS

This paper benefitted substantially from comments by Florian Hartig and Severin Hauenstein before, and from Petr Keil after submission. Particular thanks go to Daniel Fink for drawing my attention to the structurally constrained GAM as well as some of the relevant literature.

DATA ACCESSIBILITY

All data and R code are available as Supporting Information.

ORCID

Carsten F. Dormann https://orcid.org/0000-0002-9835-1794

REFERENCES

Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., … Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*, eaat4858.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*, 99–112. https://doi.org/10.1111/geb.12102

Chalcraft, D. R. (2019). To replicate, or not to replicate—that should not be a question. *Ecology Letters*, *22*, 1174–1175. https://doi.org/10.1111/ele.13286

Derville, S., Torres, L. G., Iovan, C., & Garrigue, C. (2018). Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches. *Diversity and Distributions*, *24*, 1657–1673. https://doi.org/10.1111/ddi.12782

Detto, M., Visser, M. D., Wright, S. J., & Pacala, S. W. (2019). Bias in the detection of negative density dependence in plant communities. *Ecology Letters*, *22*, 1923–1939. https://doi.org/10.1111/ele.13372

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*. in press. https://doi.org/10.1002/eap.2056

Fithian, W., & Hastie, T. (2013). Statistical models for presence-only data: Finite-sample equivalence and addressing observer bias. *The Annals of Applied Statistics*, *7*, 1917–1939.

Franklin, J. (2010). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, *16*, 321–330. https://doi.org/10.1111/j.1472-4642.2010.00641.x

Frost, C., & Thompson, S. G. (2000). Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society A*, *163*, 173–189. https://doi.org/10.1111/1467-985X.00164

Guillera-Arroita, G., Lahoz-Monfort, J., Elith, J., Gordon, A., Kujala, H., Lentini, P., … Wintle, B. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*, 276–292. https://doi.org/10.1111/geb.12268

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in R*. UK: Cambridge University Press.

Harrell, F. E. (2001). *Regression modeling strategies—with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Harrell, F. E. (2015). *Regression modeling strategies—with applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York, NY: Springer.

Hastie, T., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin, Germany: Springer.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409–2419. https://doi.org/10.1002/sim.1047

Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., … Kelling, S. (2015). Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, *25*, 1749–1756. https://doi.org/10.1890/14-1826.1

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Gutierrez, V. R., Robinson, O. J., Miller, E. T., … Fink, D. (2019). Best practices for making reliable inferences from citizen science data: Case study using eBird to estimate species distributions. *bioRxiv*, 574392.

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, *68*, 267–276. https://doi.org/10.1007/s10994-007-5018-6

Marca, W. L., Elith, J., Firth, R. S. C., Murphy, B. P., Regan, T. J., Woinarski, J. C. Z., & Nicholson, E. (2019). The influence of data source and species distribution modelling method on spatial conservation priorities. *Diversity and Distributions*, *25*, 1060–1073. https://doi.org/10.1111/ddi.12924

Martínez, B., Radford, B., Thomsen, M. S., Connell, S. D., Carreño, F., Bradshaw, C. J. A., … Wernberg, T. (2018). Distribution models predict large contractions of habitat-forming seaweeds in response to ocean warming. *Diversity and Distributions*, *24*, 1350–1366. https://doi.org/10.1111/ddi.12767

McInerny, G. J., & Purves, D. W. (2011). Fine-scale environmental variation in species distribution modelling: Regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, *2*, 248–257. https://doi.org/10.1111/j.2041-210X.2010.00077.x.

Mod, H. K., le Roux, P. C., Guisan, A., & Luoto, M. (2015). Biotic interactions boost spatial models of species richness. *Ecography*, *38*, 913–921. https://doi.org/10.1111/ecog.01129

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. http://doi.acm.org/10.1145/1102351.1102430

Pearce, J., & Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, *133*, 225–245. https://doi.org/10.1016/S0304-3800(00)00322-7

Pellissier, L., Pradervand, J.-N., Pottier, J., Dubuis, A., Maiorano, L., & Guisan, A. (2012). Climate-based empirical models show biased predictions of butterfly communities along environmental gradients. *Ecography*, *35*, 684–692. https://doi.org/10.1111/j.1600-0587.2011.07047.x.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. NJ: Princeton University Press.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schurrmans (Eds.), *Advances in large margin classifiers* (Vol. *10*, pp. 61–74). Cambridge, MA: MIT Press.

Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, *25*, 543–559. https://doi.org/10.1007/s11222-013-9448-7

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Robinson, O. J., Ruiz-Gutierrez, V., & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, *24*, 460–472. https://doi.org/10.1111/ddi.12698

Royle, J. A., Kéry, M., Gautier, R., & Schmid, H. (2007). Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs*, *77*, 465–481. https://doi.org/10.1890/06-0912.1

Sabatini, F. M., Burrascano, S., Keeton, W. S., Levers, C., Lindner, M., Pötzschner, F., … Debaive, N. (2018). Where are Europe's last primary forests? *Diversity and Distributions*, *24*, 1426–1439.

Silver, N. (2012). *The signal and the noise: The art and science of prediction*. London, UK: Penguin.

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., … Hamilton, H. (2019). Development and delivery of species distribution models to inform decision-making. *BioScience*, *69*, 544–557. https://doi.org/10.1093/biosci/biz045

BIOSKETCH

Carsten F. Dormann is a statistical ecologist with particular interests in method development and transfer. His research focusses particularly on the analysis of species distributional data as well as interaction networks.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Dormann CF. Calibration of probability predictions from machine-learning and statistical models. *Global Ecol Biogeogr.* 2020;00:1–6. https://doi.org/10.1111/geb.13070