Obtaining New Insights for Biodiversity Conservation from Broad-Scale Citizen Science Data

Daniel Fink[1], Wesley M. Hochachka[1], Marshall J. Iliff [1], Christopher L. Wood [1], Brian L. Sullivan[1], Kenneth V. Rosenberg[1], M. Arthur Munson[2], Mirek Riedewald[3], Steve Kelling[1]

Increasing public engagement in volunteer science, either through data collection or processing, is both raising public awareness of science and gathering useful information for scientists. While the payoffs of citizen science are potentially large, achieving them requires new approaches to data management and analysis that can only result from strong cross-disciplinary collaborations. This is especially true in ecology and conservation biology, where historically the understanding of species' responses to environmental change has been constrained by the limited spatial or temporal scale of available data. Here we describe collaborative research in ecology, computer science, and statistics to generate

[1] Cornell Lab of Ornithology, Ithaca, New York 14850, USA

[2] Department of Computer Science, Cornell University, Ithaca, New York 14850, USA

[3] College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115, USA

essential information for conservation management of North American birds: accurate dynamic bird distributions models based on habitat associations across much of North America. Unique is our ability to describe the broad-scale dynamics of seasonal bird distributions and the associated seasonal patterns of habitat use. Our source of bird distribution data is eBird, an online bird checklist program that currently gathers more than 74,000 checklists monthly from a large network of contributors. Our results were made possible through a data intensive scientific workflow that includes analytical methods merged from the fields of machine learning and statistics. We believe that this novel approach of data collection, synthesis, analysis, and visualization will serve as a hallmark for future research initiatives, with broad applicability across many scientific domains.

Anticipating and mitigating large scale threats to biodiversity requires a thorough understanding of species' habitat requirements. However, obtaining this knowledge of ecological systems is challenging because species’ distributions vary through time and with different environmental associations across species' ranges. Identifying systematic patterns in the face of this variability is one of the most difficult tasks in ecology for two reasons: (1) ecologically-relevant data are either

not collected across sufficiently large spatial or temporal scales, or data are heterogeneous and widely scattered; and (2) conventional expert-driven analytical methods are not effective for facilitating pattern discovery with such sparse, noisy data and highly variable ecological signals .

Meeting the first challenge has required development of a data set that contains continent-wide yet fine-resolution data on birds and environmental features, and an efficient workflow for updating these data. Avian data come from eBird, a citizen science project that gathers more observations of organisms than any other existing monitoring program (more than 10,000,000 species-date-locations observations annually). While other large-scale North American bird-monitoring programs, such as the Breeding Bird Survey and the Christmas Bird Count gather valuable data, they provide snapshots of species distributions during single seasons. eBird provides continuous, year-round data, from a continent-wide network of volunteers. We combined each of the more than 320,000 eBird locations where observations were made with over 500 environmental variables, which is freely available. Environmental predictor variables came from multiple

sources and included remote sensing (e.g., land cover), geographic, climatic, and surveys gathering human demographic information.

Overcoming the challenge of data analysis requires a new "data driven" approach wherein new information emerges from the data instead of a more traditional "hypothesis-driven" approach that examines expected patterns in the data . Nonparametric machine-learning techniques (e.g., decision tree ensembles, neural networks, support vector machines, and maximum entropy models) can detect and describe complex patterns, and are increasingly being utilized by ecologists for species distribution modeling . Although these techniques work well for analysis of static species distributions, we have found (unpublished data) that these techniques can produce highly erroneous predicted distributions when there are differences in the amounts of data available from different areas or time periods. Our solution has been to use the statistical approach of constraining the flexibility (i.e. parameterizing) the machine-learning methods. We have done this by creating an ensemble of spatiotemporal sub-models . This allows for continent-wide analysis that retains local patterns even when we do not have extensive prior knowledge of local variation in habitat associations.

Our fundamental requirement for the data and analyses is that they be able to accurately describe distributions and habitat associations throughout the year, especially for widespread and migratory species. Past attempts to design conservation landscapes across large regions or entire species' ranges have been based on models of distributions in a single season, usually the breeding season. Such models may not fully reflect the limiting factors that drive population declines, as migratory songbirds face at least 15 times greater risk of mortality during migration than during the more sedentary breeding or winter seasons . Thus, empirical knowledge about species' migration pathways, timing, and concentration areas are important new knowledge for science-based management strategies. In addition, our ability to conserve landscapes with enough resiliencies to accommodate changing bird distributions, for example due to climate change, will require an accurate understanding of species' habitat associations at all times of the year and all parts of species' ranges. Here we demonstrate how we can (1) accurately describe seasonal changes in species distributions, (2) identify regional differences in organism's migratory movements, and (3) discover seasonal differences in habitat associations.

Our combination of data and analysis techniques can produce very accurate models of species distributions. Figure 1a displays the distribution of breeding occurrence and habitat preferences of the Indigo Bunting (*Passerina cyanea*) and Chimney Swift (*Chaetura pelagica*), two species with similar breeding ranges. Informal comparisons of these predictions to prior knowledge closely match and provide evidence for the accuracy of these maps. In addition, quantitative measures of breeding season predictive performance verify this impression, with predictive accuracies of 83% and 88%, and AUC scores of 0.88 and 0.84 for indigo bunting and chimney swift respectively. (See the Methods section for more information about model validation). We emphasize that the distribution maps are habitat based, and not simple interpolations. This is illustrated in Figure 1a through the contrasting occurrence rates in urban centers, and in Figure 1b through the contrasting partial effects of human housing density.

We can also describe the timing and movement patterns of migrant species. Figure 2 illustrates our ability to describe migration by comparing the spring and fall migrations for indigo bunting and Western Wood-Pewee (*Contopus sordidus*), two neotropical migrants with very different migration paths (For more examples

see supplementary figures 1 and 2). Arrival dates are quantified as the first date when the predicted species' occurrence probability exceeds a threshold level within a specified season. This model-based approach has two fundamental advantages over analyses that rely solely on raw observations: 1) the model provides a framework to control known sources of bias (e.g. due to variation in detection rates), and 2) the model utilizes all available data, avoiding known limitations of observed "first arrival dates", which are a more variable metric of migration timing . Our data and analyses describe indigo buntings (Fig 2a) crossing the Gulf of Mexico making landfall in early April, rapidly filling their southeastern breeding range by mid-April, and after some delay arriving in northern portions of their breeding range by mid-May (see supplementary movie to visualize annual pattern of indigo bunting occurrence). Western wood-pewee arrival was also modeled (Fig 2b) to be in early April with rapid filling of their southwestern breeding range, which was followed by a later April push along the more temperate Pacific coast finally arriving in mid-May in the cooler northern Rocky Mountains. The fall departure of indigo bunting is shown (Fig 2c) to begin in mid-September, with a complete withdrawal from northern latitudes by mid-

October. Note the Mississippi River Valley harbored late indigo buntings into early November. In contrast, models for western wood-pewees (Fig 2d) indicated that they did not linger in their breeding areas, departing northern latitudes in late September and more southern latitudes by mid-October. Analyses such as estimated arrival and departure dates will facilitate our understanding of how organisms respond to broad-scale environmental variation such as changing biotic environments, and variation in weather and climate.

Few quantitative descriptions exist of season variation in habitat use for most bird species. Figure 3 demonstrates our ability to detect and describe population-level seasonal changes in habitat associations. The partial effect of the percent of deciduous forest within a 225ha neighborhood has a strong positive effect on indigo bunting occurrence rates during the breeding season and a slightly weaker positive effect during fall migration. In contrast, areas with a greater proportion of pasture appear to be preferred during the fall. Indigo buntings often nest in edges of hardwoods and insects comprise much of their diet, while they winter in more open agricultural areas where their diet shifts primarily to seeds . Our models suggests that the indigo bunting begins a shift to winter habitat

associations soon after breeding, and prefers more open habitat during fall migration.

Occurrence and habitat modeling are increasingly important tools for conservation planning and land management, and provide fundamental information for the conservation design of large landscapes . By adopting a “data intensive” approach we have been able to harness the power of a broad-scale network of citizen scientists to address the need for accurate, *year-round* predictive models of bird distributions across varied spatial extents. These models provide a framework for range-wide and full life-cycle conservation strategies, necessary to reverse population declines and implement habitat-management objectives for threatened species . As we continue to collect large volumes of eBird data we can extend these analyses to study year-to-year patterns of movement of many North American species and assess effects of environmental contamination. All of this will enable land-managers and conservation biologists to better coordinate national and international conservation efforts.

Methods Summary:

We used eBird presence-absence data collected under the traveling count protocol from January 1, 2004 – December 31, 2008 from across the conterminous United States. Throughout this paper we presented results for the indigo bunting, at times in comparison with other species. The selected species have broad distribution across much of the conterminous United States, and have fairly well-understood migrations. Thus, the quality of model predictions could be compared with expert opinion in addition to the quantitative measures of predictive performance.

Relatively little is currently known about the broad-scale migration patterns for many common North American birds. Therefore, we chose to model these migrations using an automatic, semiparametric modeling approach to facilitate the rapid exploration of migrations across a broad set of species with highly variable migration strategies. The method we used was designed specifically to discover seasonally- and regionally-varying patterns in the data. Spatiotemporal variation in distribution and habitat association is captured by combining a series of separate

submodels that each described bird occurrence within a smaller spatial region and across a roughly one-month period. Decision tree submodels were used to relate the 43 explanatory predictors to observed responses, facilitating "model-based" explorations to detect complex patterns of occurrence and uncover underlying dynamic associations between environmental features and bird distributions.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study is based on work supported by The Leon Levy Foundation, The Wolf Creek Foundation, and The National Science Foundation (Grants ITR-0427914, DBI-0542868, DUE-0734857, IIS-074826, and IIS-

0832782). In addition the authors would like to thank Benjamin Zuckerberg and Rich Caruana for comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and request for materials should be address to SK (stk2@cornell.edu).

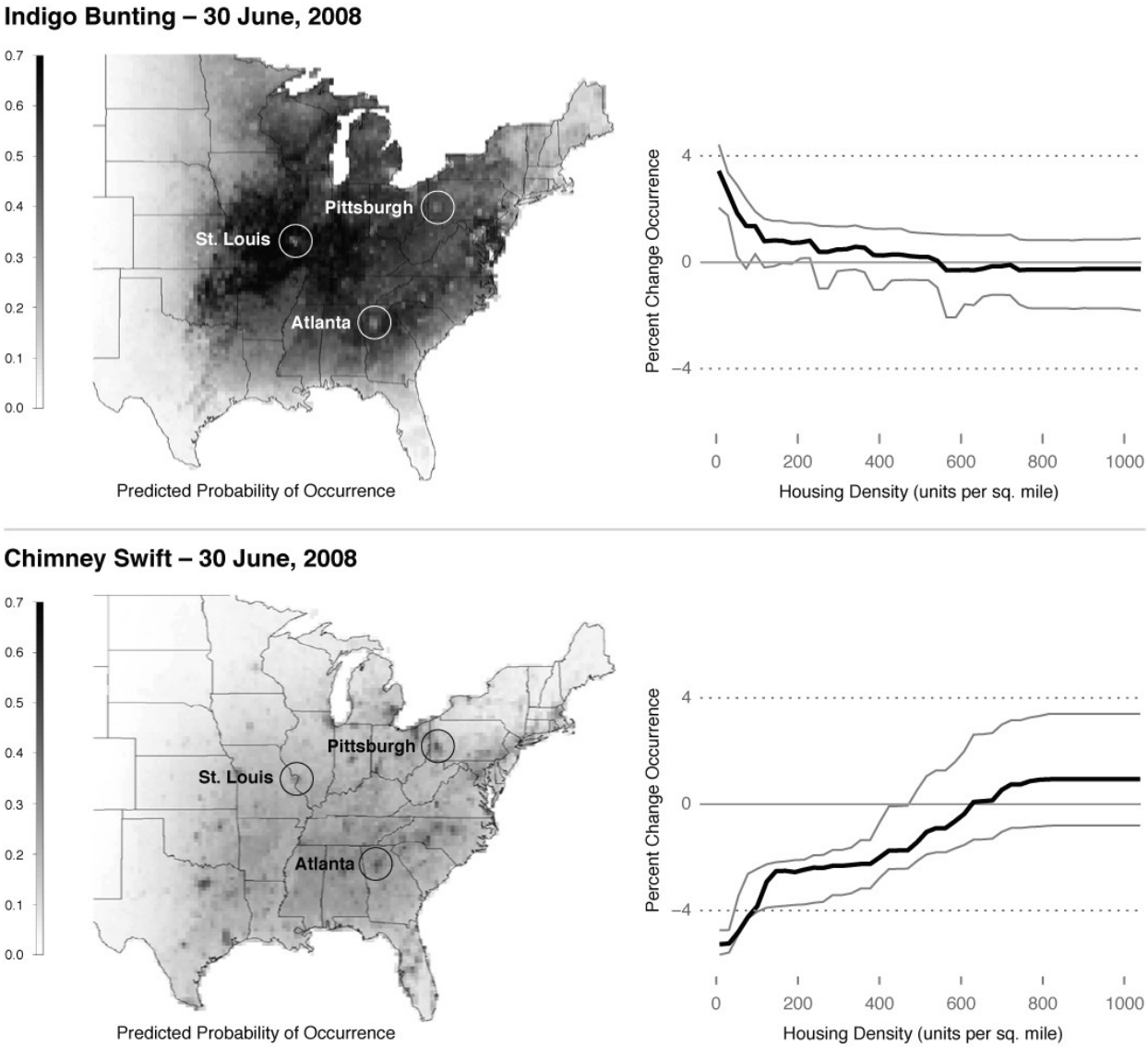

Figure 1. Predicted probability of occurrence and partial dependence on housing density for indigo bunting (top) and chimney swift (bottom) for 30 June 2008. Although both species have similar, widespread distributions across eastern U.S., the fine-scale differences around major urban centers is striking. The indigo bunting, which requires natural forest and shrub habitats, has relatively low occurrence rates in urban centers while the chimney swift, which nests in chimneys in urban and suburban areas, has relatively high occurrence rates. We note three

major urban centers (Pittsburg, St.Louis, and Atlanta) and provide the partial effect estimates of housing density to highlight these differences.

Figure 2. Comparison of timing and directional movements for two migratory bird species based on predicted arrival and departure dates. Colors indicate earliest day in 2008 when species occurrence exceeds 5% (left), or last day in 2008 when occurrence exceeds 5% (right) based on a sequence of predicted occurrence surfaces at 3 day intervals across 2008. Modeling the annual fine-scale contours of migratory arrival and departure dates facilitate tracking of species' response to long-term environmental change and will enhance our ability to identify important migration corridors or stopover sites.

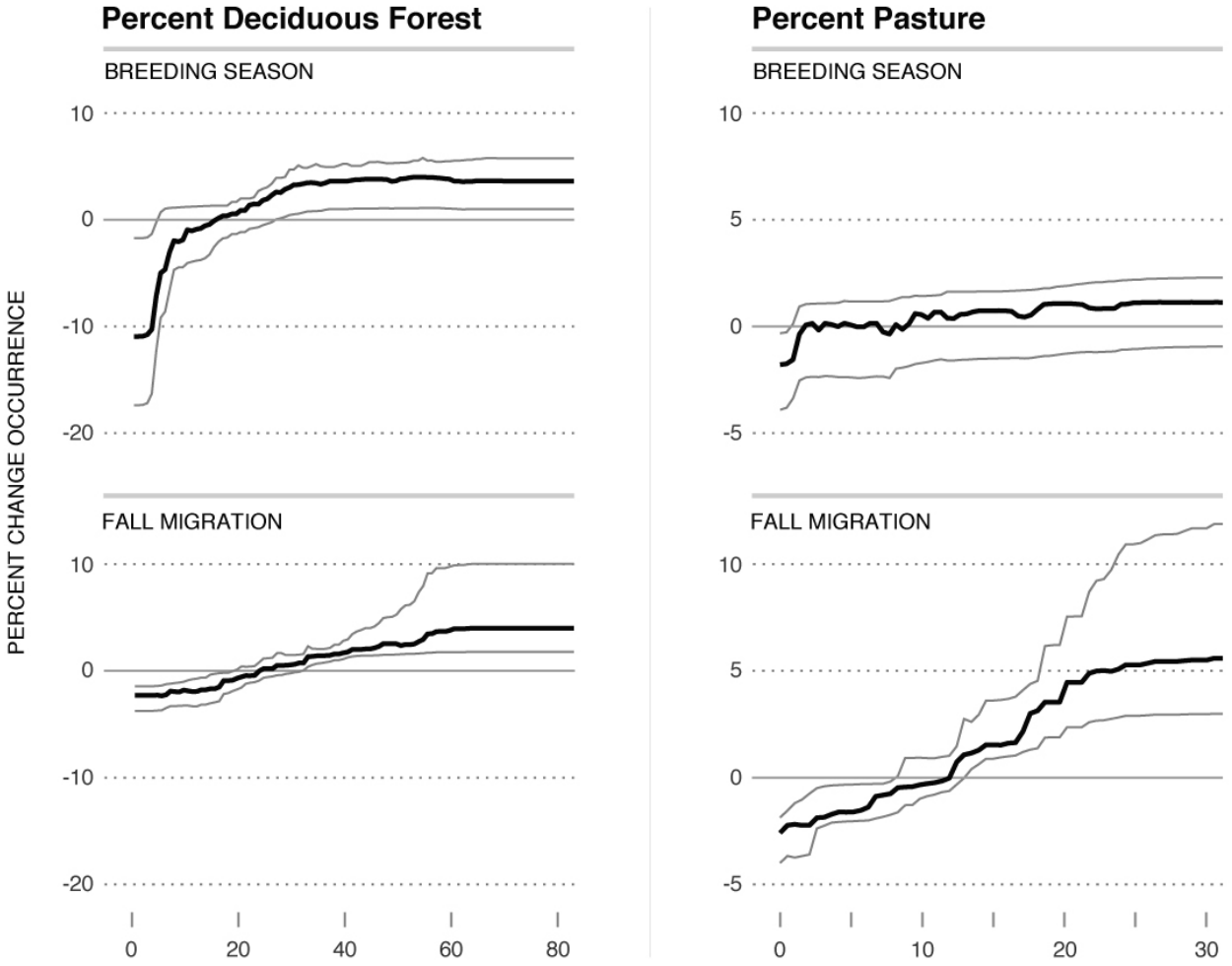

Figure 3. Differences in habitat associations for indigo bunting in breeding season and fall migration. All partial dependence effects were estimated using partial dependence functions and were estimated separately for the Breeding season (June 5 – July 31) and during the Fall migration (September 1 – October 15). Grey represents approximate 95% confidence regions. The partial effect of deciduous forest is strongly positive during the breeding season and a slightly weaker during the fall. There is no significant partial effect of pasture in spring, but a strong positive pasture effect appears during the fall migration.