Addressing data integration challenges to link ecological processes across scales

Elise F Zipkin[1,2*], Erin R Zylstra[1,2], Alexander D Wright[1,2], Sarah P Saunders[1,2,3], Andrew O Finley[2,4], Michael C Dietze[5], Malcolm S Itter[4,6], and Morgan W Tingley[7,8]

Data integration is a statistical modeling approach that incorporates multiple data sources within a unified analytical framework. Macrosystems ecology – the study of ecological phenomena at broad scales, including interactions across scales – increasingly employs data integration techniques to expand the spatiotemporal scope of research and inferences, increase the precision of parameter estimates, and account for multiple sources of uncertainty in estimates of multiscale processes. We highlight four common analytical challenges to data integration in macrosystems ecology research: data scale mismatches, unbalanced data, sampling biases, and model development and assessment. We explain each problem, discuss current approaches to address the issue, and describe potential areas of research to overcome these hurdles. Use of data integration techniques has increased rapidly in recent years, and given the inferential value of such approaches, we expect continued development and wider application across ecological disciplines, especially in macrosystems ecology.

Front Ecol Environ 2021; 19(1): 30–38, doi:10.1002/fee.2290

Data integration, or the inclusion of multiple response-variable data sources within a single statistical modeling framework, is a methodological approach that facilitates understanding of complex and interacting processes (Schaub and Abadi 2011; Michener and Jones 2012). The use of data integration (also referred to as integrated modeling, data assimilation, data fusion, integrated analysis, inverse modeling, or ensemble estimation) within ecology is rising steadily (Figure 1), reflecting advancements in computational resources and dramatic increases in the quantity of available data (LaDeau *et al.* 2017). While integrated modeling techniques are revolutionizing how analyses are conducted across an array of ecological systems, data integration can be particularly advantageous in macrosystems ecology. Macrosystems ecology is the study of ecological patterns and processes at broad spatiotemporal scales and their interactions with phenomena at other scales (Heffernan *et al.* 2014; Soranno *et al.* 2014; Fei *et al.* 2016). Some macroscale questions can be addressed with a single source or type of data and relatively simple statistics (eg spatial scaling patterns using regression in macroecology; Brown and Maurer 1989). Yet many broad- and multi-scale research questions require combining disparate datasets, especially when the focus is on understanding mechanistic processes (Levy *et al.* 2014; LaRue *et al.* 2021).

Data integration is an integral component of many investigations in macrosystems ecology. Compared to geographically or temporally restricted analyses, it can be challenging to estimate ecological parameters at macroscales using only a single data source because of interacting or nonlinear environmental, climatic, and biological processes, as well as data limitations. In macrosystems ecology, various data sources can provide information on components of the study system that operate at different scales (eg Robinson *et al.* 2018; Itter *et al.* 2019). In estimations of species distributions, for example, opportunistic records (eg from iNaturalist or museum collections) can be used to delineate the occurrence of individuals across a large spatial extent, whereas smaller-scale mechanistic studies can provide data on factors influencing density across gradients of local variables (Figure 2). Similarly, in biogeochemical modeling, eddy flux, field inventories, and remote-sensing data each contain distinct information about potential pathways of carbon dioxide (CO_2) exchange across local, regional, and even continental scales (Keenan *et al.* 2012).

In a nutshell:

- Understanding ecological processes across spatiotemporal scales can be enhanced by using multiple, independent data sources in a unified analysis via statistical data integration methods
- Although data integration can improve ecological inferences, challenges can arise during analysis
- We review the most common statistical challenges related to data integration in macrosystems ecology and discuss ways in which they can be overcome
- We provide researchers with resources from the literature to address issues that may arise during data integration and highlight avenues of future research

[1]*Department of Integrative Biology, Michigan State University, East Lansing, MI;* [2]*Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI* [*]*(ezipkin@msu.edu);* [3]*Science Division, National Audubon Society, New York, NY;* [4]*Department of Forestry, Michigan State University, East Lansing, MI;* [5]*Department of Earth and Environment, Boston University, Boston, MA;* [6]*Department of Environmental Conservation, University of Massachusetts–Amherst, Amherst, MA;* [7]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT;* [8]*Ecology and Evolutionary Biology, University of California–Los Angeles, Los Angeles, CA*

Beyond expanding the scale and scope of analysis (Isaac *et al.* 2020), data integration techniques can provide a variety of additional inferential benefits. By leveraging information from multiple sources, data integration improves the accuracy, and often the precision, of parameter estimates, enabling comprehensive assessments of processes underlying ecological responses to environmental variability (Gotway and Young 2002; Fletcher *et al.* 2016; Grace *et al.* 2016). Integrating independent datasets can also account for multiple sources of uncertainty and error in parameter estimates (Schaub and Abadi 2011; Keenan *et al.* 2013; Fithian *et al.* 2015), and allow for estimation of parameters for which no explicit data are available (ie improving parameter non-identifiability; see Panel 1 for definitions of relevant terms). The approach for data integration begins with construction of a model that describes the ecological processes of interest. Likelihood functions (Panel 1) are used to identify how each data source informs parameters in the ecological process model. The individual data sources are then linked to one another via parameters that are informed by more than one dataset (Miller *et al.* 2019). For example, a model of leaf phenology might combine data from ground-based phenocams with satellite imagery, where the different data sources inform the same ecological process model but have unique sampling errors (Viskari *et al.* 2015).

There is growing awareness and adoption of data integration techniques in macrosystems research, as well as in ecology more broadly (Figure 1), but this methodological framework is still relatively new. Although several recent papers have synthesized data integration approaches (eg Zipkin and Saunders 2018; Miller *et al.* 2019; Isaac *et al.* 2020), none have focused on describing the methodological challenges that ecologists encounter when integrating disparate data sources, nor potential solutions for overcoming those difficulties. To address this, we conducted a search of recently published peer-reviewed journal articles to identify inferential impediments to data integration in macrosystems ecology (see WebPanel 1 for search criteria). Nearly half (44%) of the articles that presented original research integrated two or more datasets, while 20% of all articles (ie research, commentaries, reviews) discussed inferential problems. The most common challenges were (in decreasing order of frequency): (1) mismatches in spatial or temporal scale of data sources, (2) differences in the quantity and/or information content of data sources, (3) sampling biases, and (4) optimization of model development and assessment. An additional challenge that we identified – nonstationarity or spatiotemporal variation in processes or covariate effects (Panel 1) – is often overlooked but is described in detail in Rollinson *et al.* (2021). Although these four challenges can occur in analyses that integrate data at any scale, they tend to be exacerbated in macrosystems ecology, where the geographic scope is large and the data tend to be "big" (Levy *et al.* 2014).

The use of complex and computationally intensive analytical approaches is growing in macrosystems ecology, and consequently development of technical skills has been identified as a key need for researchers in this field (Farrell *et al.* 2021). With this in mind, we aim to increase awareness of potential inferential pitfalls that ecologists may encounter when integrating multiple data sources and present researchers with resources that can help them avoid or ameliorate issues as they arise. We recognize that logistical constraints, such as the processing and/or management of datasets through data harmonization (a distinct informatics approach for combining similar datasets that differ only in format or origin) and computational limitations, can also hinder macrosystems ecology. However, we focus on statistical and modeling challenges of data integration because logistical issues have been recognized and discussed in greater depth previously (eg Rüegg *et al.* 2014; LaDeau *et al.* 2017). In the following sections, we review the four inferential challenges listed above by providing a general description of each problem, discussing how the problem manifests in macrosystems ecology, and offering current and potential approaches to address and resolve the issue. Although these data integration challenges can be interrelated, leading to trade-offs in modeling decisions, each are presented independently for clarity. We conclude by highlighting key steps and resources that can help researchers identify and overcome data integration challenges (WebTable 1).

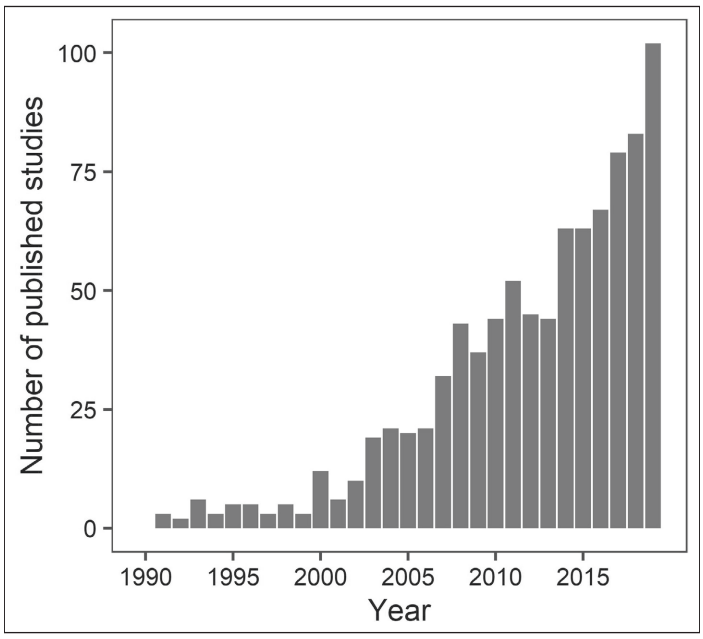

Figure 1. A search of peer-reviewed publications (see WebPanel 1 for search criteria) revealed that the use of data integration in ecological research was uncommon from 1990–1999 but increased markedly over the subsequent two decades (2000–2019).

Resolving mismatches in spatial and temporal scales of available data sources

A mismatch in the dimensions or resolution of sample units (ie grain) can arise when combining multiple data sources that have been collected at different spatial and/or temporal scales (Nguyen *et al.* 2014). Mismatches in grain are common

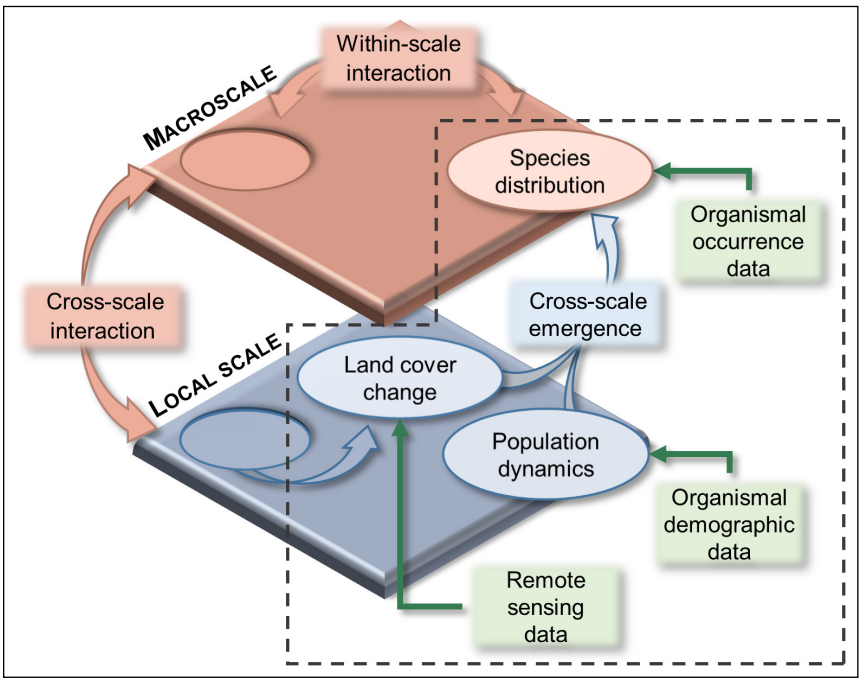

Figure 2. Schematic diagram illustrating how various data sources (green rectangles) can be used to parameterize models of ecological phenomena (ovals; linked with small green arrows) within a macrosystems framework. The large colored arrows depict relationships between ecological processes within and across scales. Using a population ecology example, three interacting processes are highlighted, along with their corresponding data sources (within the dashed polygon). Blank ovals are used to illustrate other, unnamed processes. Adapted from Heffernan *et al.* (2014).

when trying to relate remotely sensed data or geographic information system (GIS) data layers (eg data aggregated according to political units, such as human census and disease data) to each other or to field data, which tend to be collected at fine spatial resolutions (Figure 3; Nguyen *et al.* 2012; Estes *et al.* 2018). Combining data without accounting for scale mismatches (eg regridding a coarse resolution product to a fine scale, interpolating point reference data to a grid) can result in artificial inflation of the sample sizes of one or more data sources, potentially resulting in biased inferences or overstated precision (Gotway and Young 2002).

Many individual sources are used to collect data with the purpose of informing distinct processes or dynamics of a system, often at local scales, and therefore provide incomplete information to address macrosystems questions. Merging data sources that describe different components of biological and/or physical processes, portions of the geographic range of interest, or slices of a longer time series can help address macrosystems questions, but the grain and extent of data are likely to vary across available sources (Miller *et al.* 2019; Schimel *et al.* 2019). For example, data on breeding birds are available throughout North America from numerous programs, including the North American Breeding Bird Survey (BBS), eBird, and other volunteer-led monitoring efforts, but observations are reported on different spatial scales for each of these programs (eg 0.25-mile radius point counts every 0.5 miles along a 25-mile permanent transect for BBS, checklist data with variable survey area for eBird), and the time series available varies considerably among the data sources and across geographic areas (Pacifici *et al.* 2017; LaSorte *et al.* 2018).

Although mismatches in spatial and temporal scales have only recently been recognized as methodological challenges in ecological research, they have often been addressed in the statistical literature through the use of "change of support" procedures (Cressie 1996; Gotway and Young 2002). Change of support is the process by which data are either up- or down-scaled to achieve a single extent (overall geographic area or time period of interest) and grain (Panel 1). In contrast to interpolation and regridding approaches commonly used in GIS and similar software, change of support models properly account for uncertainties associated with changing scales (Cressie 1996). While naïve downscaling can artificially inflate sample sizes (eg if 100 observations are interpolated to a 1000-point grid, the effective sample size should still be 100), change of support

Panel 1. Glossary of terms related to data integration

Change of support: a class of techniques used to make inference about a variable at a different spatiotemporal extent or grain from that at which it was observed.

Information content: the extent to which data reduce uncertainty in parameter estimates (eg large volumes of data may have relatively low information content if there are strong temporal and/or spatial correlations among observations).

Likelihood: function describing the probability of observing the sample data, conditional on given parameter values from assumed probability distributions.

Non-identifiability: one or more parameters in a model are not estimable because of insufficient data or an overly complex model structure.

Nonstationarity: data that exhibit trends, cycles, or drift that result in non-constant parameters (eg mean, variance, autocorrelation) over time and/or space.

Structured data: data that are collected in a design-based framework, usually to answer pre-defined research questions.

Unstructured data: data that are collected continuously or opportunistically without a specific objective, typically occurring in higher volumes than structured data.

models preserve information content. For example, Itter *et al.* (2019) used a memory function to align water deficiency data (collected monthly) with defoliation and tree growth observation data (collected annually) to evaluate how tree growth responds to stress from water limitations. Methodological approaches to accommodate variable spatial extents and grain in ecological data have often been case-specific rather than widely applicable (eg Zipkin *et al.* 2017; Farr *et al.* 2020). Current work focused on developing a more general statistical toolbox for handling change of support across data types and spatiotemporal scales will help account for inferential uncertainties and expand data integration capabilities within macrosystems ecology (Pacifici *et al.* 2019).

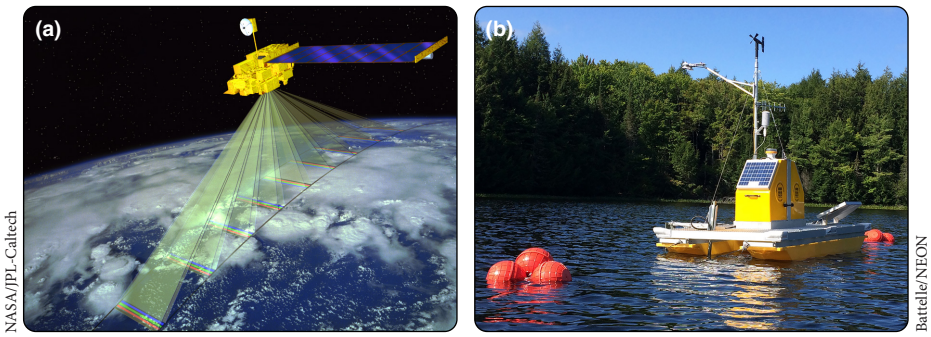

Figure 3. Studies in macrosystems ecology often integrate multiple data sources collected at different spatiotemporal resolutions. Satellites, like (a) the US National Aeronautics and Space Administration's Terra satellite, usually measure the Earth's surface and atmospheric properties at large spatial resolutions. Field data, such as that originating from (b) a weather data buoy, are collected at localized spatial resolutions. Inferences from a model integrating multiple data types can be more informative than inferences based on analyses of each data source separately, yet successful integration requires approaches to reconcile the different sample unit dimensions within each data source.

Addressing unbalanced data: uneven quantities and information content

In the context of data integration, unbalanced data refers to differences in the quantity (eg number of observations or data points) or information content (Panel 1; eg information-rich data from well-designed studies versus information-poor opportunistic data) among two or more data sources. If these differences are not accounted for, models can produce estimates biased toward abundant data sources regardless of their information content. For example, in studies of population dynamics, recaptures of marked individuals can result in precise estimates of species' survival probabilities, whereas cryptic behaviors of breeding individuals may limit data on productivity, potentially biasing estimates of reproductive output and ultimately of population growth rates (eg Campbell *et al.* 2018).

The issue of unbalanced data is particularly acute in macrosystems ecology, where the large scale of interest frequently leads to uneven quantities of data across space and time or among sources and/or components in ecological models (Levy *et al.* 2014). Macrosystems research frequently relies on unstructured data sources (Panel 1; eg opportunistic "incidental sightings" represented in museum collections or reported on iNaturalist) and on automated data sensors (eg eddy covariance, stream gauges, cameras, soundscapes, sap flux, aquatic buoys, radio telemetry). Compared to structured data (Panel 1; that is, data collected using standardized methods with the goal of addressing a specific research question), unstructured data tend to be plentiful even though they are typically of lower inferential value, providing less information per observation for parameter estimation. Similarly, automated sensors can produce a wealth of information over relatively short time periods and overwhelm field data collected manually (Williams *et al.* 2009; Figure 4). As a result, models that use a combination of sensor data and field data can produce model fits dominated by high-volume sensor data (Richardson *et al.* 2010).

Although data integration can expand the spatiotemporal scope of research, differences in data quantity and information content may affect the structure and complexity of the ecological process model (eg by assuming that mechanistic processes or covariate relationships are constant across space). Conducting preliminary analyses of independent data sources prior to integration can help identify geographic locations, temporal periods, and/or mechanistic processes in which unbalanced data could lead to biases in inferences (Kéry and Schaub 2011; Kuikka *et al.* 2014). Common approaches to address issues of unbalanced data, such as subsampling or down-weighting the larger dataset, are typically ad hoc and may lead to different conclusions based on subjective choices during model development (Maunder and Piner 2017). However, recently developed methods to weight public science data according to observer expertise (and therefore to selectively down-weight available data) have led to improved model fit and predictive performance (Johnston *et al.* 2018). More objective approaches focus on modeling factors that inflate the information content of high-volume data, such as autocorrelation in time and space, as well as systematic observation errors (Dietze 2017). Formally modeling biases within statistical likelihoods also appears to be a promising approach (Fer *et al.* 2018) and is an active area of research.

Accounting for sampling biases in one or more data source(s)

Data at any scale reflect the methods used to select sample units (eg site-selection bias) and collect individual measurements (eg observation error). Site-selection biases occur when sample units are not selected randomly from pre-defined strata or when selected units fail to adequately represent the

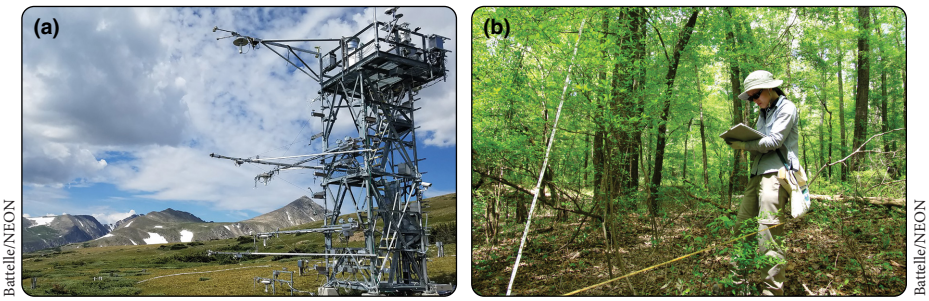

Figure 4. Merging information sources with varying quantities of data, such as those collected from automated and field-based approaches, is common in macrosystems ecology. Automated data like (a) flux tower measurements are collected nearly continuously in vast quantities, whereas (b) field-sampled data tend to be collected less frequently, resulting in sample sizes an order of magnitude smaller. A model integrating various data sources can provide a more complete picture of ecosystem processes as compared to independent analyses, yet successful integration requires that inferences be based on the information content of the data sources and not solely on the quantity of observations.

geographic area of interest. Observation biases occur when data are collected or recorded with error (eg imperfect detection by observers or instrumental noise in sensors). Although these sampling issues can materialize in analyses of a single dataset, they are likely to be more problematic in integrated analyses because biases in individual datasets can result in cumulative errors and uncertainties. Failure to account for sampling biases or errors can yield estimates that are overly precise and potentially misleading (Albert *et al.* 2010).

Although observation errors are likely to occur at similar rates across data collected at both local and broad scales, sampling biases resulting from inadequate or nonrandom site-selection may be more prevalent in data integration analyses at macrosystem scales. Limited resources and logistical constraints often prevent implementation of probabilistic sampling designs (eg stratified random sampling) at regional to continental scales. For example, macrosystems ecology increasingly uses data from volunteer-based (public science) monitoring projects to describe ecological phenomena at large spatial scales (Sullivan *et al.* 2014; LaSorte *et al.* 2018; Saunders *et al.* 2019b). Such programs can provide vast amounts of data to inform species distributions, relative abundance, and phenology, but collection efforts are often focused near urban areas, roads, or other locations with high human population densities (Figure 5; Bird *et al.* 2014). Similarly, the recent development of regional- and continental-scale research networks has facilitated the growth of macrosystems ecology through the collection of detailed and systematic data that can be used to inform models of geophysical and biological processes at large spatial scales. Yet models that incorporate these data need to account for nonrandom sampling, given that locations often reflect both the prioritization of particular ecoregions and logistical constraints (Keller *et al.* 2008). Within multi-scaled research, bias introduced at one scale via nonrandom sampling may be unintentionally propagated to inferences at other scales (Gelfand *et al.* 2012).

A wide array of strategies have been proposed to account for sampling biases, depending on the amount and type of data available, the source of bias (selection of sample units versus observation biases), and the extent or severity of the problem. For instance, when sample units are selected preferentially based on environmental features or other variables that correlate with the process of interest, biases can be reduced by including model components to describe the site-selection process (Diggle *et al.* 2010; Conn *et al.* 2017). Incorporation of spatially correlated random effects can also improve inferences (Hefley *et al.* 2017). Similarly, a state-space (ie hierarchical) framework that models the biological or physical process of interest separate from the processes used to collect data (eg de Valpine and Hilborn 2005) can account for observation biases and differences in sampling efforts among data sources, and/or be used to account for specific types of observation error mechanistically (eg Schaub and Abadi 2011). Finally, explorations of available data, independently and together, can help determine how to amend the design of ongoing data collection efforts to limit biases in model inferences or identify where and when to implement probabilistic sampling to collect auxiliary data.

Optimizing model development and assessment when incorporating multiple data sources

Balancing the complexity and realism of models with the data necessary to parameterize such models is an ongoing challenge in ecological research. Integrating data sources that are collected on different components of a system (eg various biological or physical processes, subsets of a geographic range) allows researchers to better understand spatial or temporal variation in ecological processes or mechanisms that underlie ecological patterns. However, combining disparate data sources to create increasingly complex models may not always result in improved inferences if the necessary assumptions are untenable or too restrictive, data on one or more aspects of the system are severely limited, or the model cannot be easily understood or applied to other ecological systems. Assessing the quality of inferences, or how well complex models fit both individual data sources and a suite of integrated data sources, is an active area of research (Besbeas and Morgan 2014; Carvalho *et al.* 2017).

Within macrosystems ecology, model fit may be hindered by nonstationarity (Rollinson *et al.* 2021) and cross-scale interactions (Figure 2), resulting in rejection of models that fit the data well for some (but not all) geographic locations or time periods or conversely, in acceptance of models with mediocre overall fit that fail to characterize processes in any region or time period well (Foody 2004). Moreover, standard approaches to validate model fit may not be feasible for integrated macrosystems analyses because of data limitations and/or logistical

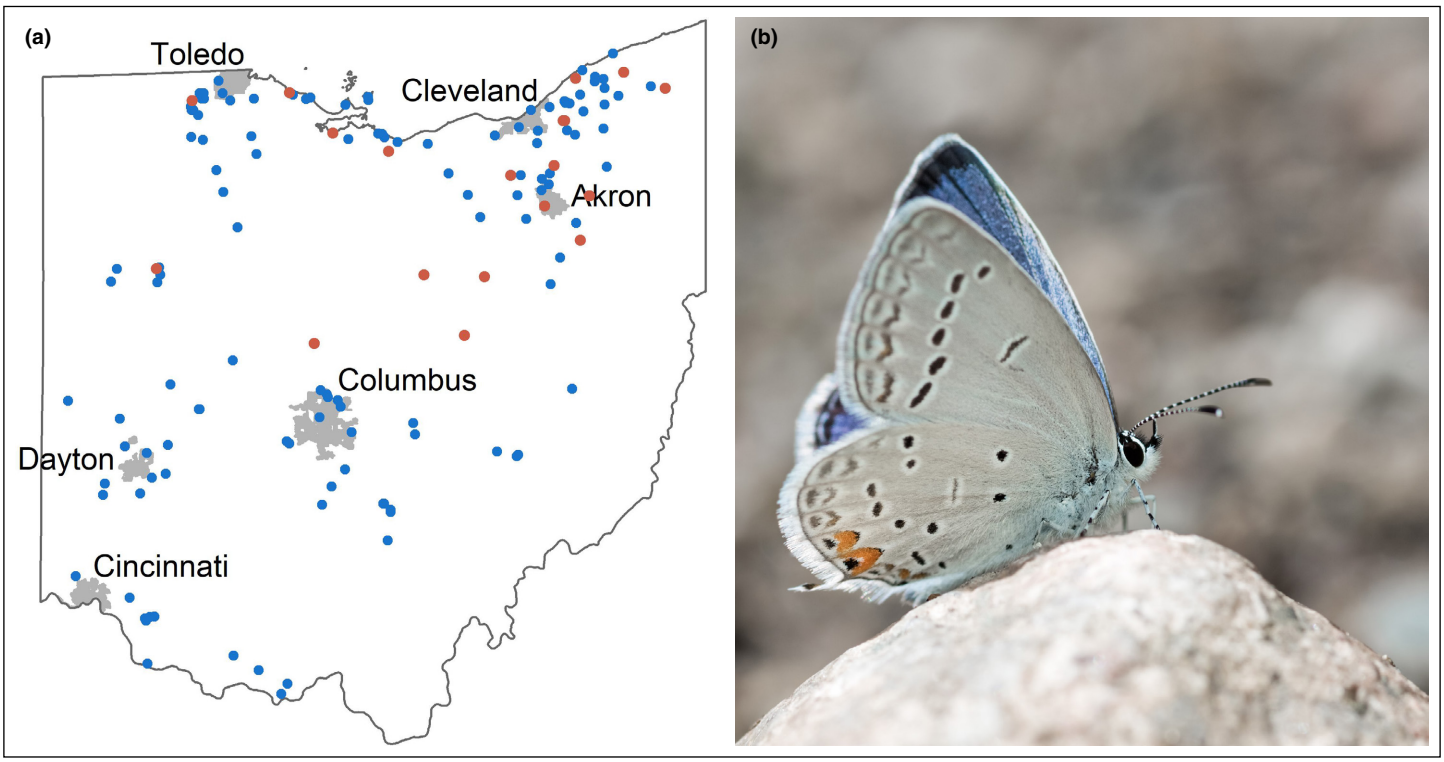

Figure 5. Macrosystems data frequently come from sources with nonrandom sampling designs. For example, (a) survey locations used for butterfly monitoring by volunteers with the North American Butterfly Association (red circles) and state organizations (blue circles) are clustered near urban areas (shown in gray for Ohio). Similarly, (b) opportunistic sightings of individual species – such as the eastern tailed-blue butterfly (*Cupido comyntas*), shown here on a gravel drive – often occur near roads and urban centers. A model that integrates these data sources could produce dynamic estimates of species' distributions, yet successful integration requires accounting for incomplete and nonrandom sampling to ensure that estimates are not biased.

constraints. For example, cross-validation methods require that some data (eg individual observations, random or geographically selected blocks of observations) be withheld from analyses to evaluate model fit (Hooten and Hobbs 2015; Roberts *et al.* 2017). In a macrosystems model that uses multiple data sources, however, it may not be clear how to select observations among disparate data sources that inform multiple ecological parameters.

In addition to challenges in assessing model fit, traditional model and variable selection approaches (eg null hypothesis testing, information–theoretic methods) may be insufficient for integrated models (Besbeas and Morgan 2014), particularly if the focus lies in evaluating the importance of various processes and drivers across multiple scales (Grueber *et al.* 2011; Levy *et al.* 2014). Model selection approaches often involve comparing models that include or exclude a given variable under the assumption that their effects on the response variable are independent. Within macrosystems ecology, however, the effects of one variable may be influenced by or co-vary with factors operating at different scales (eg Lawler and Edwards 2006). Therefore, it may not be possible to isolate the effects of a particular variable or "remove" one variable from a model without disregarding important cross-scale interactions that influence the broad-scale processes and patterns researchers seek to explain. Correlation among predictor variables that vary in scale can result in the selection of a model that includes spurious variables or excludes important predictors (Grueber *et al.* 2011).

Developing appropriate models using multiple data sources is an iterative process that generally begins by building components of the ecological process model separately to examine convergence and model fit (Kuikka *et al.* 2014; Ketz *et al.* 2018). Simulated datasets, in conjunction with goodness-of-fit discrepancy measures, can be used to compare estimated parameter values with true data values to examine whether model components systematically over- or under-predict quantities of interest (Besbeas and Morgan 2014; Zipkin and Saunders 2018). New approaches, such as multi-objective optimization (Branke *et al.* 2008), in which model selection is based on multiple criteria (Williams *et al.* 2019), may help identify appropriate models for complex multi-scale systems. Model selection choices should be based on data availability as well as objectives, such as identifying factors that have the largest effect on ecological processes of interest or minimizing uncertainty of model predictions.

Conclusions and future directions

Data integration offers ecologists an opportunity to explore complex, multi-scaled phenomena by combining available information within a single analytical framework. Integrated models improve estimation of ecological processes and

patterns because they expand the amount and scope of data available while explicitly accounting for multiple sources of error and uncertainty. The use of multiple, independent data sources can also reveal biases in parameter estimates that are hidden in analyses based on a single dataset, thereby improving the accuracy of inferences used to inform conservation and management efforts (Saunders *et al*. 2019a). In addition, ecological forecasts benefit from data integration as accurate predictions of future ecosystem states and processes require the appropriate propagation of uncertainty in parameter estimates, initial conditions, and future system states (Dietze *et al*. 2018). However, issues remain in the development, analysis, and interpretation of data integration models, particularly within macrosystems ecology, given that such analytical approaches are relatively new and often incompatible with standard software packages. We highlight statistical approaches from the literature to address common integration problems (WebTable 1). As integration challenges are often interdependent, approaches used to address one issue can affect the options available to address other issues. In many situations, more than one approach could reasonably be used to address integration issues, the most appropriate of which will depend on the specific data available, the complete set of challenges, and the ecological system and question(s) of interest.

Continued adoption and adaptation of formal approaches from the statistical literature can expand the utility of data integration analyses within ecological systems (eg Pacifici *et al*. 2019), and future research is likely to produce additional methods to resolve integration challenges. In addition to the techniques discussed in relation to the individual challenges (WebTable 1), we recommend three general considerations to help overcome analytical obstacles when integrating data at macroscales. First, it is often useful to begin by assessing the scope, grain, information content, and quantity of individual data sources to evaluate the potential structure and feasibility of an ecological model. This will identify the extent to which data sources can individually inform model estimates and help determine reasonable model complexity given available data. Second, simulating data, as well as developing and evaluating model components sequentially, can aid in determining whether inferential challenges are likely to arise prior to full model implementation. Data integration is often an iterative process, where new challenges arise as data sources are added or model structures are altered. Integrating data sources in steps using "perfect" simulated data can help pinpoint potential problems and solutions early on and determine how real datasets differ from simulated data. Finally, including random effects in one or more components of an integrated model can often ameliorate many of the inferential challenges discussed here. Random effects can be used to account for differences in the spatial and temporal extent of multiple data sources, inconsistencies in sampling effort and techniques, and variance in ecological processes not explained by available covariates (eg Pacifici *et al*. 2017).

Macrosystems ecology is emerging as a valuable and increasingly relevant field (McCallen *et al*. 2019) as society faces multifaceted, interconnected, and cross-scaled pressures from rapid and unprecedented global change (Dodds *et al*. 2021). Moreover, data integration analyses are primed to play an important role within macrosystems ecology because of the inherent need to combine data sources to obtain inferences at regional to continental scales, and the sheer volume of data that is available through automated and large-scale collection programs (LaDeau *et al*. 2017). However, determining how multiple, disparate sources of data can be used to address questions at macrosystem scales across spatial and temporal heterogeneity can be complex. Despite these challenges, data integration techniques have expanded the breadth of research focused on patterns and mechanistic processes operating at broad spatiotemporal scales (Isaac *et al*. 2020), and we expect that the continued, rapid development of data integration techniques will be crucial to advancing the growing field of macrosystems ecology.

Acknowledgements

Publication of this Special Issue was funded by the US National Science Foundation (NSF award number DEB 1928375). We thank the guest editors of the Special Issue (WK Dodds, S Fei, and S Chandra) for many useful comments that greatly improved this paper. This work was supported by the NSF, including awards EF-1702635 (EFZ), DBI-1954406 (EFZ), EF-1253225 (AOF), EF-1638577 (MCD), and EF-1703048 (MWT).

References

Albert CH, Yoccoz NG, Edwards TC, *et al*. 2010. Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography* **33**: 1028–37.

Besbeas P and Morgan BJT. 2014. Goodness of fit of integrated population models using calibrated simulation. *Methods Ecol Evol* **5**: 1373–82.

Bird TJ, Bates AE, Lefcheck JS, *et al*. 2014. Statistical solutions for error and bias in global citizen science datasets. *Biol Conserv* **173**: 144–54.

Branke J, Deb K, and Miettinen K. 2008. Multiobjective optimization: interactive and evolutionary approaches. Berlin, Germany: Springer Science & Business Media.

Brown JH and Maurer BA. 1989. Macroecology: the division of food and space among species on continents. *Science* **243**: 1145–50.

Campbell SP, Zylstra ER, Darst CR, *et al*. 2018. A spatially explicit hierarchical model to characterize population viability. *Ecol Appl* **28**: 2055–65.

Carvalho F, Punt AE, Chang YJ, *et al*. 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? *Fish Res* **192**: 28–40.

Conn PB, Thorson JT, and Johnson DS. 2017. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods Ecol Evol* **8**: 1535–46.

Cressie NA. 1996. Change of support and the modifiable areal unit problem. *Geographical Systems* **3**: 159–80.

de Valpine PD and Hilborn R. 2005. State-space likelihoods for nonlinear fisheries time-series. *Can J Fish Aquat Sci* **62**: 1937–52.

Dietze MC. 2017. Ecological forecasting. Princeton, NJ: Princeton University Press.

Dietze MC, Fox A, Beck-Johnson L, *et al.* 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. *P Natl Acad Sci USA* **115**: 1424–32.

Diggle PJ, Menezes R, and Su T. 2010. Geostatistical inference under preferential sampling. *J R Stat Soc C-Appl* **59**: 191–232.

Dodds WK, Rose KC, Fei S, and Chandra S. 2021. Macrosystems revisited: challenges and successes in a new subdiscipline of ecology. *Front Ecol Environ* **19**: 4–10.

Estes L, Elsen PR, Treuer T, *et al.* 2018. The spatial and temporal domains of modern ecology. *Nature Ecol Evol* **2**: 819–26.

Farr MT, Green DS, Holekamp KE, and Zipkin EF. 2020. Integrating distance sampling and presence-only data to estimate abundance. *Ecology*; doi.org/10.1002/ecy.3204.

Farrell KJ, Weathers KC, Sparks SH, *et al.* 2021. Training macrosystems scientists requires both interpersonal and technical skills. *Front Ecol Environ* **19**: 39–46.

Fei S, Guo Q, and Potter K. 2016. Macrosystems ecology: novel methods and new understanding of multi-scale patterns and processes. *Landscape Ecol* **31**: 1–6.

Fer I, Kelly R, Moorcroft PR, *et al.* 2018. Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences* **15**: 5801–30.

Fithian W, Elith J, Hastie T, and Keith DA. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol* **6**: 424–38.

Fletcher RJ, McCleery RA, Greene DU, and Tye CA. 2016. Integrated models that unite local and regional data reveal large-scale environmental relationships and improve predictions of species distributions. *Landscape Ecol* **31**: 1369–82.

Foody GM. 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecol Biogeogr* **13**: 315–20.

Gelfand AE, Sahu SK, and Holland DM. 2012. On the effect of preferential sampling in spatial prediction. *Environmetrics* **23**: 565–78.

Gotway CA and Young LJ. 2002. Combining incompatible spatial data. *J Am Stat Assoc* **97**: 632–48.

Grace JB, Anderson TM, and Smith MD. 2016. Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature* **529**: 390–93.

Grueber CE, Nakagawa S, Laws RJ, and Jamieson IG. 2011. Multimodel inference in ecology and evolution: challenges and solutions. *J Evol Biol* **24**: 699–711.

Heffernan JB, Soranno PA, Angilletta MJ, *et al.* 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. *Front Ecol Environ* **12**: 5–14.

Hefley TJ, Broms KM, Brost BM, *et al.* 2017. The basis function approach for modeling autocorrelation in ecological data. *Ecology* **98**: 632–46.

Hooten MB and Hobbs NT. 2015. A guide to Bayesian model selection for ecologists. *Ecol Monogr* **85**: 3–28.

Isaac NJ, Jarzyna MA, Keil P, *et al.* 2020. Data integration for large-scale models of species distributions. *Trends Ecol Evol* **35**: 56–67.

Itter MS, D'Orangeville L, Dawson A, *et al.* 2019. Boreal tree growth exhibits decadal scale ecological memory to drought and insect defoliation, but no negative response to their interaction. *J Ecol* **107**: 1288–301.

Johnston A, Fink D, Hochachka WM, and Kelling S. 2018. Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol Evol* **9**: 88–97.

Keenan TF, Davidson EA, Munger JW, and Richardson AD. 2013. Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecol Appl* **23**: 273–86.

Keenan TF, Davidson EA, Moffat AM, *et al.* 2012. Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. *Glob Change Biol* **18**: 2555–69.

Keller M, Schimel DS, Hargrove WW, and Hoffman FM. 2008. A continental strategy for the National Ecological Observatory Network. *Front Ecol Environ* **6**: 282–84.

Kéry M and Schaub M. 2011. Bayesian population analysis using WinBUGS: a hierarchical perspective. Cambridge, MA: Academic Press.

Ketz AC, Johnson TL, Monello RJ, *et al.* 2018. Estimating abundance of an open population with an N-mixture model using auxiliary data on animal movements. *Ecol Appl* **28**: 816–25.

Kuikka S, Vanhatalo J, Pulkkinen H, *et al.* 2014. Experiences in Bayesian inference in Baltic salmon management. *Stat Sci* **29**: 42–49.

LaDeau SL, Han BA, Rosi-Marshall EJ, and Weathers KC. 2017. The next decade of big data in ecosystem science. *Ecosystems* **20**: 274–83.

LaRue EA, Rohr J, Knott J, *et al.* 2021. The evolution of macrosystems biology. *Front Ecol Environ* **19**: 11–19.

LaSorte FA, Lepczyk CA, Burnett JL, *et al.* 2018. Opportunities and challenges for big data ornithology. *Condor* **120**: 414–26.

Lawler JJ and Edwards Jr TC. 2006. A variance-decomposition approach to investigating multi-scale habitat associations. *Condor* **108**: 47–58.

Levy O, Ball BA, Bond-Lamberty B, *et al.* 2014. Approaches to advance scientific understanding of macrosystems ecology. *Front Ecol Environ* **12**: 15–23.

Maunder MN and Piner KR. 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. *Fish Res* **192**: 16–27.

McCallen E, Knott J, Nunez-Mir G, *et al.* 2019. Trends in ecology: shifts in ecological research themes over the past four decades. *Front Ecol Environ* **17**: 109–16.

Michener WK and Jones MB. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol Evol* **27**: 85–93.

Miller DA, Pacifici K, Sanderlin JS, and Reich BJ. 2019. The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol Evol* **10**: 22–37.

Nguyen H, Cressie N, and Braverman A. 2012. Spatial statistical data fusion for remote sensing applications. *J Am Stat Assoc* **107**: 1004–18.

Nguyen H, Katzfuss M, Cressie N, and Braverman A. 2014. Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics* **56**: 174–85.

Pacifici K, Reich BJ, Miller DAW, and Pease B. 2019. Resolving misaligned spatial data with integrated distribution models. *Ecology* **100**: e02709.

Pacifici K, Reich BJ, Miller DAW, *et al.* 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* **98**: 840–50.

Richardson AD, Williams M, Hollinger DY, *et al.* 2010. Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia* **164**: 25–40.

Roberts DR, Bahn V, Ciuti S, *et al.* 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**: 913–29.

Robinson OJ, Ruiz-Gutiérrez V, Fink D, *et al.* 2018. Using citizen science data in integrated population models to inform conservation. *Biol Conserv* **227**: 361–68.

Rollinson CR, Finley AO, Alexander MR, *et al.* 2021. Working across space and time: nonstationarity in ecological research and application. *Front Ecol Environ* **19**: 66–72.

Rüegg J, Gries C, Bond-Lamberty B, *et al.* 2014. Completing the data life cycle: using information management in macrosystems ecology research. *Front Ecol Environ* **12**: 24–30.

Saunders SP, Farr MT, Wright AD, *et al.* 2019a. Disentangling data discrepancies and deficiencies with integrated population models. *Ecology* **100**: e02714.

Saunders SP, Ries L, Neupane N, *et al.* 2019b. Multi-scale seasonal factors drive the size of winter monarch colonies. *P Natl Acad Sci USA* **116**: 8609–14.

Schaub M and Abadi F. 2011. Integrated population models: a novel analysis framework for deeper insights into population dynamics. *J Ornithol* **152**: 227–37.

Schimel D, Schneider F, and JPL Carbon and Ecosystem Participants. 2019. Flux towers in the sky: global ecology from space. *New Phytol* **224**: 570–84.

Soranno PA, Cheruvelil KS, Bissell EG, *et al.* 2014. Cross-scale interactions: quantifying multi-scaled cause–effect relationships in macrosystems. *Front Ecol Environ* **12**: 65–73.

Sullivan BL, Aycrigg JL, Barry JH, *et al.* 2014. The eBird enterprise: an integrated approach to development and application of citizen science. *Biol Conserv* **169**: 31–40.

Viskari T, Hardiman B, Desai AR, and Dietze MC. 2015. Model-data assimilation of multiple phenological observations to constrain and predict leaf area index. *Ecol Appl* **25**: 546–58.

Williams M, Richardson AD, Reichstein M, *et al.* 2009. Improving land surface models with FLUXNET data. *Biogeosciences* **6**: 1341–59.

Williams PJ, Kendall WL, and Hooten MB. 2019. Model selection using multi-objective optimization. *Ecol Model* **404**: 21–26.

Zipkin EF and Saunders SP. 2018. Synthesizing multiple data types for biological conservation using integrated population models. *Biol Conserv* **217**: 240–50.

Zipkin EF, Rossman S, Yackulic CB, *et al.* 2017. Integrating count and detection–nondetection data to model population dynamics. *Ecology* **98**: 1640–50.

Supporting Information

Additional, web-only material may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/fee.2290/suppinfo

FrontiersEcoPics

Improved survival for an albino?

Dominated mostly by small scattered trees and a patchy canopy, savannas offer few places for bats to shelter. However, within Brazil's savanna ecoregion known as the Cerrado, riparian forests and seasonal dry forests provide an exception. Most bat species that naturally have white fur are associated with the habit of roosting under tree leaves; incidentally, the bottoms of most tree leaves are typically lighter in color than their tops. The bat pictured here was captured in a gallery forest near the Águas Emendadas Ecological Station (Planaltina, Brazil) in 2007. To the best of our knowledge, this is the first reported case of albinism not only for Dermanura cinerea, a frugivorous bat species normally distinguished by gray fur, but also among other bats in the Brazilian savanna. We believe that its white fur might better match the bottom color of the tree leaves in these forests, thereby offering greater camouflage and potentially improving its chances of survival.

Hernani FM Oliveira
Department of Zoology, University of Brasília, Brasília, Brazil
doi:10.1002/fee.2302