Integrating citizen science data with expert surveys improves the accuracy and range of inference of species distribution models

O.J. Robinson[1*], V. Ruiz-Gutierrez[1], M.D. Reynolds[2], G.H. Golet[3], M. Strimas-Mackey[1], D. Fink[1]

[1] Cornell Laboratory of Ornithology, Cornell University, 159 Sapsucker Woods Road, Ithaca, NY 14850 USA

[2] The Nature Conservancy, 201 Mission Street, Fourth Floor, San Francisco, CA 94105 USA

[3] The Nature Conservancy, 500 Orient Street Suite 150, Chico, CA 95928 USA

* Corresponding author: Email: ojr7@cornell.edu

Data Accessibility: The Nature Conservancy point count data is available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.724vn. The eBird data used for analyses in this manuscript may be downloaded from https://ebird.org/data/download. Cropland Data Layer may be downloaded from: https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php

Abstract:

Information on species' habitat associations and distributions, across a wide range of spatial and temporal scales, are a fundamental source of ecological knowledge. However, collecting biological information at relevant scales if often cost prohibitive, although it is essential for framing the broader context of more focused research and conservation efforts. Citizen-science data has been signaled as an increasingly important source of biological information needed to fill in data gaps needed to make more comprehensive and robust inferences on species distributions. However, there are perceived trade-offs of combining highly structured, scientific survey data with largely unstructured, citizen-science data. As a result, the focus of most methodological advances to combine these sources of information has been on treating these sources as independent. The degree to which each source of information is allowed to directly inform a common underlying process (e.g. species distribution) depends on the perceived quality of the data. In this paper, we explore these trade-offs by applying a simplified approach of filtering citizen-science data to resemble structured survey data, and analyze both sources of data under a common framework. To accomplish this, we explored ways of integrating high-resolution survey data on shorebirds in the northern Central Valley of California with observations in eBird for the entire region that were filtered to improve their quality. The integration of survey data with the filtered citizen-science data in eBird resulted in improved inference and predictive ability, and increased the extent and accuracy of inferences on shorebirds for the Central Valley. The structured surveys were found to improve the overall accuracy of ecological inference based only on citizen-science data, by increasing the representation of data collected from high quality habitats for shorebirds (e.g. rice fields). The practical approach we have shown for data integration can be also be used to improve the efficiency of designing biological surveys in the context of larger, citizen-science monitoring efforts, ultimately reducing the financial and time expenditures typically required of monitoring programs and focused research. The simple processing and filtering method we present can be used to integrate other types of citizen-science data (e.g. camera traps) with more localized efforts (e.g. research projects), ultimately improving our ecological knowledge on the distribution and habitat associations of species of conservation concern worldwide.

Keywords: citizen science, data filtering, data integration, data quality, species distribution model, structured survey

INTRODUCTION

Information on species' habitat associations and distributions are a fundamental source of ecological knowledge (Sofaer et al. 2019).This information is often of interest across a broad range of spatial and temporal scales, from high-resolution information that is more relevant for research on habitat selection (Matthiopoulos et al. 2011) or needed to inform management objectives (Zipkin et al. 2010) to larger-scale inferences that are useful to address broader questions (e.g. potential range-shifts with changing climatic conditions; Lyon et al. 2019). However, the process of collecting biological observations across large-spatial scales is often cost-prohibitive for most research and monitoring efforts. Under the best-case scenario, researchers and practitioners are able to monitor plant and animal communities just within their study regions, and during specific times of year. However, the need to make inferences beyond the sampled range of environmental conditions and seasons often limits our understanding of the boarder context of our results, and can limit the use of applied research to inform future monitoring efforts and effective conservation actions.

Citizen-science data has been signaled as a promising source of information to fill-in information gaps needed to model species distributions (Bradter et al. 2018, Gouraguine et al. 2019). However, there are few examples of the potential trade-offs of combining observational data collected at more localized scales with large-scale citizen-science data. This might be due in part to inherent differences that often exist between the two data types. On the one hand, you may have data that is collected at a high spatial resolution using skilled observers, sampling effort is often standardized, and sampling occurs across a habitat gradient that is representative of the region of interest. On the other hand is citizen-science data, which can be collected across a wide range of sampling conditions by observers that vary widely in their level of expertise,

collected across a wide range of spatial resolutions, and sampling effort is not standardized. This has created a perceived trade-off between data quality and quantity among data collected from structured, scientific surveys and data collected from larger-scale, volunteer-based monitoring efforts (Figure 1). The assumption being made is that an increase in quantity of citizen-science data comes at a significant cost to quality. Therefore, the logical framework to integrate these two sources of information would be one that would treat them as independent sources of information used to inform a common underlying distribution for a given species (Pacifici et al. 2017).

The integration of different data sources is a growing area of methodological development in ecological statistics, and recent advances have been made to develop ways of integrating survey data (e.g. structured) with citizen-science data (e.g. un-structured data) (Miller et al. 2019). For observational data collected at discrete locations, these methods include specifying a joint-likelihood for the two data sources to estimate the underlying species distribution (Miller et al. 2019). In cases where this is not possible, the data source that is deemed as of lower quality (e.g. citizen science data, museum observations) can be used in two ways: 1) modeled as a covariate of the underlying distribution, or 2) used to estimate a separate species distribution, where a correlation structure is specified to share information across data sources. Pacifici et al. (2017) tested these different approaches to integrate observational data from the citizen science project eBird (Sullivan et al, 2017) with more structured data from the North American Breeding Bird Survey (BBS; Sauer et al. 2017). Their results showed that the joint-likelihood approach of combining eBird and BBS data outperformed all other approaches, including using BBS data alone.

The approach used by Pacifici et al (2017) and Miller el at (2019) summarized observational data at a coarse, grid-level in order to account for differences in effort, sampling approach, and other variables that are known to influence detectability (Guillera-Arroita 2017). In addition, Pacifici et al (2017) wanted to reduce potential bias related to the degree of uncertainty about the spatial scale that observations were collected for each independent eBird checklist. This mismatch in scales between the two data sources is what often makes data integration between high-resolution survey data (e.g. point count observations) with lower-resolution citizen-science data non-trivial. However, many citizen-science programs collect high-scale resolution information (e.g. camera traps) in ways that we can infer absences, and collect additional information on effort (e.g. distance traveled, number of hours sampled) that is highly valuable for improving the accuracy of SDMs (Kelling et al. 2019).

Here, we explore a practical approach for data integration between high-quality, citizen-science data with structured survey data that builds upon existing methods for 'data pooling' (e.g. Fithian et al. 2015). We explore the trade-offs in inference of using citizen-science data alone, more localized and structured data alone, and pooling together both data sets combined. In addition, we examine specific trade-offs when combining structured and un-structured data sources by exploring the performance of increasing the quantity of citizen-science data through simulations, vs. the addition of data from more targeted survey efforts. To accomplish this, we use The Nature Conservancy's (TNC) BirdReturns project as a case study (Reynolds et al. 2017). We explore ways of combining high-resolution bird survey data collected for shorebirds on rice fields in the northern region of the Central Valley in California, with observations in eBird for the entire Central Valley that are filtered to improve their quality. The specific aim of the case

study is to provide a framework for leveraging survey data with citizen-science data to build more accurate distribution models for shorebird species across the extent of the Central Valley.

METHODS

Data

We used point counts carried out during spring surveys (February 1- May 31; n = 8,192) as part of the TNC BirdReturns project conducted in 2014-2017. This project used predicted shorebird occurrence and abundance (Johnston et al., 2015) along with predicted surface water in the Sacramento River Valley to identify times and locations that were likely to be important for migrating shorebirds. TNC used a reverse auction approach to select and incentivize rice farmers in the identified locations to flood their fields during the spring and fall, making temporary wetlands available to the migrating shorebirds. Observers made point counts at fields enrolled in the program and at unenrolled control sites, surveying a semi-circle with a 200 m fixed-radius. Each site was surveyed for at least two minutes and lasted as long as necessary to count all birds present (for more detail on count methods see Golet et al. 2018). Effort for each point count consisted of date, time started and ended, and name of observer.

We combined the point counts with data from the citizen science project eBird (Sullivan et al. 2014) collected during the same time period as the point counts. The eBird data was restricted to the Central Valley of California, USA, and to complete checklists so that non-detection could be inferred (Johnston et al. 2019). We also restricted the eBird data to stationary checklists and traveling checklists limited to 300 m. After filtering this data, we were left with 12,891 checklists. Effort variables for the eBird data set were date, time observations started,

duration of observation in minutes, survey protocol (stationary or traveling), distance traveled in meters, and number of observers.

We calculated the effort variables from the point counts to match those of the eBird data set. Observer name in the point count data set was converted into number of observers (however it was always one for this study), time started and ended for the point counts was used to calculate duration in minutes, and each point count was treated as a stationary count (distance traveled = 0). By doing this, the two data sets contained the same effort information and were identical in structure which allowed us to simply join them into one combined data set. We added a variable to the combined data set to note whether an observation was a TNC point count or an eBird checklist. We also calculated a checklist calibration index (CCI) for each checklist in the combined dataset to account for variation in expertise among observers (e.g. expertise score; Johnston et al. 2018). We attached the remotely sensed Cropland Data Layer (CDL; Boryan et al. 2011, Han et al. 2012) to the combined data set by computing the percent of each land cover or crop type in the CDL that was present within a 300 m radius centered on each point count or eBird checklist. We also similarly attached cloud-filled data from Water Tracker (Reiter et al. 2018), a high spatial and temporal resolution surface water tracking system for the Central Valley of California.

A comprehensive data set was created using all of the above eBird checklists (12,891) plus simulated eBird checklists equal to the number of TNC point counts that were included in the combined dataset (8,192). This resulted in a dataset that was the same magnitude as the combined dataset and allowed us to determine if the improvement in accuracy from combining the datasets was simply a function of increasing the sample size. The simulated eBird data was created by slightly jittering the spatial covariates of 8,192 randomly chosen checklists, similar to

the oversampling procedure in Fink et al. (2019), however we used all checklists rather than only positive observations here. This ensured that we were maintaining roughly the same prevalence rate in the dataset, and also were not simply making exact copies of the randomly chosen checklists.

Spatial Filtering and Class Imbalance

As spatial bias is always a concern when using citizen science data (Geldmann et al. 2016), we spatially subsampled the combined data set. The data was sparse for many species (proportion of detections for a species < 0.05 for all checklists in the dataset), so class imbalance was also a concern (He and Garcia 2009). We spatially undersampled the data (e.g. Robinson et al. 2017) by first creating a hexagonal grid of 3.5 km cells over the region from which our observations came via the dggridR package (Barnes et al. 2018) in R (R Core Team 2019). We then split the data for a single species into checklists on which the species was detected (positive observations) and those on which the species was not detected (negative observations). We selected one checklist from the negative observations from within each grid cell and recombined the filtered negative observations with the positive observations. As almost all of the spatial bias was from the negative observations (i.e. only a small percentage of the total number of observations were positive observations), this procedure relieves much of the spatial bias, and because only negative observations were filtered out of the dataset, class imbalance is also addressed here (King and Zeng 2001, Robinson et al. 2017).

To alleviate the effects of class imbalance, after spatially sampling eBird checklists for training distribution and population trend models, Fink et al. (2019) oversampled eBird checklists for species that had a prevalence rate of less than 25%. After spatially undersampling

our data for the current study, class imbalance was still a concern, as our undersampling improved the imbalance, but it did not improve class balance to 25% for species other than Yellowlegs. Following the recommendation of Fink et al. (2019), we oversampled the positive observations for each of our species below 25% prevalence before training the distribution models. We used the synthetic minority oversampling technique (SMOTE; Chawla et al. 2002) to create one new example of the positive class in the training dataset for every positive observation in the spatially undersampled dataset. The SMOTE algorithm does not create exact copies of the positive class as in traditional oversampling, but instead creates examples of the positive class that occupy the parameter space between a randomly chosen positive observation and a nearest neighbor. We did not randomly undersample our data as is recommended when using SMOTE (Chawla et al. 2002) because our data had already been spatially undersampled as described above. The spatial undersampling and SMOTE procedure was done to each of the datasets. As the spatial undersampling randomly chooses from negative observations within a grid cell, many negative observations may not be included in training the model. Therefore, we sampled the data using this procedure 100 times, creating 100 unique datasets.

Analysis

We selected and spatially subsampled (but did not oversample) 15% of the combined dataset to be the testing data for evaluation of each model. This test set was selected because our goal is to make accurate predictions across the entire Central Valley, and give equal importance to each location. Therefore, the test set must include observations across the spatial extent of evaluation and be spatially balanced. For the training datasets, we removed any checklist or point count that was in the test set.

We used the R package ‘ranger’ (Wright et al. 2019) to train a random forest model for each of the seven species (or combined species; Table 1) and for each of four datasets: 1) TNC point counts alone, 2) eBird checklists alone, 3) an oversampled eBird dataset, 4) the combined TNC and eBird dataset. For each species, 1000 trees were grown in the ensemble and the number of variables from which the model could select at each split for each tree was set to the square root of the number of variables included in the model ($n = 12 \approx \sqrt{142}$). We evaluated the accuracy of the models using multiple predictive performance metrics (PPMs). We did not evaluate the models using the out of bag (OOB) metric calculated by random forest internally as it favors total accuracy over correctly predicting the minority class, and is highly sensitive to class imbalance. We used the test dataset to evaluate mean squared error (MSE) between the model predictions of presence or absence and the true presence or absence in the test set. We also evaluated error using Brier score (Brier 1950), the mean squared error of the probabilistic model predictions and the true presence or absence in the test set. We evaluated each model’s ability to rank positive observations higher than negative ones using the area under the curve (AUC; Fielding and Bell 1997). We evaluated each model’s ability to predict presence or absence using Cohen’s Kappa (Kappa; Cohen 1960) and its components, sensitivity (true positive rate), and specificity (true negative rate). We produced distribution maps for each species and dataset, and we recorded the predictor importance metrics from models trained on each dataset.

RESULTS

The fields that are stored as part of the eBird project allowed us to filter the data to be of a similar protocol as the TNC point counts. This filtering reduced uncertainty in the location of eBird checklists and eliminated the need for coarse level summaries of the data for integration (Pacifici et al. 2017, Miller et al. 2019). For all species, the combined dataset had higher

predictive accuracy than either the TNC point counts or eBird checklist datasets on their own; however error (MSE, Brier score) was relatively low and AUC was relatively high for all species and all datasets, particularly for the three datasets where eBird checklists were included (Figure 1; S1). Improvement in accuracy varied among species, however; for the Kappa statistic (predicting presence and absence against the test set), the combined dataset was an improvement over all three of the other datasets evaluated (4.5% - 35.5% improvement; Figure 2). The improvement or loss in the error statistics was usually negligible (with the exception of LBCU; Figure 4; S1) for the combined dataset versus the next best dataset; however, error metrics were already very low for most species, so great improvement here was not likely. Likewise, AUC was relatively high for most species and for the three datasets containing eBird checklists, therefore the gain was negligible for many species (improvement of -0.60% - 7.5%; S1). For the few species/metrics combinations where the combined dataset was not the best performing model, it was the second best, with the best being the eBird checklists with simulated eBird checklists added. The TNC point counts, nor eBird checklists alone performed better for any metric than the dataset where the two were combined.

We produced distribution maps for each species (Figure 5; S1) to determine if there was a difference in the overall pattern of distribution estimated by models trained on the different datasets. For most species, there was greater contrast between the presence estimates and absence estimates for the combined dataset when compared to the other three. Visually, this means hotter ‘hotspots’ and darker regions where absence is predicted (Figure 5; S1). We also collected the important variables for the model when run with each of the datasets. For all species, the importance of rice and water was apparent as it was among the top variables for each of the datasets, however, it was not until the datasets were combined that each rice and the Water

Tracker layer had a high importance score (Figure 6; S1). The combination of the data allowed the models to hone in on rice and surface water as highly important, where they had only moderate importance comparatively when using the other datasets. This is likely because only 1.5 % of eBird checklists in our study come from locations where the percent land-cover is >50% rice. Conversely, almost 60% of the TNC point counts come from locations where the percent land-cover is >50% rice.

DISCUSSION

Our results lend further support to efforts looking to combine data from multiple sources to improve the inference and/or predictive ability of distribution models (Pacifici et al. 2017, Miller et al. 2019). We have shown that citizen-science data can be filtered to generate a high-quality data set that can closely match the resolution and sampling approach of structured surveys, supporting the call for current and future citizen-science projects to collect essential information related to location and effort, as well as complete surveys (Kelling et al, 2019). The integration of survey data with the filtered citizen-science data in eBird resulted in improved inference, predictive ability, and ultimately increased the extent of inference of the structured surveys. In turn, the structured surveys were able to improve the ecological inference of the citizen-science data, by improving the representation of sampled habitats that are key for shorebird species. Most importantly, the practical approach we have shown for data integration is an improvement on simpler ‘data pooling’ approaches for data integration, and can be used to improve the efficiency of designing biological surveys to collect distribution information in the

context of larger, citizen-science monitoring efforts, ultimately reducing the financial and time expenditures typically required of monitoring programs and focused research (Reich et al. 2018).

The combined dataset resulted in improved accuracy across all metrics relative to the TNC survey data or eBird data alone, for all of the species considered in this study. Our combined approach also predicted presence/absence via agreement with a test data set more accurately than the different permutations of datasets considered (e.g. TNC alone, eBird alone, eBird plus simulated eBird). The observed improvement in accuracy is not likely a function of simply increasing sample size, since augmenting eBird data with simulated data did not show the same levels of improvement. The addition of the TNC survey data to eBird data improved the coverage of the data both spatially and in targeted habitats. Previous work has shown that migrating shorebirds heavily use flooded rice fields in the Central Valley, including unflooded rice fields (Elphick and Oring 1998, Golet et al. 2018). Given that rice fields are the main focal habitat of the TNC BirdReturns program, ~60% of the survey data from this project was carried out at sites where at least 50% of the landcover is rice. In contrast, the majority of the rice fields in the Central Valley are not accessible to regular eBird participants, as they are often on private lands. This is likely why we see so few (~1.5%) eBird checklists from the Central Valley in locations where at least 50% of the landcover is rice. The addition of more data from high quality habitat is what provided the improvement in accuracy of the combined eBird and TNC data set. This highlights the importance and value of more targeted research and survey efforts within the context of large-scale citizen-science monitoring efforts. Given that private lands make up more than half of the land in the United States, supporting wildlife monitoring efforts on privately held lands that are linked into large-scale efforts such as eBird can greatly improve inferences on

species distributions and habitat associations across scales of interest for all stakeholders (Hilty and Merenlender 2003).

Interestingly, models based only on the TNC point counts struggled to learn where each species was likely to be absent across the entire Central Valley because the data came largely from "good" shorebird habitat in the northern portion of the Central Valley. The original intent of this project was not focused on species distribution modeling, so this does not come as a surprise. However, given that absence information allows for more accurate distribution models (Brotons et al. 2004), the addition of eBird data was able to provide information on where species are likely to be absent, and improved inferences on what habitat types most benefit shorebirds, but were not surveyed as part of the TNC monitoring efforts. On the other hand, the models using the eBird data alone were able to predict absences well, and had relatively high accuracy when predicting presence/absence, but overall ecological inference was improved when combined with the point count dataset. The TNC point counts acted as targeted surveys in under-surveyed habitat, which previous work has shown can improve the accuracy of distribution models using eBird checklists (e.g. Xue et al. 2016). The complementary nature of the datasets is also shown when examining the important predictors. For Dunlin (Figure 6), the rice landcover and the Water Tracker layer are important predictors for each dataset, however their importance values more than double when the combined dataset is used to train the model. While other species do not show such a drastic shift in the importance value for these two habitat variables, the pattern is similar for most in that these variables become more important to the model's predictions when the datasets are combined.

The approach we present for combining data is applicable for other conservation monitoring programs and ongoing research efforts, given that a significant amount of research is

conducted over small spatial and temporal scales (Heidorn 2008), and often difficult to scale-up without similarly structured data (Poisot et al. 2019). The data fields that exist in eBird data facilitated further processing and filtering to match the structure of the individual dataset of interest, and allowed us to leverage the strengths of both data sources. However, even citizen-science datasets that do collect effort information are often lacking information on sampling locations, although incentivizing participants to collect data in these data-poor locations has been shown to improve the accuracy of distribution models (e.g. avicaching; Xue et al. 2016). Similarly, data from smaller-scale, individual research projects can also help fill in these gaps in citizen-science data. Given that inferences from small-scale studies cannot be extrapolated to larger spatial extents (Sandel and Smith 2009), the approach we present here for combining data from small-scale studies with citizen-science data filtered to match the existing data structure will increase the overall extent of inference, and improve our ability to conceptualize conservation actions within the larger context of the target population(s) of interest.

Information on species distributions across large scales is one of the most fundamental information needs for basic and applied research fields in ecology. However, this level of information often requires large-scale, coordinated surveys that can be time consuming and costly to manage. In addition, models used to estimate species distributions are often data-hungry, and are often unable to generate information at the spatial and temporal scales that are most relevant for research and conservation efforts. We have shown the utility of combining survey data with semi-structured citizen science data (Kelling et al, 2019) for improving accuracy in species distribution models, which can result in more efficient and cost effective surveys (Pacifici et al. 2017, Reich et al. 2018, Miller et al. 2019). The simple processing and filtering method we present for citizen science data allows for the integration of a small-scale

point count data set with eBird checklists, can be used to integrate similar types of data being collected by citizen-scientists (e.g. camera traps) with more localized efforts (e.g. patrolling by park rangers), ultimately improving our ecological knowledge on the distribution and habitat associations of species of conservation concern worldwide.

REFERENCES

Barnes, R., K. Sahr, G. Evenden, A. Johnson, F. Warmerdam, and] Maintainer. 2018. Package "dggridR" Type Package Title Discrete Global Grids.

Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. Geocarto International 26:341–358.

Bradter, U., L. Mair, M. Jönsson, J. Knape, A. Singer, and T. Snäll. 2018. Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models of less common species? Methods in Ecology and Evolution 9:1667–1678.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78:1–3.

Brotons, L., W. Thuiller, M. B. Araújo, and A. H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437–448.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20:37–46.

Elphick, C. S., and L. W. Oring. 1998. Winter management of Californian rice fields for waterbirds. Journal of Applied Ecology 35:95–108.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.

Fink, D., T. Auer, A. Johnston, V. Ruiz-Gutierrez, W. M. Hochachka, and S. Kelling. 2019. Modeling Avian Full Annual Cycle Distribution and Population Trends with Citizen Science Data. bioRxiv:251868.

Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424–438.

Geldmann, J., J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Diversity and Distributions 22:1139–1149.

Golet, G. H., C. Low, S. Avery, K. Andrews, C. J. McColl, R. Laney, and M. D. Reynolds. 2018. Using ricelands to provide temporary shorebird habitat during migration. Ecological Applications 28:409–426.

Gouraguine, A., J. Moranta, A. Ruiz-Frau, H. Hinz, O. Reñones, S. C. A. Ferse, J. Jompa, and D. J. Smith. 2019. Citizen science in data and resource-limited areas: A tool to detect long-

term ecosystem changes. PLoS ONE 14:1–14.

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography 40:281–295.

Haibo He, H., and E. A. Garcia. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21:1263–1284.

Han, W., Z. Yang, L. Di, and R. Mueller. 2012. CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Computers and Electronics in Agriculture 84:111–123.

Heidorn, B. 2008. Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57:280–299.

Hilty, J., and A. M. Merenlender. 2003. Studying Biodiversity on Private Lands. Page Conservation Biology.

Johnston, A., D. Fink, W. M. Hochachka, and S. Kelling. 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods in Ecology and Evolution 9:88–97.

Johnston, A., D. Fink, M. D. Reynolds, W. M. Hochachka, B. L. Sullivan, N. E. Bruns, E. Hallstein, M. S. Merrifield, S. Matsumoto, and S. Kelling. 2015. Abundance models improve spatial and temporal prioritization of conservation resources. Ecological Applications 25:1749–1756.

Johnston, A., W. Hochachka, M. Strimas-Mackey, V. R. Gutierrez, O. Robinson, E. Miller, T. Auer, S. Kelling, and D. Fink. 2019. Best practices for making reliable inferences from

citizen science data: case study using eBird to estimate species distributions. bioRxiv:574392.

Kelling, S., A. Johnston, A. Bonn, D. Fink, V. Ruiz-Gutierrez, R. Bonney, M. Fernandez, W. M. Hochachka, R. Julliard, R. Kraemer, and R. Guralnick. 2019. Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity. BioScience 69:170–179.

King, G., and Zeng. 2001. Logistic Regression in Rare Events Data. Political Analysis 9:137–163.

Lyon, N. J., D. M. Debinski, and I. Rangwala. 2019. Evaluating the utility of species distribution models in informing climate change-resilient grassland restoration strategy. Frontiers in Ecology and Evolution 7:1–8.

Matthiopoulos, J., M. Hebblewhite, G. Aarts, and J. Fieberg. 2011. Generalized functional responses for species distributions. Ecology 92:583–589.

Miller, D. A. W., K. Pacifici, J. S. Sanderlin, and B. J. Reich. 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution 10:22–37.

Pacifici, K., B. J. Reich, D. A. W. Miller, B. Gardner, G. Stauffer, S. Singh, A. McKerrow, and J. A. Collazo. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion*. Ecology 98:840–850.

Poisot, T., A. Bruneau, A. Gonzalez, D. Gravel, and P. Peres-Neto. 2019. Ecological Data Should Not Be So Hard to Find and Reuse. Trends in ecology & evolution 34:494–496.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reich, B. J., K. Pacifici, and J. W. Stallings. 2018. Integrating auxiliary data in optimal spatial design for species distribution modelling. Methods in Ecology and Evolution.

Reiter, M. E., N. K. Elliott, B. Barbaree, and D. Moody. 2018. Water Tracker An Automated Open Surface Water Tracking System for California's Central Valley Water Tracker An Automated Open Surface Water Tracking System for California's Central Valley Point Blue Conservation Science.

Reynolds, M. D., B. L. Sullivan, E. Hallstein, S. Matsumoto, S. Kelling, M. Merrifield, D. Fink, A. Johnston, W. M. Hochachka, N. E. Bruns, M. E. Reiter, S. Veloz, C. Hickey, N. Elliott, L. Martin, J. W. Fitzpatrick, P. Spraycar, G. H. Golet, C. McColl, and S. A. Morrison. 2017. Dynamic conservation for migratory species. Science Advances 3:e1700707.

Robinson, O. J., V. Ruiz-Gutierrez, and D. Fink. 2017. Correcting for bias in distribution modelling for rare species using citizen science data. Diversity and Distributions.

Sandel, B., and A. B. Smith. 2009. Scale as a lurking factor: incorporating scale-dependence in experimental ecology. Oikos 118:1284–1291.

Sauer, J. R., K. L. Pardieck, D. J. Ziolkowski, A. C. Smith, M.-A. R. Hudson, V. Rodriguez, H. Berlanga, D. K. Niven, and W. A. Link. 2017. The first 50 years of the North American Breeding Bird Survey. The Condor 119:576–593.

Sofaer, H. R., C. S. Jarnevich, I. S. Pearse, R. L. Smyth, S. Auer, G. L. Cook, T. C. Edwards, G. F. Guala, T. G. Howard, J. T. Morisette, and H. Hamilton. 2019. Development and Delivery of Species Distribution Models to Inform Decision-Making. BioScience 69:544–557.

Sullivan, B. L., J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. A. Dhondt, T. Dietterich, A. Farnsworth, D. Fink, J. W. Fitzpatrick, T. Fredericks, J. Gerbracht, C. Gomes, W. M. Hochachka, M. J. Iliff, C. Lagoze, F. A. La Sorte, M. Merrifield, W. Morris, T. B. Phillips, M. Reynolds, A. D. Rodewald, K. V. Rosenberg, N. M. Trautmann, A. Wiggins, D. W. Winkler, W. K. Wong, C. L. Wood, J. Yu, and S. Kelling. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation 169:31–40.

Wright, M. N., S. Wager, P. Probst, and] Maintainer. 2019. Package “ranger” Type Package Title A Fast Implementation of Random Forests.

Xue, Y., I. Davies, D. Fink, C. Wood, and C. P. Gomes. 2016. Behavior identification in two-stage games for incentivizing citizen science exploration. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9892 LNCS:701–717.

Zipkin, E. F., J. Andrew Royle, D. K. Dawson, and S. Bates. 2010. Multi-species occurrence models to evaluate the effects of conservation and management actions. Biological Conservation 143:479–484.

Table 1. List of species for which we modeled spring distribution in the Central Valley of California with each of the four datasets.

Common Name	**Latin Name**	**Abbreviation in Tables & Figures**
American Avocet	*Recurvirostra americana*	AMAV
Dunlin	*Calidris alpina*	DUNL
Greater Yellowlegs*	*Tringa melanoleuca*	YLEG*
Least Sandpiper	*Calidris minutila*	LESA
Lesser Yellowlegs*	*Tringa flavipes*	YLEG*
Long-billed Curlew	*Numenius americanus*	LBCU
Long-Billed Dowitcher**	*Limnodromus scolopaceus*	DOWI**
Short-billed Dowitcher**	*Limnodromus griseus*	DOWI**
Western Sandpiper	*Calidris mauri*	WESA

* Combined in analysis and referred to collectively as “Yellowlegs” in this study

** Combined in analysis and referred to collectively as “Dowitcher” in this study

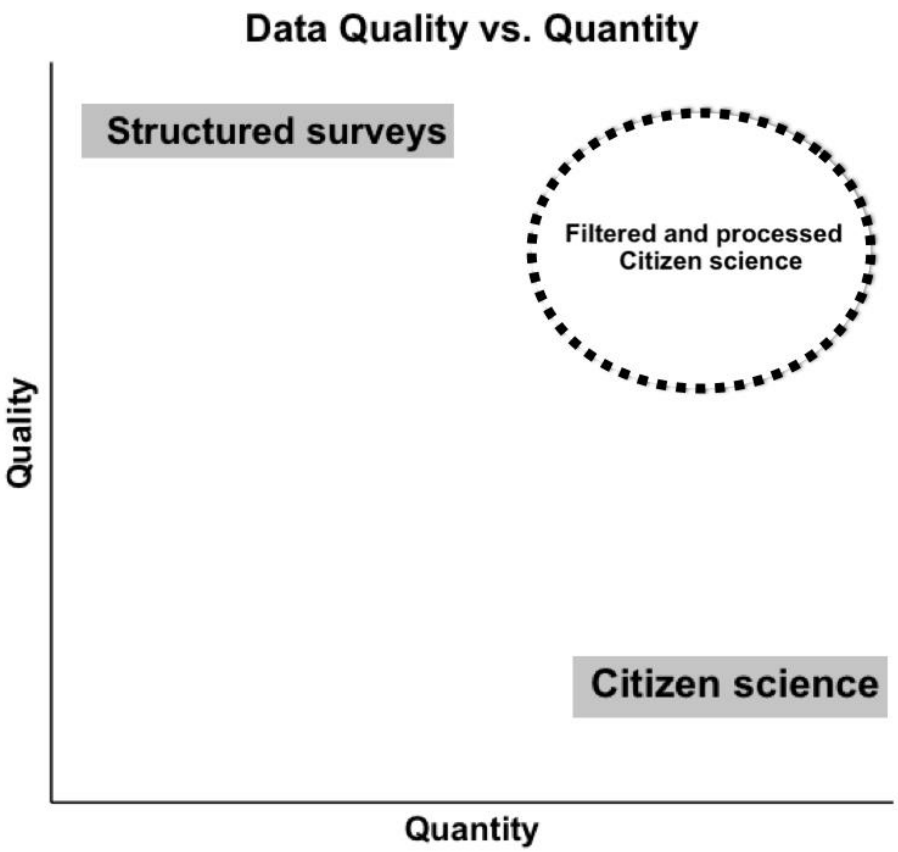


Figure 1. The perceived trade-off between data quality and quantity of data collected using structured surveys and citizen-science efforts. You can bypass this tradeoff by processing and filtering citizen-science data using criteria such as count type (e.g. stationary vs. traveling), duration of counts (e.g. all observations < 30mins), and other measures that align well with the existing data set. This approach is recommended as the most flexible data integration method for the application of a broad-range of species distribution models.

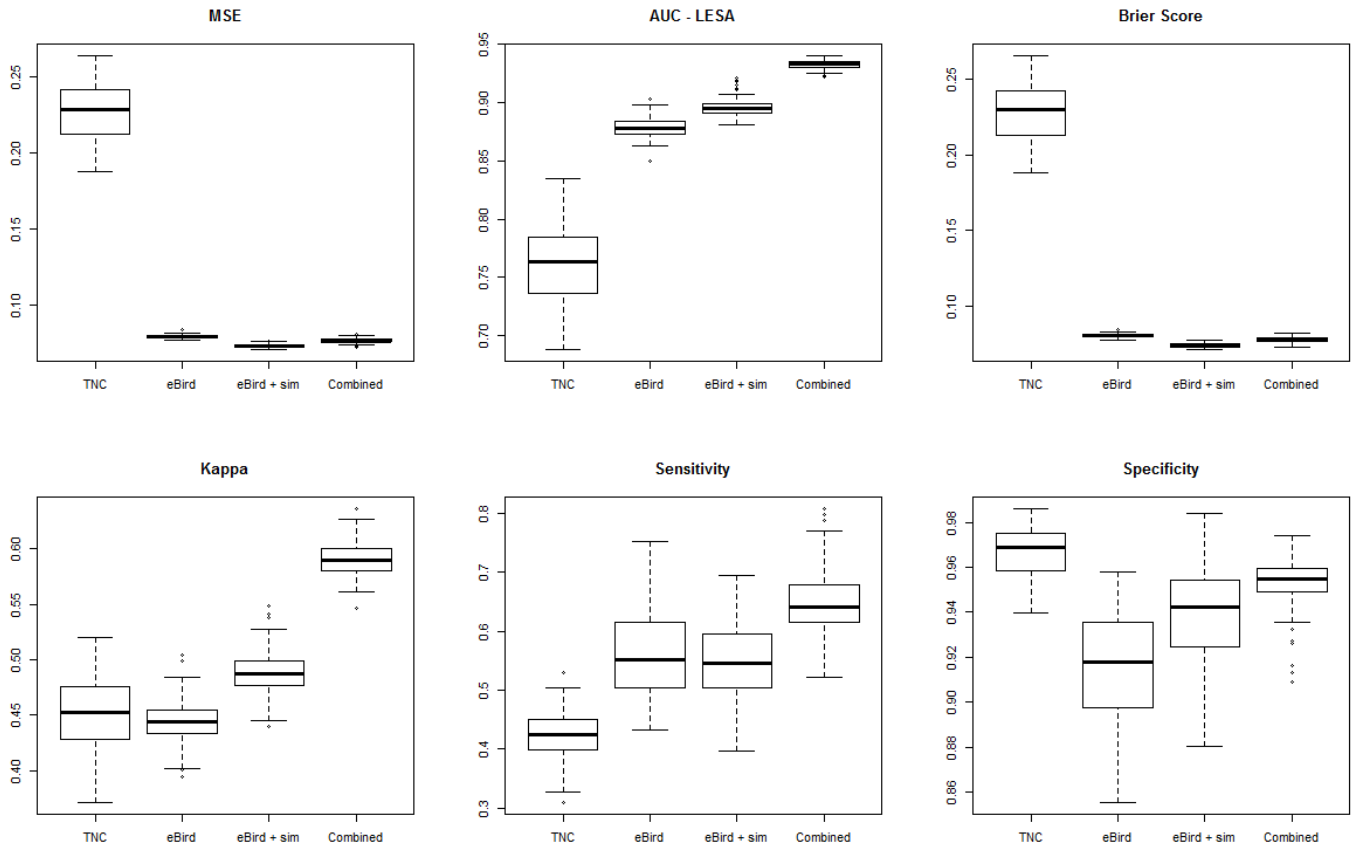


Figure 2. Boxplots for MSE, AUC, Brier score, Cohen's Kappa, Sensitivity and Specificity for models run for Least Sandpiper (LESA) with each dataset; TNC point counts (TNC), eBird checklists (eBird), eBird checklists with additional simulated eBird data (eBird + sim), and the dataset of TNC point counts and eBird checklists combined (Combined).

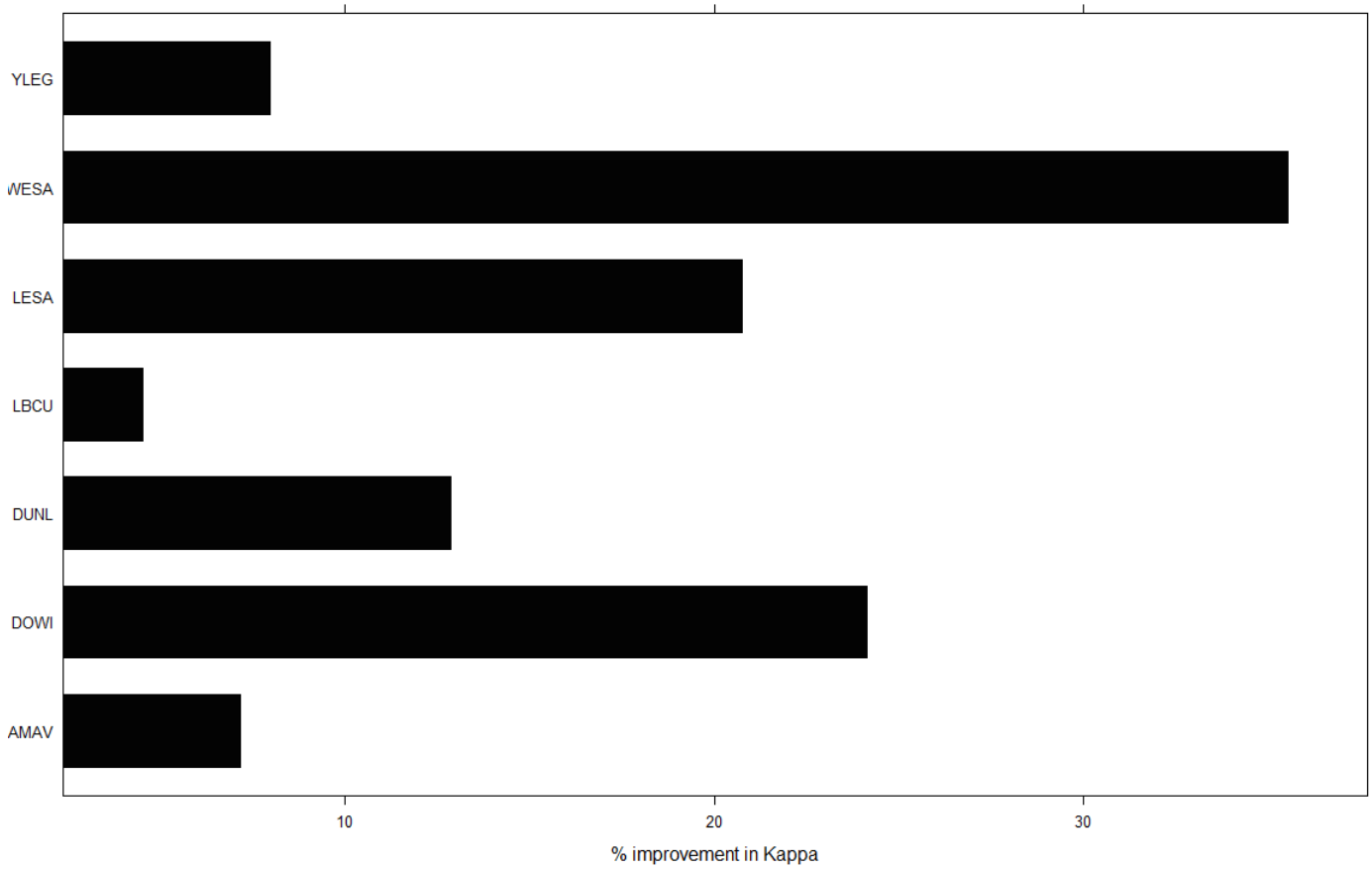


Figure 3. Average improvement in Cohen's Kappa for distribution models of each species when using the combined dataset versus the dataset that provided the next highest Kappa value.

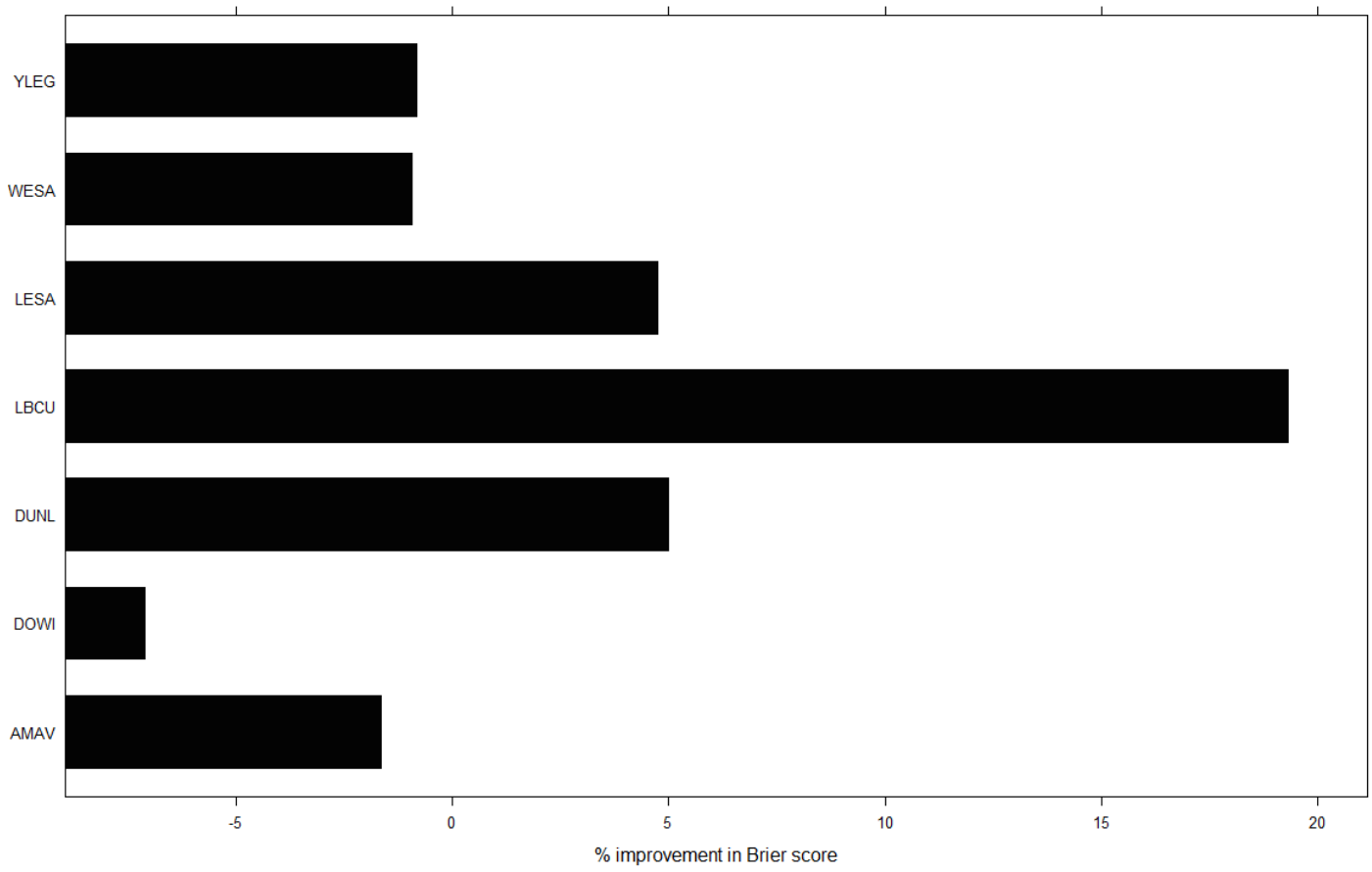


Figure 4. Average improvement or loss in Brier score for distribution models of each species when using the combined dataset versus the dataset that provided the next lowest value.

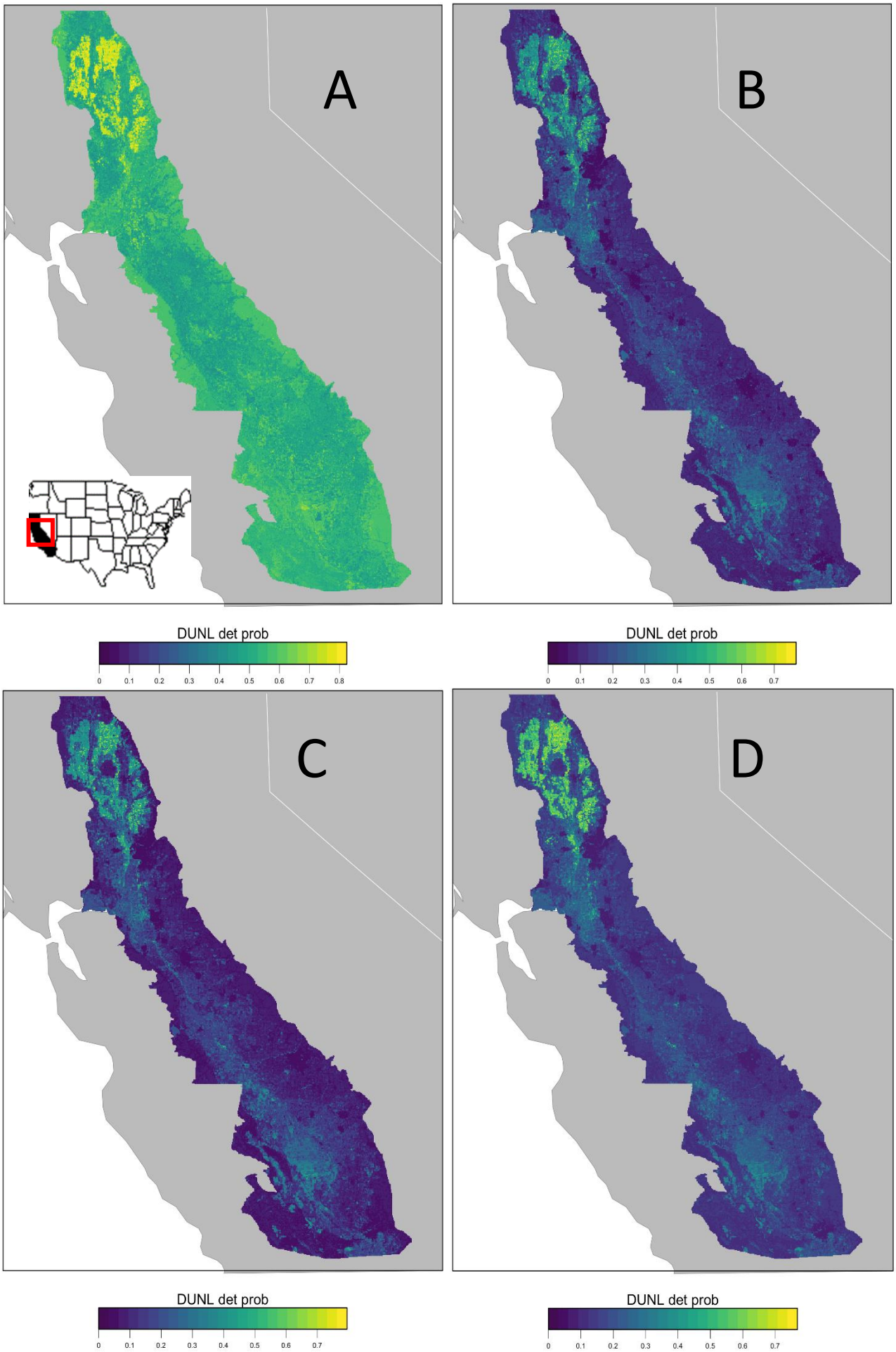
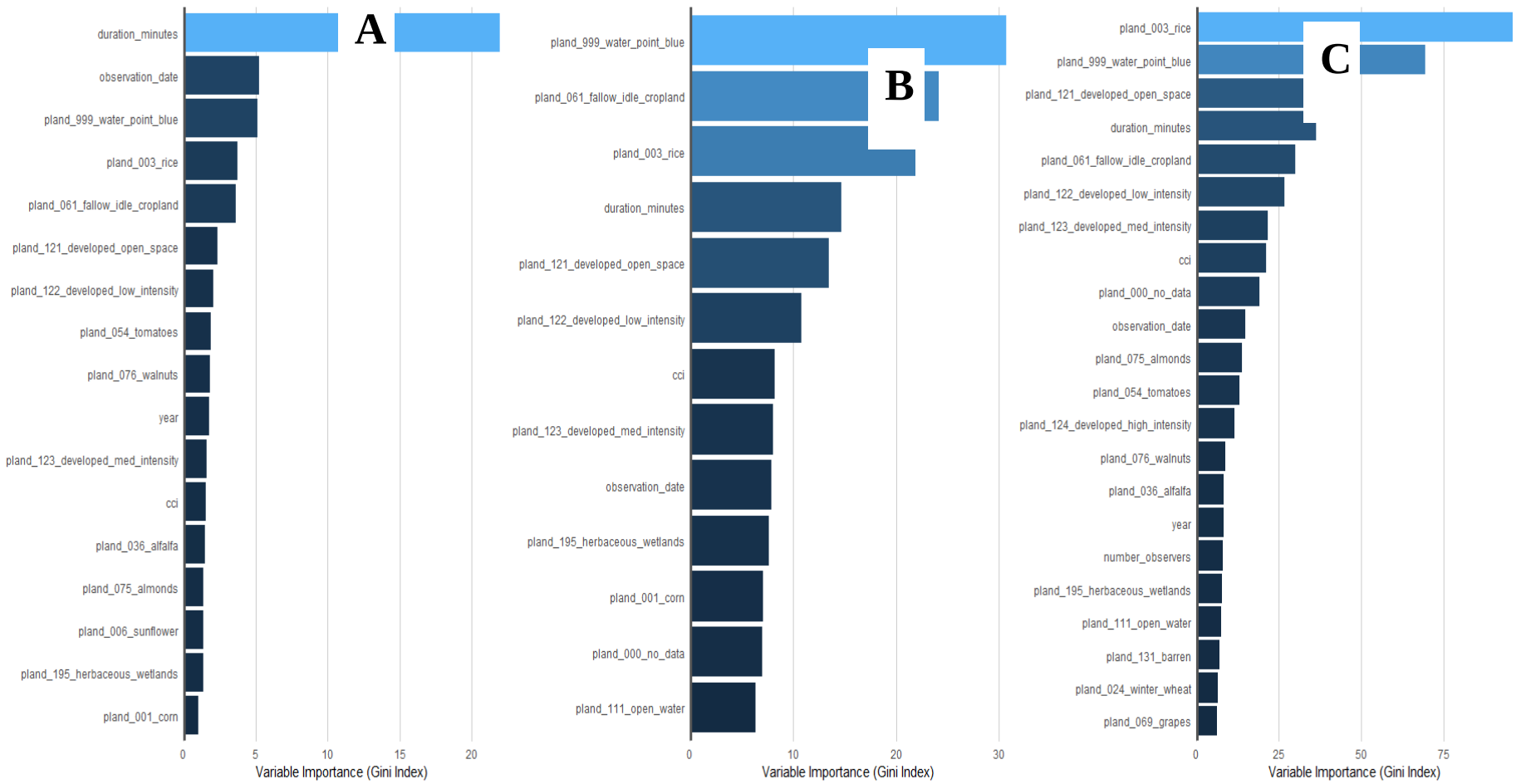
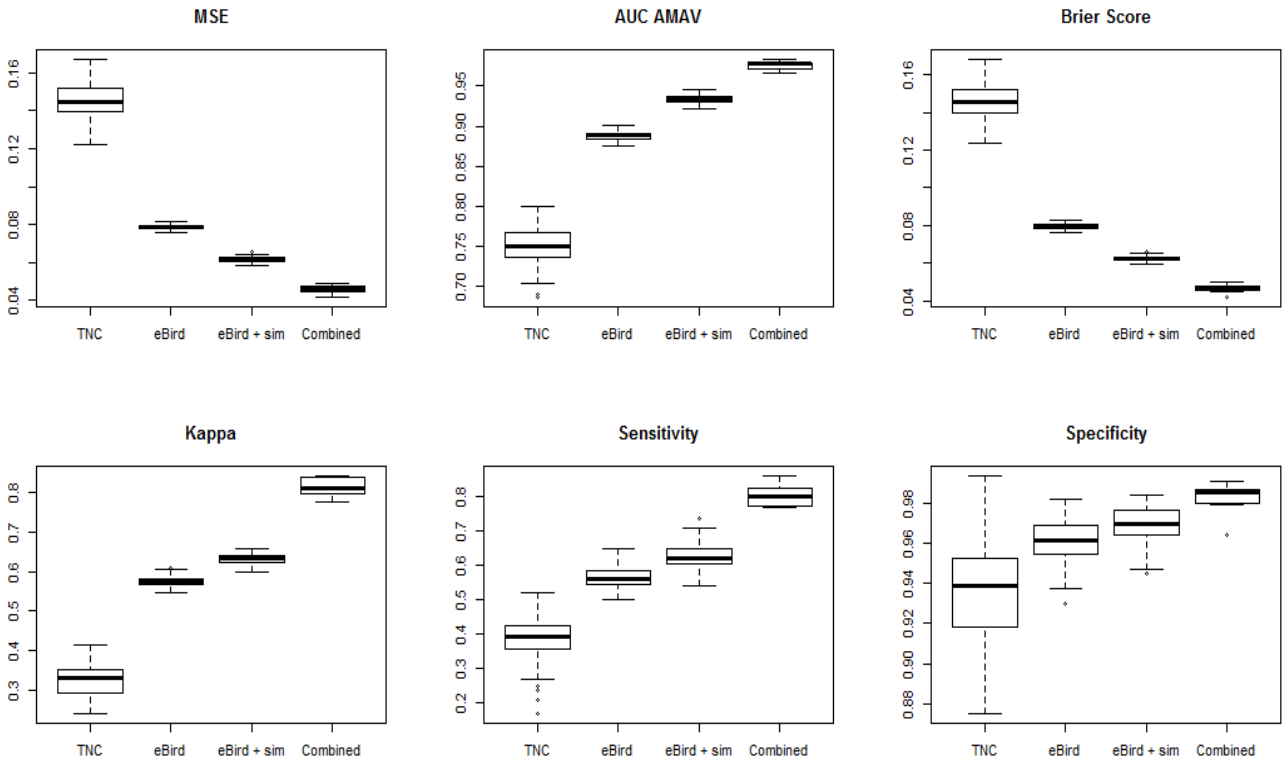

A
B
C
D
DUNL det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
DUNL det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
DUNL det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
DUNL det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

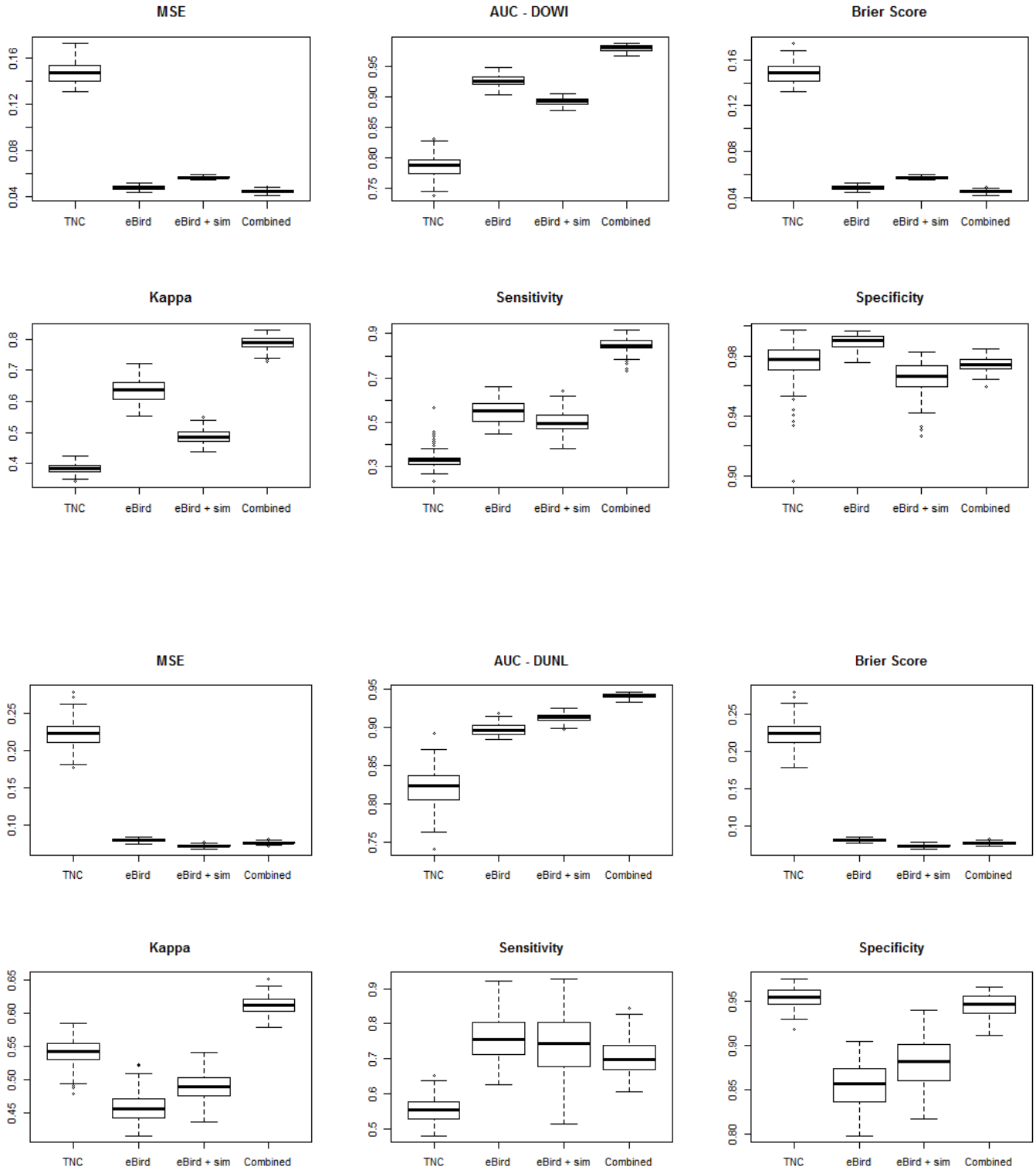
Figure 5. Predicted probability of an expert surveyor detecting a Dunlin in the spring of 2016 on a one-hour long checklist where the distance traveled was < 300m., as estimated by the model using TNC point counts alone (A), eBird checklists alone (B), eBird checklists with added simulated eBird checklists (C), and the combined dataset of TNC point counts and eBird checklists.

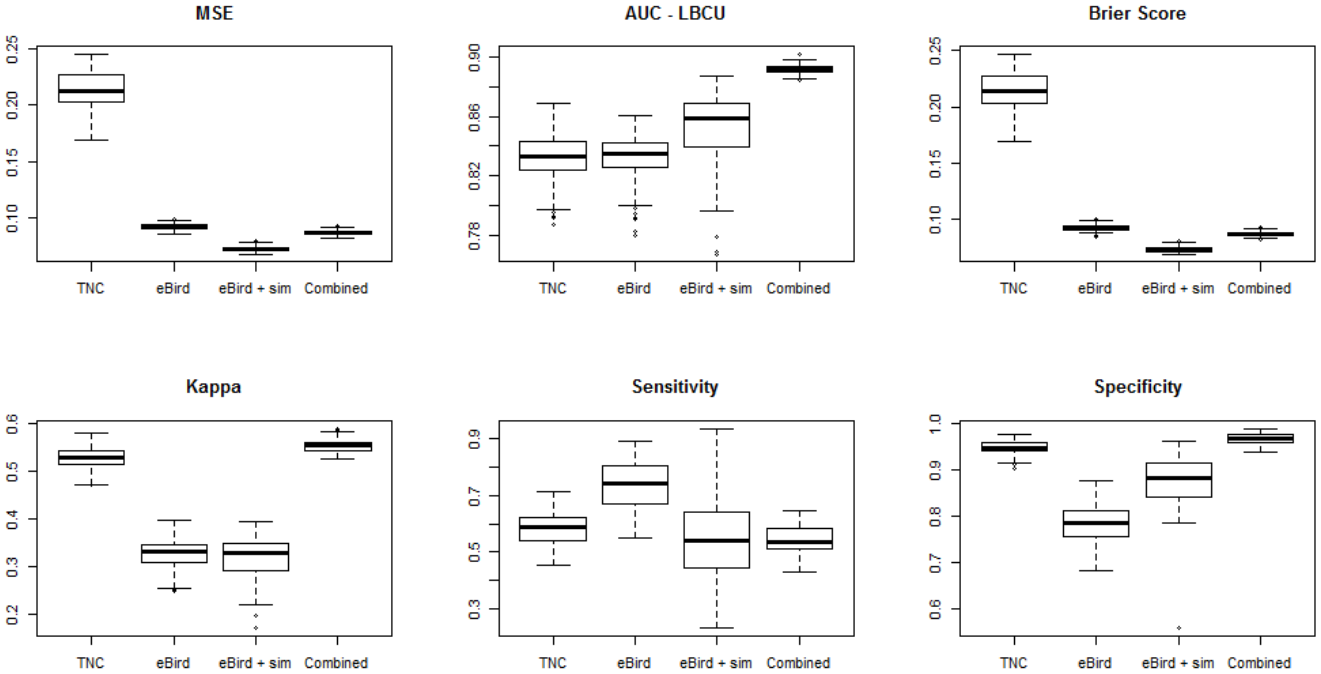

A
duration_minutes
observation_date
pland_999_water_point_blue
pland_003_rice
pland_061_fallow_idle_cropland
pland_121_developed_open_space
pland_122_developed_low_intensity
pland_054_tomatoes
pland_076_walnuts
year
pland_123_developed_med_intensity
cci
pland_036_alfalfa
pland_075_almonds
pland_006_sunflower
pland_195_herbaceous_wetlands
pland_001_corn
0
5
10
15
20
Variable Importance (Gini Index)
B
pland_999_water_point_blue
pland_061_fallow_idle_cropland
pland_003_rice
duration_minutes
pland_121_developed_open_space
pland_122_developed_low_intensity
cci
pland_123_developed_med_intensity
observation_date
pland_195_herbaceous_wetlands
pland_001_corn
pland_000_no_data
pland_111_open_water
0
10
20
30
Variable Importance (Gini Index)
C
pland_003_rice
pland_999_water_point_blue
pland_121_developed_open_space
duration_minutes
pland_061_fallow_idle_cropland
pland_122_developed_low_intensity
pland_123_developed_med_intensity
cci
pland_000_no_data
observation_date
pland_075_almonds
pland_054_tomatoes
pland_124_developed_high_intensity
pland_076_walnuts
pland_036_alfalfa
year
number_observers
pland_195_herbaceous_wetlands
pland_111_open_water
pland_131_barren
pland_024_winter_wheat
pland_069_grapes
0
25
50
75
Variable Importance (Gini Index)

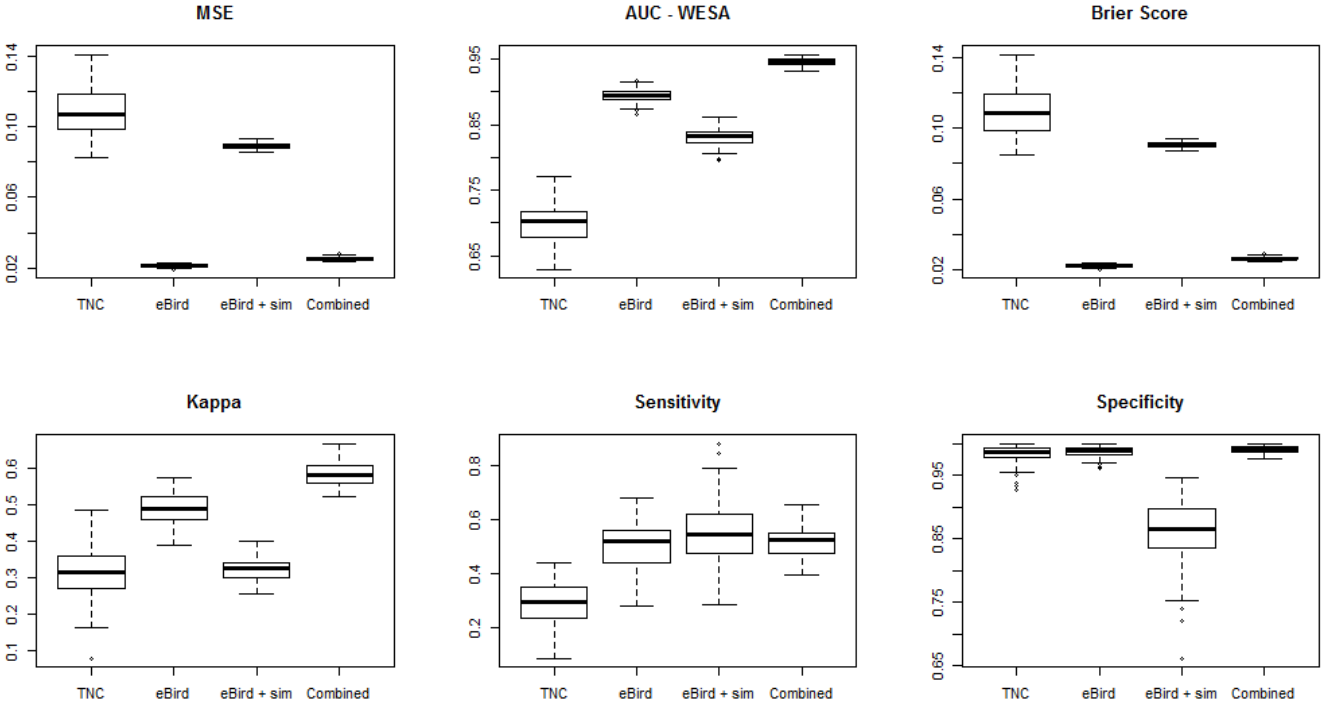
Supplementary Information 1

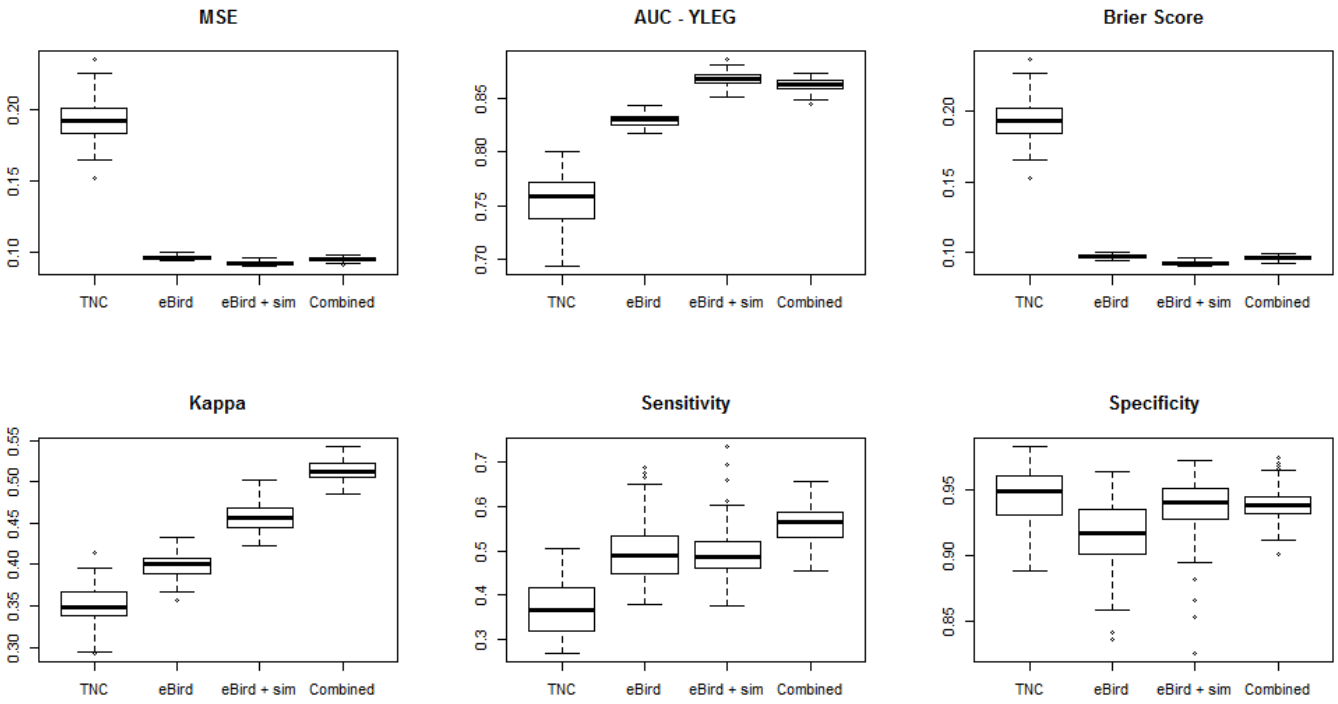
Boxplots for MSE, AUC, Brier score, Cohen's Kappa, Sensitivity and Specificity for models run for American avocet (AMAV), long- and short-billed dowitcher combined (DOWI), dunlin (DUNL), long-billed curlew (LBCU), western sandpiper (WESA) and greater and lesser yellowlegs combined (YLEG) with each dataset; TNC point counts (TNC), eBird checklists (eBird), eBird checklists with additional simulated eBird data (eBird + sim), and the dataset of TNC point counts and eBird checklists combined (Combined). Accuracy metrics for least sandpiper (LESA) are in the main text.




MSE
AUC - DOWI
Brier Score
Kappa
Sensitivity
Specificity
MSE
AUC - DUNL
Brier Score
Kappa
Sensitivity
Specificity
TNC
eBird
eBird + sim
Combined

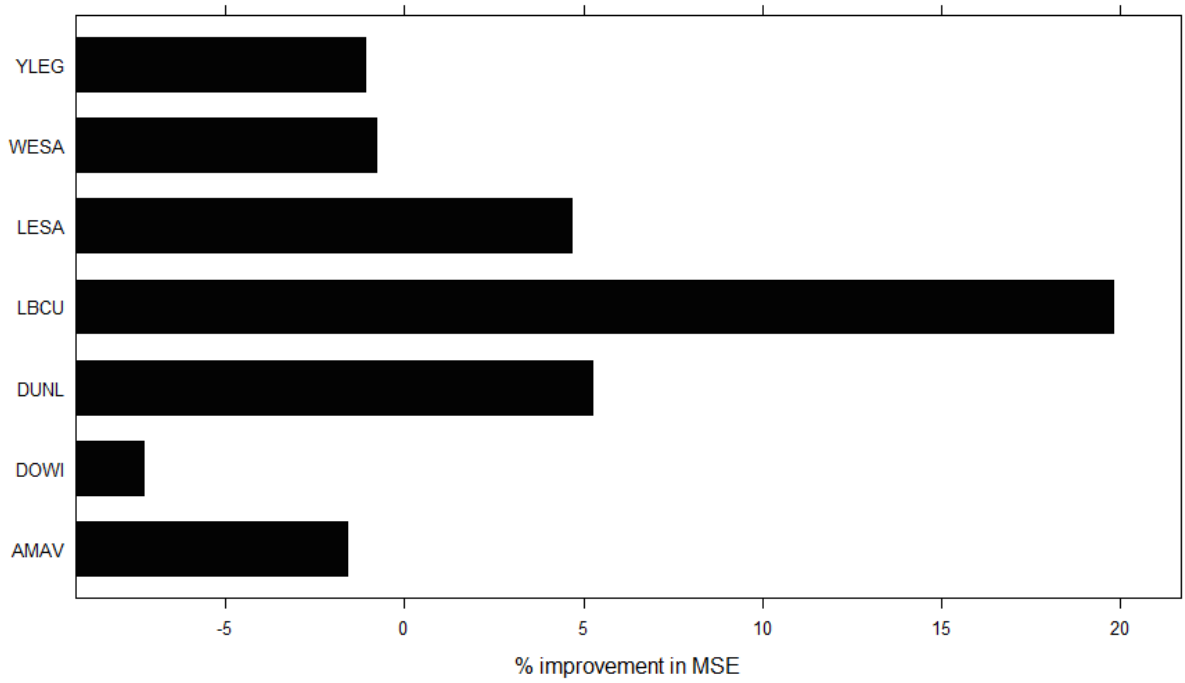

MSE
AUC - LBCU
Brier Score
Kappa
Sensitivity
Specificity
TNC
eBird
eBird + sim
Combined

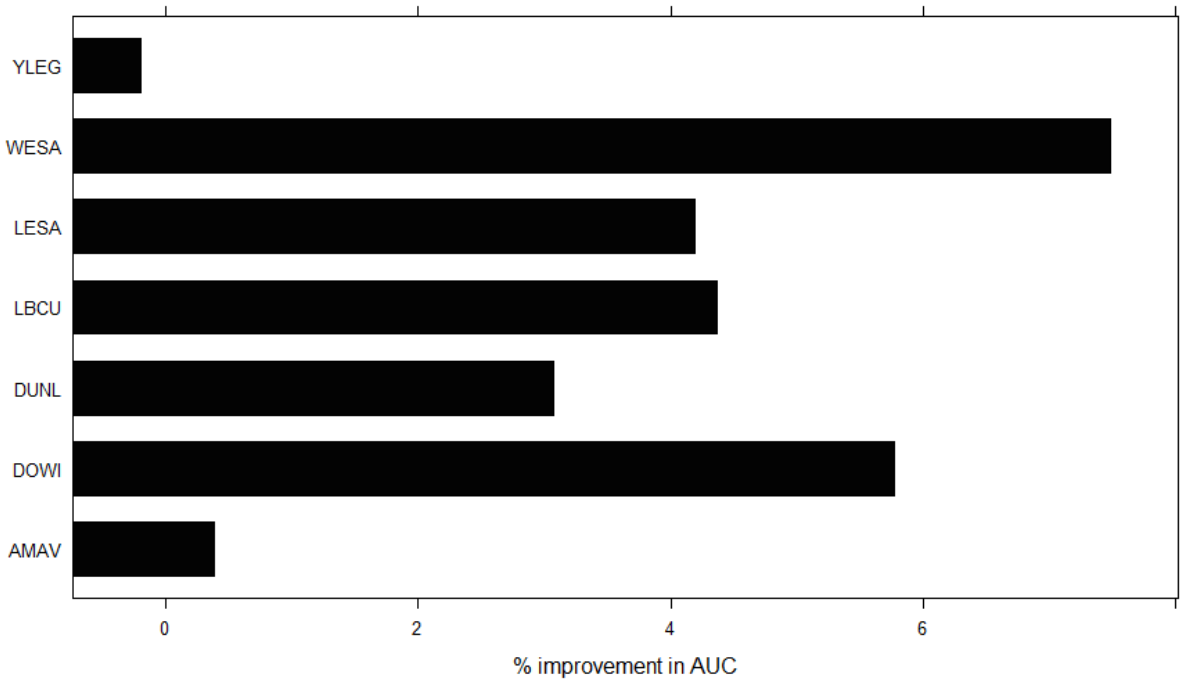

MSE
AUC - WESA
Brier Score
Kappa
Sensitivity
Specificity
TNC
eBird
eBird + sim
Combined


MSE
0.10
0.15
0.20
TNC
eBird
eBird + sim
Combined
AUC - YLEG
0.70
0.75
0.80
0.85
TNC
eBird
eBird + sim
Combined
Brier Score
0.10
0.15
0.20
TNC
eBird
eBird + sim
Combined
Kappa
0.30
0.35
0.40
0.45
0.50
0.55
TNC
eBird
eBird + sim
Combined
Sensitivity
0.3
0.4
0.5
0.6
0.7
TNC
eBird
eBird + sim
Combined
Specificity
0.85
0.90
0.95
TNC
eBird
eBird + sim
Combined

Supplementary Information 2

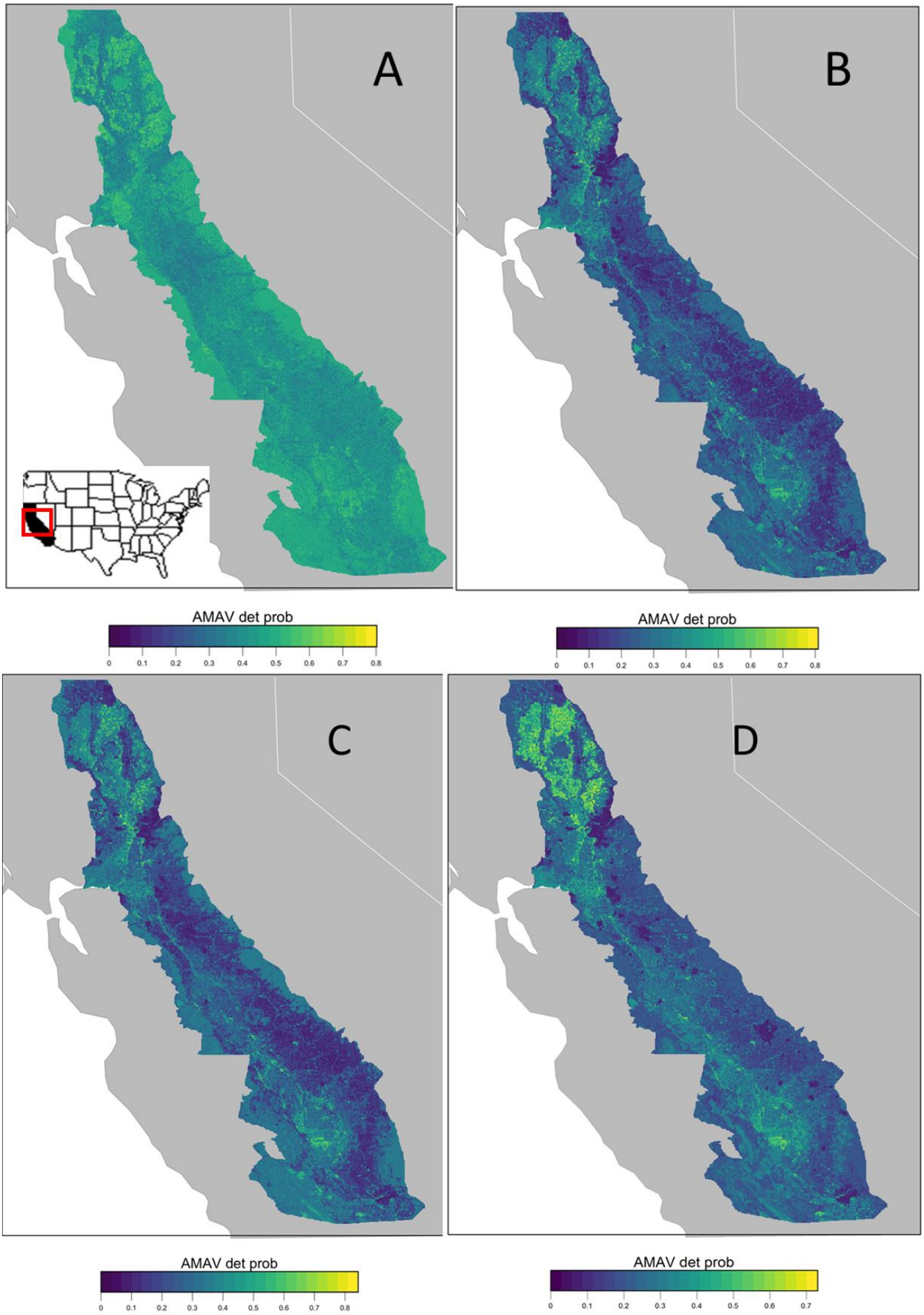
Average improvement or loss in mean squared error and AUC for distribution models of each species when using the combined dataset versus the dataset that provided the next lowest value. Recall that the values for these metrics were already very low (for MSE) and very high (for AUC) to begin with. So even a somewhat large change in the percent improvement or loss represents a rather small change for most species. The plots showing improvement or loss for Brier score and Kappa are in the main text.

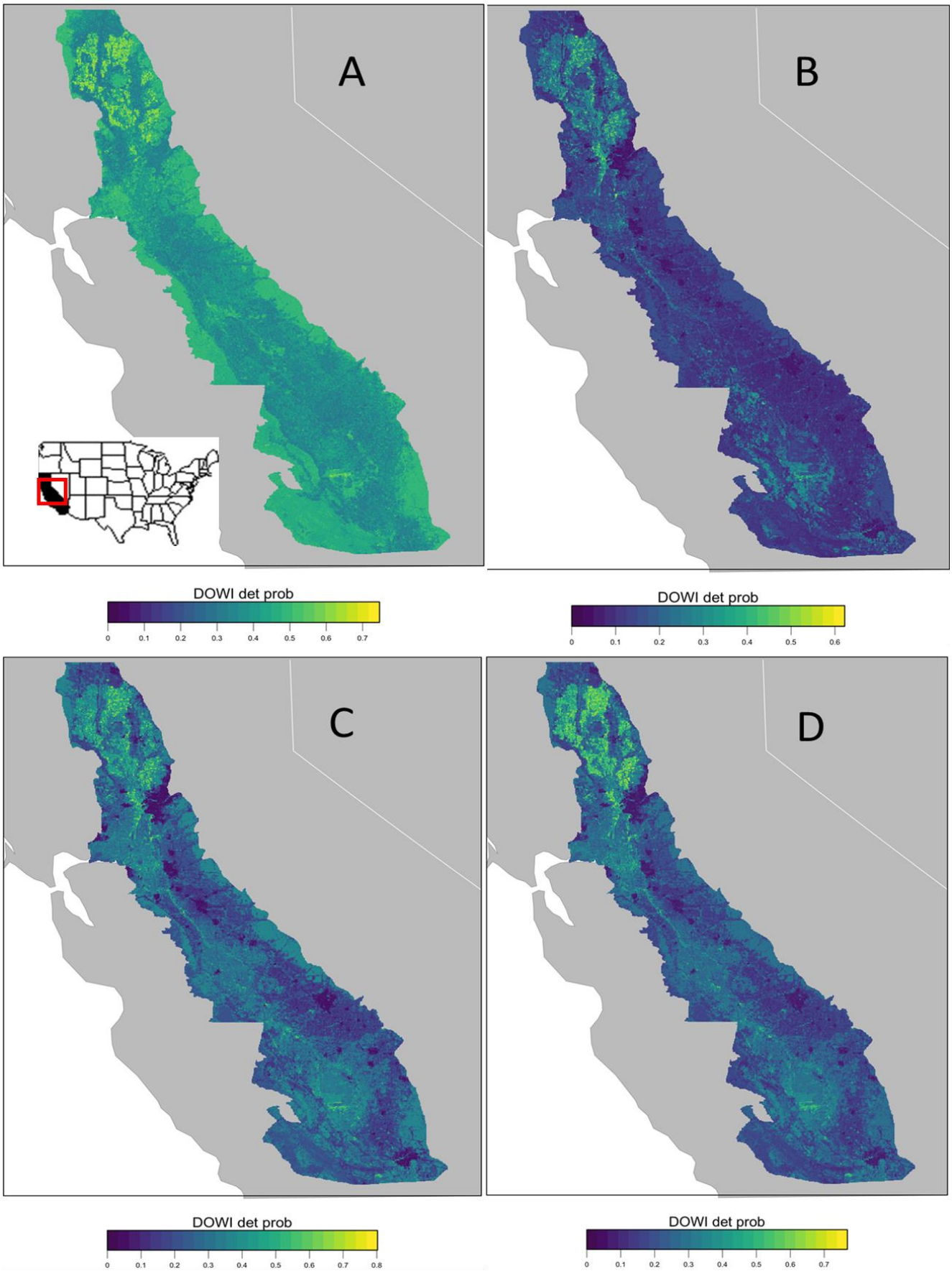


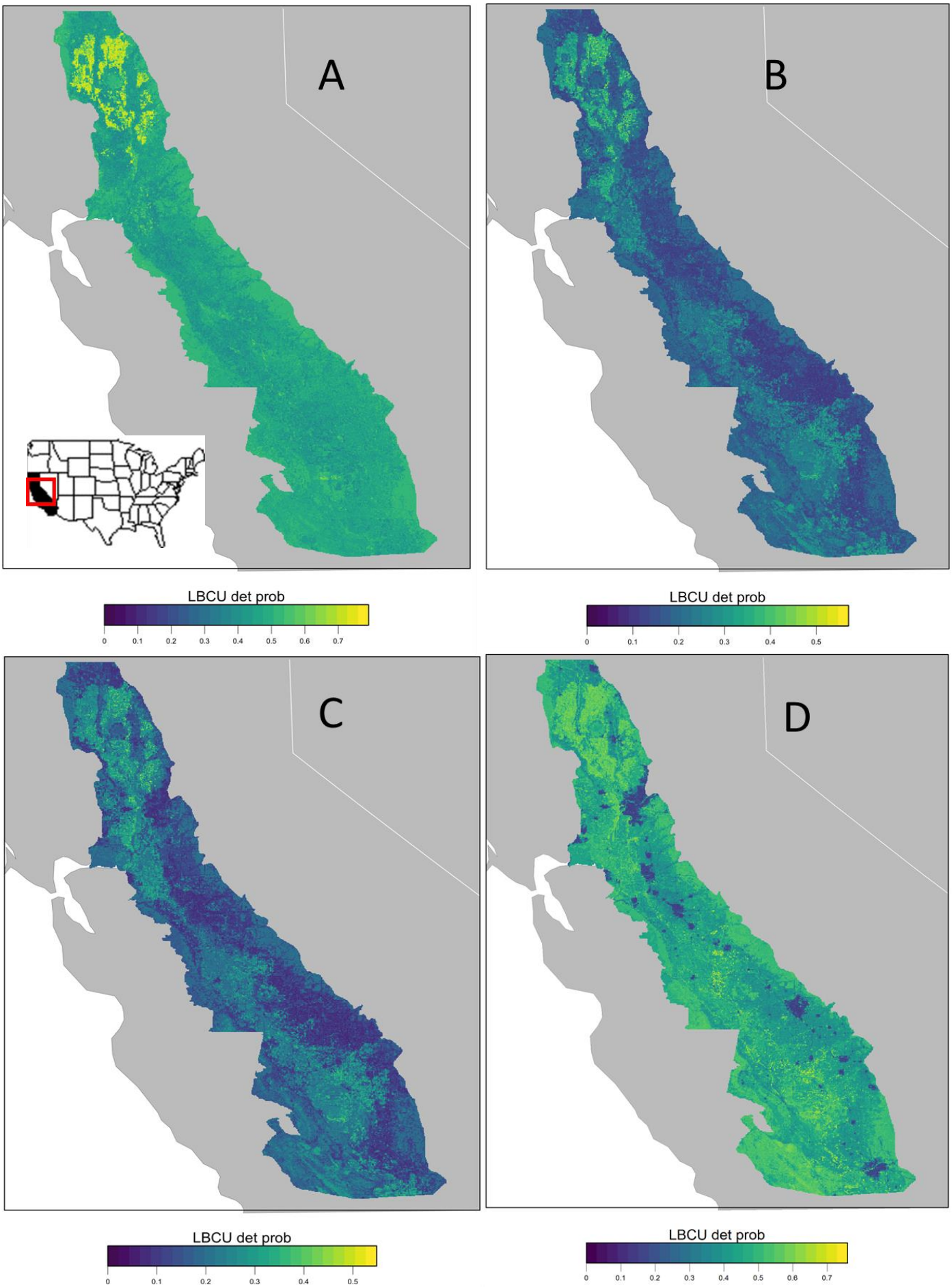

YLEG
WESA
LESA
LBCU
DUNL
DOWI
AMAV
0
2
4
6
% improvement in AUC

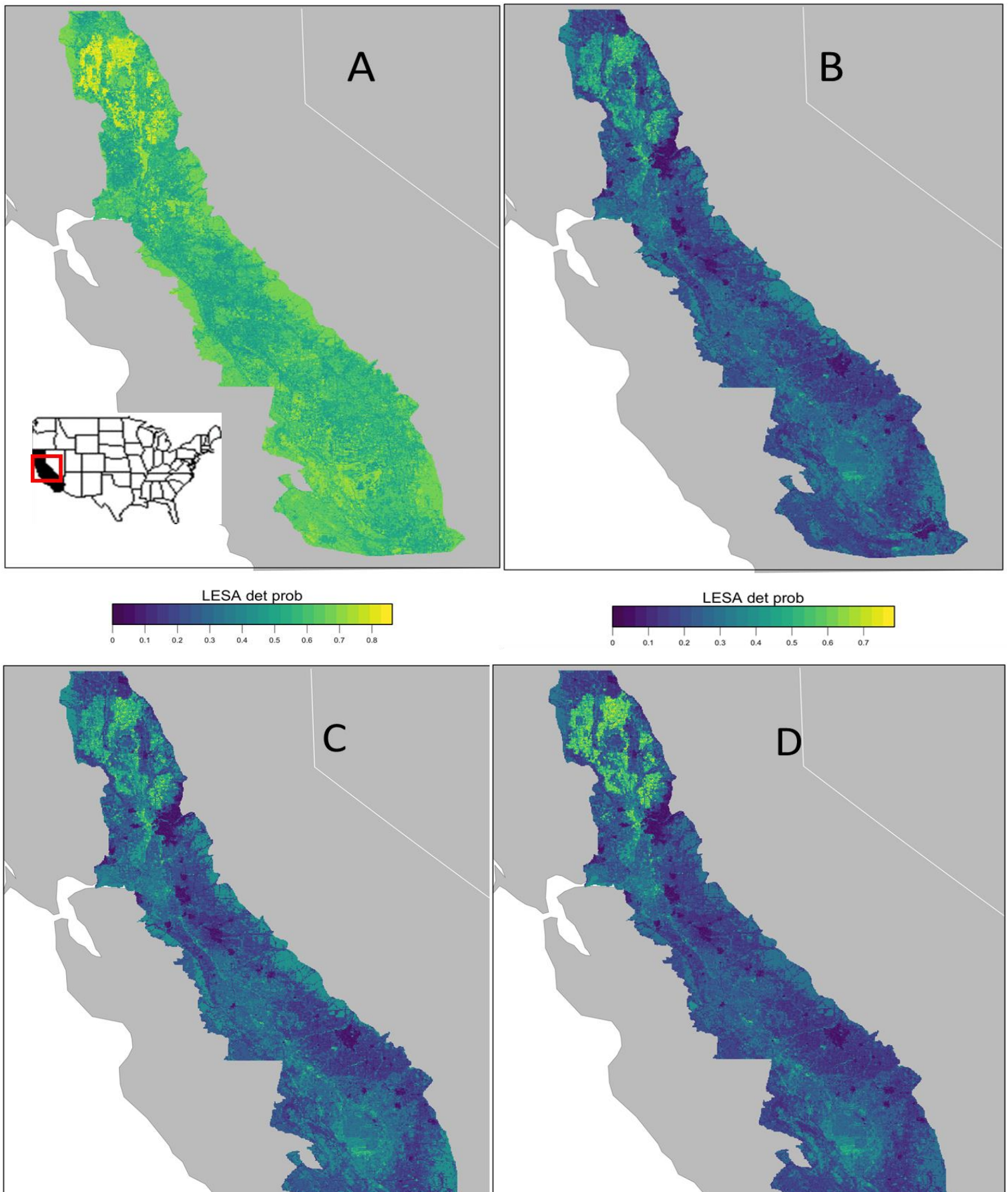
Supplementary Information 3

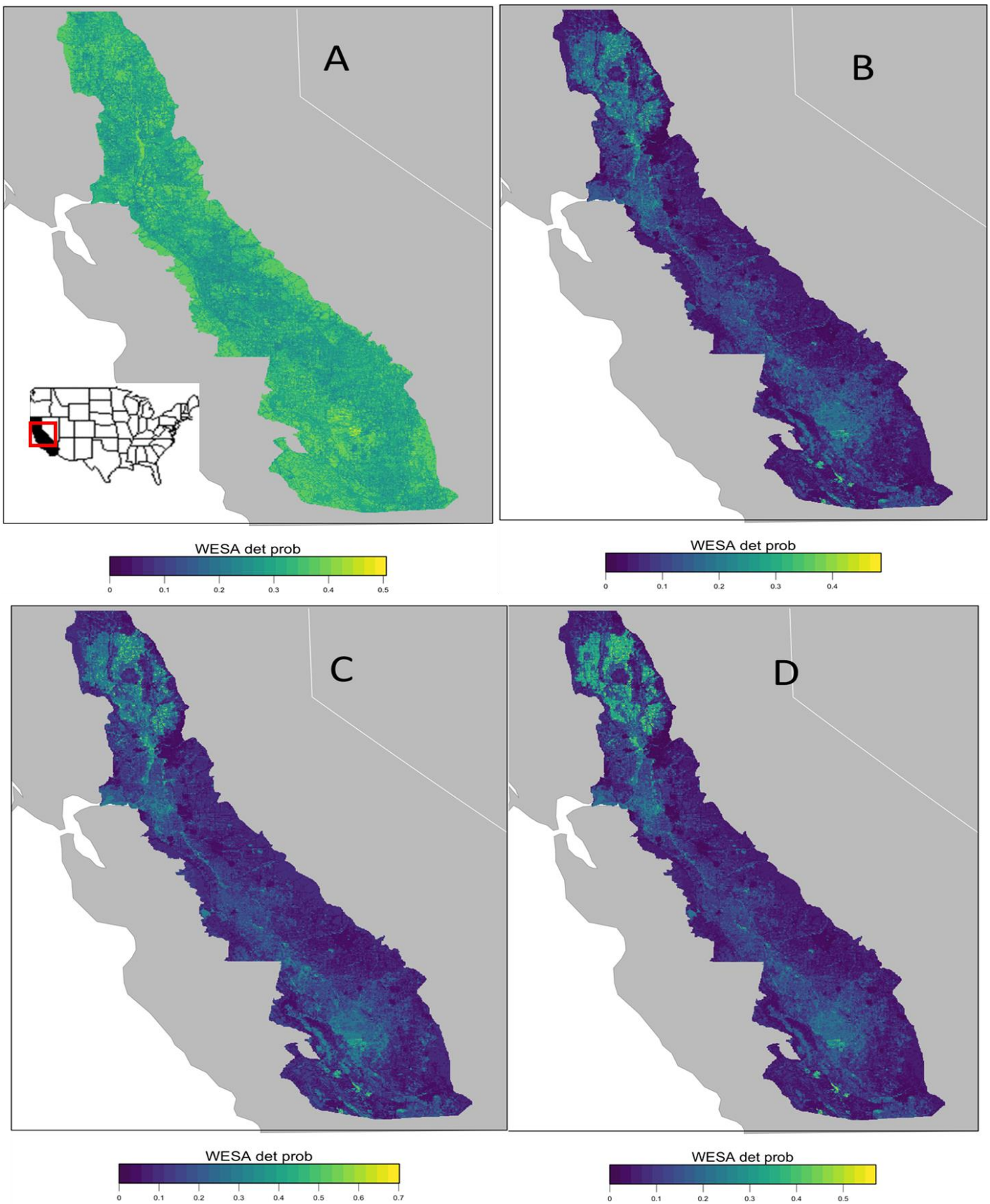
Predicted probability of an expert surveyor detecting an American avocet (AMAV), long- and short-billed dowitcher combined (DOWI), dunlin (DUNL), long-billed curlew (LBCU), least sandpiper (LESA), western sandpiper (WESA) and greater and lesser yellowlegs combined (YLEG) in the spring of 2016 on a one-hour long checklist where the distance traveled was < 300m., as estimated by the model using TNC point counts alone (A), eBird checklists alone (B), eBird checklists with added simulated eBird checklists (C), and the combined dataset of TNC point counts and eBird checklists (D). Maps for dunlin (DUNL) are in the main text.

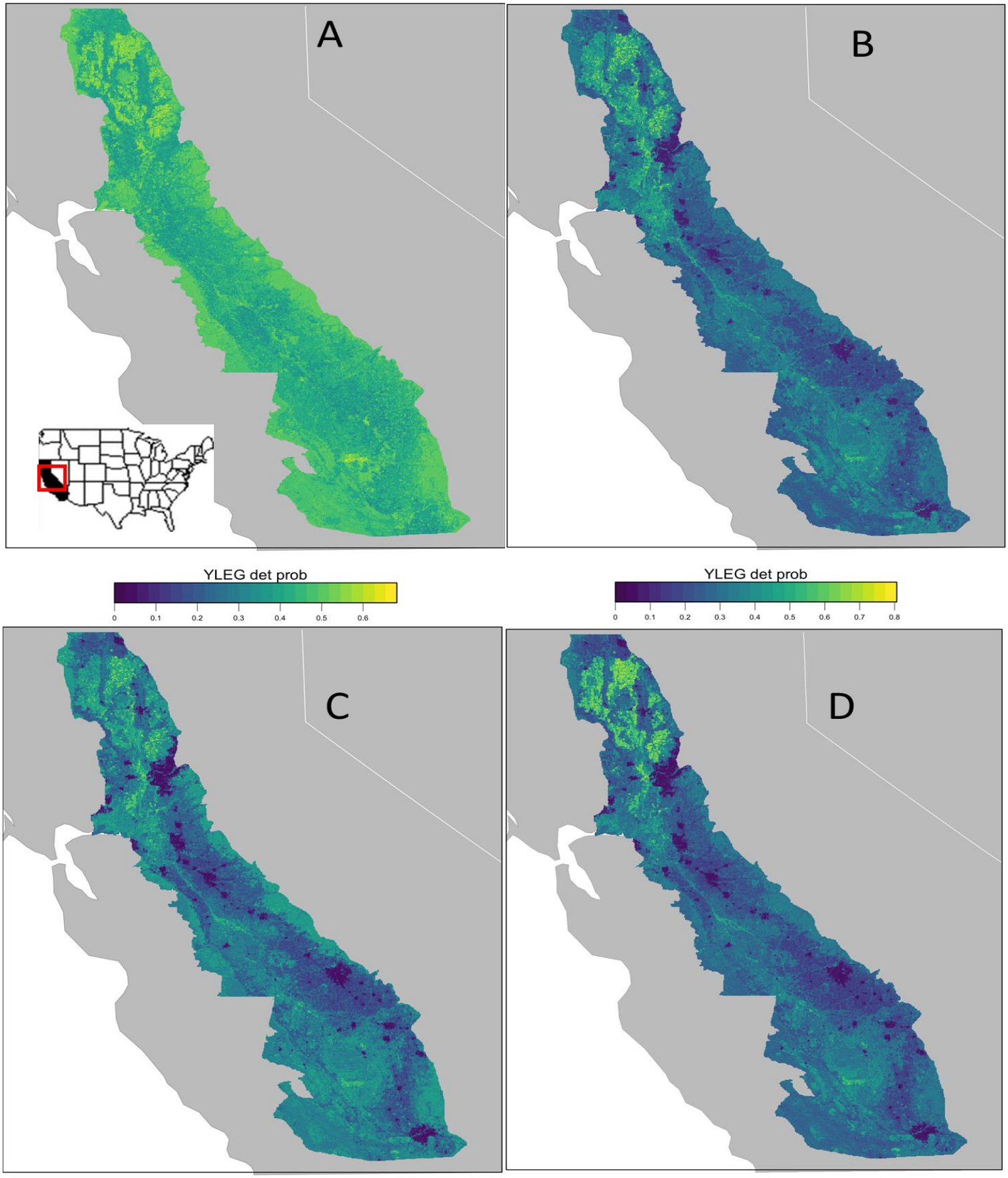

A
B
C
D
AMAV det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
AMAV det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
AMAV det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
AMAV det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


A
B
C
D
DOWI det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
DOWI det prob
0 0.1 0.2 0.3 0.4 0.5 0.6
DOWI det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
DOWI det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


A
B
C
D
LBCU det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
LBCU det prob
0 0.1 0.2 0.3 0.4 0.5
LBCU det prob
0 0.1 0.2 0.3 0.4 0.5
LBCU det prob
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7


A
B
C
D
LESA det prob
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
LESA det prob
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


A
B
C
D
WESA det prob
0
0.1
0.2
0.3
0.4
0.5
WESA det prob
0
0.1
0.2
0.3
0.4
WESA det prob
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
WESA det prob
0
0.1
0.2
0.3
0.4
0.5

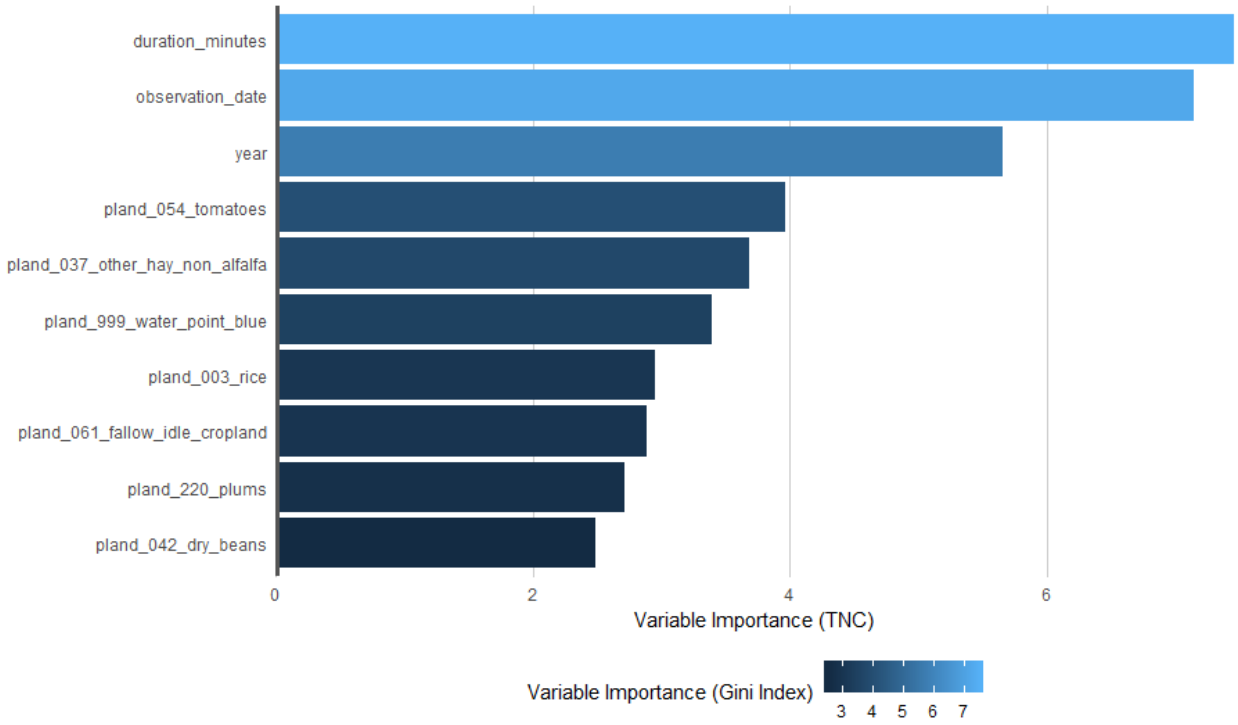

A
B
C
D
YLEG det prob
0
0.1
0.2
0.3
0.4
0.5
0.6
YLEG det prob
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Supplementary Information 4

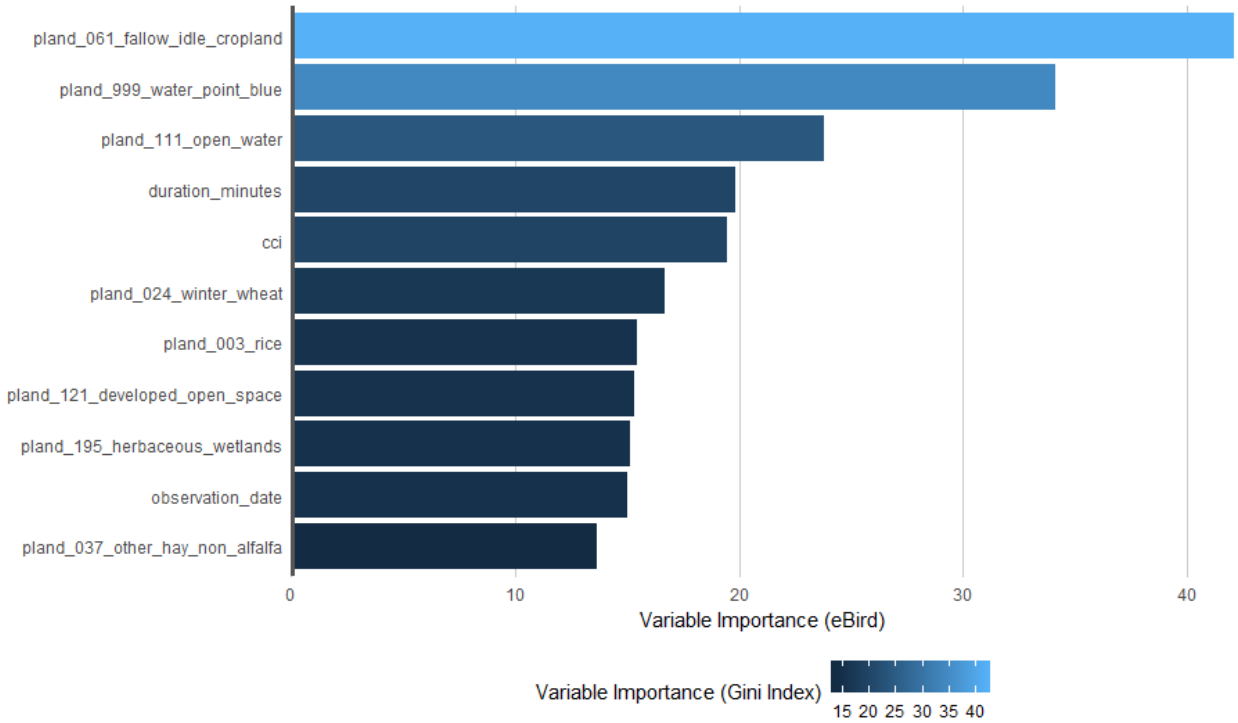
Variable Importance for American Avocet, Dowithcer sp., Long-billed Curlew, Least Sandpiper, Western Sandpiper, and Yellowlegs sp. These figures for Dunlin appear in the main text.

Important variables selected by the model trained on each data set for American Avocet as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

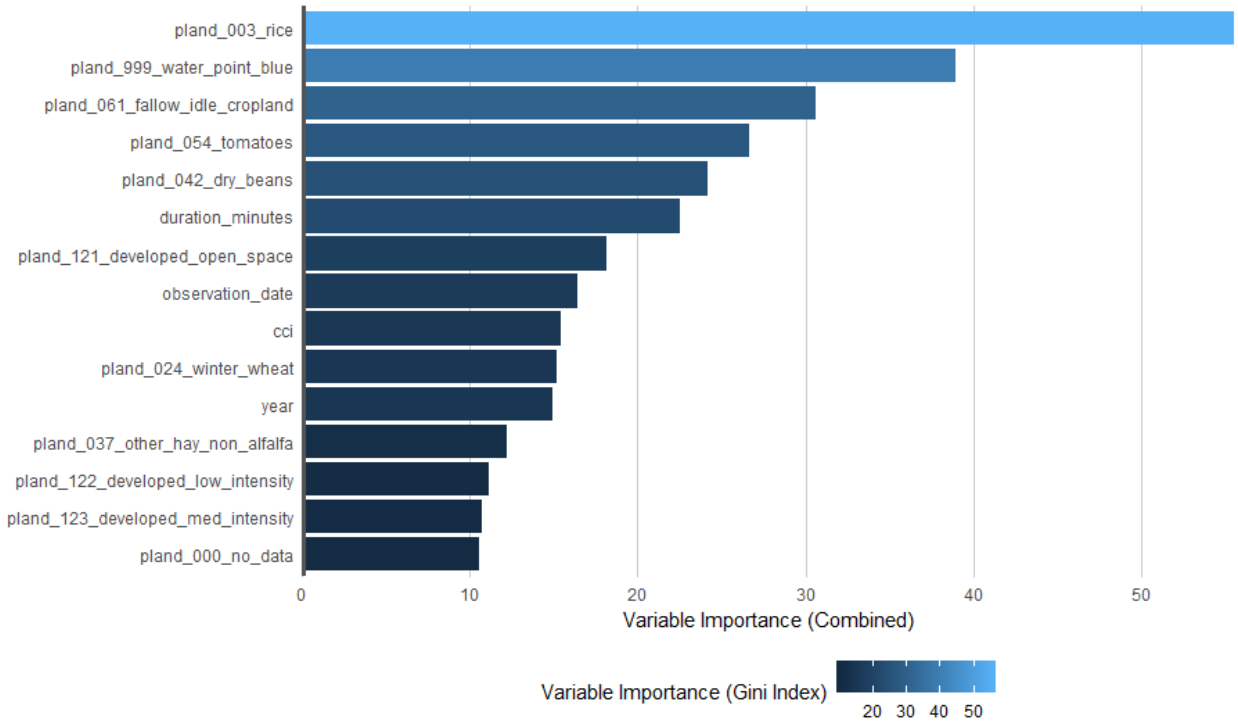
A



B

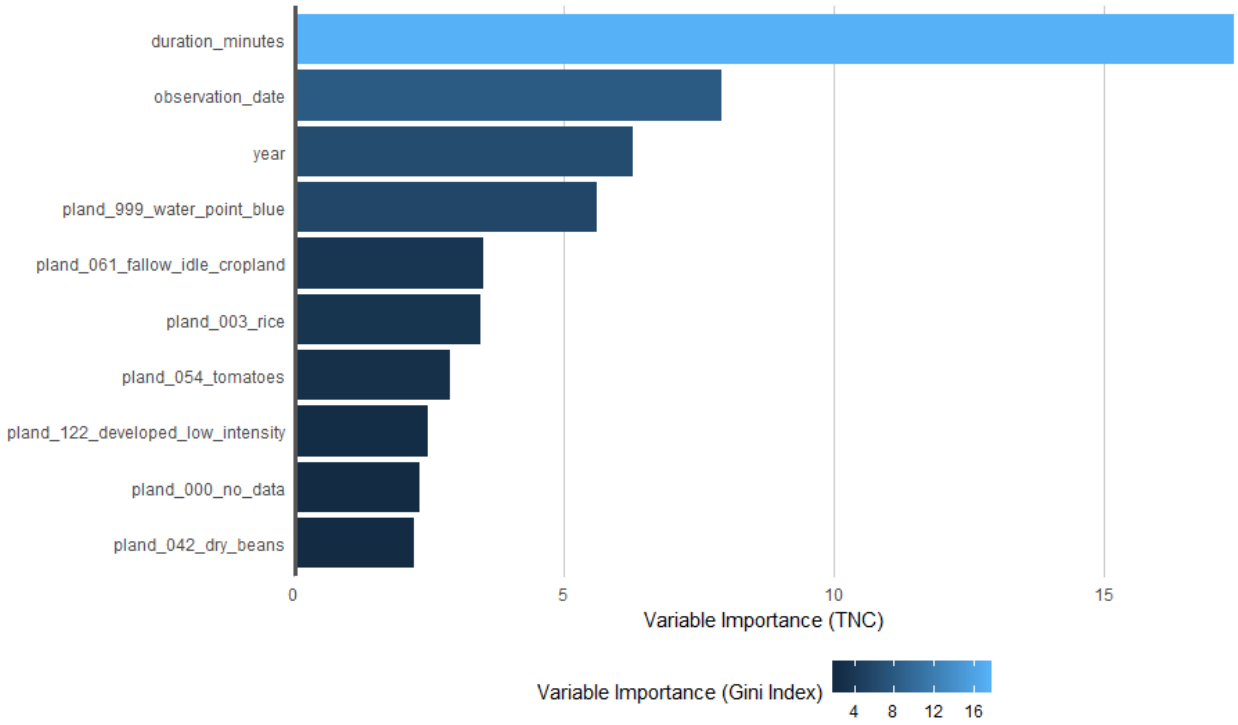

pland_061_fallow_idle_cropland
pland_999_water_point_blue
pland_111_open_water
duration_minutes
cci
pland_024_winter_wheat
pland_003_rice
pland_121_developed_open_space
pland_195_herbaceous_wetlands
observation_date
pland_037_other_hay_non_alfalfa
0
10
20
30
40
Variable Importance (eBird)
Variable Importance (Gini Index)
15 20 25 30 35 40

C

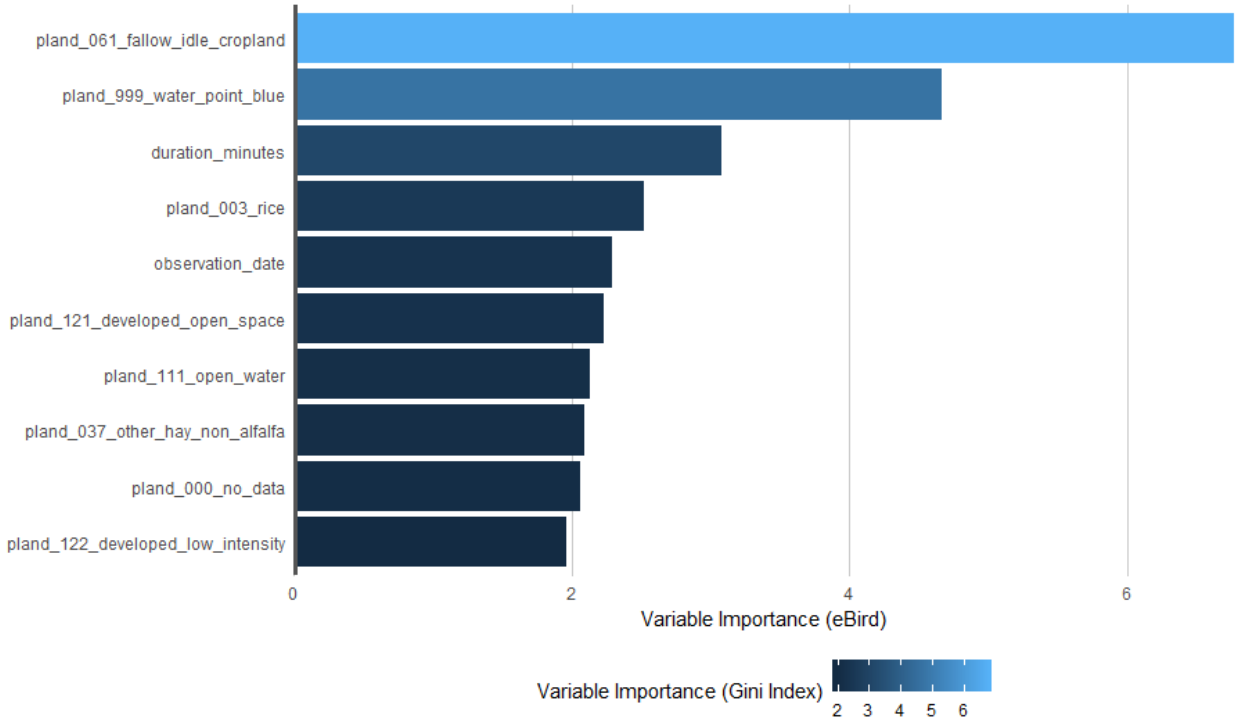

pland_003_rice
pland_999_water_point_blue
pland_061_fallow_idle_cropland
pland_054_tomatoes
pland_042_dry_beans
duration_minutes
pland_121_developed_open_space
observation_date
cci
pland_024_winter_wheat
year
pland_037_other_hay_non_alfalfa
pland_122_developed_low_intensity
pland_123_developed_med_intensity
pland_000_no_data
0
10
20
30
40
50
Variable Importance (Combined)
Variable Importance (Gini Index)
20
30
40
50

Important variables selected by the model trained on each data set for short- and long-billed dowitcher combined as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

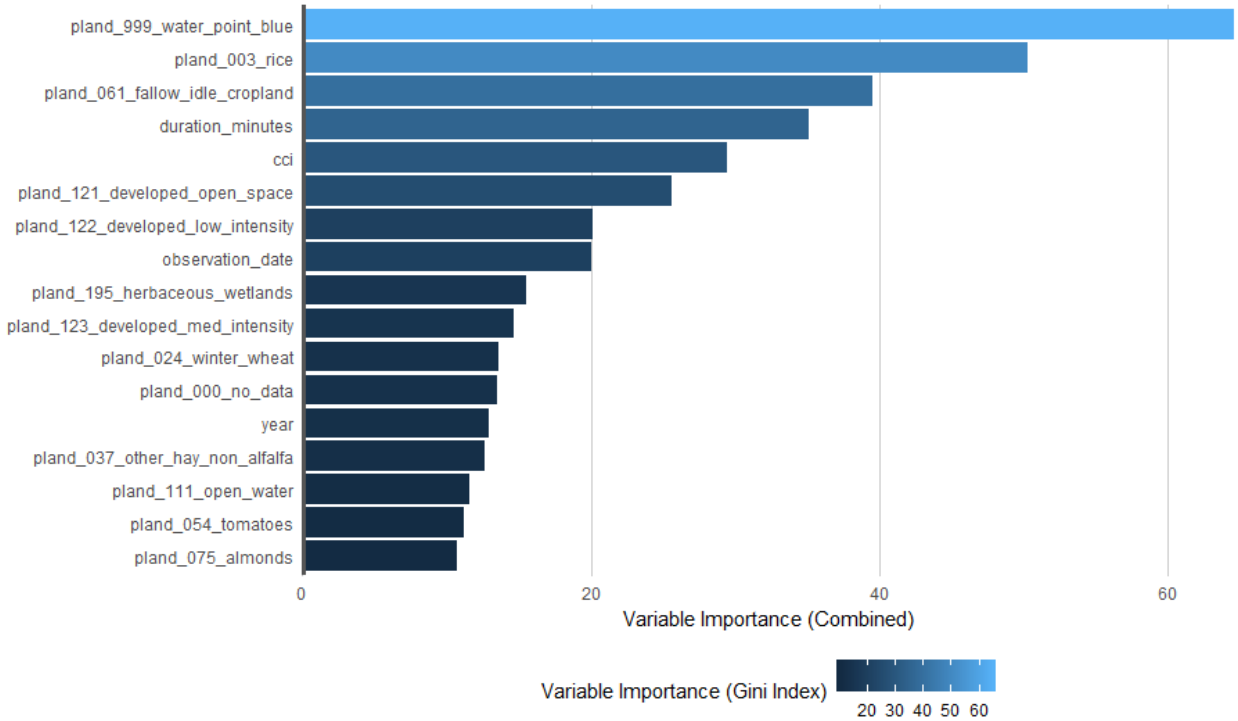
A



B

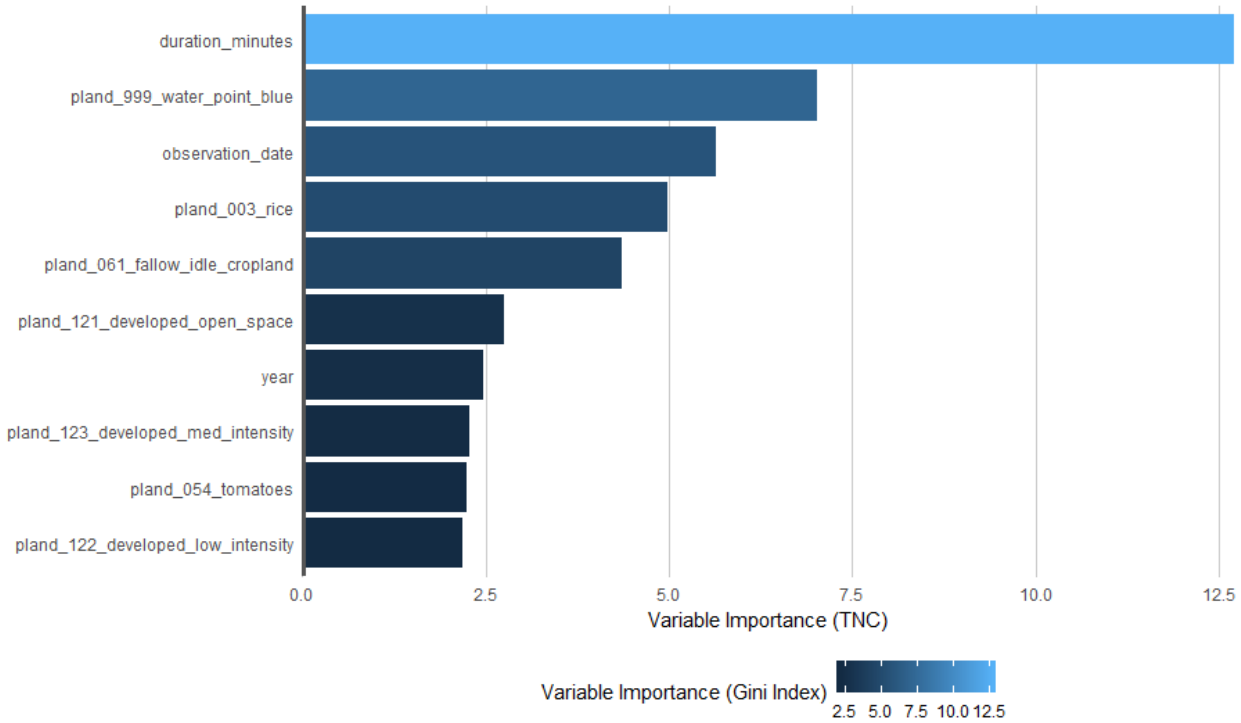

pland_061_fallow_idle_cropland
pland_999_water_point_blue
duration_minutes
pland_003_rice
observation_date
pland_121_developed_open_space
pland_111_open_water
pland_037_other_hay_non_alfalfa
pland_000_no_data
pland_122_developed_low_intensity
0
2
4
6
Variable Importance (eBird)
Variable Importance (Gini Index)
2 3 4 5 6

C

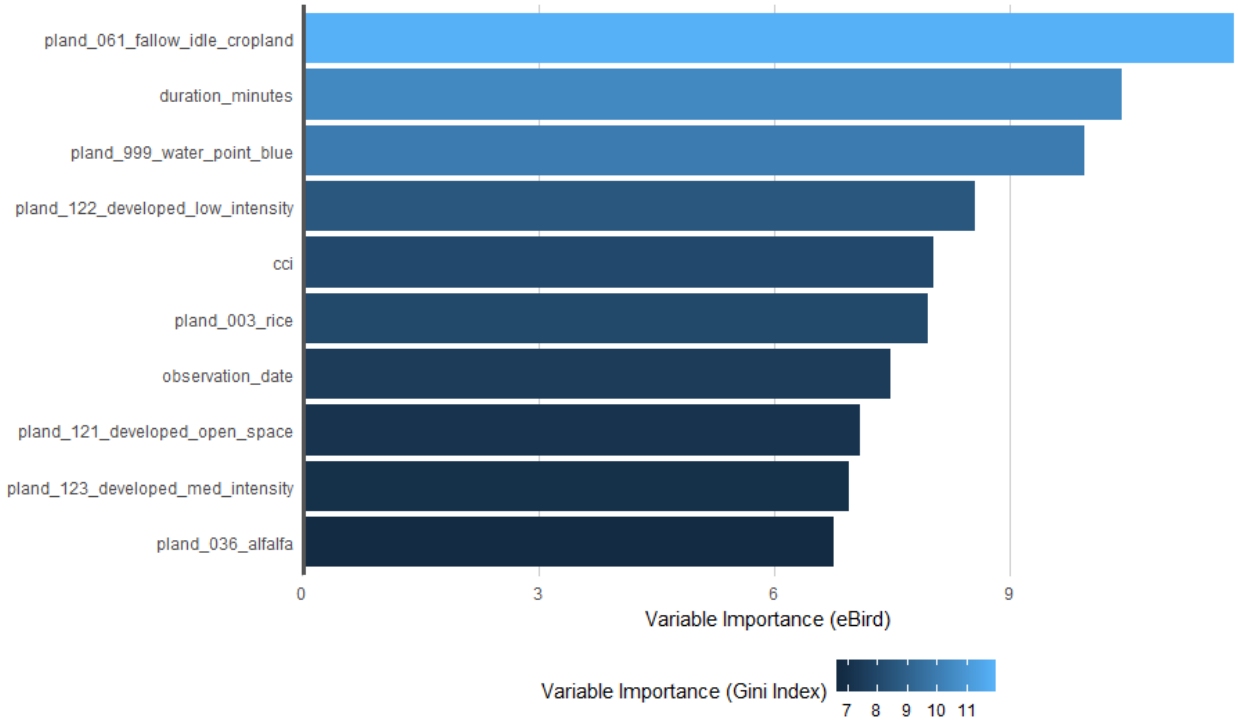

pland_999_water_point_blue
pland_003_rice
pland_061_fallow_idle_cropland
duration_minutes
cci
pland_121_developed_open_space
pland_122_developed_low_intensity
observation_date
pland_195_herbaceous_wetlands
pland_123_developed_med_intensity
pland_024_winter_wheat
pland_000_no_data
year
pland_037_other_hay_non_alfalfa
pland_111_open_water
pland_054_tomatoes
pland_075_almonds
0
20
40
60
Variable Importance (Combined)
Variable Importance (Gini Index)
20 30 40 50 60

Important variables selected by the model trained on each data set for long-billed curlew as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

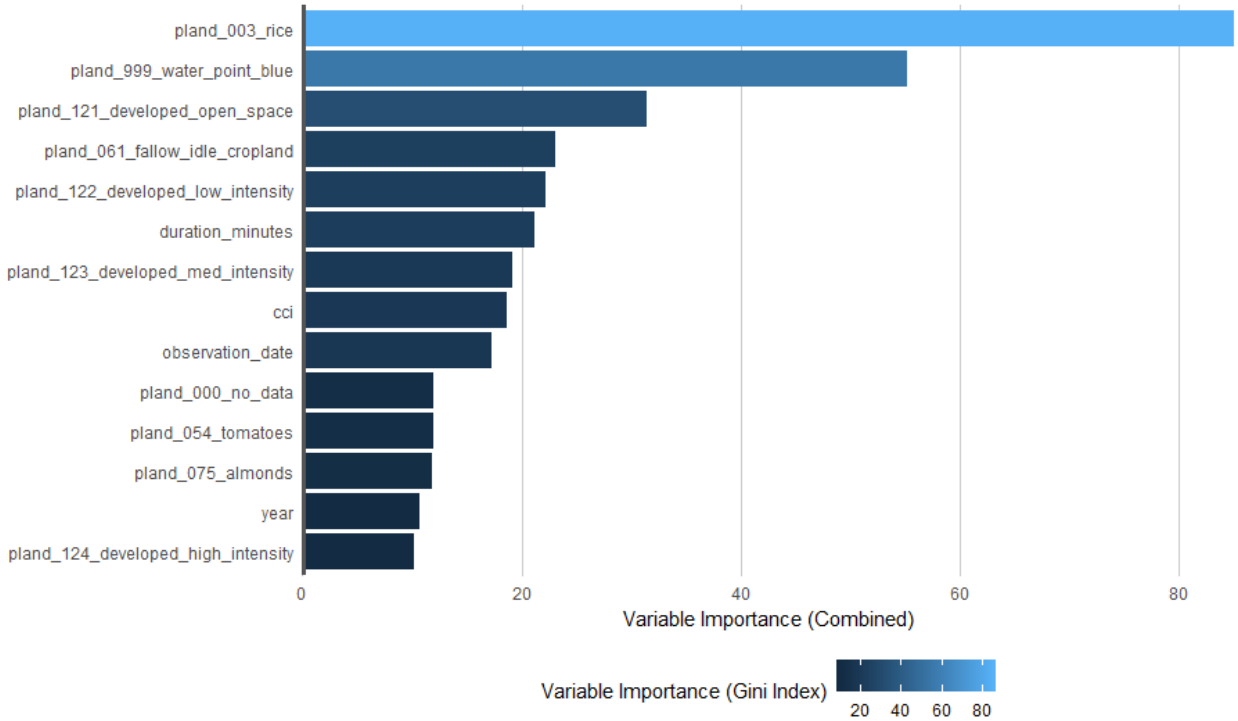
A



B

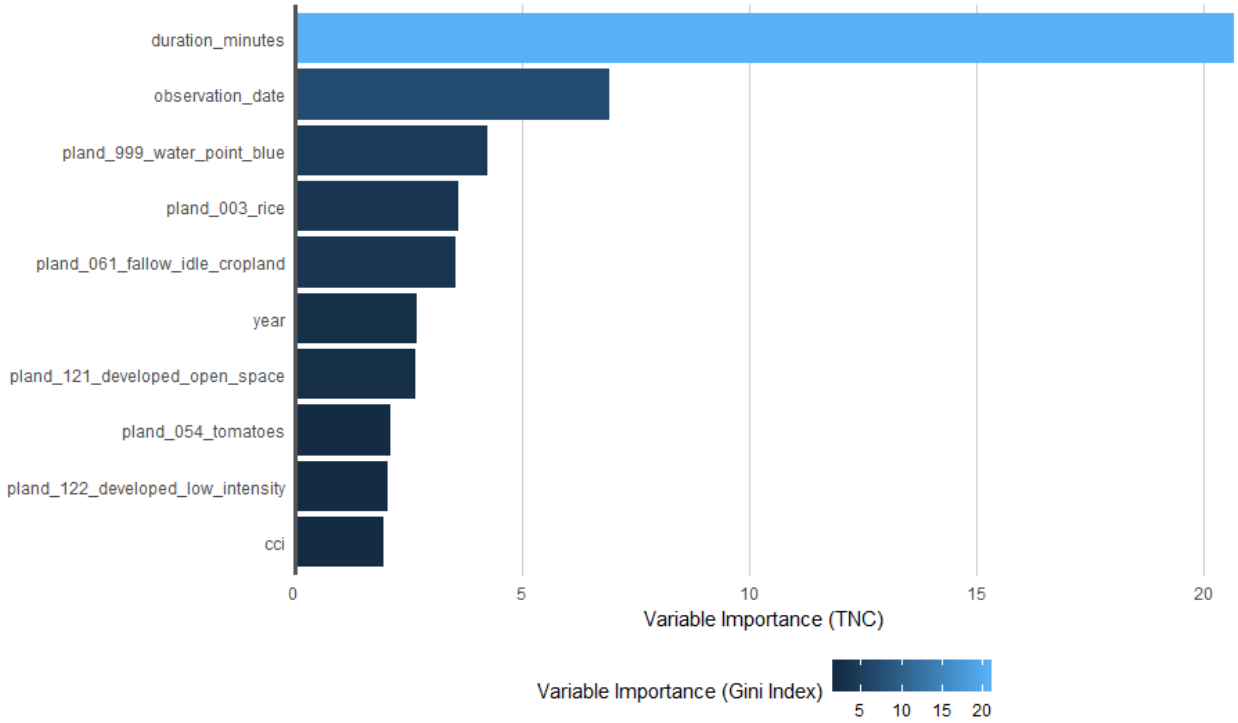

pland_061_fallow_idle_cropland
duration_minutes
pland_999_water_point_blue
pland_122_developed_low_intensity
cci
pland_003_rice
observation_date
pland_121_developed_open_space
pland_123_developed_med_intensity
pland_036_alfalfa
0
3
6
9
Variable Importance (eBird)
Variable Importance (Gini Index)
7 8 9 10 11

C

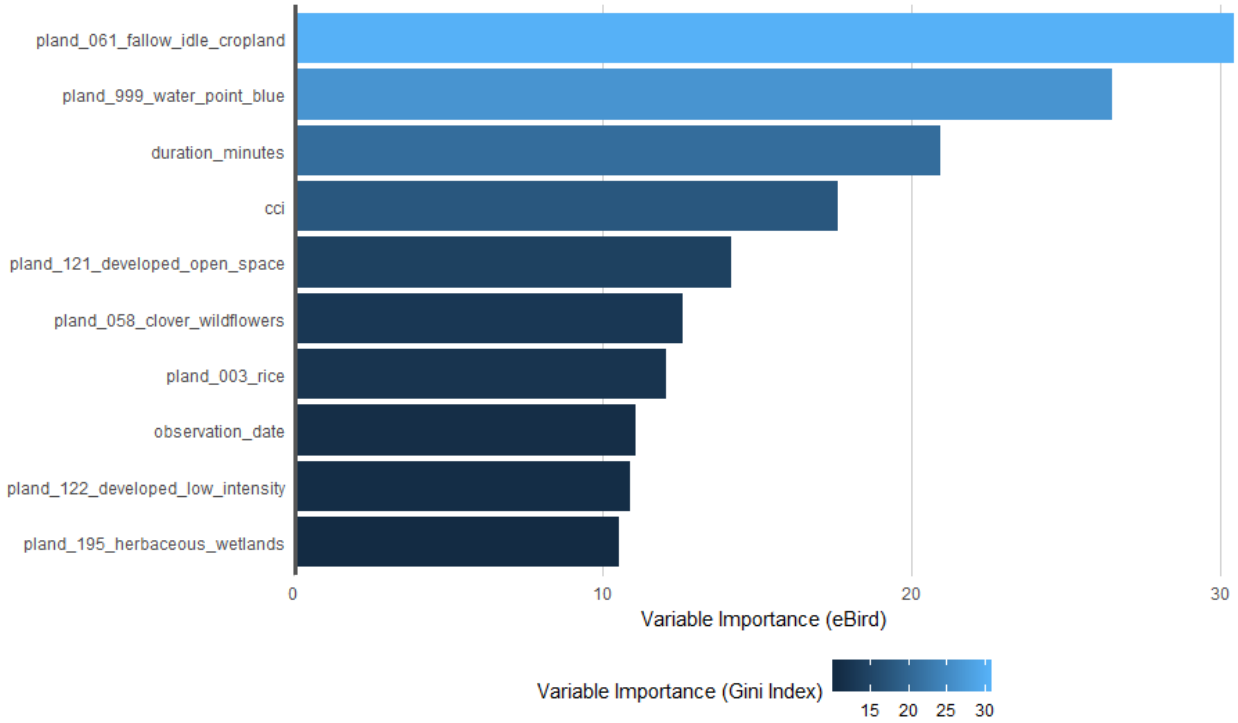

pland_003_rice
pland_999_water_point_blue
pland_121_developed_open_space
pland_061_fallow_idle_cropland
pland_122_developed_low_intensity
duration_minutes
pland_123_developed_med_intensity
cci
observation_date
pland_000_no_data
pland_054_tomatoes
pland_075_almonds
year
pland_124_developed_high_intensity
0
20
40
60
80
Variable Importance (Combined)
Variable Importance (Gini Index)
20 40 60 80

Important variables selected by the model trained on each data set for least sandpiper as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

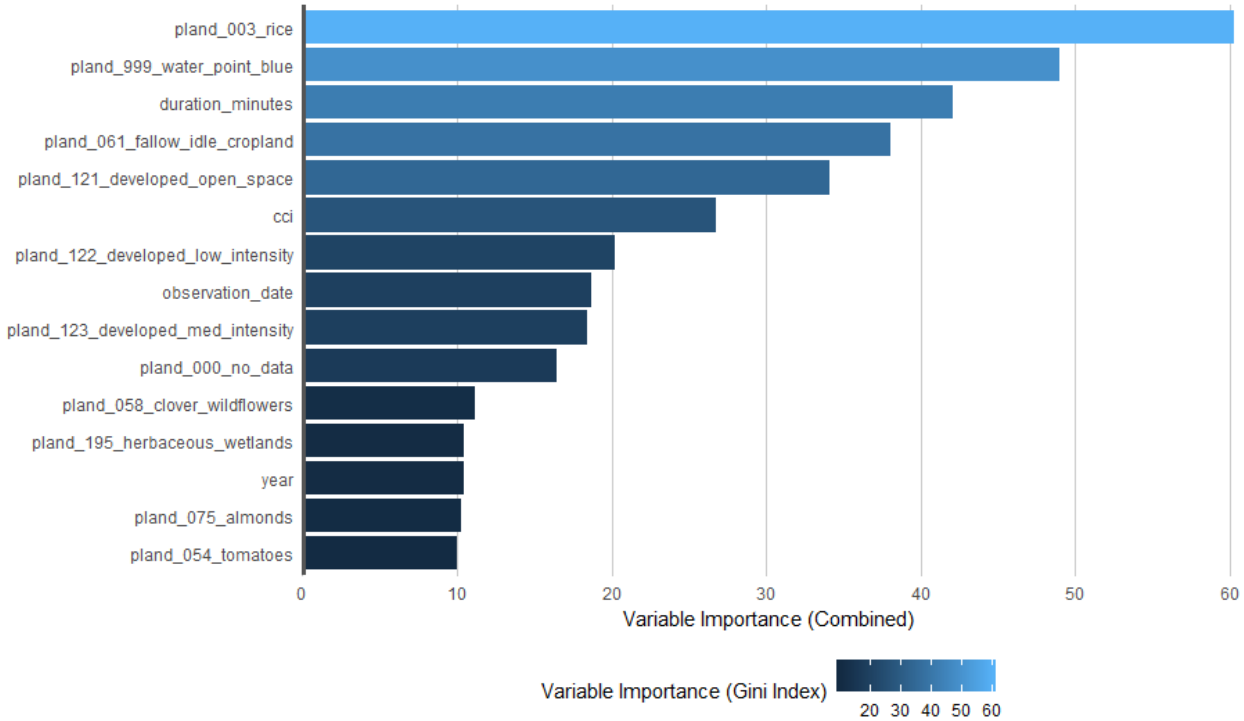
A



B

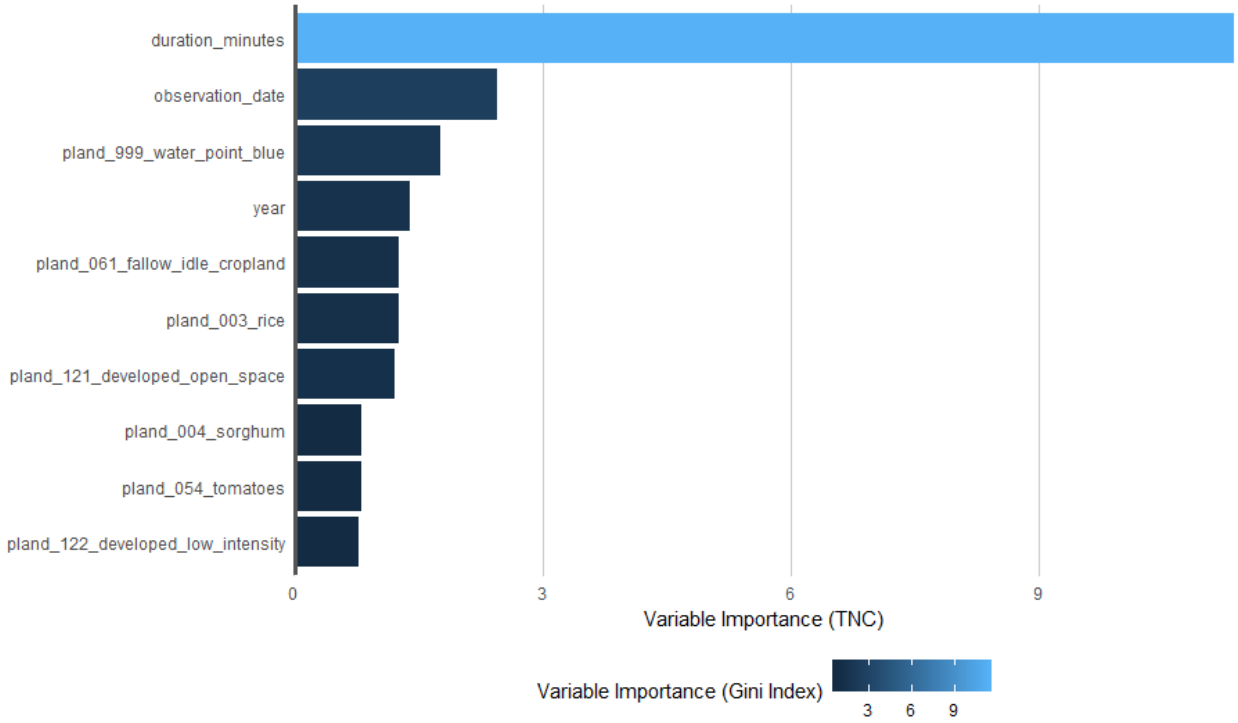

pland_061_fallow_idle_cropland
pland_999_water_point_blue
duration_minutes
cci
pland_121_developed_open_space
pland_058_clover_wildflowers
pland_003_rice
observation_date
pland_122_developed_low_intensity
pland_195_herbaceous_wetlands
0
10
20
30
Variable Importance (eBird)
Variable Importance (Gini Index)
15 20 25 30

C

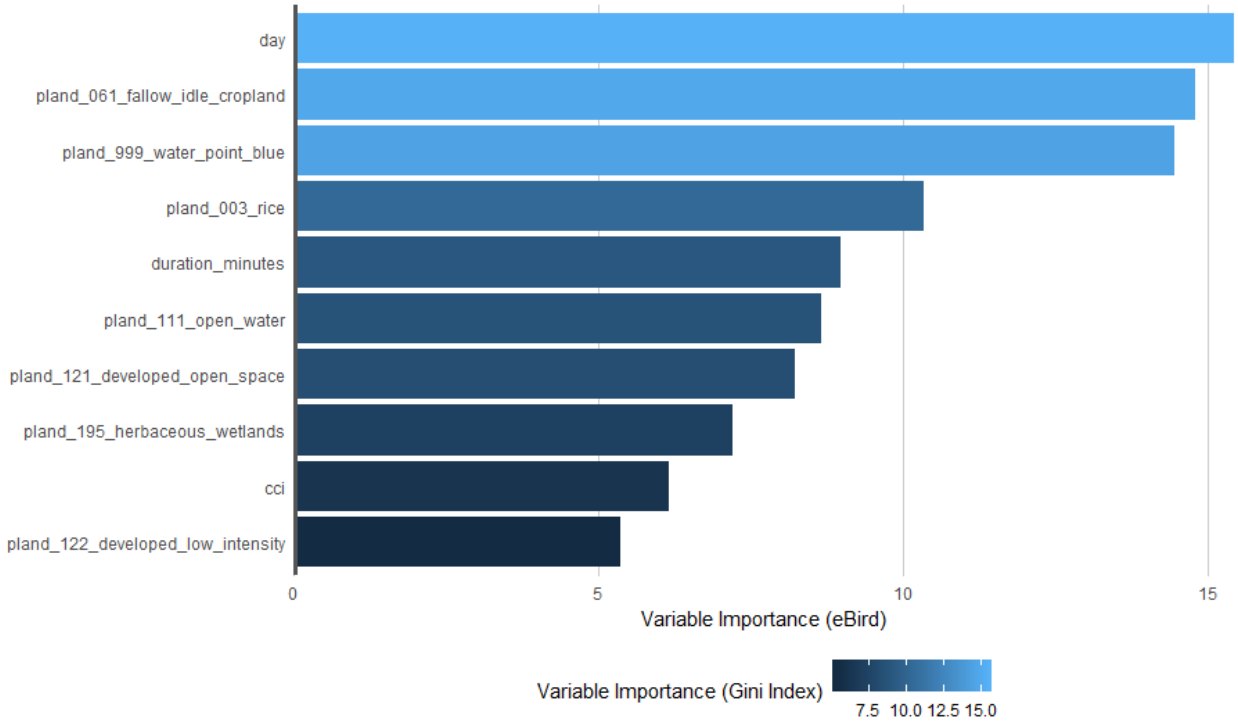

pland_003_rice
pland_999_water_point_blue
duration_minutes
pland_061_fallow_idle_cropland
pland_121_developed_open_space
cci
pland_122_developed_low_intensity
observation_date
pland_123_developed_med_intensity
pland_000_no_data
pland_058_clover_wildflowers
pland_195_herbaceous_wetlands
year
pland_075_almonds
pland_054_tomatoes
0
10
20
30
40
50
60
Variable Importance (Combined)
Variable Importance (Gini Index)
20 30 40 50 60

Important variables selected by the model trained on each data set for western sandpiper as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

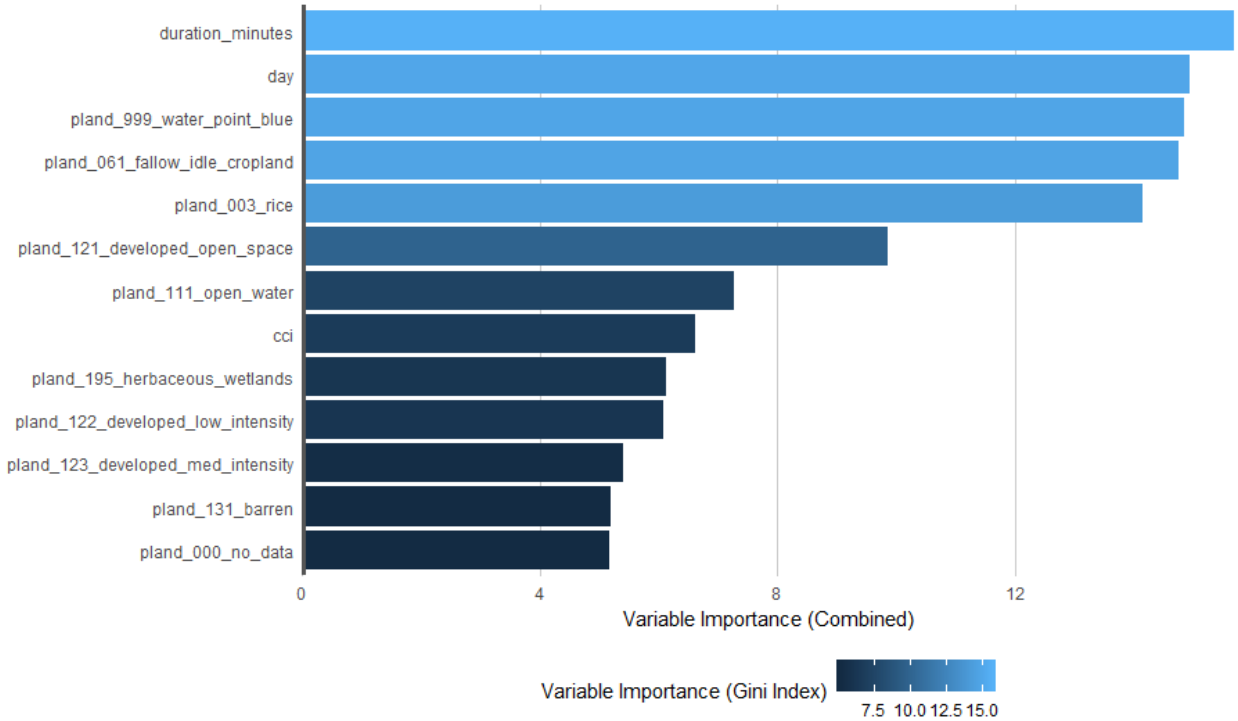
A



B

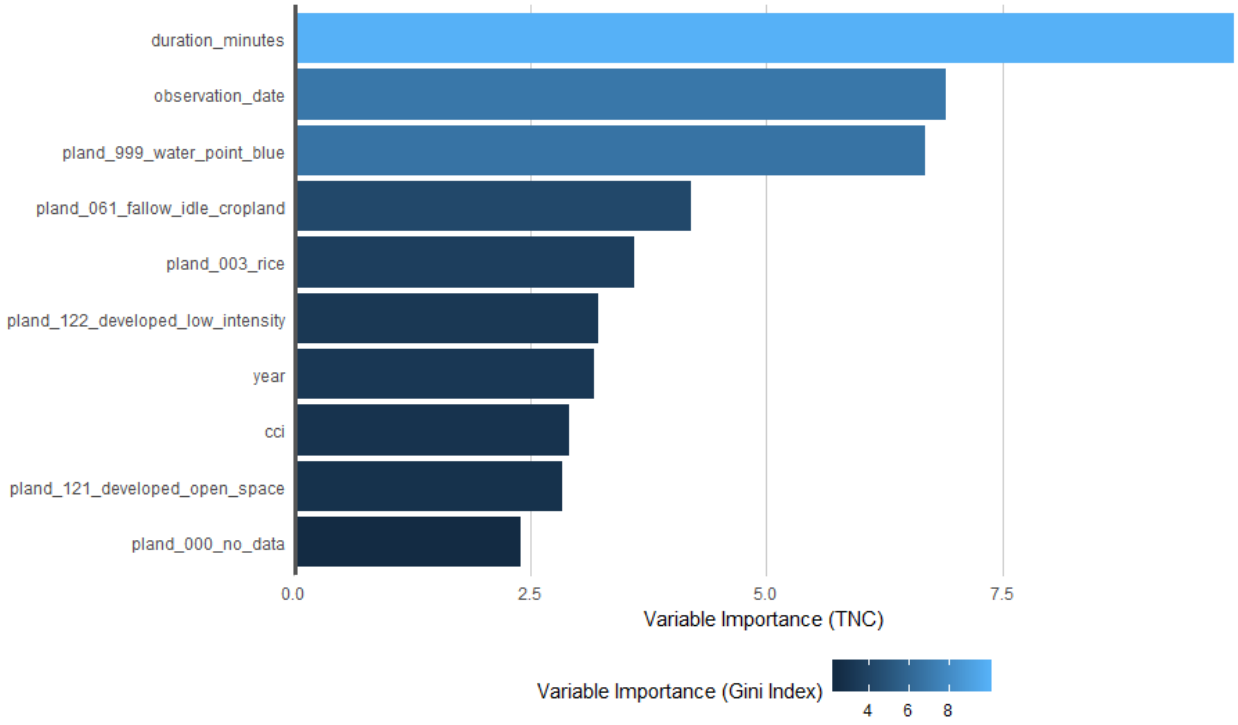

day
pland_061_fallow_idle_cropland
pland_999_water_point_blue
pland_003_rice
duration_minutes
pland_111_open_water
pland_121_developed_open_space
pland_195_herbaceous_wetlands
cci
pland_122_developed_low_intensity
0
5
10
15
Variable Importance (eBird)
Variable Importance (Gini Index)
7.5 10.0 12.5 15.0

C

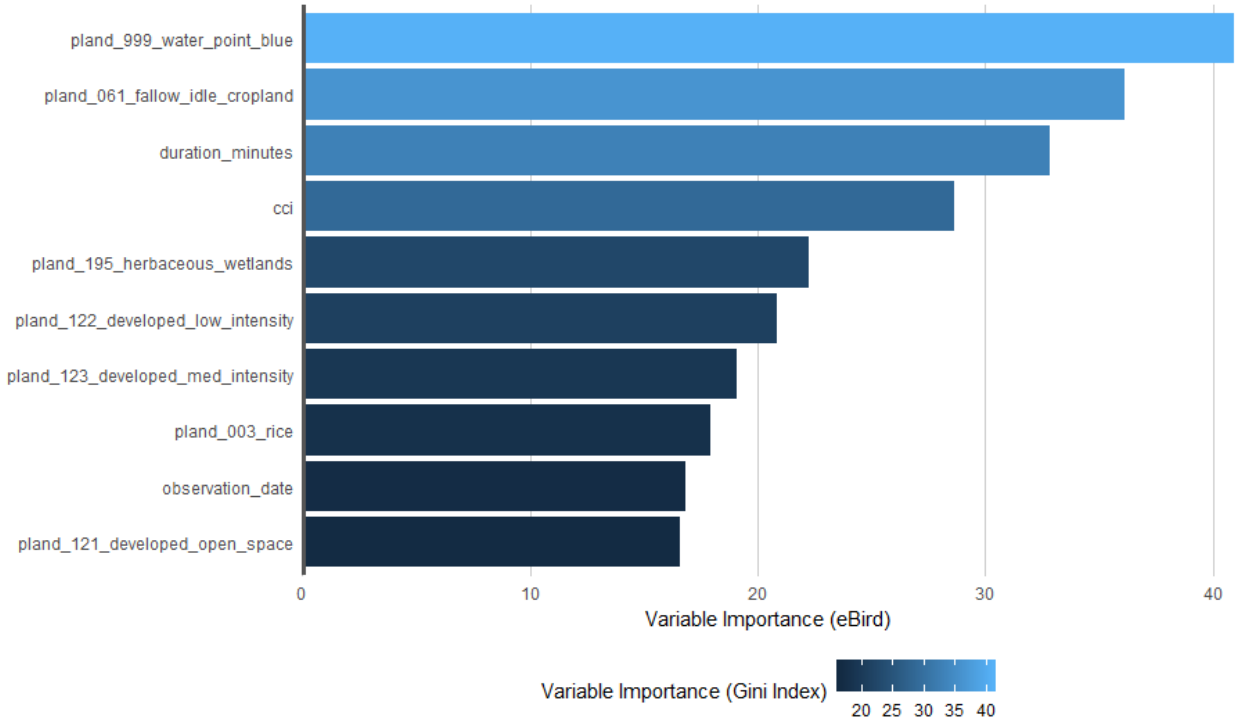

duration_minutes
day
pland_999_water_point_blue
pland_061_fallow_idle_cropland
pland_003_rice
pland_121_developed_open_space
pland_111_open_water
cci
pland_195_herbaceous_wetlands
pland_122_developed_low_intensity
pland_123_developed_med_intensity
pland_131_barren
pland_000_no_data
0
4
8
12
Variable Importance (Combined)
Variable Importance (Gini Index)
7.5 10.0 12.5 15.0

Important variables selected by the model trained on each data set for greater and lesser yellowlegs combined as selected by Gini importance index; TNC point counts alone (A), eBird checklists alone (B), and the combined dataset of TNC point counts and eBird checklists (C). Please note the different scales for each panel, and that this only represents the importance of each variable, not the direction of its relationship with probability of detection.

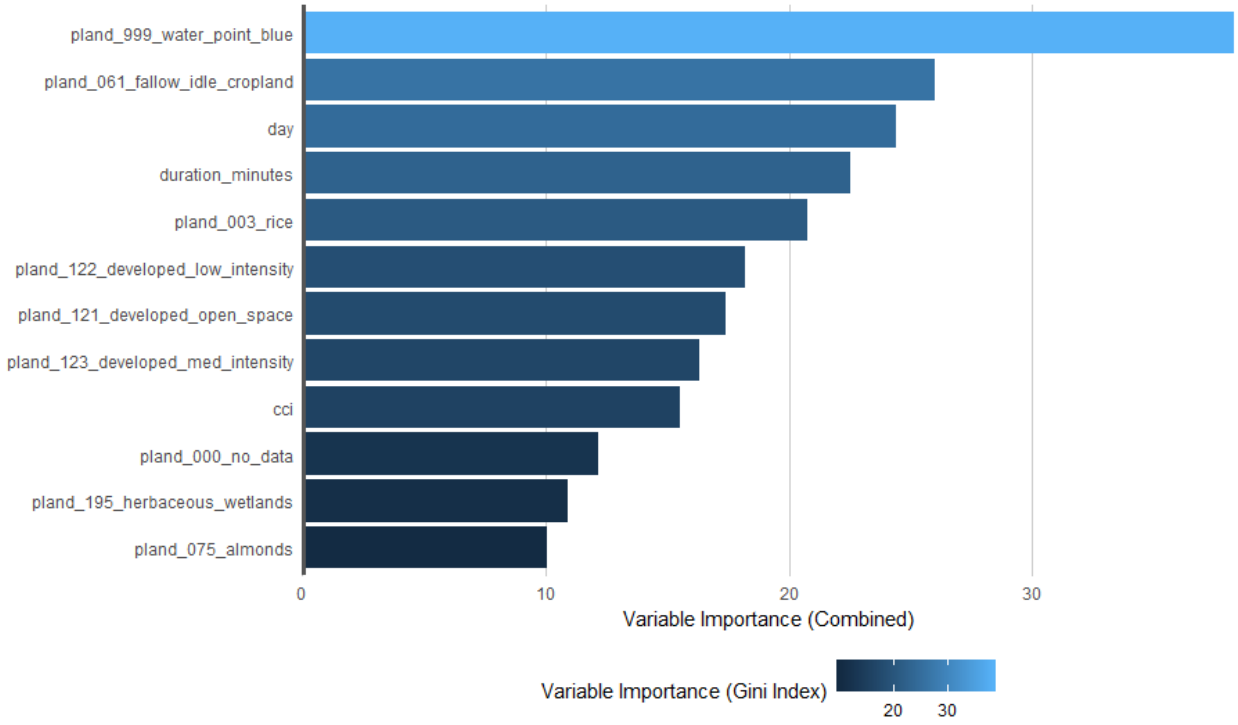
A


duration_minutes
observation_date
pland_999_water_point_blue
pland_061_fallow_idle_cropland
pland_003_rice
pland_122_developed_low_intensity
year
cci
pland_121_developed_open_space
pland_000_no_data
0.0
2.5
5.0
7.5
Variable Importance (TNC)
Variable Importance (Gini Index)
4
6
8

B


pland_999_water_point_blue
pland_061_fallow_idle_cropland
duration_minutes
cci
pland_195_herbaceous_wetlands
pland_122_developed_low_intensity
pland_123_developed_med_intensity
pland_003_rice
observation_date
pland_121_developed_open_space
0
10
20
30
40
Variable Importance (eBird)
Variable Importance (Gini Index)
20
25
30
35
40

C


pland_999_water_point_blue
pland_061_fallow_idle_cropland
day
duration_minutes
pland_003_rice
pland_122_developed_low_intensity
pland_121_developed_open_space
pland_123_developed_med_intensity
cci
pland_000_no_data
pland_195_herbaceous_wetlands
pland_075_almonds
0
10
20
30
Variable Importance (Combined)
Variable Importance (Gini Index)
20
30