Realising the promise of large data and complex models

Citation for published version:
McCrea, R, King, R, Graham, L & Börger, L 2023, 'Realising the promise of large data and complex models', *Methods in ecology and evolution*, vol. 14, no. 1, pp. 4-11. https://doi.org/10.1111/2041-210X.14050

Digital Object Identifier (DOI):
10.1111/2041-210X.14050

Editorial: Realising the promise of large data and complex models

Rachel McCrea[1], Ruth King[2], Laura Graham[3,4] and Luca Börger[5,6]

[1] Department of Mathematics and Statistics, Lancaster University, Lancaster UK.
[2] School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK.
[3] Geography, Earth & Environmental Sciences, University of Birmingham, Birmingham, UK.
[4] Biodiversity, Ecology & Conservation Group, International Institute for Applied Systems Analysis, Vienna, Austria.
[5] Department of Biosciences, Swansea University, Swansea, UK.
[6] Centre for Biomathematics, Swansea University, Swansea, UK.

1 Motivation for this Special Feature

In an era of rapid change, ecologists are increasingly asked to provide answers to big, urgent questions of global concern (Solé and Levin, 2022; Yates et al., 2018; Sutherland et al., 2013). Concurrently, technological advances allow ecological data to be collected at increasingly higher resolutions (e.g. temporal and/or spatial scales), leading to both new types of data and larger datasets becoming available (Farley et al., 2018). These data provide the opportunity to investigate new, and even previously unanswerable, questions, e.g. from those concerning animal movements (Nathan et al., 2022) to those addressing conservation and sustainability issues (Runting et al., 2022). Increasingly realistic models need to be developed and fitted to these data (Fer et al., 2018), pushing the boundaries of the type and intricacy of questions that can be explored (Niu et al., 2020). However, big data and big models can lead to big troubles, across multiple aspects, from storing and processing the data to the fitting of complex models to data and interpreting the output.

Close collaborations between ecologists, statisticians, mathematical modellers, computer scientists and other disciplines offer exciting ways forward to solve these problems, leading to mutually beneficial advancements. For example, computer scientists may aid in the efficient storage/extraction of data and development of new algorithms; statisticians may help and guide ecologists in the analysis of data, fitting complex models to the data via efficient computational algorithms and propagating or quantifying uncertainties throughout the process; mathematicians can ensure models are constructed in the most suitable fashion for the specific questions asked and demonstrate suitable properties (such as, realistic territorial ranges; or population predictions); and ecologists can guide mathematical scientists on the biological characteristics of the systems studied and ecological interpretation of the corresponding results, thus informing future models and influencing policy decisions. The need to answer important ecological questions is unprecedented, due to declines in biodiversity and ecosystem services which will impact our ability to meet Sustainable Development Goals (Reyers and Selig, 2020), and it is through interdisciplinary collaborations that the biggest steps forward will be able to be made.

Data analysis challenges arise across the full data analytic pipeline, including processing and visualising the data, developing ecologically-relevant and interpretable models to fit to the data, adapting the associated algorithms to fit the models to the data efficiently and obtaining meaningful interpretations of the output. In practice, there are often many trade-offs between these different aspects due to the challenges that arise during the data analysis pipeline. For example, within the initial processing of the data, decisions may need to be made regarding cleaning the data (e.g. to remove recorded data errors) or the summarised form of the processed data to report (e.g. the temporal and/or spatial scale). This itself can be challenging and there will often be uncertainty within the process, leading to potential new errors being introduced. The decisions made will typically impact the model fitted to these data. For example, for motion-sensor camera trap data, there may be a trade-off between the level of initial data processing (i.e. the level of advanced tools that may be used for uniquely identifying individuals via, say machine learning techniques) and associated models that may be fitted to incorporate the amount of uncertainty in the pre-processed data (e.g. from assuming no error in the matches; to incorporating matching uncertainty; to allowing for both marked and unmarked individuals). Alternatively, complex models often require computationally intensive algorithms for them to be fitted to the data, which may not scale as datasets increase in size. This may lead to the consideration of a simpler model that can be more easily fitted, thus reducing the level of fine-detail that may be extracted from the data; or adaptations to the model-fitting process such as using some form of approximate model-fitting approach that aims to be robust to the approximations used, but potentially could lead to biased parameter estimates.

This Special Feature provides a combination of review papers and scientific articles that address one or more of the challenges of modern day analyses of large and/or complex ecological data. Echoing the challenges facing the discipline we present these in the natural statistical cycle, starting with the challenges of new types of data, to the limitations of statistical models and associated algorithms (and computer packages) used to fit the models to the data to the interpretation and presentation of the corresponding model outputs.

2 Broad themes

We consider each of the themes identified in turn relating to (i) data; (ii) statistical models and model-fitting; and (iii) visualisation and interpretation. However we also emphasise that these are very closely interlinked and although we have used these coarse "pigeon holes" there are many overlapping aspects and challenges.

2.1 Data

Ecology, like environmental sciences and other branches of biology, has entered into an era of big data, with enormous possibilities for a better understanding of environmental state (Runting et al., 2022). Data can be "big" due to different characteristics. The "Four Vs Framework" (see discussion in Farley et al. (2018) and references therein) discuss four distinct aspects: (1) volume: quantity of data (2) velocity: time-varying data; (3) variety: multiple data types with complex relationships; and (4) veracity: trustworthiness of the data. These different aspects often do not occur in isolation, leading to multiple intricate data challenges when analysing ecological data. We highlight just some of the problems and approaches to address specific associated "V" challenges that authors of the papers within this Special Feature have encountered and discussed.

Biologging sensor technologies have been at the forefront of creating large volumes of available data, frequently at a range of different scales. Thus, the analysis of biologging data is often pioneering within ecology in relation to big data, with the potential to rapidly transform our understanding of the ecology, particularly in their application to animal movements (Williams et al., 2020; Nathan et al., 2022). A key limitation of most current systems is however the trade-off between collecting ultra-fine sub-second scale movement and behaviour data over shorter periods of time vs. more coarse but longer-term movement and space use data. Wild et al. (2022) take advantage of rapid developments in the field of the Internet of Things, (i.e. methods for attaching electronic sensor devices, connected to a network, to everyday objects) to overcome key limitations in current biologging data networking systems and present new Wi-Fi solutions, combined with smart embedded software, for big biologging data. The authors are able demonstrate orders of magnitude of improvement in data retrieval efficiency, which is the biggest limitation of animal biologging systems. In particular, Wild et al. (2022) discuss in detail challenges and solutions concerning software architecture, onboard processing of biologging sensor data, difficulties of time synchronisation, and the data transmission concept and the pros and cons of different Wi-Fi infrastructures.

Advances in technology has also led to (perhaps less foreseen) forms of data gathering mechanisms gaining momentum, and associated build-up of large quantities of data, with the rise of citizen (or community) science initiatives. The resulting data from such initiatives are typically very varied in nature, often involving multiple data collection protocols with more limited/reduced structure than compared to traditional survey methods, including data arising from opportunistic events. Whilst analysing citizen science data from designed surveys requires carefully developed methods, difficulties increase markedly with data from semi-structured projects, e.g. without fixed data collection protocols or data collected by observers of any degree of observer knowledge. This leads to new challenges across the whole spectrum of the 4 "V"s. Whilst these challenges have some commonality in terms of similar issues to address and overcome, due to the large expanse of types of data collection techniques, the specific challenges and associated data analytic approaches will vary. Johnston et al. (2022) summarise four overarching categories of challenges: (i) observer behaviour, including, for example, spatial bias, observer or reporting differences, and false positive errors; (ii) data structures, relating to both measures of detectability and procedures for validation; (iii) statistical models, including the opportunities provided by data integration and multi-species models, but also sources of bias and computational limitations; and (iv) communication, motivated by the application of citizen science within biodiversity monitoring.

The veracity of data within biodiversity also arises in less obvious ways, outside the sphere of data collection protocols "in the field", most commonly considered as the reason for querying the trustworthiness of the data. In particular, there is a wealth of information contained with many ecological and biodiversity databases. However, to combine this information, data must typically be uniquely associated with specific species and taxa. This in itself raises methodological challenges, due to, for example, dynamic species names, the discovery of new species, changing biological attributes etc. As a result, homonyms, synonyms, and errors may accumulate while for many taxa a general consensus on an accepted name and taxonomic and phylogenetic relationships may not have been reached so that taxonomy itself may resemble a confusingly intricate tangled bank. To address such issues Grenié et al. (2022) provide an extensive review of the tools, databases and best practices for harmonising taxon names in biodiversity studies. In particular, they categorise the "wild world" of existing publicly available taxonomic databases and resources, along the axes of taxonomic breadth and spatial scope, and discuss the associated strengths and caveats of each database. In addition, on the practical computation side, they review the existing computational tools provided in different `R` packages for taxonomic harmonisation, and, perhaps rather fittingly, provide a "taxonomy" of the `R` packages, classifying them according to their associated functions.

2.2 Models and model fitting

A vast array of different statistical models have been developed and fitted to ecological data in the last decade or so (Royle et al., 2014; McCrea and Morgan, 2015; Kery and Royle, 2016; Guisan et al., 2017; Hooten et al., 2017; MacKenzie et al., 2018; Schaub and Kéry, 2021), often with limited critical review of the characteristics and associated disadvantages and challenges of each. The advancement in models and associated model-fitting tools reflect the changing quantity of the data (as highlighted above), quality of the data (e.g. increased spatial/temporal resolution), emerging forms of data from new technologies (e.g. earth observation and/or drone data, eDNA) and advanced computational techniques (and associated computational power). Thus, summary overviews of these emerging and advancing areas are important and timely for ecologists and statisticians to be able to understand what can, and often importantly, what cannot (or should not), be done and also provide tools for fitting such models to different data. These models encompass all areas of ecology from population and community ecology to landscape and ecosystem ecology. Interrogation of the associated modelling ideas motivates further advances in addressing the challenges and model development to account for additional data complexities or efficient model-fitting tools, for example. We briefly summarise here some of the types of models and associated challenges that arise across a range of different types of models, and data, within this Special Feature.

Developing, or adapting, general statistical models that can be applied to different forms of data can be very efficient scientifically. Such approaches also often permit the use of readily available software packages, for example, NIMBLE (de Valpine et al., 2017), R-INLA Lindgren and Rue (2015) and inlabru (Bachl et al., 2019) as well as specific application focused packages, such as MARK/RMARK (for capture-recapture models; (Laake, 2013); momentuHMM (for hidden Markov models applied to movement data; (McClintock and Michelot, 2018)) and Distance (for distance sampling; (Thomas et al., 2010)). Areas which have accessible software are witnessing substantial statistical development, enhanced by the flexibility of the computational tools provided. For example, R-INLA and inlabru have been used by both Laxton et al. (2022) and Torney et al. (2022), whilst Newman et al. (2022) discusses the relative merits of available software tools for fitting models. However, Barros et al. (2022) take one step further from the issue of readily accessible computer packages, suggesting that model fitting is not the primary challenge, rather that the models being used by ecologists need to be considered as predictive models, which can be used transparently and easily adapted following updated data sets or statistical methodology. Their proposal of the PERFICT workflow provides a framework by which these important challenges can be aligned.

Understanding the relationship between such general statistical models and specific ecological models can be challenging, as well as structuring the data into the required general form. Two particular "umbrella" models that have been applied extensively within ecological models are the closely related hidden Markov models (HMMs) and state-space models (SSM). Both of these types of models are widely used in ecological settings in the presence of longitudinal data (McClintock et al., 2021; Auger-Methe et al., 2021). One attraction of these models within the ecological applications, is that they both directly separate out the distinct ecological and/or sampling processes. This often simplifies the model specification, permitting the consideration of the separate components independently. A common distinction between these models relates to whether the latent processes are defined to be discrete-valued (for HMMs) or continuous-valued (SSMs); although we note that this distinction is not universally used. Specific ecological areas where these models have been extensively applied, include, but are far from limited to, for example, fisheries stock assessment (Aeberhard et al., 2018); population dynamics (Newman et al., 2014); animal movement (Langrock et al., 2012; Hooten et al., 2017; Patterson et al., 2017); and capture-recapture-type surveys (King, 2014; McCrea and Morgan, 2015). Glennie et al. (2022) and Newman et al. (2022) provide a methodological (and practical) review of HMMs and SSMs, respectively.

In particular, Glennie et al. (2022) highlight the potential difficulties that may be encountered when specifying HMMs for different systems, including issues which arise when model assumptions are not valid and the challenges of defining and fitting a suitable model in an HMM framework when the underlying hidden process increases in complexity. Providing descriptions of these general statistical models that can be applied to a variety of different forms of ecological data and associated discussion of issues to be aware of are a very useful resource for practitioners, particularly when describing the pitfalls that may arise. The rapid growth of the application of HMMs has also been aided by associated efficient model-fitting algorithms, due to the Markovian structure of the model (Zucchini et al., 2016).

The practical issues of fitting general and flexible SSMs, assuming a continuous-valued ecological (latent) process, is highlighted and addressed by Newman et al. (2022). Importantly, they discuss and contrast a wide-range of model-fitting techniques, dependent on the underlying assumptions of the specified model. In particular, they describe model-fitting algorithms that can accommodate more complex modelling dynamics, such as nonlinear processes and/or non-Gaussian stochasticity. Such models are less familiar/used within the ecological community, most likely due to the associated model-fitting challenges, however such adaptations of SSMs have great potential for the modelling of ecological data. The important aspect of what software can be used to fit such complex model is also highlighted in the paper.

The challenges of fitting models to data can concern both the associated algorithms required (as for SSMs), but also the increase in computational expense, particularly as the complexity of the model increases. With increasingly large datasets, for example, as routinely collected in bioacoustics or biologging studies (see (Wild et al., 2022)), many standard methods break down and cannot be practically applied. There is hence a necessity to identify and develop suitable modifications to improve computational efficiency and scalability, adapting traditional (and developing new) methods to big data. Providing successful examples, and the associated strategies that were most successful, including for example, computational efficiencies (Newman et al., 2022) and as demonstrated in King et al. (2022), as well as model simplifications that retain the signal within the data, are promising avenues forward. The challenges that arise regarding scalability due to large (and new) datasets are, however, also an opportunity for the development and use of machine learning algorithms. Off-the-shelf algorithms may however not be sufficient or be too limiting, as described by Wang et al. (2022), such that additional developments may be required for ecological applications. For example, it will generally be important to incorporate known ecological processes within the data analysis.

There are numerous opportunities, risks and trade-offs in building structurally complex models to increase insight on the underlying ecological processes. For example, Laxton et al. (2022) use the very popular species distribution models (SDMs) to highlight the importance of increasing model complexity based on ecological theory. The authors showcase the usefulness of a marked point process approach, which permits the inclusion of key population dynamic processes linked to ecological covariates (relating to landscape structure and the range of movements of the study species), and highlight the importance of maintaining an understanding of the roles and effects of each model component, to ensure interpretability and useful ecological insight. Alternatively, Torney et al. (2022) show that, in relation to the study of movement behaviour, including complex mechanisms driving animal distributions into the statistical models can substantially increase model performance and predictive ability. Further, they demonstrate that the relationship between model complexity and model performance is non-monotonic, highlighting the importance of robust procedures for checking models.

2.3 Interpretability and Visualisation

It is now possible to fit a wealth of complex models to data sets; however where does the line get drawn between fitting a model for complexity's sake and because it is actually required for an understanding of the dynamics exhibited by the data? In many cases can a simple model actually be more useful/informative? Such questions are long standing in many areas, including ecology (Murtaugh, 2007). Statistical models continue to be developed to represent the underlying data generating ecological processes - but these will always be a simplification of reality - with more complex models aiming to extract meaningful and useful interpretable ecological insight. In general, there is a trade-off between the complexity of the model being fitted and the associated intricacy of the information that can be extracted (given suitable and available data). Further, statistical learning (or machine learning) techniques are rapidly increasing in their prominence and usage within ecology (Pichler and Hartig, 2022; Ho and Goethals, 2022), with such techniques often demonstrating good predictive performance, but at the lack of ecologically interpretable parameters. Extracting interpretable and meaningful results/output from appropriate models fitted to real data, combined with intelligent visualisations, is becoming increasingly important, not least within the wider scientific community and policy-makers, for example.

One particular area of ecology in which increasing model complexity leads to further interpretability challenges is that of species' distribution modelling. Traditionally, such models have been used to establish a correlation between a single species and the environment that it occupies in order to gain an understanding of habitat suitability, or to predict the impacts of environmental change. However, there has been increasing interest for these models to go beyond a single species in isolation and to include interactions between species (Kissling et al., 2012; Pollock et al., 2014) and/or the underlying mechanisms (Buckley et al., 2010) in order to improve predictability of multi-species models. However, in increasing the complexity of the model, the associated interpretability of the model parameters can become more difficult. To address this issue Powell-Romero et al. (2022) use a feature-based approach to describing community structure within ensemble modelling approaches to improve the practical interpretability of multi-species models. Through the inclusion of simple features to describe communities, it is possible to obtain insight of not only which models outperform others, but also why this is the case. Further, within more complex dynamic SDMs, Laxton et al. (2022) argue that any increased complexity in the model needs to be grounded in ecological theory. This in turn permits greater interpretability since the different mechanisms or patterns of each component of the model can be identified leading to increased interpretable ecological insight.

As models and data become more complex and high-dimensional, obtaining meaningful and useful *visualisations* of the data and/or model outputs for improved insight also becomes more challenging. Traditional methods, such as dimension reduction and considering pair-wise correlations, may lead to more nuanced and/or intricate ecological insights to be masked, or even lead to biases in their presentation (McInerny et al., 2014; McInerny and Krzywinski, 2015). This is particularly challenging in more complex data/model structures, such as networks or graphs structures. For example, food web visualisation should allow us to gain an understanding of the structure of foodwebs, and provide insight into the detail of the complexity, however, current approaches tend to simplify the structure and therefore cannot provide the insight needed. To address some of these challenges, Pawluczuk and Iskrzyński (2022) propose methods for visualising increasingly complex foodweb (and other network) structures by combining heatmaps, interactive and animated graphs. Alternatively, Van Moorter et al. (2022) have developed the package ConScape (in Julia) which allows users to efficiently analyse and visualise landscape and habitat connectivity more simply. Further issues arise when attempting to analyse objects that contain multiple distinct (non-independent) parts that make up the complete object (e.g. when analysing skeletons rather than individual bones). With this focus, Thomas et al. (2022) propose a method based on regularised consensus principal components analysis to be able to summarise and compare shape variation in multi-part morphospaces. Importantly, they also provide an accompanying R package, to permit wider usage and impact within the large scientific community.

3 Concluding comments and Future Outlook

The opportunities for gaining an understanding of ecological systems from the range of different forms of available data (and new emerging data) are immense. However, to fully capitalise on these opportunities, addressing the associated challenges and achieving academic and societal impact, a multi-disciplinary approach considering the whole data analytic pipeline is required. We discuss a number of important aspects that will contribute to advancing ecological knowledge and address important societal issues (though we note that this is far from an exhaustive list):

Interdisciplinarity: Immersive interdisciplinarity in the ecological community's research approach has the largest potential for achieving research step-changes within the discipline. The cross-fertilisation of knowledge from, for example, ecologists, engineers (designing data collection devices), statisticians (developing advanced modelling techniques to fully exploit the available data and designing survey sampling strategies) and computer scientists (offering expertise in machine learning and automation) provides the opportunity for the co-creation of new and exciting approaches to address challenging ecological problems. Close collaboration with mathematical ecologists allows a better realistic connection of models to ecological theory; equally important is the collaboration with ecologists at the model output stage, to build confidence that the results are biologically realistic.

Data-centric methodological innovation: It is important to ensure that data analytic methods are being developed to make the most of the diverse and sizeable amounts of ecological data now being efficiently collected at increasing scale and quantity (Zipkin et al., 2021). However, the advancement of data collection technology continues at a rapid pace, and, necessarily the associated data analytic tools develop at a lagged timescale (there is no point in developing analytic tools for data that do not exist and/or cannot be collected). Again, an interdisciplinary outlook will help identifying novel data collection tools and methods not used yet in ecology.

Robust data integration: There has been a natural development towards integrating data sets within a single model in recent years (Frost et al., 2023), spanning both multile data types of a single species (Isaac et al., 2020) and data from multiple species (Barraquand and Gimenez, 2019). This means that one of the biggest challenges facing statistical ecologists is to think about whether the types of data being combined in an analysis are indeed comparable – do they have differing quality, and will this affect the model performance? For example, will combining small structured datasets with large unstructured data, for example from the Global Biodiversity Information Facility (GBIF), help to limit the bias in the latter, or the context dependency in the former? (Isaac et al., 2020)

"All models are wrong, but some are useful": This phrase attributed to the statistician George Box continues to provide useful insight. In particular, we apply this reasoning to the idea that the ability of being able to fit complex statistical models to data (accessible through advances in associated software) does not mean that the models are appropriate (or useful) for the data. There is a need to consider the philosophy of "should we" fit a model to a given data set, and ask whether it is necessary and/or appropriate given the particular ecological question of interest and available data. Gain in knowledge should trump model complexity or methods sophistication per se.

Machine learning and artificial intelligence: Such approaches are likely to have an important role in the future direction of methods in the ecological domain (Pichler and Hartig, 2022), particularly when prediction is a primary objective. However such methods should not simply be blindly applied to align with popular analytical trends - it is important that there is a methodological driver underpinning their usage. The interpretability of such models is more challenging due to the "black-box" nature of the algorithms and lack of ecological constraints or input, for example. Considerable debate and uncertainty remains in the validity and best practices of these approaches particularly in relation to generalisability, conceptual simplicity, robustness and transparency. There is a need to increase research efforts into machine learning and artificial intelligence approaches so that their power can be appropriately harnessed for ecology and evolution. For example, novel understanding from carefully fitted and interpreted machine learning methods could be more often also used to guide the development of new likelihood-based methods.

Software: This is an increasingly prominent feature of statistical analyses. The type of software ranges from general statistical packages to which ecological models and data analyses can be conducted (such as inlabru (Bachl et al., 2019) or NIMBLE (de Valpine et al., 2017)), to specialised packages for very specific problems (Van Moorter et al., 2022). However, the variety of computer packages (and in different languages, such as `R` or Python or Julia) leads to additional challenges of identifying the most relevant and/or efficient for the given problem at hand. Clear guidance regarding the advantages and disadvantages of different approaches is a particularly useful resource, though often difficult as there may be many different data and question dependent decisions in practice.

Communication: The importance of improved communication for addressing and solving the inherent challenges of citizen science data are highlighted in Johnston et al. (2022). In particular, the authors focus on the importance of disseminating new statistical methods beyond the limited circle of technical groups. This requires moving beyond code sharing, investing also in software development and teaching activities and resources. They also conclude that a 'democratisation' of data analysis may emulate the progress brought by the democratisation of data collection through citizen science and help make the most of these data, which has to be one of the most pressing issues facing statistical ecologists at this current time.

The papers in this Special Feature only scratch the surface of the challenges present with large data and complex models, and propose some possible approaches for dealing with different issues and advance our ecological understanding. These areas of research will continue to provide a rich and diverse set of challenges for ecological researchers. However, it is through recognising the challenges, building interdisciplinary data analytic pipelines, and providing interpretable results, that will ensure the research produced by this cross disciplinary academic community will reach its full potential, leading to step-changes in our ecological understanding, and be a firm basis for informed policy decision-making.

4 Acknowledgements

This special feature arose from discussions and interactions at the National Centre for Statistical Ecology meeting in Edinburgh in 2018, and the joint BES Quantitative and Movement Ecology Special Interest Group Meeting in Sheffield in 2018.

References

Aeberhard, W. H., J. M. Flemming, and A. Nielsen (2018). Review of state-space models for fisheries science. *Annual Review of Statistics and Its Application 5*, 215–235.

Auger-Methe, M., K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. M. Flemming, A. Nielsen, G. Petris, and L. Thomas (2021). A guide to state–space modeling of ecological time series. *Ecological Monographs 91*, 1–38.

Bachl, F. E., F. Lindgren, D. L. Borchers, and J. B. Illian (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution 10*(6), 760–766.

Barraquand, F. and O. Gimenez (2019). Integrating multiple data sources to fit matrix population models for interacting species. *Ecological Modelling 411*, 108713.

Barros, C., Y. Luo, A. Chubaty, I. Eddy, T. Micheletti, C. Boisvenue, D. Andison, S. Cumming, and E. McIntire (2022). Empowering ecological modellers with a PERFICT workflow: seamlessly linking data, parameterisation, prediction, validation and visualisation. *Methods in Ecology and Evolution*.

Buckley, L. B., M. C. Urban, M. J. Angilletta, L. G. Crozier, L. J. Rissler, and M. W. Sears (2010). Can mechanism inform species' distribution models? *Ecology Letters 13*(8), 1041–1054.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics 26*(2), 403–413.

Farley, S. S., A. Dawson, S. J. Goring, and J. W. Williams (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience 68*(8), 563–576.

Fer, I., R. Kelly, P. R. Moorcroft, A. D. Richardson, E. M. Cowdery, and M. C. Dietze (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences 15*(19), 5801–5830.

Frost, F., R. S. McCrea, R. King, O. Giminez, and E. Zipkin (2023). Integrated population models: Achieving their potential. *Journal of Statistical Theory and Practice 17.*

Glennie, R., T. Adam, V. Leos-Barajas, T. Michelot, T. Photopoulou, and B. T. McClintock (2022). Hidden markov models: Pitfalls and opportunities in ecology. *Methods in Ecology and Evolution n/a*(n/a).

Grenié, M., E. Berti, J. Carvajal-Quintero, G. M. L. Dädlow, A. Sagouis, and M. Winter (2022). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution n/a*(n/a).

Guisan, A., W. Thuiller, and N. E. Zimmermann (2017). *Habitat Suitability and Distribution Models: with Applications in R.* Cambridge University Press.

Ho, L. and P. Goethals (2022). Machine learning applications in river research: Trends, opportunities and challenges. *Methods in Ecology and Evolution 13*(11), 2603–2621.

Hooten, M. B., D. S. Johnson, B. T. McClintock, and J. M. Morales (2017). *Animal Movement: Statistical Models for Telemetry Data.* CRC Press: Boca Raton.

Isaac, N. J. B., M. A. Jarzyna, P. Keil, L. I. Dambly, P. H. Boersch-Supan, E. Browning, S. N. Freeman, N. Golding, G. Guillera-Arroita, P. A. Henrys, S. Jarvis, J. Lahoz-Monfort, J. Pagel, O. L. Pescott, R. Schmucki, E. G. Simmonds, and R. B. O'Hara (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution 35*(1), 56–67.

Johnston, A., E. Matechou, and E. B. Dennis (2022). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution n/a*(n/a).

Kery, M. and J. A. Royle (Eds.) (2016). *Applied Hierarchical Modeling in Ecology*. Boston: Academic Press.

King, R. (2014). Statistical ecology. *Annual Review of Statistics and its Application 1*, 410–426.

King, R., B. Sarzo, and V. Elvira (2022). When ecological individual heterogeneity models and large data collide: An importance sampling approach. Technical report, University of Edinburgh. https://arxiv.org/abs/2205.07261.

Kissling, W. D., C. F. Dormann, J. Groeneveld, T. Hickler, I. Kühn, G. J. McInerny, J. M. Montoya, C. Römermann, K. Schiffers, F. M. Schurr, A. Singer, J.-C. Svenning, N. E. Zimmermann, and R. B. O'Hara (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography 39*(12), 2163–2178.

Laake, J. (2013). RMark: An R interface for analysis of capture-recapture data with MARK. AFSC Processed Rep. 2013-01, Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., Seattle, WA.

Langrock, R., R. King, J. Matthiopoulos, L. Thomas, D. Fortin, and J. M. Morales (2012). Flexible hidden Markov-type models for animal telemetry data. *Ecology 93*, 2336–2342.

Laxton, Megan, R., O. Rodriguez de Rivera, A. Soriano-Redondo, and J. B. Illian (2022). Balancing structural complexity with ecological insight in spatio-temporal species distribution models. *Methods in Ecology and Evolution n/a*(n/a).

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software 63*(19), 1–25.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines (2018). *Occupancy Estimation and Modeling* (Second Edition ed.). Boston: Academic Press.

McClintock, B. T., R. Langrock, O. Gimenez, E. Cam, D. L. Borchers, R. Glennie, and T. A. Patterson (2021). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters 23*, 1878–1903.

McClintock, B. T. and T. Michelot (2018). momentuHMM: R package for generalized hidden Markov models of animal movement. *Methods in Ecology and Evolution 9*(6), 1518–1530.

McCrea, R. S. and B. J. T. Morgan (2015). *Analysis of capture-recapture data.* Boca Raton: Chapman and Hall/CRC Press.

McInerny, G. and M. Krzywinski (2015). Unentangling complex plots. *Nature Methods 12*(7), 591–591. Number: 7 Publisher: Nature Publishing Group.

McInerny, G. J., M. Chen, R. Freeman, D. Gavaghan, M. Meyer, F. Rowland, D. J. Spiegelhalter, M. Stefaner, G. Tessarolo, and J. Hortal (2014). Information visualisation for science and policy: Engaging users and avoiding bias. *Trends in Ecology & Evolution 29*(3), 148–157.

Murtaugh, P. A. (2007). Simplicity and complexity in ecological data analysis. *Ecology 88*(1), 56–62.

Nathan, R., C. T. Monk, R. Arlinghaus, T. Adam, J. Alós, M. Assaf, H. Baktoft, C. E. Beardsworth, M. G. Bertram, A. I. Bijleveld, T. Brodin, J. L. Brooks, A. Campos-Candela, S. J. Cooke, K. Gjelland, P. R. Gupte, R. Harel, G. Hellström, F. Jeltsch, S. S. Killen, T. Klefoth, R. Langrock, R. J. Lennox, E. Lourie, J. R. Madden, Y. Orchan, I. S. Pauwels, M. Říha, M. Roeleke, U. E. Schlägel, D. Shohami, J. Signer, S. Toledo, O. Vilk, S. Westrelin, M. A. Whiteside, and I. Jarić (2022). Big-data approaches lead to an increased understanding of the ecology of animal movement. *Science 375*(6582), eabg1780.

Newman, K., R. King, V. Elvira, P. de Valpine, R. S. McCrea, and B. J. T. Morgan (2022). State-space models for ecological time-series data: Practical model-fitting. *Methods in Ecology and Evolution n/a*(n/a).

Newman, K. B., S. T. Buckland, B. J. T. Morgan, R. King, D. L. Borchers, D. J. Cole, P. Besbeas, O. Gimenez, and L. Thomas (2014). *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-space Methods.* Springer: New York.

Niu, S., S. Wang, J. Wang, J. Xia, and G. Yu (2020). Integrative ecology in the era of big data—from observation to prediction. *Science China Earth Sciences 63*, 1429–1442.

Patterson, T. A., A. Parton, R. Langrock, P. G. Blackwell, L. Thomas, and R. King (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *Advances in Statistical Analysis 101*, 399–438.

Pawluczuk, and M. Iskrzyński (2022). Food web visualisation: Heat map, interactive graph and animated flow network. *Methods in Ecology and Evolution n/a*(n/a).

Pichler, M. and G. Hartig (2022). Machine learning and deep learning – A review for ecologists. Technical report. https://arxiv.org/abs/2204.05023.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution 5*(5), 397–406.

Powell-Romero, F., N. M. Fountain-Jones, A. Norberg, and N. J. Clark (2022). Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles. *Methods in Ecology and Evolution n/a*(n/a).

Reyers, B. and E. R. Selig (2020). Global targets that reveal the social–ecological interdependencies of sustainable development. *Nature Ecology and Evolution 4*, 1011–1019.

Royle, J., R. B. Chandler, R. Sollmann, and B. Gardner (2014). *Spatial Capture-recapture.* Boston: Academic Press.

Runting, R. K., S. Phinn, Z. Xie, O. Veter, and J. E. M. Watson (2022). Opportunities for big data in conservation and sustainability. *Nature Communications 11*, 2003.

Schaub, M. and M. Kéry (2021). *Integrated Population Models.* Academic Press.

Solé, R. and S. Levin (2022). Ecological complexity and the biosphere: The next 30 years. *Philosophical Transactions of the Royal Society B: Biological Sciences 377*(1857), 20210376.

Sutherland, W. J., R. P. Freckleton, H. C. J. Godfray, S. R. Beissinger, T. Benton, D. D. Cameron, Y. Carmel, D. A. Coomes, T. Coulson, M. C. Emmerson, R. S. Hails, G. C. Hays, D. J. Hodgson, M. J. Hutchings, D. Johnson, J. P. G. Jones, M. J. Keeling, H. Kokko,

W. E. Kunin, X. Lambin, O. T. Lewis, Y. Malhi, N. Mieszkowska, E. J. Milner-Gulland, K. Norris, A. B. Phillimore, D. W. Purves, J. M. Reid, D. C. Reuman, K. Thompson, J. M. J. Travis, L. A. Turnbull, D. A. Wardle, and T. Wiegand (2013). Identification of 100 fundamental ecological questions. *Journal of Ecology 101*(1), 58–67.

Thomas, D. B., A. M. T. Harmer, S. Giovanardi, E. J. Holvast, C. M. McGoverin, and A. Tenenhaus (2022). Constructing a multiple-part morphospace using a multiblock method. *Methods in Ecology and Evolution n/a*(n/a).

Thomas, L., S. T. Buckland, E. A. Rexstad, J. L. Laake, S. Strindberg, J. S. L. Hedley, R. B. Bishop, T. A. Marques, and K. P. Burnham (2010). Distance software: Design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology 47*, 5–14.

Torney, C. J., M. Laxton, D. J. Lloyd-Jones, E. M. Kohi, H. L. Frederick, D. C. Moyer, C. Mrisha, M. Mwita, and J. G. C. Hopcraft (2022). Estimating the abundance of a group-living species using multi-latent spatial models. *Methods in Ecology and Evolution n/a*(n/a).

Van Moorter, B., I. Kivimäki, A. Noack, R. Devooght, M. Panzacchi, K. R. Hall, P. Leleux, and M. Saerens (2022). Accelerating advances in landscape connectivity modelling with the conscape library. *Methods in Ecology and Evolution n/a*(n/a).

Wang, Z., H. Gong, M. Huang, F. Gu, J. Wei, Q. Guo, and W. Song (2022). A multi-model random forest ensemble method for an improved assessment of chinese terrestrial vegetation carbon density. *Methods in Ecology and Evolution n/a*(n/a).

Wild, T. A., M. Wikelski, S. Tyndel, G. Alarcón-Nieto, B. C. Klump, L. M. Aplin, M. Meboldt, and H. J. Williams (2022). Internet on animals: Wi-fi-enabled devices provide a solution for big data transmission in biologging. *Methods in Ecology and Evolution n/a*(n/a).

Williams, H. J., L. A. Taylor, S. Benhamou, A. I. Bijleveld, T. A. Clay, S. de Grissac, U. Demšar, H. M. English, N. Franconi, A. Gómez-Laich, R. C. Griffiths, W. P. Kay, J. M. Morales, J. R. Potts, K. F. Rogerson, C. Rutz, A. Spelt, A. M. Trevail, R. P. Wilson, and

L. Börger (2020). Optimizing the use of biologgers for movement ecology research. *Journal of Animal Ecology 89*(1), 186–206.

Yates, K. L., P. J. Bouchet, M. J. Caley, K. Mengersen, C. F. Randin, S. Parnell, A. H. Fielding, A. J. Bamford, S. Ban, A. M. Barbosa, C. F. Dormann, J. Elith, C. B. Embling, G. N. Ervin, R. Fisher, S. Gould, R. F. Graf, E. J. Gregr, P. N. Halpin, R. K. Heikkinen, S. Heinänen, A. R. Jones, P. K. Krishnakumar, V. Lauria, H. Lozano-Montes, L. Mannocci, C. Mellin, M. B. Mesgaran, E. Moreno-Amat, S. Mormede, E. Novaczek, S. Oppel, G. O. Crespo, A. T. Peterson, G. Rapacciuolo, J. J. Roberts, R. E. Ross, K. L. Scales, D. Schoeman, P. Snelgrove, G. Sundblad, W. Thuiller, L. G. Torres, H. Verbruggen, L. Wang, S. Wenger, M. J. Whittingham, Y. Zharikov, D. Zurell, and A. M. Sequeira (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology Evolution 33*(10), 790–802.

Zipkin, E. F., E. R. Zylstra, A. D. Wright, S. P. Saunders, A. O. Finley, M. C. Dietze, M. S. Itter, and M. W. Tingley (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment 19*(1), 30–38.

Zucchini, W., I. MacDonald, and R. Langrock (2016). *Hidden Markov Models for Time Series: An Introduction Using R* (2nd ed.). Chapman and Hall/CRC.